CVWW 2026

Proceedings of the 29th Computer Vision Winter Workshop

Ondřej Chum (ed.)

February 9 – February 11, 2026

Jindřichův Hradec, Czech Republic

ISBN 978-80-11-08247-5

Editor

Ondřej Chum
Visual Recognition Group
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2, 166 27 Prague 6, Czech Republic

Publisher

Czech Society for Cybernetics and Informatics
(Czech Pattern Recognition Society group)
Pod vodárenskou věží 4, 182 00 Prague 8,
Czech Republic

ISBN 978-80-11-08247-5

Preface

Dear colleagues,

The Computer Vision Winter Workshop is an annual international event supported by leading research groups from Graz, Ljubljana, Vienna, and Prague. It serves as a platform for researchers and PhD students to connect, exchange ideas, and foster collaboration, driving innovation in the field of computer vision. Topics of interest include, among others, pattern recognition, machine learning, image analysis, 3D vision, biometrics, human-computer interaction, vision for robotics, and applied computer vision.

This year, the winter workshop was organized by the Czech Technical University in Prague. The venue was the lovely town of Jindřichův Hradec in South Bohemia. There were 41 papers submitted to CVWW 2026 from various countries and institutions, including 12 submissions into the contributed papers track. Each contribution has received three independent reviews conducted by the Program Committee, comprising 52 esteemed experts in computer vision and machine learning. As a result of this double-blind review process, 8 original contributed papers were accepted for publication and presented at oral sessions in the workshop. In addition to the contributed presentations, the program included 28 invited talks. These were carefully selected by the Chairs in consultation with the Program Committee. Besides the standard human reviews, the authors of 34 papers opted in for an additional AI review. This experimental AI feedback was not part of the decision process. The details of the process and statistics of authors' perception of it were reported at the workshop, followed by a heated debate.

The highlight of this year's program is the keynote by Prof. Josef Šivic from Czech Technical University in Prague. We are grateful to Josef for his inspiring presentation.

We would like to thank the reviewers for their high-quality feedback, which provided valuable insights to the authors and contributed significantly to the success of CVWW 2026. We extend our gratitude to the local organizers, Eva Matysková, Petra Ivaničová, and Míra Purkrábek, who made the organization of the workshop easier.

Václav Hlaváč and Ondřej Chum, CVWW 2026 Chairs

Table of Contents

Keynote talk

Learning for physical interaction: from manipulating objects to designing protein interfaces

Josef Šivic
Czech Institute of Informatics and Robotics
Czech Technical University in Prague, Czech Republic

Abstract: Large-scale neural networks have enabled major progress in several areas of artificial intelligence, including natural language processing and computer vision, demonstrating remarkable performance on complex tasks such as writing computer programs or creating images. These impressive results are powered by Internet-scale datasets, transformer-based neural architectures, self-supervised learning techniques, and supercomputer infrastructures. However, the progress has been limited so far in areas that require interactions with physical environments, where collecting large-scale datasets is challenging and slow. Examples include robotic manipulation, where data collection is limited by the speed of the robot, or designing protein-protein interactions, where data collection is limited by the speed of laboratory experiments. In this talk, I will show our recent progress in this area.

Short speaker biography:

Josef Sivic holds a distinguished researcher position at the Czech Institute of Robotics, Informatics and Cybernetics (CIIRC) at the Czech Technical University in Prague, where he heads the Intelligent Machine Perception team and the ELLIS Unit Prague. He received the habilitation degree from Ecole Normale Superieure in Paris and PhD from the University of Oxford. After PhD, he was a post-doctoral associate at the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology and then spent more than 10 years at Inria Paris where he received an ERC Starting Grant. He was awarded the British Machine Vision Association Sullivan Thesis Prize, three test-of-time awards at major computer vision conferences, and, in 2023, an ERC Advanced Grant. From 2019 to 2025 he served on the board of the European Laboratory of Learning and Intelligent Systems (ELLIS).

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hradec, Czech Republic, February 9–12, 2026

Multi-Label Cardinality-Incremental Learning

Laurenz A. Farthofer
KAI GmbH.
Graz University of Technology
laurenz.farthofer@k-ai.at

Marc Masana
Institute of Visual Computing
Graz University of Technology
mmasana@tugraz.at

Code: https://github.com/LaurenzBeck/Cardinality-Incremental-Learning

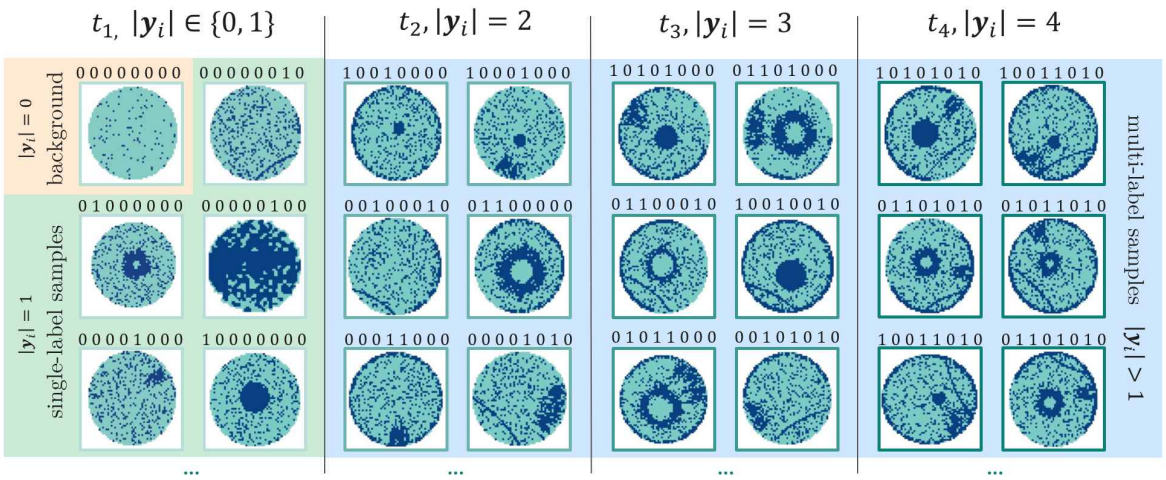


Figure 1. A *multi-label cardinality-incremental* scenario based on the MIXEDWM38 dataset [69], which consists of 38 000 wafer maps containing 0 to 4 of 8 possible defect patterns. It is constructed by splitting the dataset into four sequential tasks t_i with increasing cardinality $|y|$ (i.e., the number of patterns per sample). This continual multi-label learning scenario facilitates the study of how well a model generalises to new combinations of previously encountered classes, while not losing its ability to detect previous combinations.

Abstract

Real-world classification systems must be updated continually as new classes and new co-occurrences of classes emerge. Most work in class-incremental learning builds streams by splitting single-label datasets into disjoint tasks, a construction that omits concept repetition and under-represents core challenges of continual multi-label classification – a generalisation to new combinations of previously learned classes. To address this gap, we introduce a continual learning scenario which starts with single-label samples and incrementally introduces samples with increasing label cardinality. We analyse why standard class-incremental methods struggle in this setting and provide principled adaptations. Furthermore, we analyse how task-recency bias *manifests in this new scenario and propose an evaluation strategy with a focus on knowledge transfer. Our scenario provides a foundation for developing continual learners that scale to realistic multi-label settings. Furthermore, experiments on* PASCAL VOC *and* MS COCO *reveal that the natural repetition of classes acts as an implicit mechanism against catastrophic forgetting, aligning with existing literature.*

1. Introduction

A prime example of the adoption of deep learning in industrial applications is the semiconductor manufacturing context, where automated detection of process patterns in wafer maps can greatly enhance the effectiveness of the production [61]. Early work has focused on single-label classification [73]. While this paradigm simplifies the labelling process and allows the utilisation of established classifiers, the single-label assumption limits the reliability in interesting production events, where co-occurring process deviations cause multiple patterns to be present in wafer maps. Automatically detecting these rare events and notifying engineers to take counter measures is a main motivation for classification-based detection systems. This motivated the introduction of multi-label classification methods and datasets [1, 16, 17, 31, 41, 72].

Recent research has emphasised the importance of increasing the robustness of classifiers against data drifts inherent to most productive environments, thus decreasing the costly need for regular monitoring and re-training [79]. Continual Learning (CL) – the study of learning over a se-

quence of tasks – promises a more scalable solution compared to traditional learning on static i.i.d. datasets [9, 47, 70, 80]. Steps towards continual learning scenarios that are applicable to a wider range of real-world use-cases have recently been taken by removing the restriction of having a single class per sample [12, 63]. These *multi-label class-incremental* scenarios conform to the prevailing paradigm from *class-incremental* literature of having disjoint tasks. They are constructed by splitting datasets into tasks, where only a subset of the classes are annotated, even though objects from previous and future tasks are present. Therefore, methods focus on avoiding performance decline on classes from earlier tasks (*catastrophic forgetting* [48]) in the context of missing annotations.

In many real-world use-cases, classes frequently *reappear* throughout a stream, which acts as an implicit mechanism against *catastrophic forgetting* [23]. The focus of *domain incremental* scenarios [18, 25, 27] lies on dealing with changes in the environment instead of the class-space. We argue that research on *domain-incremental* settings will lead to methods better suited for real-world applications. Our primary contributions are:

1. proposing a novel CL scenario, where the data stream is constructed from tasks with increasing cardinality,
2. discussing aspects of adapting single-label class-incremental methods to the multi-label setting, and
3. the description of a new form of the *task-recency bias* which emerges in the proposed scenario. The *cardinality bias* describes the tendency of a continual learner to make predictions with cardinality biased towards values from more recently encountered tasks.

This work is an initial exploration of challenges such as *intra-class generalisation* and *cardinality bias* in continual multi-label learning. Our primary objective is to identify and discuss current limitations and challenges in this area, but we also provide recommendations in Sec. 5, laying the groundwork for future research on solutions.

2. Preliminaries

Multi-label classification. The goal of a classification task is to assign to each input instance $\mathbf{x}_i \in \mathcal{X}$ a corresponding class label $\mathbf{y}_i \in \mathcal{Y}$. The main difference between multi- and single-label classification is the number of possible class assignments per instance. While in single-label classification, each instance is assigned to exactly one of C possible classes, in multi-label classification, each sample can be assigned to zero, one or multiple classes, such that $\mathcal{Y} = \{0,1\}^C$.

The cardinality $|\mathbf{y}_i|$ of a multi-label sample is defined as the number of positive class assignments [13, 53, 75]:

$$|\mathbf{y}_i| = \sum_{c=1}^{C} \mathbf{y}_{i,c} \tag{1}$$

Multi-label targets and prediction vectors do not represent a single probability distribution over classes, but a collection of independent distributions. Adapting single-label methods to work with multi-label targets generally revolves around handling this difference [26].

Incremental learning scenarios. We consider an offline incremental learning scenario defined by a stream $\mathcal{S} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_T\}$ of T sequential tasks. At each incremental step, we have access to data $\mathcal{D}_t = \{(\mathbf{x}_i^t, \mathbf{y}_i^t)\}_{i=1}^{N_t}$ containing a set of N_t samples and their corresponding labels [63]. Depending on the stationary and evolving parts of the stream, van de Ven *et al*. [67] differentiate between *domain-incremental* learning, where a learner tries to solve the same classification problem ($\mathcal{Y}_t = \mathcal{Y}, \forall\, t \in T$) under different contexts (*e.g*. applying different image corruptions [25], or capturing objects in different environments [27, 44]), and *class-incremental* learning, where a learner needs to discriminate between disjoint sets of classes ($\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset, \forall\, i \neq j$) in a static environment.

In their categorisation of scenarios, Hemati *et al*. [23] focus on concept and instance repetition, allowing the reappearance of previously seen samples and the addition of samples from previously seen classes in new tasks, respectively. The *class-incremental with repetition* scenario [24, 66] lies within the continuum between the extreme cases of aforementioned class-incremental without any concept repetition and domain-incremental scenarios with full concept repetition in every task.

3. Cardinality-Incremental Scenarios

We propose a new continual multi-label scenario which focuses on the exploration of different label combinations (the sub-distributions within the class distributions). Similar to domain-incremental scenarios, all classes and pure backgrounds are introduced in the first task through samples with a cardinality of zero or one. Each subsequent task contains samples with new combinations of known classes with increasing cardinality, facilitating the study of intra-class generalisation (see Fig. 1 and Eq. (2)). This reflects the way humans are often introduced to new concepts: initially in isolation through dedicated instruction or examination, and subsequently encountered within a richer context alongside other concepts [5, 64]. We refer to this as the CARDINALITY-INCREMENTAL scenario.

$$|\mathbf{y}_i^t| = \begin{cases} \in \{0,1\} & \text{if } t = 1 \\ t & \text{if } t > 1 \end{cases} \tag{2}$$

Analogous to how the unrealistic stream constraints in the class-incremental setting have allowed a very detailed investigation of the phenomenon of *catastrophic forgetting* [48] by emphasising its stream-related causes, our scenario emphasises the problem of partial and incremental coverage of the sub-distributions of the different label combinations. Specifically, any approach that proves effective in this extreme construction should generalize seamlessly to more realistic settings, where the presence of repetition of classes and different cardinality values

would only serve to alleviate the challenges associated with continual learning.

3.1. Continual learning, baselines and methods

Researchers in the field of continual learning have established a set of common baselines [9, 21, 47, 54, 70, 82]. During *Finetuning*, the model parameters θ are naïvely optimised to solve the current task without measures to protect previously acquired knowledge, which makes it the most *plastic* but least *stable* baseline. During *Joint* training (also known as *cumulative* training), data from each task is accumulated $\mathcal{D}_t = \cup_{s=1}^{t} \mathcal{D}_s$. This serves as an upper bound for the classification performance. Unfortunately, its compute and memory requirements do not scale well to large data streams [68]. In *Freezing*, the parameters θ_f of the feature backbone f are frozen after the first task and only the parameters θ_g of the classification head g are updated for the rest of the tasks $t > 1$. In addition, we report results for *Static* (also known as *Source-Only*), where a model is only optimised on the first task [32], which simulates a model usage without updates.

From the early class-incremental works, *learning without forgetting* (*LwF*) [40] belongs to the family of approaches that apply *functional regularisation* to penalise activation drift in the predictions. *LwF* adds a *knowledge distillation* loss ℓ_{dis} [28] to the classification loss ℓ_{cls} which ensures that the temperature-scaled predictions $\hat{\mathbf{y}}^{\theta_t}$ of the current model stay close to the scaled predictions $\hat{\mathbf{y}}^{\theta_{t-1}}$ of a frozen copy of the model from the previous task:

$$\ell_{\text{dis}} = -\sum_{c=1}^{C} \frac{(\hat{\mathbf{y}}_c^{\theta_{t-1}})^{1/\tau}}{\sum_{j=1}^{C}(\hat{\mathbf{y}}_j^{\theta_{t-1}})^{1/\tau}} \log \frac{(\hat{\mathbf{y}}_c^{\theta_t})^{1/\tau}}{\sum_{j=1}^{C}(\hat{\mathbf{y}}_j^{\theta_t})^{1/\tau}} \quad (3)$$

We also report results for a variant that applies L_2 knowledge-distillation on a feature-level (*LwF-F*) [33].

Another prominent family of methods stores a fraction of the samples from previous tasks in a memory buffer $\mathcal{M}$ and combines those *exemplars* with the data from the current task ($\mathcal{D}'_t = \mathcal{D}_t \cup \mathcal{M}$) to approximate the joint distribution [6, 8, 56]. We chose the *Finetuning with exemplars* approach from [47] and refer to it as the *Replay* method.

Recently, class-incremental research has increasingly focused on exemplar-free methods, which often utilise class prototypes – the mean of the feature representations $\mathbf{h}$ – as a light-weight alternative to storing exemplars [54, 62, 82]. Orthogonal to the usage of class prototypes, exemplar-free methods often include representation learning tasks like a rotation prediction, where the label space $\mathcal{Y}$ is expanded to include rotated versions of the original classes. *Prototype augmentation and self-supervision* (*PASS*) [82] uses a three part loss consisting of a self-supervised loss ℓ_{ssl} based on that rotation expansion, a knowledge distillation loss ℓ_{dis} applied on the representations $\mathbf{h}$ (as in *LwF-F*), and a task specific classification loss ℓ_{cls} applied to samples of the current task and features sampled from a Gaussian distribution centred around the corresponding class prototypes with an adaptive uncertainty based on the variance of the representations of each class from the first task. They refer to the addition of noise to the stored features as explicit prototype augmentation.

The final family of methods included in our analysis specifically tackles the scaling of compute and memory to large streams by utilising product-quantised (PQ) [34] latent replay in conjunction with efficient online training schemes, where samples are only used once for model updates instead of being repeatedly iterated over throughout many epochs. *REMIND* [22] was among the first methods to utilize this strategy. They freeze an image backbone after the first task, fit a PQ model on the features and store the quantised features for later tasks. After the first task, they perform online training, where for every sample they construct a combined, class-balanced batch with stored, quantised features. *SIESTA* [21] extends *REMIND* with a two-phase learning cycle for online learning. An efficient *wake* phase adds incoming samples to the memory buffer and updates a cosine-similarity classification head using a backpropagation-free, online update step involving a running-mean calculation of each class-embedding. The *sleep* phase consists of multiple update steps on replayed class-balanced batches.

Some recent methods rely on the usage of pre-trained foundation models as backbones for continual learning, which are not adapted during training. This freezing of the backbone has many advantages like avoidance of representation drift by design [74], or the possibility to use closed-form, analytic updates for the classifier [49, 52, 83]. We do not consider those methods for our experiments for two reasons:

1. there are often no pre-trained foundation models available for many industrial tasks and use-cases, and
2. by design, the freezing of the backbone inhibits any form of backward transfer of knowledge, which we posit to be of great value in realistic scenarios, where concepts can re-occur later in a data stream.

3.2. Method adaptations

After the initial interest in the permuted MNIST domain-incremental scenario [18], a lot of continual learning research shifted its attention to class-incremental scenarios [37, 40, 76]. Fortunately, only some components of existing methods rely on clear class identities ($|\mathbf{y}_i| = 1$). Therefore, many strategies can be adapted to our setting.

A necessary adaptation for all methods is on the classification loss, as the cross-entropy loss ℓ_{CE} is not suitable for multi-label targets [26]. The binary-cross-entropy loss ℓ_{BCE} treats every class independently and is a standard (lower) baseline in multi-label classification. Recently, new losses have been proposed, including the asymmetric (ASL) [60] and the two-way loss [38]. We report results using the ASL loss, as it outperforms the other losses by a small margin (see Tab. S2).

Many class-incremental methods utilise cross-entropy-based knowledge distillation (Eq. (3)), which assumes that the predictions represent a single probability distribution. To adapt these methods, we use an alternative based on

the L_2 distance, proposed by Hinton *et al*. [28], which is already used by many continual learning methods [54, 82] that apply distillation on hidden representations:

$$\ell_{\text{dis}}^{(z)} = \left\|\mathbf{z}^{\theta_{t-1}} - \mathbf{z}^{\theta_t}\right\|_2 \quad \triangleleft \text{ on logits (\textit{LwF})}, \tag{4a}$$

$$\ell_{\text{dis}}^{(h)} = \left\|\mathbf{h}^{\theta_{t-1}} - \mathbf{h}^{\theta_t}\right\|_2 \quad \triangleleft \text{ on features (\textit{LwF-F})}. \tag{4b}$$

Another component that needs to be adapted is the exemplar selection, since most strategies from class-incremental literature rely on class indices [22, 56, 81]. Possible alternatives for the multi-label setting include the selection of samples regardless of the presence of other classes, or a selection based on the power set problem transformation of the label space $\wp(\mathcal{Y})$, where every combination of classes is treated as its own class, and $\wp(\mathcal{Y})$ is the set of all subsets of $\mathcal{Y}$, including the empty set $\emptyset$ and $\mathcal{Y}$ itself [11]. While this problem transformation allows utilising any (single-label) class-incremental method on a multi-label dataset, it comes with serious drawbacks:

1. label relationships are not efficiently encoded in the label space and need to be learned from data,
2. splitting the samples of each class into a set of 2^C disjoint combinations introduces a serious data imbalance problem with a long-tailed distribution where there are often no samples for many combinations, and
3. predicting the presence of a specific class (computing $p(y_c = 1|x)$) becomes complex due to the need to aggregate predictions for all combinations with y_c.

All these factors make the learning process less data-efficient. Thus, in the reported experiments, only pure samples with clear class identities ($|y| = 1$) from the first task are selected as exemplars for the reference buffer $\mathcal{M}$.

A prominent component of many methods is the nearest class mean classifier [2, 46, 56] or its differentiable approximation, the cosine similarity classifier [21], where it enables very efficient, backpropagation-free, online updates of the weights of the final classifier layer. We investigate a multi-label generalisation of the cosine-similarity classifier that was proposed in the hyper-spherical learning theory of Ke *et al*. [35]. In our preliminary experiments (see Fig. S7), we identified a significant difference in the magnitude of features from samples with multiple classes being present w.r.t. the ones where only one class is present. This causes the running mean used to update the class-embeddings (rows of θ_g) in the final layer to slowly drift towards large values when naïvely using the equation from [21]. We propose to normalise the computed features $\mathbf{h}$ to stay close to the hyper-sphere throughout the training:

$$\theta_{g,c} \leftarrow \frac{k_c \theta_{g,c} + \frac{\mathbf{h}}{||\mathbf{h}||}}{k_c + 1}, \tag{5}$$

where k_c is the running sample count for class c.

Very related to the scaling problems of exemplar selection strategies from class-incremental literature is the usage of class-prototypes of exemplar-free methods such as *PASS* [82], where the mean of the representations $\mathbf{h}$ would include signals from other classes in multi-label samples. To cover as many adaptation options as possible, we explore the use of the power-set problem transformation. For this we store a class-prototype for each encountered label combination. While the self-supervised rotation prediction task of *PASS* is applicable to multi-label data without adaptation, in the specific use-case of wafer maps, their patterns tend to have either a spherical symmetry or can appear in any position with the same distance to the centre of the wafer. Thus, discrimination of the extra classes generated through rotation is unfeasible. Therefore, we use the rotation from *PASS* exclusively as data augmentation and not label augmentation in all experiments.

4. Experiments

We compare the performance of different baselines and adapted methods, that were originally proposed for the single-label class-incremental scenario, on our proposed *cardinality-incremental* scenarios constructed from the MIXEDWM38 [69], PASCAL VOC 2012 [14] and MS COCO 2017 [42] datasets. See Sec. A and Fig. S5 in the supplements for more details.

4.1. Training protocol

We adopt the implementation, training protocol (see Algorithm 2) and most hyper-parameters (see Tab. S3) from both the FACIL framework [47] and the work from Liu *et al*. [43]. The latter provides a very strong upper bound for (non-incremental) multi-label learning on VOC and COCO. For improved performance and throughput [60], a TRESNET-M [58] model f (IMAGENET-1K [19] pre-trained for VOC and COCO, randomly initialized for MIXEDWM38) and a linear classification head g with parameters $\theta = \{\theta_f, \theta_g\}$ are optimized on the ASL loss [60] with focusing parameters $\gamma^- = 2$, $\gamma^+ = 0$ using stochastic gradient descent. Following Masana *et al*. [47], we use an adaptive stepped learning rate schedule with early stopping. This ensures that methods which converge slower receive more update steps, leading to a fairer comparison. Each experiment is repeated with the same five seeds to provide 95% confidence intervals.

4.2. Evaluation

In class-incremental learning, the performance of a model is measured after each task on the test data from the current and previous tasks [9, 47]. In domain- and cardinality-incremental scenarios, the label space $\mathcal{Y}$ stays constant throughout the stream, allowing the measurement of the performance on future (yet unseen) data as well, which is often used as a measure of generalisation.

In practice, evaluation usually involves complex indexing and metric aggregation schemes that use both the index of the task the model was last trained on and the index of the task on which the model is evaluated. We propose to shift this indexing complexity to the construction of several test streams, which aggregate the test data of different tasks, simplifying the implementation of the

evaluation to a simple indexing into every test stream using the current task index t (see Algorithm 1). The PRESENT stream $\{\mathcal{D}_1^{\text{test}}, \mathcal{D}_2^{\text{test}}, \cdots, \mathcal{D}_T^{\text{test}}\}$ contains the current in-distribution test set for each task and can either be given by a dataset, or split from the train stream. The PAST and FUTURE streams are used to estimate the backward and forward transfer of knowledge [45] respectively, and the ALL stream, contains all test data in every task $\mathcal{D}_t^{\text{all}} = \cup_{s=1}^{T} \mathcal{D}_s^{\text{test}}$, providing a global performance overview throughout the training.

$$\mathcal{D}_t^{\text{past}} = \begin{cases} \emptyset & \text{if } t = 1 \\ \cup_{s=1}^{t-1} \mathcal{D}_s^{\text{test}} & \text{if } t > 1 \end{cases} \tag{6a}$$

$$\mathcal{D}_t^{\text{future}} = \begin{cases} \cup_{s=t+1}^{T} \mathcal{D}_s^{\text{test}} & \text{if } t < T \\ \emptyset & \text{if } t = T \end{cases} \tag{6b}$$

A detailed discussion of the benefits of this evaluation approach is included in the supplementary material Sec. E.

In addition to metrics like *precision*, *recall* or *mean average-precision* (mAP), we also use cardinality confusion matrices to investigate the relation between the predicted and true cardinality of the samples.

5. Discussion

Differences to class-incremental learning. Typical class-incremental curves of metrics computed on past data differ mostly in their rate of decay after a common starting point. In comparison, the mAP-scores for each method from the PAST test stream have very different offsets, which stay relatively constant across tasks (see Fig. 2). This indicates that the natural repetition of classes present in this scenario acts as an implicit measure against the forgetting of classes, which is in line with the class-incremental learning with repetition literature [23, 66].

The curves on the PRESENT stream exhibit a general downward trend, which is in line with the class-incremental scenario [3, 47]. There, the performance of a continual learner on the current task usually decreases over time since: **1**. free model capacity for learning the new tasks is decreasing throughout training, and **2**. the task complexity grows with an increasing number of classes. Both factors appear in our proposed scenario, albeit in different forms: **a**. even though a model does not encounter new classes throughout the stream, it still incrementally encounters the different sub-distributions of the label combinations, which have semantic differences and thus require model capacity, and **b**. there is a strong correlation between the cardinality and the complexity of a scene, which explains the increasing task complexity of our setting, as also observed by Hajimirsadeghi *et al*. [20]. We explore the relationship between cardinality and task complexity in Sec. A and Fig. S6 in the supplements.

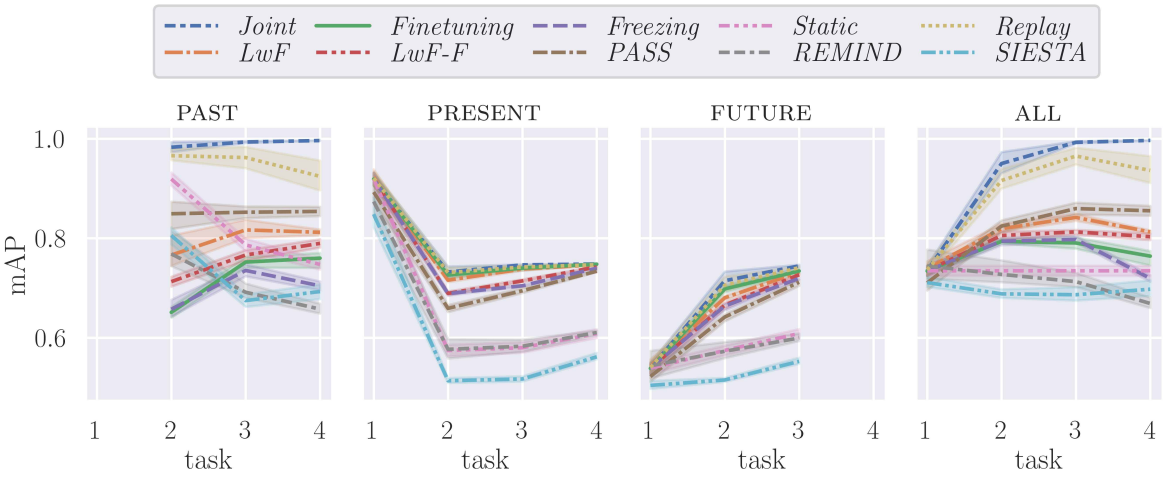
The presence of positive forward transfer in the FUTURE stream for all datasets suggests that the model benefits from the increasing semantic similarities between later tasks. Moreover, the significant improvement in overall performance on ALL test data after the introduction of real multi-label samples ($|\mathbf{y}_i| > 1$) in the second task underlines the importance of these samples for learning to discriminate between classes within samples.

A final difference is the lower variance of results when experiments are repeated over several random seeds, which can be attributed to the deterministic stream construction process as opposed to a random task split, which is a common practice in class-incremental learning [67].

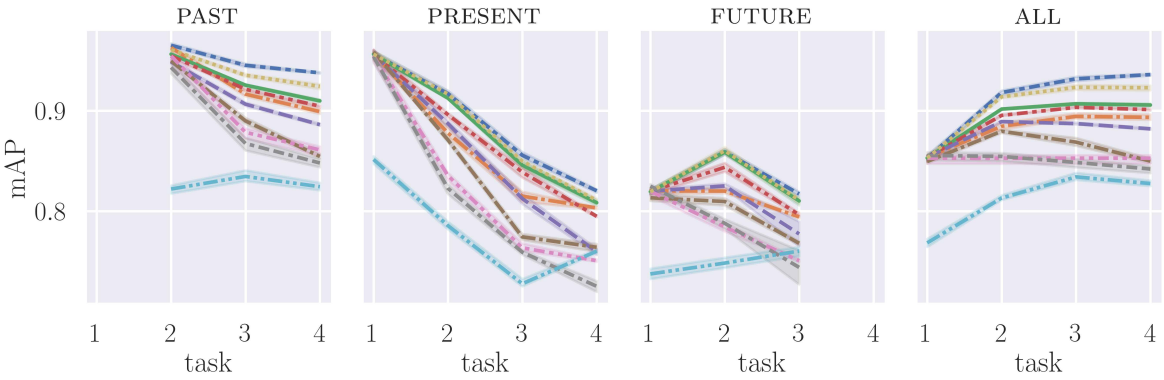
Differences between the datasets. Before comparing the different methods, we want to highlight a few differences between the three evaluated datasets. The MIXEDWM38 dataset [69] has a relatively balanced distribution of the labels regarding both classes and cardinality. All wafer maps are centred and have uniform dimensions of 52×52 pixels. Furthermore, combinations of different patterns do not change the locations of the individual patterns, which removes the need to learn spatial label relationships, for which transformer decoders as classification heads have shown promising results [7, 39, 43, 59]. All these factors combined lead to a relatively simple *cardinality-incremental* scenario, which is ideal for prototyping.

By contrast, natural-image benchmarks (*e.g*. COCO and VOC) exhibit greater variability in object scale and composition. This diversity tends to reduce forward transfer and increase forgetting (see Sec. A in the supplements). This semantic gap between samples of higher and lower cardinality provides an explanation for why the *Finetuning* baseline performs on par with, or better than, all evaluated methods except *Replay*: when the feature distributions of successive tasks are sufficiently distinct, updates from later tasks are less likely to interfere with parameters critical for earlier tasks, thereby reducing forgetting [37]. Rather than being discouraging, this opens the door to revisiting the design goals of continual-learning algorithms – specifically, to better preserve knowledge tied to the sub-distributions of the different label combinations that we hypothesize to be paramount for robust multi-label generalization. In line with this perspective, Prabhu *et al*. [55] showed that a simple baseline (*GDumb*) can outperform specialized solutions, questioning the efficacy of recent continual learning algorithms. Originally, we expected the presence of samples without any objects to be of great value in the process of learning to distinguish the objects from the background. However, PASCAL VOC does not have such background samples, and the results do not differ significantly from the other two datasets, which contradicted our expectations.

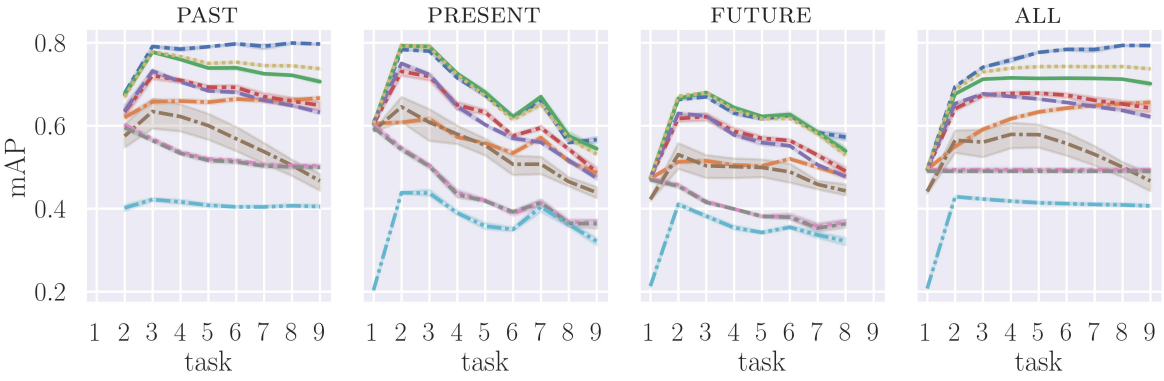
Differences between the methods. As in the class-incremental setting, *Joint* training is the upper performance bound for all three datasets. Its performance is almost matched by *Replay* with its simple rehearsal scheme of adding 100 exemplars per class from the first task. Research by Hemati *et al*. [23] and Hess *et al*. [27] in scenarios involving some form of either concept or instance repetition align with our observations for the *cardinality-incremental* scenario: **1**. given enough concept repetition,



(a) MIXEDWM38 [69]



(b) PASCAL VOC 2012 [14]



(c) MS COCO 2017 [42]

Figure 2. A comparison of different continual learning baselines and methods [21, 22, 40, 47, 82] evaluated after training on every task of three cardinality-incremental scenarios. The PAST stream contains all the test data from previously seen tasks and provides insight into how much each method suffers from catastrophic forgetting. The FUTURE stream contains test data with yet unseen cardinality values from future tasks and provides insight into forward transfer of knowledge. The PRESENT stream contains in-distribution test data for the current task, and the ALL stream tests the global performance of the methods on all test data.

even *Finetuning* can accumulate knowledge with minor forgetting, **2**. regularisation that protects acquired knowledge against catastrophic forgetting can degrade classification performance whenever there is a need for plasticity to accommodate new sub-distributions within the class-distributions, and **3**. implementing some form of replay,

Table 1. Results of the *cardinality-incremental* baselines and methods adapted from class-incremental literature. Metrics are calculated on ALL test data after the final task and include precision (🎯), recall (🔎), mean average precision (mAP), and total runtime (⏱).

	Method	MIXEDWM38 [69]				PASCAL VOC 2012 [14]				MS COCO 2017 [42]			
		🎯↑	🔎↑	mAP↑	⏱↓	🎯↑	🔎↑	mAP↑	⏱↓	🎯↑	🔎↑	mAP↑	⏱↓
Baselines	*Joint*	95	99	100	27 m	78	93	94	77 m	66	78	79	75.7 h
	Finetuning	59	78	76	12 m	47	97	91	45 m	30	89	70	5.8 h
	Freezing	69	90	72	11 m	46	97	88	44 m	34	82	62	2.9 h
	Static	66	51	73	5 m	86	76	85	17 m	53	36	49	1.3 h
Adaptations	*Replay* [47]	**75**	**99**	**94**	14 m	72	93	**92**	70 m	54	80	**74**	5.5 h
	LwF [40]	71	91	81	16 m	63	93	89	24 m	**60**	64	66	6.7 h
	LwF-F [40]	69	92	80	15 m	50	**97**	90	33 m	37	84	64	14.0 h
	PASS [82]	67	97	86	49 m	47	92	85	93 m	31	68	47	25.8 h
	REMIND [22]	60	69	67	**6 m**	**82**	75	84	**42 m**	51	37	49	**1.7 h**
	SIESTA [21]	40	98	70	10 m	53	91	83	42 m	5	**99**	41	1.9 h

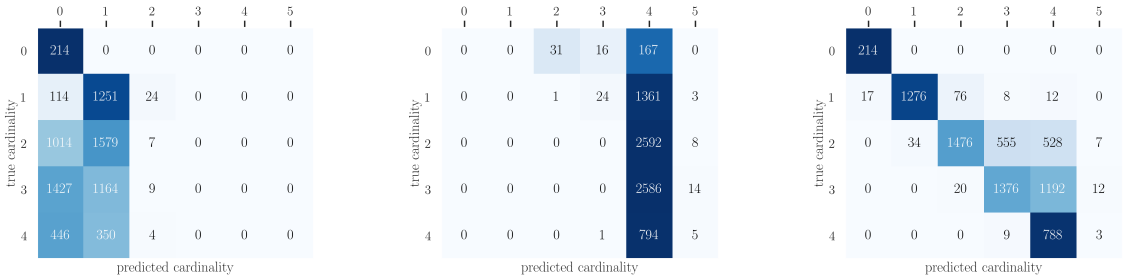
whether through exemplars or prototypes, is highly effective in achieving performance close to *Joint* training.

Given that scaling issues and privacy concerns are less pertinent in small-scale company-internal use cases, approaches that utilize *Replay* represent both effective and efficient solutions for industrial applications.

Finetuning is the most *plastic* baseline, which generally leads to the highest performance on the current task, while suffering the most from *task-recency bias* [50]. Zhao *et al*. [78] attribute this bias to a misalignment of the weights of old and new classes, where the weights of the new classes exhibit higher L_2 norms. In Fig. S7, we provide a visualisation of the distribution of the L_2 norms of the weights of the classifier g. They confirm a clear upwards trend throughout the whole incremental learning sequence, which is not as pronounced as in previous work [29, 47]. Intriguingly, this increase in the L_2 norms of the weights does not merely bias the model towards the class distributions from recent tasks but also affects the cardinality of the predictions, as evidenced by the cardinality confusion matrices in Fig. 3. We term this new aspect of *task-recency bias* as the *cardinality bias*. This finding challenges the multi-label assumption that the classification of each class is independent of the presence of other classes and underlines the necessity for an explicit learning of de-correlated class and label relationships from data, as these are not an innate outcome of multi-label learning.

The *Static* and *Freezing* baselines represent the most *stable* and least *plastic* methods. Notably, *Freezing* still allows to update the weights of the final linear layer, which consistently improves overall performance, without sacrificing performance on the PAST data. This suggests that the final linear layer is the place where most of the adaptations for the different sub-distributions within a class happens. This is in line with the argumentation from Petit *et al*. [54], who stressed the importance of the ability of a model to reorganise the feature space to encode new information in existing class regions. The performance gains of *LwF* and *LwF-F* [40], which are more *plastic* versions of the aforementioned baselines, further support this idea.

PASS [82] is the method with the most notable downward trend on the ALL stream of the natural image datasets, which indicates that the (noise) augmentation of the prototypes of every label combination using the variance estimates from the first task is detrimental to the per-



(a) *Finetuning* and *Replay* after task 1 (b) *Finetuning* after task 4 (c) *Replay* after task 4

Figure 3. Cardinality confusion matrices of the *Finetuning* baseline after training on the first (3a) and last (3b) tasks, evaluated on ALL test data of the MIXEDWM38 dataset [69]. They display a clear bias towards the cardinality values present at those tasks, indicating that the model has *forgotten* how to correctly determine them. *Replaying* a few samples from the first task (3c) reduces the bias significantly.

MAJOR FINDINGS OF OUR PERFORMANCE EVALUATION ON CARDINALITY-INCREMENTAL LEARNING

- Using multi-label losses is not enough for learning to discriminate between classes within samples. Access to samples with high cardinality is crucial.
- We observed a distinct *cardinality bias* towards recent tasks in all rehearsal-free methods.
- Methods that are designed to leverage high class-separability struggle in multi-label settings.
- Learning the sub-distributions of the different label combinations requires sufficient capacity and happens mostly in the final linear layer. (see Fig. S8)
- When there is a strong semantic difference between samples of different cardinality (*e.g.* in scale or composition), *Finetuning* performs better than all other methods except *Replay*, raising concerns as to whether they protect the knowledge from the class-conditional sub-distribution of different label combinations.
- The hyper-spherical learning generalisation of the cosine-similarity classifier from Ke *et al.* [35] leads to a collapse in precision caused by an over-exploitation of the label-relationships.

formance. However, estimating the covariance matrices for every label combination would further increase the scaling issues of the method. A feature translation approach as suggested by Petit *et al.* [54] would just turn the problem on its head: instead of forcing a covariance structure of the single-label samples on the multi-label ones, we would use the covariance structure from multi-label data from later tasks to update the distributions of the exemplars from single-label data from the first task. The absolute difference in mAP between *Finetuning* and *PASS* on COCO of 23% (see Tab. 1) suggests that methods designed for single-label distributions with a high class-separability do not fit the needs of multi-label datasets.

With its very efficient online training scheme after the first task, *REMIND* almost reaches the training time of the *Static* baseline, which just trains on the first task. Unfortunately, the class-balanced batch-construction, where the majority of the samples within a batch come from the product-quantised features from the memory buffer, is not plastic enough to learn generalisable features, consistently leading to worse performance than the *Static* baseline.

Unfortunately, the hyper-spherical learning generalisation of the cosine-similarity classifier from Ke *et al.* [35] that we used for the adaptation of *SIESTA* [21] leads to a collapse in precision caused by an over-exploitation of the label-relationships, which can be seen as a more drastic version of the *cardinality bias*. We suspect that making the individual classifiers more expressive than a single linear layer (*e.g.* the label-wise embeddings proposed by Yang *et al.* [75]) might help against this collapse in precision. Given the interesting applications of the cosine-similarity classifier and its frequent use in class-incremental literature, it is a component worth improving for the multi-label setting.

6. Recommendations

In this paper, we wanted to focus on a thorough introduction of the *cardinality-incremental* scenario and the challenges of continual multi-label learning it reveals. To maintain that as the main focus of our work, we did not propose a dedicated continual learning method that tackles the investigated challenges like the *cardinality bias*. Nevertheless, we want to provide recommendations, that should improve a continual multi-label learner:

1. Balancing not only the classes in a replay strategy (*e.g.* [23]), but also the distribution of the cardinality values (*e.g.* [65]) to counteract *cardinality bias*.
2. Using a multi-label adaptation of *MixUp* regularization [71, 77] to artificially explore label combinations for a better forward transfer of knowledge. Mi *et al.* [51] already argued that *MixUp* regularization counteracts many challenges related to continual learning with concept repetition.
3. Using the multi-label version of knowledge distillation and the label-wise embedding classifier from Yang *et al.* [75] to improve all methods that rely on distillation.

7. Conclusions

Multi-label learning has already shown substantial improvements in semiconductor manufacturing contexts, where it enhances the effectiveness of the production through automated inspection of wafer maps [61, 69]. However, there is a notable gap in research on studying multi-label classification methods in realistic continual learning settings [79]. Existing multi-label class-incremental scenarios [12, 63] do not adequately represent real-world data streams, as they lack concept repetition – a vital property inherent to most natural streams. The *multi-label cardinality-incremental scenario* presents a novel and valuable setting to explore a critical aspect of continual multi-label learners: their ability to detect classes reliably in the presence of other classes. This ability is crucial for achieving greater generalisation to new combinations of previously learned classes.

Our preliminary experiments, with methods from class-incremental literature that were adapted to the multi-label context, have uncovered new learning dynamics and characteristics unique to this scenario, which we summarize in the grey box at the top of this page. Particularly, the *cardinality bias* highlights a critical area for further investigation and method development, for which we provide recommendations in Sec. 6. It becomes evident that continual learning in multi-label contexts is not merely an extension of single-label or class-incremental learning. The unique challenges it presents, such as managing the overlap and inter-dependencies between classes while avoiding catastrophic forgetting, require dedicated research.

Acknowledgements

The authors would like to thank Gianluca Guglielmo and Benedikt Tscheschner for the fruitful discussions during the development of this research. We would also like to thank Thomas Pock for his academic supervision at Graz University of Technology.
This work was funded by the Austrian Research Promotion Agency (FFG, Project No. 931130), and by the Austrian Science Fund (FWF) 10.55776/COE12.

References

[1] Insung Baek, Sung Jin Hwang, and Seoung Bum Kim. Cowssl: contrastive open-world semi-supervised learning for wafer bin map. *J. Intell. Manuf.*, 36(3):2163–2175, 2025. 1

[2] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *ICCV*, pages 583–592, 2019. 4

[3] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 135, 2021. 5

[4] Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215, 2022. 13

[5] Jerome S. Bruner. *The Process of Education*. Harvard University Press, 2009. 2

[6] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*, 33: 15920–15930, 2020. 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020*, pages 213–229. Springer International Publishing, 2020. 5

[8] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019. 3

[9] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, 44(7): 3366–3385, 2021. 2, 3, 4, 15

[10] Nicki Detlefsen, Jiri Borovec, Justus Schock, Ananya Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics - Measuring Reproducibility in PyTorch. *JOSS*, 7(70):4101, 2022. 14

[11] Keith J. Devlin. *Fundamentals of Contemporary Set Theory*. Springer US, 1979. 4

[12] Songlin Dong, Haoyu Luo, Yuhang He, Xing Wei, Jie Cheng, and Yihong Gong. Knowledge restore and transfer for multi-label class-incremental learning. In *ICCV*, pages 18711–18720, 2023. 2, 8

[13] Kaile Du, Yifan Zhou, Fan Lyu, Yuyang Li, Chen Lu, and Guangcan Liu. Confidence self-calibration for multi-label class-incremental learning. In *ECCV*, pages 234–252. Springer, 2024. 2

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 4, 6, 7, 12, 13, 14, 16

[15] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018. 15

[16] Luca Frittoli, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, and Giacomo Boracchi. Deep open-set recognition for silicon wafer production monitoring. *PR*, 124: 108488, 2022. 1

[17] Hao Geng, Fan Yang, Xuan Zeng, and Bei Yu. When wafer failure pattern classification meets few-shot learning and self-supervised learning. In *IEEE/ACM Int. Conf. Computer Aided Design*, pages 1–8. IEEE, 2021. 1

[18] Ian J. Goodfellow, Mehdi Mirza, Xia Da, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *ICLR*, 2014. 2, 3

[19] Julia Grabinski. ImageNet-1k and ImageNet-100, 2024. 4

[20] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2596–2605, 2015. 5

[21] Md Yousuf Harun, Jhair Gallardo, Tyler L. Hayes, Ronald Kemker, and Christopher Kanan. SIESTA: Efficient online continual learning with sleep. *Transactions on Machine Learning Research*, 2023. 3, 4, 6, 7, 8, 17

[22] Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. REMIND Your Neural Network to Prevent Catastrophic Forgetting. In *Computer Vision – ECCV 2020*, pages 466–483. Springer International Publishing, 2020. 3, 4, 6, 7, 17

[23] Hamed Hemati, Andrea Cossu, Antonio Carta, Julio Hurtado, Lorenzo Pellegrini, Davide Bacciu, Vincenzo Lomonaco, and Damian Borth. Class-incremental learning with repetition. In *Conference on Lifelong Learning Agents*, pages 437–455. PMLR, 2023. 2, 5, 8

[24] Hamed Hemati, Lorenzo Pellegrini, Xiaotian Duan, Zixuan Zhao, Fangfang Xia, Marc Masana, Benedikt Tscheschner, Eduardo Veas, Yuxiang Zheng, Shiji Zhao, et al. Continual learning in the presence of repetition. *Neural Networks*, 183:106920, 2025. 2

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018. 2

[26] Francisco Herrera, Francisco Charte, Antonio J. Rivera, and María J. Del Jesus. *Multilabel Classification*. Springer International Publishing, 2016. 2, 3

[27] Timm Hess, Martin Mundt, Iuliia Pliushch, and Visvanathan Ramesh. A procedural world generation framework for systematic evaluation of continual learning. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2, 5

[28] Geoffrey E. Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop in Conjunction with NIPS*, 2014. 3, 4

[29] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via

rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 7

[30] Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018. 15

[31] Hanbin Hu, Chen He, and Peng Li. Semi-supervised wafer map pattern recognition using domain-specific data augmentation and contrastive learning. In *IEEE ITC*, pages 113–122. IEEE, 2021. 1

[32] M. Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An Efficient Domain-Incremental Learning Approach to Drive in All Weather Conditions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3000–3010, 2022. 3

[33] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetful learning for domain expansion in deep neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 3358–3365. AAAI Press, 2018. 3

[34] Herve Jégou, Matthijs Douze, and Cordelia Schmid. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. 3

[35] Bo Ke, Yunquan Zhu, Mengtian Li, Xiujun Shu, Ruizhi Qiao, and Bo Ren. Hyperspherical Learning in Multi-Label Classification. In *Computer Vision – ECCV 2022*, pages 38–55. Springer Nature Switzerland, 2022. 4, 8

[36] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 3390–3398. AAAI Press, 2018. 14

[37] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3, 5

[38] Takumi Kobayashi. Two-way multi-label loss. In *CVPR*, pages 7476–7485, 2023. 3, 15, 17

[39] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General Multi-label Image Classification with Transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16473–16483. IEEE, 2021. 5

[40] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 40(12):2935–2947, 2017. 3, 6, 7, 15, 16, 17

[41] Qi Liang, Jian Zhou, and Yonglin Wang. Masked autoencoder with dynamic multi-loss adaptation mechanism for few shot wafer map pattern recognition. *Engineering Applications of Artificial Intell.*, 137:109070, 2024. 1

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014. 4, 6, 7, 12, 13, 14

[43] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2Label: A Simple Transformer Way to Multi-Label Classification, 2021. 4, 5, 13, 17

[44] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR, 2017. 2

[45] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 5, 15

[46] Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *CVPR*, pages 3589–3599, 2021. 4

[47] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE TPAMI*, 45(5):5513–5533, 2022. 2, 3, 4, 5, 6, 7, 13, 14, 16, 17

[48] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, pages 109–165. Elsevier, 1989. 2

[49] Mark McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random Projections and Pre-trained Models for Continual Learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 3

[50] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The Stability-Plasticity Dilemma: Investigating the Continuum from Catastrophic Forgetting to Age-Limited Learning Effects. *Frontiers in psychology*, 4:504, 2013. 7

[51] Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized Class Incremental Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 970–974. IEEE, 2020. 8

[52] Saleh Momeni, Sahisnu Mazumder, and Bing Liu. Continual learning using a kernel-based method over foundation models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pages 19528–19536. AAAI Press, 2025. 3

[53] Manjunath Mulimani and Annamaria Mesaros. Class-incremental learning for multi-label audio classification. In *ICASSP*, pages 916–920. IEEE, 2024. 2

[54] Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *WACV*, pages 3911–3920, 2023. 3, 4, 7, 8

[55] Ameya Prabhu, Philip H. S. Torr, and Puneet K. Dokania. GDumb: A Simple Approach that Questions Our Progress

in Continual Learning. In *Computer Vision – ECCV 2020*, pages 524–540. Springer International Publishing, 2020. 5

[56] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 3, 4

[57] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91, 2021. 15, 17

[58] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben, Baruch Gilad Sharir, and Itamar Friedman. TResNet: High Performance GPU-Dedicated Architecture. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1399–1408, 2021. 4

[59] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 32–41, 2023. 5

[60] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. ML-Decoder: Scalable and Versatile Classification Head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–41, 2023. 3, 4

[61] Stefan Schrunner. Pattern recognition in snalog wafer test data. *Ph. D. dissertation, Graz University of Technology, Graz, Austria*, 2019. 1, 8

[62] Wuxuan Shi and Mang Ye. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *ICCV*, pages 1772–1781, 2023. 3, 13

[63] Xiang Song, Kuang Shu, Songlin Dong, Jie Cheng, Xing Wei, and Yihong Gong. Overcoming catastrophic forgetting for multi-label class-incremental learning. In *WACV*, pages 2389–2398, 2024. 2, 8

[64] John Sweller. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2):257–285, 1988. 2

[65] Piotr Szymański and Tomasz Kajdanowicz. A Network Perspective on Stratification of Multi-Label Data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 22–35. PMLR, 2017. 8

[66] Benedikt Tscheschner, Eduardo Veas, and Marc Masana. Incremental learning with repetition via pseudo-feature projection. *Computer Vision Winter Workshop*, 2025. 2, 5

[67] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intell.*, 4(12):1185–1197, 2022. 2, 5

[68] Eli Verwimp, Rahaf Aljundi, Shai Ben-David, Matthias Bethge, Andrea Cossu, Alexander Gepperth, Tyler L Hayes, Eyke Hüllermeier, Christopher Kanan, Dhireesha Kudithipudi, et al. Continual learning: Applications and the road forward. *Transactions on Machine Learning Research*, 2024. 3

[69] Junliang Wang, Chuqiao Xu, Zhengliang Yang, Jie Zhang, and Xiaoou Li. Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition. *IEEE TSM*, 33(4):587–596, 2020. 1, 4, 5, 6, 7, 8, 12

[70] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE TPAMI*, 46(8):5362–5383, 2024. 2, 3

[71] Lei Wang, Yibing Zhan, Leilei Ma, Dapeng Tao, Liang Ding, and Chen Gong. SpliceMix: A Cross-Scale and Semantic Blending Augmentation Strategy for Multi-Label Image Classification. *IEEE Transactions on Multimedia*, 27:3251–3265, 2025. 8

[72] Zihu Wang, Hanbin Hu, Chen He, and Peng Li. Recognizing wafer map patterns using semi-supervised contrastive learning with optimized latent representation learning and data augmentation. In *IEEE ITC*, pages 141–150. IEEE, 2023. 1

[73] Ming-Ju Wu, Jyh-Shing R Jang, and Jui-Long Chen. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE TSM*, 28(1):1–12, 2014. 1, 15

[74] Shipeng Yan, Jiangwei Xie, and Xuming He. DER: Dynamically Expandable Representation for Class Incremental Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3013–3022. IEEE, 2021. 3

[75] Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Multi-label knowledge distillation. In *ICCV*, pages 17271–17280, 2023. 2, 8

[76] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, pages 3987–3995. PMLR, 2017. 3, 15

[77] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, 2018. 8

[78] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *CVPR*, 2020. 7

[79] Zeyun Zhao, Jia Wang, Qian Tao, Andong Li, and Yiyang Chen. An unknown wafer surface defect detection approach based on incremental learning for reliability analysis. *Reliability Engineering & System Safety*, 244:109966, 2024. 1, 8, 15

[80] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE TPAMI*, 2024. 2

[81] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *NeurIPS*, 34:14306–14318, 2021. 4

[82] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *CVPR*, pages 5871–5880, 2021. 3, 4, 6, 7, 13, 15, 17

[83] Huiping Zhuang, Yizhu Chen, Di Fang, Run He, Kai Tong, Hongxin Wei, Ziqian Zeng, and Cen Chen. GACL: Exemplar-Free Generalized Analytic Continual Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hradec, Czech Republic, February 9–12, 2026

Instantaneous Monocular Camera Tilt Stabilization for Autonomous Student Formula

Erik Doležal and Jan Čech
Faculty of Electrical Engineering
Czech Technical University in Prague
{dolezeri,cechj}@fel.cvut.cz

Abstract

A novel method for camera tilt stabilization using two parallel homographies is proposed. The method is evaluated in the context of autonomous student formula racing, where the vehicle visually localizes traffic cones that delineate the race track. The mapping between the camera image and the ground plane is modeled by a planar homography, as the ground plane is flat. This homography is determined offline through calibration. However, when the vehicle drives dynamically, the camera undergoes tilt (due to roll and pitch motions), which introduces distortion. The proposed method estimates the instantaneous camera tilt from the discrepancy between homographies—the ground plane (spanning cone bottoms) and a parallel plane (spanning cone tops)—and provides undistorted measurements of the traffic cones. Simulations and real-world experiments demonstrate that the method is accurate, yields significant benefits for the autonomous pipeline, and enables reliable mapping, localization, and navigation. Moreover, the method is lightweight and operates at approximately 300 Hz.

1. Introduction

Autonomous driving in the context of Student Formula competitions presents a unique set of challenges. The race track is typically flat and bounded by traffic cones that serve as the primary landmarks for navigation. Reliable perception of these ground-plane objects is essential for accurate localization and subsequent trajectory planning and control. A common approach is to employ planar homography to map image coordinates into the ground plane. The homography mapping is determined offline during the calibration, see Fig. 1. However, this mapping is highly sensitive to camera tilt (roll and pitch), which can vary dynamically due to vehicle motion, suspension travel, and uneven track conditions. Even small deviations in orientation can lead to significant localization errors, undermining the stability of autonomous navigation.

To address this issue, we propose a monocular stabilization method that exploits the known setup of Student

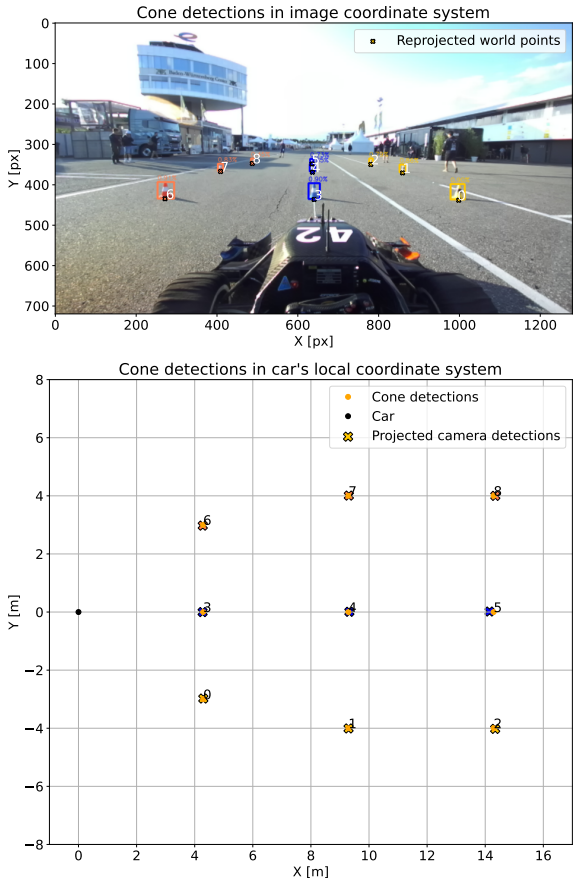


Figure 1. Homography maps cone detections on the ground plane into the car frame and is calibrated offline using cones at known positions. Our method corrects distortions in the homography that arise from camera tilt during dynamic vehicle motion.

Formula tracks. Specifically, the cones placed along the track are objects of known height, and their appearance in the image can be detected using a lightweight bounding-box detector. By leveraging these detections, our method estimates the instantaneous camera tilt and corrects the homography at the frame level. This approach avoids the need for temporal tracking, thereby eliminating error ac-

cumulation over time, and requires only a basic object detector without reliance on complex segmentation or depth estimation models.

The advantages of the proposed method are threefold. First, it provides real-time stabilization of the homography, ensuring consistent localization of cones on the ground plane. Second, it is computationally efficient, relying solely on monocular vision and simple bounding-box detections, making it suitable for embedded deployment in autonomous race cars. Third, it directly addresses the practical problem of camera tilt during dynamic driving, offering a solution that is both lightweight and robust.

The contributions of this paper are as follows:

- We introduce a monocular, vision-only method for estimating camera tilt (roll and pitch) using objects of known height.
- We demonstrate how this tilt estimation is used to stabilize planar homography, improving localization accuracy for ground-plane objects.
- We show that the proposed method is not merely a theoretical concept but has been implemented and deployed in a Student Formula autonomous vehicle, where it significantly enhances measurement accuracy in real-world racing conditions.
- Our approach achieves super real-time performance without requiring large-scale models such as monocular depth networks, making it practical for resource-constrained racing platforms.

Through this work, we aim to bridge the gap between theoretical homography stabilization and its practical application in autonomous racing, providing a method that is both simple and effective in its deployment.

2. Related Work

The tilt of the camera with respect to the ground plane is a well-known challenge in automotive research, as even small deviations in roll or pitch can significantly affect perception and localization accuracy.

One line of work relies on vehicle-mounted sensors. Suspension measurements such as damper extension have been used to infer body roll and pitch, exploiting the rigid connection between the camera and the vehicle chassis [1, 18]. A drawback of such an approach is that the damper extensions do not capture the tilt perfectly due to tire compression, which is quite significant for the autonomous formula. To maximize grip, only low tire pressure is used, such as 0.6 bar. More advanced systems employ lidar to reconstruct the road surface and estimate the camera orientation relative to the ground plane [5]. Similarly, stereo vision setups [6, 9] can recover depth and road profile, while recent methods based on monocular depth estimation (monodepth) leverage large neural models to infer scene geometry and camera tilt from single images [3, 8, 14]. However, while impressive results have been reported recently, monodepths require running heavy neural models [4, 15, 17]. Moreover, metric depth is still not reliable for unstructured outdoor scenes [13].

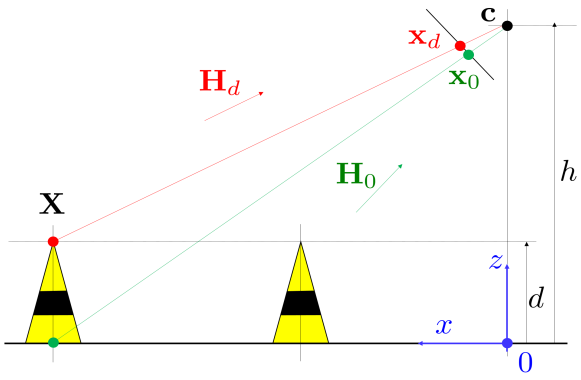


Figure 2. Overview of the setup.

A classical vision-based approach is to estimate the Essential matrix between consecutive frames through point tracking. By decomposing the epipolar geometry, the camera motion parameters can be recovered, including orientation [10]. For structured road scenes, orientation can also be inferred from vanishing points derived from lane markings or motion cues [2, 11].

In the context of the autonomus student formula, an orthogonal approach [7] extends the object detector to localize landmarks on objects of known geometry, *i.e.* traffic cones. By associating 2D image detections with a 3D model of the object, camera pose can be estimated using the Perspective-n-Point (PnP) algorithm. This approach provides direct orientation estimates, but typically requires precise landmark localization.

In contrast to these methods, our work proposes a lightweight monocular technique that leverages only bounding-box detections of cones with known height. By estimating camera tilt at the frame level and correcting the homography accordingly, we achieve stabilized localization of ground-plane objects without the need for additional sensors, complex depth models, or temporal tracking.

3. Method

The core idea of our approach is to exploit the geometric structure of the Student Formula track, where traffic cones define two parallel planes in the scene: the ground plane at $Z = 0$ and a plane at the cone tops $Z = d$. During calibration, we establish homographies that map the camera image into each of these planes, ensuring that detections of cone bottoms and cone tops correspond to the same scene point in the car frame. However, when the vehicle undergoes dynamic maneuvers, the camera orientation with respect to the ground plane changes, and the calibration homographies become inaccurate, leading to distortions in localization. Our method addresses this by parametrizing the camera motion and optimizing these parameters to minimize the offset between cone bottom and top projections. This yields an estimate of the instantaneous camera tilt and enables correction of the homography, resulting in

stabilized and precise localization of cones on the ground plane in the car frame.

Formally, let $\mathbf{P} = \mathbf{K}\mathbf{R}[\mathbf{I}, -\mathbf{c}]$ be the camera projection matrix that maps 3D scene points of the car frame to the image. The car frame is chosen such that x-axis points forward, z-axis is the normal of the grund plane and $\mathbf{c} = [0, 0, h]^T$ is the camera center, see Fig. 2.

The homography that maps scene points on the plane $Z = d$ to the camera image is

$$\tilde{\mathbf{x}}_d = \mathbf{H}_d \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \quad \mathbf{H}_d = \mathbf{K}\mathbf{R} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d-h \end{bmatrix}, \quad (1)$$

where $\tilde{\mathbf{x}}_d = [x, y, 1]^T$ is camera image point in homogeneous coordinates. And the inverse mapping is

$$\tilde{\mathbf{X}}_d = \mathbf{H}_d^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad \mathbf{H}_d^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{d-h} \end{bmatrix} \mathbf{R}^T\mathbf{K}^{-1}. \quad (2)$$

After an offline calibration, we have two homographies $\mathbf{H}_d^{-1}$ and $\mathbf{H}_0^{-1}$ (for $d = 0$), which maps the top and the bottom detections, respectively, into the same scene point

$$(\mathbf{H}_d^{-1}\tilde{\mathbf{x}}_d)_E = (\mathbf{H}_0^{-1}\tilde{\mathbf{x}}_0)_E = [X, Y]^T. \quad (3)$$

The homographies are estimated from known cone positions in the car frame and their corresponding detection in the image. The notation $(\tilde{\mathbf{X}})_E = ([\lambda X, \lambda Y, \lambda]^T)_E = [X, Y]^T = \mathbf{X}$ denotes conversion to Euclidean coordinates.

When the vehicle performs dynamic maneuvers, the camera is rotated with respect to the ground-plane. This is modelled by a 2-DOF camera projection matrix

$$\mathbf{P}'(\theta, \phi) = \mathbf{K}\mathbf{R}\mathbf{R}'(\theta, \phi)^T[\mathbf{I}, -\mathbf{c}'(\theta, \phi)]. \quad (4)$$

The rotation is expressed in angle-axis representation as $\theta\,(-\sin\phi, \cos\phi, 0)$, which captures tilt motion of the vehicle. Dynamic rotation matrix $\mathbf{R}'$ is given by Rodrigues' formula. The camera center becomes

$$\mathbf{c}'(\theta, \phi) = \mathbf{R}'(\theta, \phi)\mathbf{c} = \begin{bmatrix} h\sin\theta\cos\phi \\ h\sin\theta\sin\phi \\ h\cos\theta \end{bmatrix}. \quad (5)$$

The homography induced by the new camera matrix $\mathbf{P}'$ is

$$\mathbf{H}'_d(\theta, \phi) = \mathbf{K}\mathbf{R}\mathbf{R}'(\theta, \phi)^T[\mathbf{i}_1,\ \mathbf{i}_2,\ d\mathbf{i}_3 - \mathbf{c}'(\theta, \phi)], \quad (6)$$

where $\mathbf{i}_i$ are standard basis vectors, so the identity matrix is $\mathbf{I} = [\mathbf{i}_1,\ \mathbf{i}_2, \mathbf{i}_3]$.

Now, we introduce composed homography mapping

$$\mathbf{G}_d(\theta, \phi) = \mathbf{H}_d^{-1}\mathbf{H}'_d(\theta, \phi), \quad (7)$$

that maps distorted to correct scene points. By substituting Eq. (2) and Eq. (6), we get

$$\mathbf{G}_d(\theta, \phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{d-h} \end{bmatrix} \mathbf{R}'(\theta, \phi)^T \begin{bmatrix} 1 & 0 & h\sin\theta\cos\phi \\ 0 & 1 & h\sin\theta\sin\phi \\ 0 & 0 & d - h\cos\theta \end{bmatrix}. \quad (8)$$

which does not depend on either intrinsic camera matrix $\mathbf{K}$ or static rotation $\mathbf{R}$.

In case of dynamic motion, *i.e.* non-zero θ, ϕ, using the original static homographies will result in a discrepancy between projections of top and bottom detections (center of the top and bottom sides of the bounding boxes, respectively) into the scene. The projections are corrected by Eq. (8), so the error is

$$\mathbf{e}(\theta,\phi) = \big(\mathbf{G}_d(\theta, \phi)\mathbf{H}_d^{-1}\tilde{\mathbf{x}}_d\big)_E - \big(\mathbf{G}_0(\theta, \phi)\mathbf{H}_0^{-1}\tilde{\mathbf{x}}_0\big)_E. \quad (9)$$

Minimization of $\sum_i \mathbf{e}_i^T\mathbf{e}_i$ is not the best option, since the localization accuracy of individual cones depends on their position, *e.g.* more distant cones are less accurate. Therefore, we introduce weights in the least squares problem. Let $\mathbf{J}_i = \frac{d(\mathbf{H}_0^{-1}\tilde{\mathbf{x}})_E}{d\mathbf{x}}$ be the 2×2 Jacobian of the inverse homography mapping. Then $\mathbf{W}_i = (\mathbf{J}_i\mathbf{J}_i^T)^{-1}$. Note that term $\mathbf{J}_i\mathbf{J}_i^T$ is the unit covariance matrix warped by the homography mapping. See an example in Fig. 3. Additionally, the Cauchy loss is used for increased robustness and better outlier rejection. Finally, the optimization problem is the following:

$$\min_{\theta,\phi} \sum_{i=1}^{N} \ln\big(1 + \mathbf{e}_i(\theta, \phi)^T\mathbf{W}_i\mathbf{e}_i(\theta, \phi)\big). \quad (10)$$

The minimum is found by the Trust Region Reflective algorithm[1], where the initialization is zero for the first frame, and the subsequent frames are initialized by the previous solution to support continuity. The results are clipped by the end of each iteration to avoid sudden parameter explosion with poor input data conditioning. The motion parameters θ, ϕ determine the instantaneous tilt. Finally, the cone positions in the car frame is found as

$$\mathbf{X} = \big(\mathbf{G}_0(\theta, \phi)\mathbf{H}_0^{-1}\tilde{\mathbf{x}}_0\big)_E. \quad (11)$$

3.1. Runtime details

Dominant rotational distortions in driving datasets allow the 2-DoF model to converge rapidly, averaging 11 iterations per frame. On an Apple M4 Pro CPU, the complete stabilization process, including point selection and final correction, executes in 3.3 ms (300 Hz). This performance confirms that the method is computationally efficient and suitable for real-time applications without introducing significant latency.

4. Experiments

To validate the proposed method, we conducted both simulations and real-world experiments under static and dynamic scenarios.

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.least_squares.html

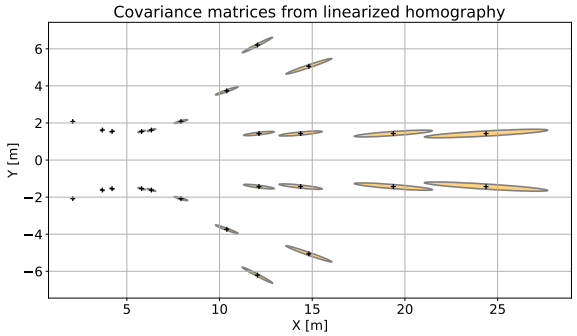


Figure 3. Isotropic Gaussian distribution of camera detections (having unit covariance) are projected into the car frame using a linearized homography. The car is located at point $(0, 0)$. The resulting covariances increase with distance.

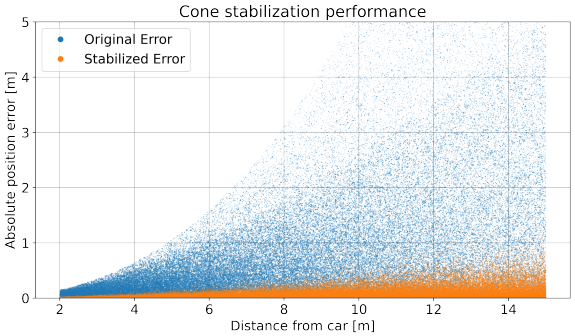


Figure 4. Result of a simulation epoch for 10000 trials with image pixel noise at level $\sigma = 0.3$ comparing the original and stabilized projection errors.

4.1. Simulations

The simulated scenarios provide the most reliable basis for comparison against the known ground truth, and they further enable an examination of the method's sensitivity to noise.

The simulation setup replicates the behavior observed during tilt distortion when driving on a flat plane. Accordingly, all cone bases lie on the ground at zero height $d = 0$, while the cone tops are consistently placed at $d = 0.1515$ m. Two static homographies are computed according to Eq. (2), with the rotation matrix configured such that the camera is oriented forward along the world x-axis.

The ground-truth cone positions are uniformly sampled over a predefined range corresponding to the distances observed during driving and field of view of the used camera. For the experiment ten cones per sample were chosen, as it represents the usual number of observed cones during driving. The distortion transformation described in Eq. (6) is then applied to obtain the image-space positions of these points, with parameters uniformly distributed in range $\theta \in [-2, 2]$ deg and $\phi \in [-180, 180)$ deg. Gaussian noise is added to the image coordinates to approximate the inaccuracy of cone detections, that are found by YOLO detector in practice[2]. The stabilization method described above is then executed on 10,000 random trials for each noise level, ranging from 0 to 1 px in increments of 0.1 px.

For low noise levels, the stabilizer performs very well, consistently reducing distortion-induced errors across the entire distance range as can bee seen in Fig. 4. As the noise level increases, the sensitivity of the homography becomes more apparent, as seen in Fig. 5. The improvement in positional accuracy becomes less pronounced, although the error is still reduced by approximately half. This can also be said about the distance relation of the error in Fig. 6, where it increases for the more distant points even for low noise, but such behavior is expected, as it quite well follows projected covariances in Fig. 3.

[2] YOLOv8 [12] was trained on a public dataset of of traffic cones with tightly annotated bounding boxes [16].

In low-noise scenarios, the optimized parameters closely match the simulated rotation. However, their accuracy degrades significantly for noise levels exceeding 0.4 px. Although the applied noise is zero-mean, the number of points available for stabilization is insufficient to fully eliminate bias introduced by the noise. Such bias may lead to an estimated rotation that differs substantially from the true distortion. Nevertheless, the stabilization procedure continues to operate as intended, resulting in only a modest increase in the final position error.

4.2. Static scenario

In the first set of real-world experiments, the accuracy of the stabilization method was evaluated using known ground-truth cone positions. The camera was initially calibrated with zero tilt, as shown in Fig. 1. To introduce tilt-induced errors, the vehicle was subsequently lifted using individual wheels or entire axles, generating pure pitch, pure roll, and combined pitch–roll conditions. The resulting accuracy of the stabilization algorithm is illustrated in Fig. 7. The corresponding distance error of the stabilized cone positions is presented in Fig. 8. In contrast to the original projected positions, the stabilized error remains approximately constant with respect to distance from the vehicle, improving the mean absolute error from 1.00 ± 1.33 m to only 0.15 ± 0.12 m. The residual error is likely attributable to neural-network detection noise as well as possible ground-truth measurement inaccuracies.

4.3. SLAM

The primary objective of the stabilization procedure is to improve the autonomous pipeline, given that vision constitutes the key perceptual input and plays a decisive role in overall performance. The component most influenced is the SLAM algorithm, and the subsequent effects on the pipeline stem from it.

To evaluate the impact of stabilization on SLAM performance, an ICP-based SLAM algorithm was employed, with odometry used solely for prediction during data association or to compensate for motion in scenarios where association fails. As the approach is primarily vision-based,

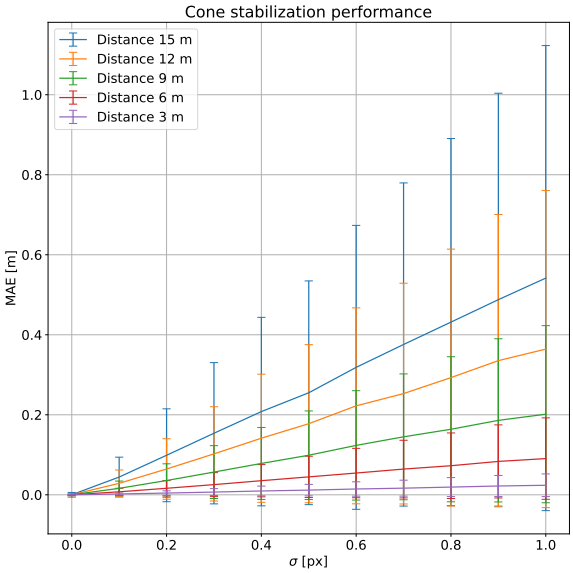


Figure 5. Stabilized position errors as a function of the image pixel noise σ for multiple distance bands with radius 1 m around the mean, resulting from the simulated scenarios.

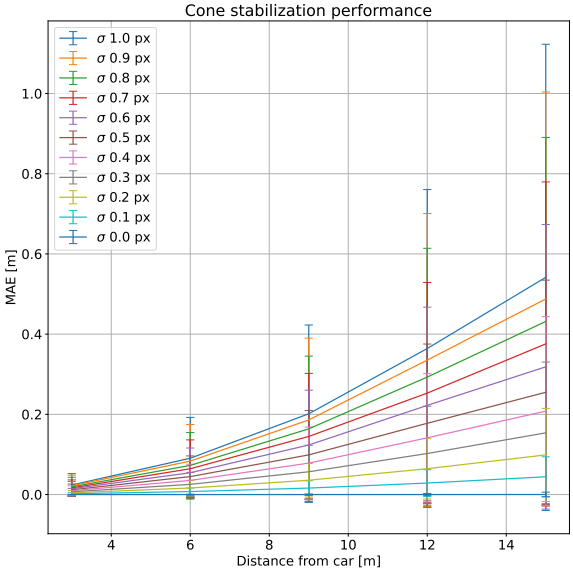


Figure 6. Stabilized position errors as a function of the distance from the car for multiple noise levels, resulting from the simulated scenarios.

it effectively highlights how inaccuracies in cone position estimates can destabilize the SLAM solution and subsequently affect the remainder of the perception and localization pipeline. For the experiment, the algorithm was configured to operate as a pure localizer, thereby avoiding failures caused by incorrectly constructed maps.

During the run, the car achieved speeds up to 20 m/s on the straights with 1 G lateral acceleration during cornering

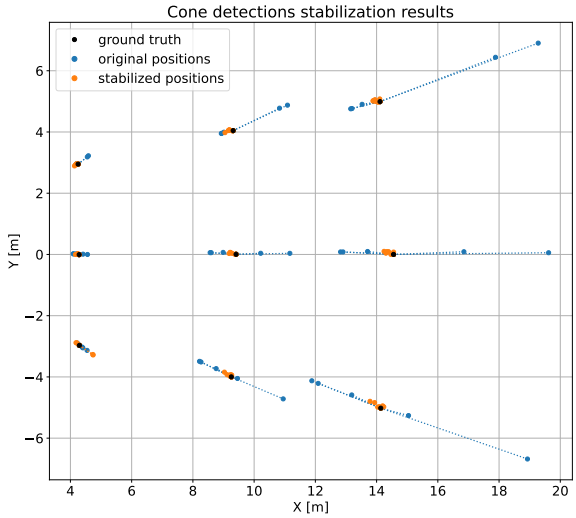


Figure 7. Static tilt test with a real vehicle. Comparison between original (unstabilized) and stabilized cone localizations.

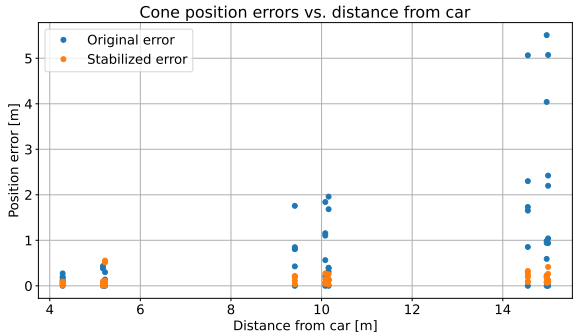


Figure 8. Static tilt test with a real vehicle. Comparison of distance error between original (unstabilized) and stabilized cone localizations.

and 1 G peak longitudinal acceleration when braking, the tilt of the car was up to 2 deg in relation to the gravitational vector. Note that some of the observed distortions in this run originated from the bumpiness of the track.

The resulting position-tracking performance is shown in Fig. 9. When using stabilized detections, the SLAM system successfully completed the entire course, ten laps, while maintaining accurate localization without significant position spikes. In contrast, the SLAM system operating on unstabilized cone detections exhibits noticeable spikes in the estimated vehicle position. These spikes lead to poor localization, wrongly placing the vehicle off the track despite it remaining on course for most of the experiment. The system eventually became lost on the lower straight section and only intermittently reconnected to the track through odometry updates and later lost the position information completely. This behavior is attributed to data association aliasing caused by distortion at the beginning of the straight. If such localization were used directly for

vehicle control, these errors would likely cause the vehicle to leave the track rapidly, with limited ability to recover, or induce oscillatory behavior due to noise, leading to a similar outcome.

SLAM can serve to evaluate the localization accuracy of cones. Given the known map positions and observed cone detections at each iteration, the accuracy of the stabilization method can be directly assessed. One limitation of this evaluation is the slightly curved surface of the track; therefore, only cones located within 15 m in the x-direction and 12 m in the y-direction were considered to preserve the assumption of local surface flatness.

Figure 10 reveals a pronounced tendency for unstabilized detections to disperse significantly in cornering sections of the track. In contrast, the stabilized detections are considerably more concentrated around the SLAM map positions used for subsequent error evaluation. A small number of outliers remain, which can be attributed to abrupt changes in track surface curvature that violate the local flat-plane assumption.

The overall absolute distance error of stabilized and unstabilized cone positions relative to the SLAM map is shown in Fig. 11. The results clearly show a significant reduction in error when stabilization is applied, aligning with the patterns observed in the simulated experiments. The slightly higher error at shorter distances is probably caused by differences in real-world noise characteristics compared to the simulated noise model, as well as increased perspective skew of cones in the image, which affects the accuracy of the selected image points.

Quantitatively, the mean absolute error of the unstabilized cone positions is 0.913 ± 1.165 meters. After applying stabilization, this error is reduced to 0.362 ± 0.418 meters, representing an overall error reduction by factor of 2.5.

4.4. Validation of estimated parameters

While driving on a flat horizontal plane, the estimated rotation is expected to correspond to the rotation measured by the vehicle's Inertial Navigation System (INS) unit. Accordingly, this experiment was conducted on a high-speed track featuring strong braking zones and sharp corners in order to induce significant vehicle tilt. The resulting roll and pitch rotations were measured by the SBG Ellipse-D INS unit and independently estimated by the stabilization algorithm.

To enable a direct comparison, both rotation estimates were first converted into rotation matrices and subsequently transformed into an angle-axis representation using the inverse Rodrigues transform. The parameter θ therefore corresponds to the rotation magnitude. The angle ϕ is computed as the $\mathrm{atan2}$ of the first two components of the rotation vector and is subsequently normalized to the interval $[0, 2\pi)$ for better continuity in the plot.

The comparison of the two rotation estimates is shown in Fig. 12. In most cases, the stabilization algorithm produces estimates θ that closely match the INS measurements. However, this correlation is less consistent for ϕ, as this parameter is considerably more sensitive to noise. Consequently, the estimated ϕ does not always align with the corresponding INS value. We are aware, the INS was not precisely set up and calibrated (aligned with the camera).

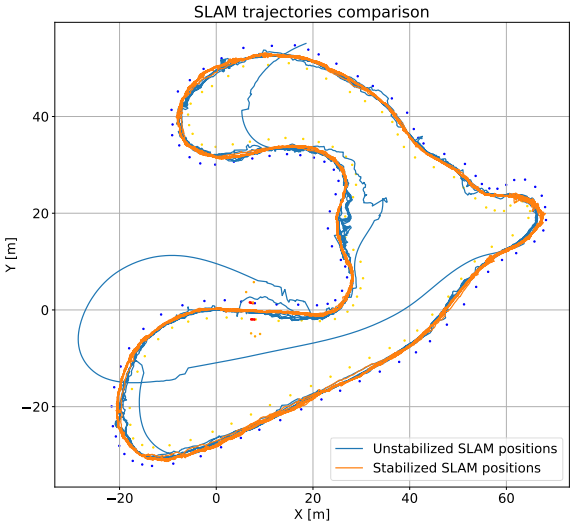


Figure 9. Comparison of the trajectories of ICP-based SLAM with and without stabilization during a ten lap run on the Formula Student Germany 2025 trackdrive track. The coloured dots around the trajectories represent SLAM-estimated boundaries of the track.

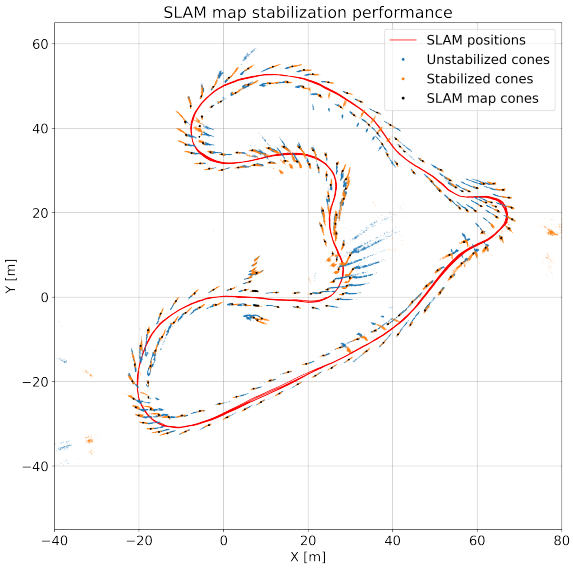


Figure 10. Instantaneous local estimates of cones projected into the global frame, given by SLAM map, together with car positions throughout the race.

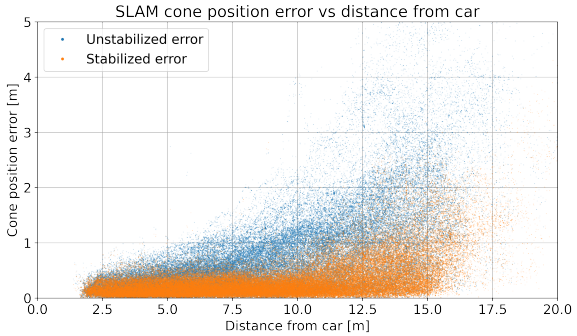


Figure 11. Absolute distance error of stabilized and unstabilized cone positions from the SLAM-map cones in local car frame in relation with distance from the car.

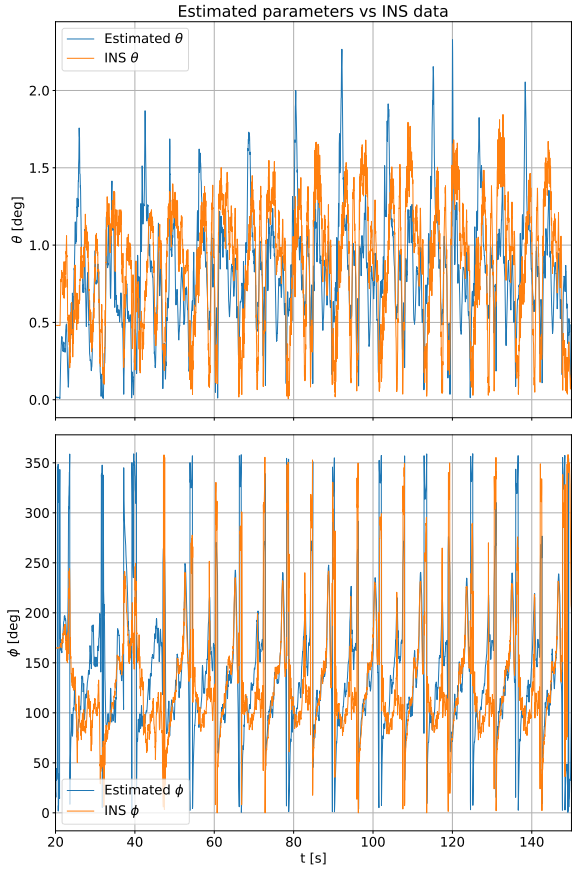


Figure 12. Comparison between the estimated camera rotation by the proposed stabilization method and inertial measurement by SBG Ellipse-D INS unit.

4.5. Ablation study

The SLAM experiment provides a strong basis for validating the choice of loss function used in the optimization problem, as real-world driving encompasses a wide range of scenarios that could potentially challenge the algorithm. Figure 13 compares the unstabilized case with

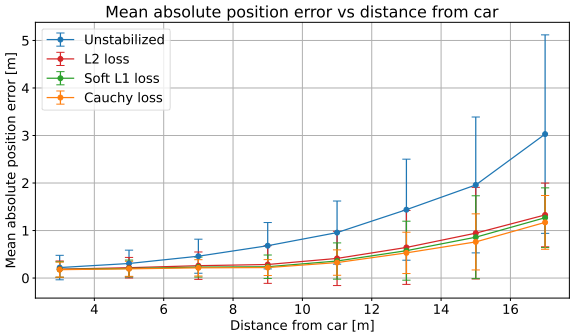


Figure 13. Comparison of mean absolute error compared to SLAM cones for multiple loss functions – Cauchy, soft L1 and L2 loss.

the stabilization by the selected Cauchy loss, L2 and soft L1 loss functions. The Cauchy loss slightly outperforms the other two approaches across all distances in terms of mean absolute error, while also exhibiting the smallest standard deviation. The L2 loss performs worst due to its poor robustness to outliers, whereas the soft L1 loss represents a compromise between robustness and sensitivity. The mean absolute error aggregated over all distances for each loss function is reported in Table 1.

Loss	MAE [m]
Unstabilized	0.913 ± 1.165
L2	0.437 ± 0.628
Soft L1	0.393 ± 0.528
Cauchy	$\mathbf{0.362 \pm 0.418}$

Table 1. Mean of the measured absolute errors with standard deviations for tested loss functions.

Subsequently, we assess the importance of the weighting in Eq. (10). This experiment was conducted using the same dataset as the SLAM experiments. As shown in Fig. 14, the stabilization without weighting still reduces the error compared to the original cone positions. With uniform weights, the resulting mean absolute error is 0.563 ± 0.779 m. Compared with the values reported in Table 1, it is evident that the inclusion of weighting has a more significant impact on overall performance than the specific choice of loss function.

5. Conclusions

We presented a monocular stabilization method for planar homography that addresses the critical issue of camera tilt in autonomous Student Formula racing. By exploiting the geometric relationship between cone bottoms on the ground plane and cone tops on a parallel plane, the proposed method estimates instantaneous roll and pitch and corrects distortions in real time. Unlike sensor-based or heavy neural depth models, our method relies solely on lightweight bounding-box detections, making it computationally efficient and suitable for embedded deployment.

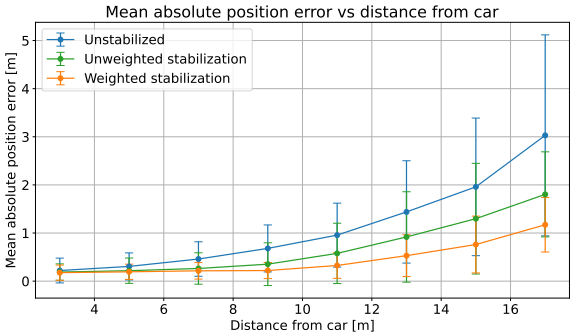


Figure 14. Comparison of the mean absolute error of the estimated cone positions with unstabilized, stabilized without and with the weighting.

Extensive simulations and real-world experiments demonstrated that the proposed technique consistently reduces localization errors, stabilizes cone detections across varying noise levels and tilt conditions, and significantly improves downstream modules such as SLAM. The method achieved super real-time performance at approximately 300 Hz, ensuring that stabilization does not become a bottleneck in the perception pipeline. Validation against INS measurements confirmed the accuracy of the estimated tilt parameters, while SLAM experiments highlighted the substantial benefits for reliable mapping and navigation.

A possible limitation of our approach is the assumption of a flat ground plane. This assumption holds across all races in which the formula vehicle competed, and minor deviations from planarity have a negligible impact compared to the errors introduced by ignoring camera tilt during dynamic driving. Another limitation arises from the motion model assumptions. We assume pure rotation, whereas in practice the camera motion can be more complex: the camera center may be slightly displaced due to suspension non-rigidity or when the vehicle traverses a sudden bump. We experimented with an extended model that incorporated small translations and displacement of the rotation axis; however, this more complex model proved less accurate overall. Increasing the model flexibility by even a single degree of freedom for the shift tended to overfit the noisy data.

Overall, this work is not only a theoretical concept, but it has been practically deployed in an autonomous racing vehicle. The proposed approach is simple, robust in dynamic driving scenarios, and impactful for the broader autonomy stack.

Acknowledgements

This research was supported by the CTU Student Grant SGS23/173/OHK3/3T/13 and by the project for promoting and supporting studies at FEE CTU in Prague "EFORCE driverless".

References

[1] Anton Albinsson and Christoffer Routledge. The damper levels influence on vehicle roll, pitch, bounce and cornering behaviour of passenger vehicles-A study in cooperation with Volvo Car Corporation. Master's thesis, Chalmers University of Technology, 2013. 2

[2] Mohamed Aly. Real time detection of lane markers in urban streets. In *IEEE intelligent vehicles symposium*, pages 7–12, 2008. 2

[3] Vasileios Arampatzakis, George Pavlidis, Nikolaos Mitianoudis, and Nikos Papamarkos. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2396–2414, 2023. 2

[4] Aleksei Bochkovskii, Amael Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 2

[5] Carolin Bösch, Fabian Arzberger, and Andreas Nüchter. Real-time lidar based calculation of the ground plane's normal vector on spherical mobile mapping systems. In *European Conference on Mobile Robots (ECMR)*, pages 1–7. IEEE, 2025. 2

[6] Nikolay Chumerin and Marc Van Hulle. Ground plane estimation based on dense stereo disparity. In *International conference on Neural Networks and Artificial Intelligence, Date: 2008/05/27-2008/05/30, Location: Minsk, Belarus*, 2008. 2

[7] Ankit Dhall, Dengxin Dai, and Luc Van Gool. Real-time 3d traffic cone detection for autonomous driving. In *IEEE intelligent vehicles symposium (IV)*, pages 494–501, 2019. 2

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2

[9] Paul Herghelegiu, Adrian Burlacu, and Simona Caraiman. Robust ground plane detection and tracking in stereo sequences using camera orientation. In *2016 20th International Conference on System Theory, Control and Computing (ICSTCC)*, pages 514–519. IEEE, 2016. 2

[10] Petr Jahoda and Jan Čech. Predicting road surface anomalies by visual tracking of a preceding vehicle. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 1795–1800, 2025. 2

[11] Jinbeum Jang, Youngran Jo, Minwoo Shin, and Joonki Paik. Camera orientation estimation using motion-based vanishing point detection for advanced driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 22(10):6286–6296, 2020. 2

[12] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. https://github.com/ultralytics/ultralytics. 4

[13] Ondrej Kodejs and Jan Cech. Evaluation of monocular depth predictors. Master's thesis, Czech Technical University in Prague, 2025. 2

[14] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–96, 2014. 2

[15] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 2

[16] Niclas Vödisch, David Dodel, and Michael Schötz. Fsoco: The formula student objects in context dataset. *SAE International Journal of Connected and Automated Vehicles*, 5 (12-05-01-0003), 2022. 4

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2

[18] Tianyi Zeng, Tianyi Wang, Zimo Zeng, Feiyang Zhang, Jiseop Byeon, Yujin Wang, Yajie Zou, Yangyang Wang, Junfeng Jiao, Christian Claudel, et al. Damper-b-pinn: Damper characteristics-based bayesian physics-informed neural network for vehicle state estimation. *arXiv preprint arXiv:2502.20772*, 2025. 2

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hradec, Czech Republic, February 9–12, 2026

SAM-pose2seg: Pose-Guided Human Instance Segmentation in Crowds

Constantin Kolomiiets Miroslav Purkrabek Jiri Matas

Visual Recognition Group
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
kolomcon@fel.cvut.cz

Input image SAM 2.1 SAM-pose2seg (our)

Figure 1. **SAM-pose2seg is superior to SAM 2** [21] for human instance segmentation, especially in crowded scenes. Both examples are generated from the same set of predicted keypoints from [19]. Notice the noise and incorrect masks in the middle of the crowd for SAM 2.1. Different prompts are used for SAM 2 and SAM-pose2seg since each model works best with different prompts.

Abstract

Segment Anything (SAM) provides an unprecedented foundation for human segmentation, but may struggle under occlusion, where keypoints may be partially or fully invisible. We adapt SAM 2.1 for pose-guided segmentation with minimal encoder modifications, retaining its strong generalization. Using a fine-tuning strategy called PoseMaskRefine, we incorporate pose keypoints with high visibility into the iterative correction process originally employed by SAM, yielding improved robustness and accuracy across multiple datasets. During inference, we simplify prompting by selecting only the three keypoints with the highest visibility. This strategy reduces sensitivity to common errors, such as missing body parts or misclassified clothing, and allows accurate mask prediction from as few as a single keypoint. Our results demonstrate that pose-guided fine-tuning of SAM enables effective, occlusion-aware human segmentation while preserving the generalization capabilities of the original model. The code and pretrained models will be available at the project website [1].

[1] MiraPurkrabek.github.io/BBox-Mask-Pose/

1. Introduction

A standard approach to human instance segmentation is to use a detector that predicts instance masks directly. However, in crowded scenes with heavy occlusion, these detectors often fail to separate overlapping instances. In contrast, human pose estimators are more robust in these conditions and produce structured keypoints that are easier to annotate and more stable under clutter.

Keypoint annotations are also significantly cheaper than per-pixel segmentations, making pose a practical intermediate representation for instance segmentation. In particular, human keypoints can serve as effective prompts for segmentation models. The task of *pose-guided human instance segmentation* takes either Ground-Truth or detected keypoints as input and outputs a segmentation mask for each instance.

This task was introduced with the OCHuman dataset [32] and the pose2seg model and was explored in many models [1, 2, 6, 27, 32, 33] since then. More recently, it has been used in the self-improving BMP loop (BBox-MaskPose, [19]), where pose-guided segmentation plays a key role in resolving multi-body ambiguity in heavily occluded scenes.

The Segment Anything Model (SAM, [11]) introduced a general segmentation framework trained on a large

and diverse dataset of masks and images. SAM 2 [21] extended this with large-scale video training. While SAM shows strong generalization and has revolutionized prompted segmentation, it lacks semantic understanding and is not specialized for any class, such as humans.

We build on SAM's generalization and introduce a method to adapt it for pose-guided human instance segmentation. Our model, SAM-pose2seg, incorporates semantic information into SAM through two main modifications. First, we fine-tune the decoder on human-only segmentation masks to specialize the predictions. Second, we replace SAM's random point prompts with body keypoints during training. This aligns the prompt encoder and mask decoder more closely with the target task.

SAM-pose2seg is a pose-guided variant of SAM 2.1, fine-tuned for human instance segmentation. It introduces semantic specialization through decoder fine-tuning and aligns the prompt encoder with the task by training on human keypoints. This design enables robust segmentation from both the predicted pose and the Ground-Truth. In the remainder of the paper, we show that SAM-pose2seg achieves state-of-the-art performance on pose-guided human instance segmentation benchmarks and analyze design choices through ablation studies.

2. Related Work

Detection-Based Segmentation: The most direct approach to human instance segmentation uses a detector [9, 13, 16, 17, 22, 26, 26, 34]. A detector takes an image as input and outputs instance masks and class labels. Detectors work well in scenes similar to the training data, but struggle when instances overlap heavily. Under severe occlusion, they often merge multiple people into one mask, or assign different body parts to different instances. Tuning hyperparameters (e.g., non-maximum suppression) can reduce this, but increases false positives.

General Prompted Segmentation requires an image and a human or automatically generated prompt. A prominent example is the SAM family [11, 21], trained on large image and video datasets with human-in-the-loop supervision. These models generalize well, but lack semantic understanding, making the task inherently ambiguous. As a result, they often segment skin, face, hair, clothing, or body parts instead of the full human instance.

Semantics-Aware Prompted Segmentation: Several works [8, 23, 28, 30] inject semantics into SAM. They take an image and a text prompt as input and detect and segment instances. This setting is often referred to as *open-vocabulary* or *zero-shot* segmentation. While these models capture semantics and segment entire instances, they do not provide localized cues and offer no control over which instance is chosen. Moreover, they are unsuitable for iterative loops such as BMP [19].

Pose-Aware Instance Segmentation is the most specific form of prompted segmentation. The model takes an image and a detected or Ground-Truth human pose and segments the corresponding person. These methods specialize in humans, sacrificing the generality of open-vocabulary or generic segmenters for robustness and precise, localized control. Test-time optimization methods [5, 6] refine predictions at inference without training. While training-free, they are computationally expensive and slow at test time. Standard pose-guided human segmentation methods [1, 2, 14, 27, 31–33] rely on small training datasets and thus generalize poorly, especially in crowded scenes. The closest related work is CrowdSAM [7], which builds a framework around SAM to automatically annotate bounding boxes in crowds. CrowdSAM uses SAM and DINO [18] as external tools and optimizes prompts, similar to [19]. In contrast, our SAM-pose2seg is end-to-end and, when used in an iterative loop, outperforms CrowdSAM, as shown in [19].

Datasets: There are not many datasets with annotated human instance segmentation masks and human poses. One of the first to offer this kind of data was COCO [15] and it became a standard for pose-guided segmentation training. COCO has two problems. First, it does not focus on overlapping instances, which is now the most challenging scenario (see [19]). Second, the human pose is annotated only for "large" instances, so the models are not trained or evaluated on small background persons. Overlapping instances were addressed in OCHuman dataset [32] but it is too small for training and contains only validation and testing sets. There are datasets such as [3, 10, 12, 25, 29] focusing on analysis of human body in crowded scenes, but none of these offer both segmentation and pose annotations. In this work, we worked with COCO [15] and CIHP [10] for training and OCHuman [32] for evaluation. For details about used datasets, see Sec. 4.

Iterative Methods: Pose-guided instance segmentation has been used in iterative pipelines [19, 24]. Our work builds on BMP [19], where instance segmentation is prompted by detected human keypoints. They use unmodified SAM 2 with a complex prompt selection strategy. We also build on SAM 2, but adapt it to this task. Our SAM-pose2seg is fine-tuned for human instances, removing the need for complex prompt selection. Compared to SAM 2 in BMP, SAM-pose2seg achieves better performance, with higher robustness and lower complexity.

3. Method

SAM-pose2seg builds on SAM2 [21], introducing task-specific modifications and a dedicated prompting strategy driven by pose keypoints. Our approach consists of three main components: adapting the base SAM architecture, fine-tuning with a pose-aware iterative sampling strategy, and designing an effective keypoint prompting mechanism for inference.

3.1. SAM

Segment Anything (SAM) is a powerful, robust tool with great generalization capacity. Its robustness to noise and point prompting capability offers a highly optimal foundation for our task. However, standard prompting tech-

niques often struggle in occluded scenarios, where detected keypoints lack unequivocal visibility information, making SAM unreliable.

Our goal is to adapt SAM (specifically, *SAM 2.1*) for pose-guided segmentation while retaining its ability to generalize. We introduce minimal modifications to the encoder structure to tailor the model specifically for this task.

3.2. Fine-tuning

We fine-tuned the *SAM 2.1 Hiera Base Plus* model using the official Meta training script, focusing specifically on the point selection mechanism. To preserve the generalization power of the backbone—trained on massive, diverse dataset—and to accelerate training, we froze the image encoder. Only the prompt encoder and mask decoder were optimized. No architectural changes were made to the mask decoder; however, it was also a part of the training process, as the prompt encoder's only learned parameters are the weights for positive, negative, and bounding box embeddings. We adopted the sampling strategy *PoseMaskRefine*.

Method MaskRefine (Default Training Strategy). Meta's default training procedure serves as a strong baseline, which only builds on the Ground-Truth (GT) masks. It does not rely on pre-defined semantic or pose-based points, instead, they are sampled dynamically: the first is drawn uniformly from the Ground-Truth mask, while the remaining seven are sampled from the error region between the GT mask and the previous prediction (with a small probability, sampling is done from the GT mask instead). This totals eight points per iteration. By conditioning the model on both the previous logits and the newly sampled point, this method enables iterative refinement without dependence on explicit semantic cues.

Method PoseMaskRefine (Pose-Guided Refinement). Building on the strong performance of MaskRefine, we introduce PoseMaskRefine to meet our specific task requirements. Here, the initial point is not sampled uniformly but is selected as an available pose keypoint with the highest visibility, unless no keypoint is available. In subsequent iterations, the remaining seven points are preferentially sampled from pose keypoints located within the current error region; with a small probability, sampling follows the original MaskRefine strategy and selects points uniformly from the Ground-Truth mask instead. If no such keypoints exist, sampling reverts to uniform selection from the error region. By incorporating pose cues while preserving the iterative, error-driven nature of MaskRefine, this strategy consistently outperformed the default approach (see Tab. 1).

The MaskRefine strategy effectively acts as hard negative mining; because samples are selected based on local error, boundary and occluded regions receive strong supervision. We attempted to simplify this by eliminating the iterative correction process in favor of a pure keypoint-based prediction to reflect the inference prompting (see Sec. 4.3), but this yielded no significant improvement.

Fine-tuning with PoseMaskRefine makes the model significantly less sensitive to common errors, such as segmenting clothing only or omitting body parts (see Fig. 7).

3.3. Prompting

Effective prompting is crucial for interacting with SAM, yet pose keypoints do not perfectly align with the correction-based prompting concept inherent to SAM's training. Therefore, our goal was to find the most effective way to pose prompting while sticking to the SAM's structure.

Firstly, we incorporate the ProbPose [20] *visibility* metric as a reliable criterion for keypoint selection. This metric serves as a confidence measure for each keypoint, indicating whether the respective body part is observable in the picture and enabling the exclusion of keypoints that are likely occluded or noisy.

For the base model, selecting six keypoints based on both visibility and distance proved most stable (see Fig. 3), as this provided sufficient variability to recognize the whole body without the detriment caused by exceeding eight points.

However, across all fine-tuning experiments, PoseMaskRefine consistently yielded the best generalization and enabled a significant simplification of our inference strategy. Its flexibility allows us to replace complex heuristics with a simpler approach, reducing the risk of selecting occluded points (see Fig. 8). We now select only the **3 keypoints with the highest visibility scores**. This simplified strategy improved accuracy across all observed datasets using detected keypoints.

We employ different selection methods for the base SAM 2.1 and SAM-pose2seg in figures and tables, unless stated otherwise, to compare the optimal performance of each version.

4. Experiments

4.1. Implementation Details

Data. For training, we used the COCO and CIHP datasets together with ProbPose [20] keypoints to ensure a wide range of human poses and occlusion conditions, mirroring their main use at inference time. Training on COCO alone is not sufficient due to the pronounced domain shift relative to CIHP and OCHuman, both of which emphasize multi-person and heavily occluded scenarios. We generated detected pose keypoints for all instances (including small ones that are usually ignored for pose estimation) so that our model is robust to localization noise.

Architecture and Training Details. We fine-tuned the *SAM 2.1 Hiera Base Plus* checkpoint. We trained the prompt encoder and the mask decoder only, while the image encoder (backbone)

COCO	CIHP	train selection	test selection	# points	COCO AP	OCH AP	CIHP AP
✗	✗	mV	mV	6	37.7	29.4	66.4
✗	✗	mS	mS	6	41.2	29.5	71.6
✓	✗	mV	mV	6	41.2	27.0	62.2
✓	✗	mS	mS	6	42.6	26.1	64.3
✓	✓	mS	mS	6	43.3	29.3	67.7
✓	✗	mR	mV	3	43.7	29.8	68.9
✓	✓	mR	mV	3	43.7	34.1	71.7
✓	✓	P1mR	mV	3	44.5	34.5	72.7
✓	✓	PmR	mV	3	**44.6**	**34.7**	**72.7**

Table 1. **Ablation study: fine-tuning recipes.** Keypoint selection methods are MaxVis (mV), MaxSpread (mS), MaskRefine (mR), PoseMaskRefine (PmR) and Pose1MaskRefine (P1mR). First two rows are baselines SAM 2.1 without any fine-tuning. For mV and mS methods, the optimal # points is 6, for mR-based methods it is 3. The best point method is PmR with 3 points. Training on both COCO and CIHP is crucial for generalization to unseen OCHuman.

remained frozen. The configuration file, `SAM 2.1_hiera_b+_MOSE_finetune.yaml`, was taken from the SAM 2 repository. All hyperparameters not mentioned here were left unchanged. The number of frames was set to 1. We trained for 15 epochs, though the model converged earlier; performance changes were marginal when varying epochs between 10 and 30. The probability of bounding box usage (`prob_to_use_box_input_for_train`) was set to 0.0, while the probability of point usage (`prob_to_use_pt_input_for_train`) was set to 1.0.

4.2. SOTA Comparison

The proposed *SAM-pose2seg* model represents a strong all-round solution for the pose-to-segmentation task. Across a wide range of datasets and evaluation settings, it consistently achieves competitive or superior performance. In particular, SAM-pose2seg demonstrates robust behavior under both detected and Ground-Truth pose inputs, making it well suited for practical deployment scenarios.

Using detected ProbPose [20] keypoints, we achieve **44.6 AP** on COCO val2017, **60.3 AP** on COCOPersons (COCO val2017 excluding small instances), **34.7 AP** on OCHuman test, and **72.7 AP** on CIHP val (for comparison with SOTA, see Tab. 2). These results demonstrate strong generalization across datasets with varying levels of occlusion and pose complexity, with particularly strong performance on the challenging OCHuman benchmark.

When evaluated with Ground-Truth poses, SAM-pose2seg further improves performance, reaching **61.6 AP** on COCOPersons, **70.0 AP** on OCHuman test, and **69.5 AP** on OCHuman val (for comparison with SOTA, see Tab. 3). For Ground-Truth keypoints, we apply spread-based keypoint selection, as their binary visibility annotations do not allow for visibility-based ranking.

Model (Prompting)	COCO val AP	COCOPersons val AP	OCHuman test AP
HQNet R-50	-	-	31.1
Pose2Seg	-	55.5	23.8
ExPoSeg$^+$	-	61.9	26.8
Occlusion C&P$^+$	-	-	28.3
Crowd-SAM$^+$	22.0	-	31.4
MultiPoseSeg	-	56.3	-
SAM 2.1	41.2	56.0	29.5
SAM-pose2seg	44.6	60.3	34.7

Table 2. **Evaluation of the base SAM model and SAM-pose2seg** on **detected ProbPose [20] keypoints**. The improvement is most visible on the OCHuman test set. Prompting methods are explained in Sec. 4.3.2. COCOPersons is a subset of COCO without Small category persons (no Ground-Truth poses are available there) for comparison with Pose2Seg. Models labeled with $^+$ estimate either masks or report detection AP. Every result except for SAM 2.1 was achieved on a different set of detected keypoints.

Model (Prompting)	COCOPersons val AP	OCHuman test AP	OCHuman val AP
Pose2Seg	58.2	55.2	54.4
base SAM 2.1	57.1	70.1	70.2
SAM-pose2seg	61.6	70.0	69.5

Table 3. **Evaluation of the base SAM model and SAM-pose2seg** AP on the **Ground-Truth pose keypoints**. MaxVis strategy does not make sense as the visibility classification is binary here, we therefore use selection by distance to optimize, while sorting out all keypoints with visibilty set to zero. COCOPersons is a subset of COCO without Small category persons (no Ground-Truth poses are available there). MaxSpread$_3$ keypoint selection is used for SAM-pose2seg.

4.3. Ablation Study

4.3.1. Correction Points In Training

We further review the correction process to prove why our model converges to the optimal usage of three keypoints during inference. Crucially, as the model adapts to human segmentation, it rarely requires all seven correction points. In practice, the fine-tuned *SAM-pose2seg* model often predicts a correct mask after the first sampled keypoint. When comparing Ground-Truth masks to masks in each of the correction iteration, it was discovered that the differences changes between the mask in each iteration and their IoU with the Ground-Truth mask are small – mean IoU on a small subset during before any correction iterations is already 81.0 % and only goes up by 6.4 percentage points after the last iteration (see Fig. 2). This is corroborated by our *Pose1MaskRefine* experiment, where only the first point is pose-derived, and subsequent correction points are sampled uniformly as in the default method. The marginal difference in AP between

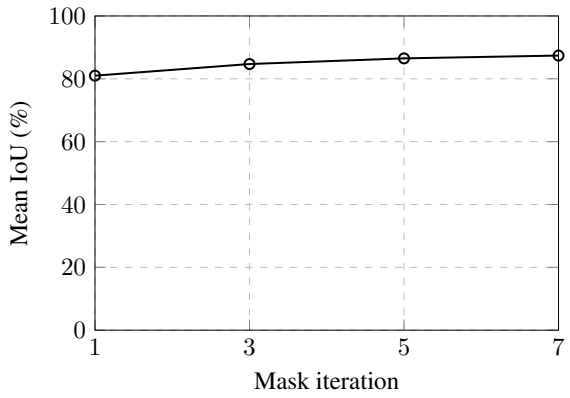


Figure 2. **Importance of correction points in the default SAM training method MaskRefine.** Comparison on a small COCO + CIHP training subset of mean IoU of the ground-truth mask. The first iteration contains only one point, any other i-th iteration is prompted by i keypoints and a correction mask. It seems that the refinement mostly causes minor changes in the overall mask shape.

Pose1MaskRefine and the full PoseMaskRefine confirms that the first keypoint is the primary driver of performance (see Tab. 1).

4.3.2. Prompting – Keypoint Selection

We evaluate two prompting strategies, MaxVis and MaxSpread, which differ in keypoint selection and interaction with Ground-Truth (GT) masks during inference and training (yet we stuck with PoseMaskRefine in the case of training).

Method MaxVis$_n$ (Visibility-Based Keypoint Selection). This method selects the top n keypoints solely based on *visibility* scores, imposing no additional spatial, semantic, or structural constraints.

Method MaxSpread$_n$ (Distance- and Visibility-Based Keypoint Selection). Originally introduced in Bbox-Mask-Pose [19] (specifically MaxSpread$_6$), this method extends visibility-based selection with a spatial diversification heuristic inspired by *k-means++* [4]. The first keypoint is chosen for maximum visibility; subsequent points are selected to maximize distance from previous ones. To reduce redundancy, number of used eye and nose keypoints is limited to one, and low-visibility points are excluded. This strategy generally outperforms MaxVis on the default model and is preferred for binary visibility classification.

MaxSpread$_6$ proved most effective for pose-to-segmentation, remaining consistent after incorporating ProbPose [20] visibility. For the base model, spatially distributed keypoints provide richer cues than simple visibility selection, which often overrepresents facial points and causes segmentation ambiguity (e.g., face vs. full body).

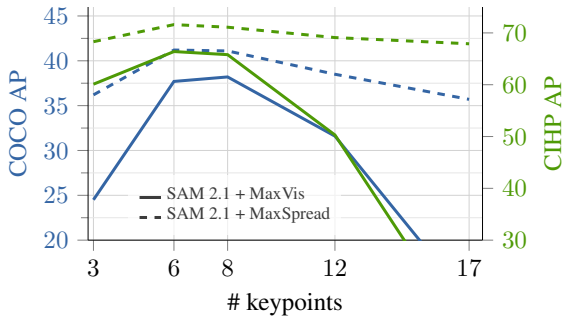


Figure 3. **Keypoint selection methods on SAM 2.1**. Prompting methods MaxVis (full) and MaxSpread (dashed) on COCO and CIHP datasets. 6 keypoints is the best for both methods. MaxSpread outperforms MaxVis as shown in [19].

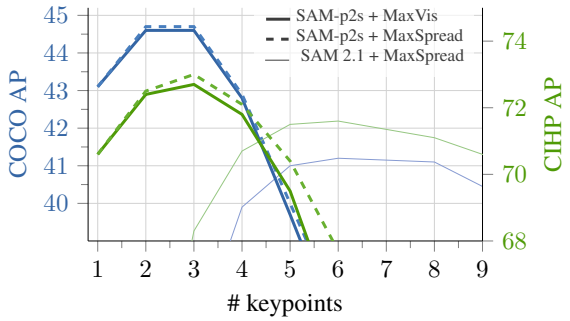


Figure 4. **Keypoint selection methods for SAM 2.1 and SAM-pose2seg** on COCO and CIHP datasets. SAM 2.1 with MaxSpread (thin line) peaks at 6 keypoints for both datasets. SAM-pose2seg with MaxSpread (dashed) and MaxVis (full) behaves the same on both datasets and peaks at 3 keypoints. We selected MaxVis method due to its simplicity. SAM-pose2seg outperforms SAM 2.1 with both selection methods on both datasets.

However, after fine-tuning with MaskRefine and PoseMaskRefine, the performance gap between MaxSpread$_n$ and MaxVis$_n$ becomes negligible (see Fig. 4).

We also briefly explored negative keypoint prompting; however, as it did not yield consistent gains and introduces strong context dependence, we defer a detailed analysis to the Supplementary.

4.3.3. Bounding Boxes

Bounding boxes are not part of our final pipeline, but we analyze their impact as an ablation to better understand SAM's behavior under pose-guided prompting.

The SAM model does support the use of bounding boxes as prompts, and since ProbPose [20] provides bounding box predictions, it would be natural to incorporate them into our pipeline. Ground-Truth bounding boxes were found to be beneficial, as they define precise instance boundaries (see Tab. 4). Even in this case, difficult settings, where a correctly identified bounding box contains large parts of multiple people, might pose a challenge (see Fig. 5). Nevertheless, bounding boxes predicted by vari-

Input SAM 2.1 SAM 2.1 + BBox

Figure 5. **Problematic bounding box usage** in SAM 2 when multiple people are included. Note: a part of the other person's hand is recognized incorrectly in both cases due to an incorrect pose keypoint.

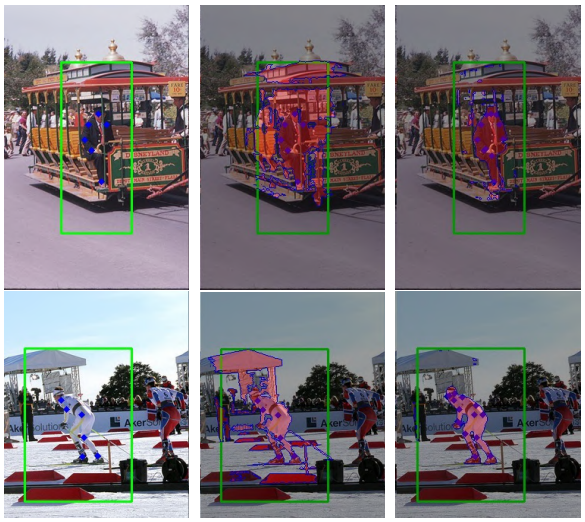
Model	Bbox type	COCO AP	CIHP AP
SAM 2.1	GT	50.5	76.4
SAM 2.1	inflated GT	17.5	52.4
SAM 2.1	none	41.2	71.6
SAM 1	GT	52.5	72.2
SAM 1	inflated GT	45.2	59.2
SAM 1	none	43.0	65.6

Table 4. **Usage of bounding boxes in SAM 1 and SAM 2.1.** on COCO val and CIHP val. Performance comparison of the base SAM 1 and SAM 2.1 models if ProbPose [20] keypoints are accompanied by bounding boxes. We use the keypoint selection method MaxSpread_6. While Ground-Truth bounding boxes seem to be helpful in both scenarios, if both dimensions are enlarged to simulate possible detector noise (by 50 % in each direction – 400 % area increase in total), SAM 2.1 is not reliable.

ous detectors (that would be a part of iterative pose refinement) may not be as precise [19]. As a result, segmentation precision degrades significantly when using SAM 2.1: the model may incorrectly include parts of other people or background regions in order to fill the provided bounding box (see Fig. 6).

4.3.4. Backbone Choice: SAM 1 vs. SAM 2

As we tried to analyze whether we could improve our prompting method, we also turned our attention to the older SAM 1 model. In both cases, the second largest models were used (*Hiera Base Plus* for SAM 2 and *Vit-L* for SAM 1). We conducted several tests showing its strengths and weaknesses, and it seems that for our task, SAM 2.1 is more suitable. SAM 1 performs slightly better



Input SAM 2.1 + BBox SAM 1 + BBox

Figure 6. **Inflated bounding box usage** in base SAM 2.1 (middle) and SAM 1 (right). SAM 2.1 often incorporates unnecessary noise when the prompted bounding box does not fit the person exactly. In contrast, SAM 1 performance generally does not drop drastically.

on COCO dataset, where the poses are more visible, yet lags behind in OCHuman and CIHP datasets.

It also seems that SAM 1 is able to make better use of negative keypoints in the case of OCHuman and CIHP datasets. When sampling the closest pose keypoints, it serves better as an option to remove superficial parts. Imprecise bounding boxes also do not harm the prediction process significantly, which could not be said for SAM 2.1 that often includes other objects just to fill the boundaries.

However, in general, SAM 2.1 turned out to be a stronger base for our task, as the model itself is lighter (and, therefore, much faster during inference), the provided training code serves well for the fine-tuning and the experiments generally yielded worse results.

4.3.5. SAM-pose2seg Performance In Challenging Scenes

Incomplete segmentation. In contrast to the base SAM 2.1, incomplete segmentation occurs far less frequently. Since SAM-pose2seg is fine-tuned specifically for human segmentation, it learns to recover the full human body more consistently. This addresses a common ambiguity of the base SAM model, which in some cases produced masks corresponding only to clothing, skin regions, or isolated body parts rather than complete human instances.

Partially occluded pose. SAM-pose2seg demonstrates precise segmentation in scenarios where pose information is unreliable or heavily occluded. By prioritizing keypoints with the highest visibility scores, the model effectively leverages strong, reliable cues, reducing the impact of missing or ambiguous points. In cases where a

Input SAM 2.1 SAM-pose2seg

Figure 7. **Improved handling of incomplete segmentation**, with SAM-pose2seg correctly segmenting full human instances. More examples in Supplementary.

Input Base SAM 2.1 SAM-pose2seg

Figure 8. **Improved handling of partially occluded poses.** More examples in Supplementary.

single additional misdetected keypoint could compromise the segmentation, SAM-pose2seg maintains accuracy and produces consistent, complete masks.

5. Conclusions

To sum up, we achieved substantial improvements in the pose-to-segmentation task for humans by adapting and fine-tuning SAM. We identified and validated the most effective keypoint selection strategies, demonstrating that leveraging *visibility* scores allows the model to reliably choose strong, informative points. Simple and clear prompting methods can be particularly useful in occluded or crowded scenarios.

Our proposed SAM-pose2seg model shows strong generalization and robustness. It excels at recovering full human instances even when most keypoints are missing or unreliable, and is less sensitive to errors that would otherwise harm segmentation in the base SAM model. This makes it suitable both for iterative refinement in pose estimation pipelines and as a standalone tool for human instance segmentation. The deployment of SAM-pose2seg in the BMP loop [19] improves the segmentation accuracy from 31.8 AP to 33.7 AP on OCHuman.

Despite these improvements, some limitations remain. SAM-pose2seg can struggle when multiple people are closely overlapping and even high-*visibility* keypoints carry ambiguous semantic meaning, occasionally merging visually similar regions across instances. Future work could explore adaptive keypoint weighting, explicitly modeling the semantic meaning of each keypoint, or integrating the approach into SAM3 to leverage its language-based reasoning capabilities.

Overall, our results demonstrate that SAM-pose2seg provides a practical and reliable approach for human pose-guided segmentation, with potential applications as a pseudo-annotation tool for large-scale datasets or as a robust component in downstream vision pipelines.

Acknowledgements. This work was supported by the Ministry of the Interior of the Czech Republic project No. VJ02010041, the Technology Agency of the Czech Republic project CEDMO 2.0 No. FW10010387, the European Union's Digital Europe Programme under Contract No. 101158609, and the Czech Technical University student grant SGS23/173/OHK3/3T/13.

References

[1] Niaz Ahmad, Jawad Khan, Jeremy Yuhyun Kim, and Youngmoon Lee. Multiposeseg: Feedback knowledge transfer for multi-person pose estimation and instance segmentation. *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2086–2092, 2022. 1, 2

[2] Niaz Ahmad, Jawad Khan, Jeremy Yuhyun Kim, and Youngmoon Lee. Joint human pose estimation and instance segmentation with poseplusseg. In *AAAI Conference on Artificial Intelligence*, 2022. 1, 2

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 5

[5] Kambiz Azarian, Debasmit Das, Hyojin Park, and Fatih Porikli. Test-time adaptation vs. training-time generalization: A case study in human instance segmentation using keypoints estimation, 2022. 2

[6] Kambiz Azarian, Debasmit Das, Hyojin Park, and Fatih Murat Porikli. Test-time adaptation vs. training-time generalization: A case study in human instance segmentation using keypoints estimation. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 411–420, 2022. 1, 2

[7] Zhi Cai, Yingjie Gao, Yaoyan Zheng, Nan Zhou, and Di Huang. Crowd-sam: Sam as a smart annotator for object detection in crowded scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[8] Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3395–3405, 2025. 2

[9] Z Ge. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[10] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–785, 2018. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 2

[12] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 2

[13] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 2

[14] Zhong Li, Xin Chen, Wangyiteng Zhou, Yingliang Zhang, and Jingyi Yu. Pose2body: Pose-guided human parts segmentation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 640–645, 2019. 2

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 2

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[17] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *ArXiv*, abs/2212.07784, 2022. 2

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[19] Miroslav Purkrabek and Jiri Matas. Detection, pose estimation and segmentation for multiple bodies: Closing the virtuous circle. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9004–9013, 2025. 1, 2, 5, 6, 7

[20] Miroslav Purkrabek and Jiri Matas. ProbPose: A probabilistic approach to 2d human pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 27124–27133, 2025. 3, 4, 5, 6

[21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2

[22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[23] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2

[24] Danila Rukhovich, Konstantin Sofiiuk, Danil Galeev, Olga Barinova, and Anton Konushin. Iterdet: iterative scheme for object detection in crowded environments. In *Structural, syntactic, and statistical pattern recognition: Joint IAPR international workshops, s+ SSPR 2020, padua, Italy, January 21–22, 2021, proceedings*, pages 344–354. Springer, 2021. 2

[25] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *ArXiv*, abs/1805.00123, 2018. 2

[26] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[27] Subarna Tripathi, Maxwell D. Collins, Matthew A. Brown, and Serge J. Belongie. Pose2instance: Harnessing keypoints for person instance segmentation. *ArXiv*, abs/1704.01152, 2017. 1, 2

[28] Zhaoyang Wei, Pengfei Chen, Xuehui Yu, Guorong Li, Jianbin Jiao, and Zhenjun Han. Semantic-aware sam for point-prompted instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3585–3594, 2024. 2

[29] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. Ai challenger : A large-scale dataset for going deeper in image understanding. *ArXiv*, abs/1711.06475, 2017. 2

[30] Shiting Xiao, Rishabh Kabra, Yuhang Li, Donghyun Lee, Joao Carreira, and Priyadarshini Panda. Openworld-sam: Extending sam2 for universal image segmentation with language prompts. *arXiv preprint arXiv:2507.05427*, 2025. 2

[31] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[32] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 1, 2

[33] Desen Zhou and Qian He. Poseg: Pose-aware refinement network for human instance segmentation. *IEEE Access*, 8:15007–15016, 2020. 1, 2

[34] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023. 2

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hradec, Czech Republic, February 9–12, 2026

Dynamic Ensemble of Deepfake Detectors Conditioned on CLIP Features

Patricie Petrilakova and Jan Cech
Faculty of Electrical Engineering
Czech Technical University in Prague
{petripat, cechj}@fel.cvut.cz

Abstract

Deepfake image detection framework that combines the outputs of multiple pre-trained detectors using dynamically predicted importance weights is proposed. The final prediction is computed as a convex combination of detector scores, with weights derived from CLIP image embeddings, enabling the ensemble to adapt to different types of synthetic content. To support evaluation, we construct a dataset of approximately 50k images, including four real-image sources and fake images generated by ten modern synthesis tools spanning GANs, diffusion models, and commercial systems. We set up fourteen publicly available detectors and benchmark them individually and within our ensemble. Extensive experiments show that the proposed model outperforms individual detectors and simpler baselines, particularly under image degradations and when generalizing to unseen generators. Ablation studies further confirm the benefits of dynamic weighting over static combinations or CLIP-only predictors.

1. Introduction

The recent developments in generative models have undermined the credibility of visual media. Images once assumed to be trustworthy can now be effortlessly manipulated using modern synthesis techniques. Contemporary image-generation models, such as generative adversarial networks and diffusion-based architectures, can produce photorealistic face images or manipulate real ones with such precision that the results are often indistinguishable from authentic images, even for trained experts.

While this progress has opened new opportunities in creative expression, education, and personalised content creation, it also introduces considerable societal and security risks. Face-image generation can be weaponised to fuel misinformation, interfere with democratic processes, or facilitate identity deception in financial systems. Moreover, the widespread availability of these tools enables even unskilled users to produce realistic synthetic media at scale. Text-to-image generation, in particular, allows adversaries to construct highly specific imagery aligned with their intended narrative.

Consequently, there is a high demand for tools that help governments, fact-checkers, or journalists establish authenticity even when images have been deliberately altered to evade detection. This work aims to support experts in answering a question that is becoming increas-

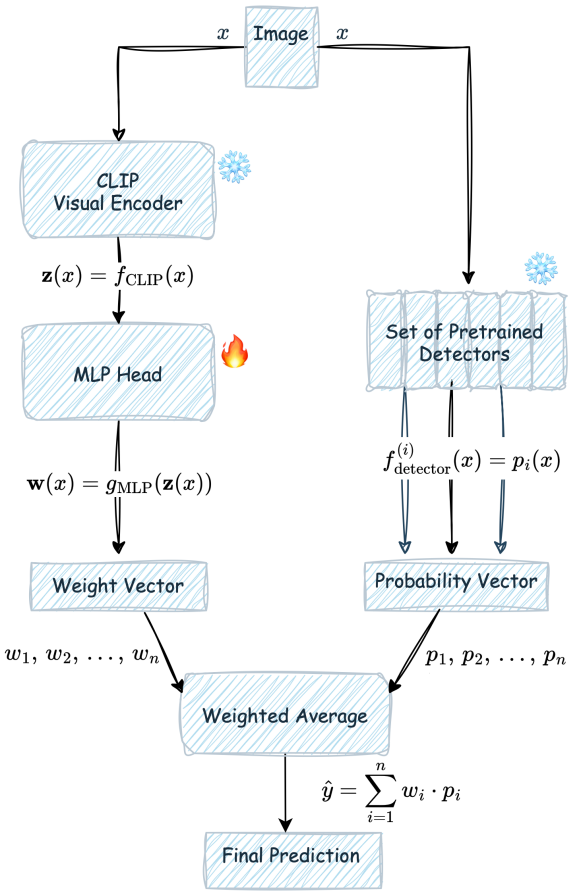


Figure 1. **The proposed architecture of the ensemble model.** For a given image, a bank of pre-trained deepfake detectors is executed, producing a probability vector. The final "fake/real" prediction is computed as a weighted average of the individual detector probability scores. The model is trained to estimate the importance weights dynamically from image embeddings produced by the CLIP visual encoder. Models following this architecture are denoted `Ens-Small` and `Ens-Final`.

ingly common in social media contexts or journalism: Is this real or not?

The research landscape of deepfake image detection is highly active, with a continual emergence of new detector models. However, their accuracy varies substantially. A key challenge is poor generalization: many detectors fail when confronted with fake content that differs from the data seen during training. Our approach is to aggregate a diverse set of available detectors and construct an ensemble that leverages all of them simultaneously. Since many detectors appear to specialize in particular families of generators (*e.g.*, GANs or diffusion models), we propose the ensemble architecture shown in Fig. 1, where the final probability score is computed as a convex combination of individual detector outputs. The corresponding weights are predicted dynamically from image embeddings produced by a foundation model, namely CLIP [22].

Our contributions are as follows:

- We constructed a diverse evaluation dataset made up of about 50k images, containing 4 different real datasets and realistic face images manipulated or fully generated by 10 modern synthesis tools, including generative adversarial networks, diffusion models, and commercial image generators.
- We set up 14 deepfake image detection models sourced from publicly available GitHub repositories.
- We proposed and developed an ensemble detection model that combines outputs of a set of pre-trained detectors with image features.
- We performed an ablation study comparing the proposed ensemble architecture with models that either rely solely on detector outputs combined using static weights or use only image embeddings without detector outputs.
- We conducted extensive experiments evaluating the detectors' robustness to common image transformations (degradations that occur in social media) and their ability to generalize to previously unseen generators.

The remainder of the paper is structured as follows. The related work is summarized in Sec. 2, the proposed method is presented in Sec. 3, the experiments are given in Sec. 4, and finally Sec. 5 concludes the paper.

2. Related Work

2.1. Deepfake production techniques

Advanced face manipulation began with techniques such as Face2Face, which used 3D models to transfer expressions from one person to another in real time [29]. The field then shifted significantly with the rise of deep learning and deepfakes, where autoencoders and GAN-based models enabled face swapping with much higher realism [17]. For example, SimSwap improved face swapping by injecting source identity into the target's features while preserving target attributes such as expression and gaze direction [4]. More recently, diffusion models have been introduced into face manipulation tasks, achieving improved performance compared to GAN-based methods [14].

2.2. Deepfake detectors

The detection of fake images has become a critical research area in recent years due to rapid advances in generative technologies. These methods can modify existing images or synthesise them entirely. As generative models continue to improve in visual quality, fake images have become increasingly difficult for regular, untrained observers to identify [10]. Notably, even experts often struggle, sometimes misclassifying authentic images as generated and attributing imperfections made by human artists to AI [10].

Classifiers trained in-distribution (*i.e.*, on data from a specific image-generator class) achieve high accuracy but often fail when encountering images from unseen generative models [20]. Since new generative models emerge frequently, there is a growing need for detectors capable of identifying fake images produced by previously unseen methods. This property is referred to as *generalization*, and it reflects a detector's ability to perform well across a range of generation techniques and manipulation types. Another important property is *robustness*, which denotes a model's ability to maintain accuracy under transformations such as resizing, compression, or adversarial noise. To improve generalization and robustness, researchers have introduced heavy augmentations, including JPEG compression and blurring [30].

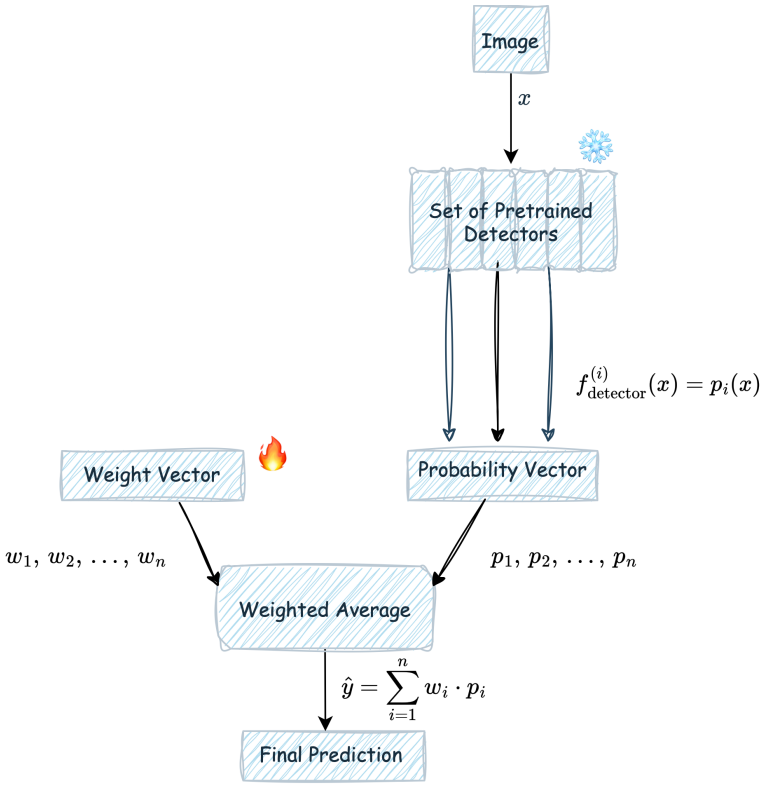
Wang et al. [30] showed that standard CNN classifiers trained on images generated by models such as ProGAN can generalise effectively across different GAN-based images, likely due to distinctive and persistent low-level artifacts shared across many GAN generations. However, detectors trained on GAN-generated images often fail when applied to diffusion-generated images [23]. These findings laid the groundwork for exploring more generalisable detection methods.

There are approaches that do not rely on training on large datasets of fake images in order to avoid overfitting. Shiohara et al. [27] focus on detecting blending artifacts that are common in face-swap deepfakes, but absent in fully synthetic images. Another method, DIRE [32], reports strong generalization based on the observation that synthetic images are easily invertible by diffusion models, whereas real images are not.

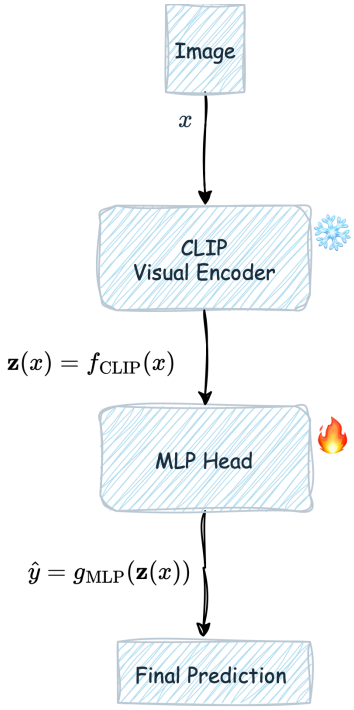
In recent research, leveraging a large foundation model as an image encoder, such as CLIP, has demonstrated notable generalisation across multiple deepfake-production techniques and strong robustness, highlighting the advantages of large-scale pretraining [7, 34].

2.3. Ensemble models

Ensemble models are a well-established strategy for improving classification accuracy by combining the outputs of multiple classifiers and reducing the overall bias of the final prediction.



(a) **Ensemble model with static weights.** The final prediction probability score is calculated as a weighted average of the individual detectors' probability outputs. The weights are constant and do not depend on the input image, unlike in the full architecture shown in Fig. 1. The model is denoted as `Ens-Probs`.



(b) **Image only model using CLIP embeddings.** The model utilises image embeddings from a frozen CLIP visual encoder and passes them to an MLP head, which produces the final prediction. The model is denoted `Mono-CLIP`.

Figure 2. Diagrams of two baseline detection models explored in this work.

An early study regarding deepfake detection by Panigrah et al. [19] employed an ensemble of 13 CNN architectures including Xception, DenseNet121, and ResNet variants with a majority voting mechanism for identifying manipulated images created via splicing and copy-move operations. This approach demonstrated significant improvements over single-model solutions.

More recently, computationally efficient ensemble methods suitable for mobile deployment have been proposed, combining lightweight CNN architectures like EfficientNetB0 and MobileNetV2, showing promising generalisation capabilities across various GAN-based generators [37]. However, these results primarily focused on GAN-generated images, known for their distinctive and detectable artifacts.

Another ensemble-based approach, MaskCLIP [31], integrates pretrained models such as CLIP and MAE (Masked Autoencoders) to detect high-level semantic inconsistencies and low-level reconstruction anomalies simultaneously. Furthermore, MaskCLIP employs learnable continuous prompts and attention mechanisms to combine diverse feature sources.

3. Method

The primary motivation for the ensemble model is to enhance the robustness and generalization of existing fake-image detectors while leveraging their learned knowledge. Individual detectors may exhibit limited generalization due to their training data or architectural constraints. Nevertheless, their outputs contain valuable information, and our goal is to exploit this by combining the predictions of individual detectors into a collective decision mechanism.

To achieve this, we introduce an additional degree of freedom by incorporating semantic image understanding through the CLIP visual encoder [22] to condition the contribution of each detector. The learned weighting mechanism adapts dynamically, assigning more or less trust to individual detectors depending on the input image. Conditioning the weighting system on CLIP features enables the model to recognize when a detector is unreliable or confident with respect to specific artifacts and to accordingly decrease or increase that detector's influence in the final prediction. This approach allows us to leverage transfer learning in two complementary ways: by exploiting CLIP's visual understanding and by utilizing the outputs

of existing fake-image detectors. Moreover, the architecture provides interpretability by producing per-image detector weights.

3.1. Proposed architecture

The architecture of the proposed model is shown in Figure 1. It is designed to weight the outputs of pre-trained detectors according to visual cues extracted from the image by the CLIP visual encoder. For a given input image, all pre-trained detectors are executed, producing scores that form a probability vector. In parallel, semantic features using a frozen pre-trained CLIP visual encoder are extracted. These features are further processed by a small multi-layer perceptron head, which produces the weight vector. The final prediction score is computed as the dot product between the weight and probability vectors. The output is therefore a convex combination of the individual detector outputs, as the weights are non-negative and sum to one.

Model Formulation Let $x \in \mathbb{R}^{H\times W\times 3}$ be an input image. The CLIP visual encoder extracts a feature representation $f_{\text{CLIP}}(x) \in \mathbb{R}^d$, which is passed to an MLP head g_{MLP} that maps it to a weight vector

$$\mathbf{v}(x) = [w_1(x), w_2(x), \ldots, w_n(x)] = g_{\text{MLP}}(f_{\text{CLIP}}(x)),$$

which is normalised with the softmax function

$$\mathbf{w}(x) = \text{softmax}(\mathbf{v}(x)),$$

then $w_i(x) \geq 0, \sum_{i=1}^n w_i(x) = 1$.

Simultaneously, the image is passed to a fixed set of n pretrained detectors, each defined as a function $f^{(i)}_{\text{detector}}\colon x \to \langle 0, 1\rangle$. Each detector outputs a probability score (indicating the likelihood of belonging to the "fake" class) for the input image:

$$p_i(x) = f^{(i)}_{\text{detector}}(x), \quad \text{for } i = 1, \ldots, n.$$

These individual outputs form a probability vector:

$$\mathbf{p}(x) = [p_1(x),\ p_2(x),\ \ldots,\ p_n(x)] \in [0, 1]^n.$$

The final prediction is computed as the dot product between the weight vector and the vector of detector probabilities, producing the final probability score

$$\hat{y}(x) = \mathbf{w}(x)^\top \mathbf{p}(x) = \sum_{i=1}^n w_i(x) \cdot p_i(x) \in [0, 1],$$

where $\hat{y} \approx 1.0$ indicates full confidence in the fake class and $\hat{y} \approx 0.0$ indicates full confidence in the real class.

3.2. Baseline architectures

The proposed model is ablated in two ways, resulting in the following baseline variants.

Static Ensemble Detector. We implemented an ensemble with static weights shown in Figure 2a using only detector outputs (without image features) to assess whether the image-based features are necessary for optimal weighting.

Let $\mathbf{w} = [w_1, w_2, \ldots, w_n] \in \mathbb{R}^n$ be a fixed weight vector, where the weights are constrained that $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$. The rest is the same as the full model, so

$$\hat{y}(x) = \mathbf{w}^\top \mathbf{p}(x) = \sum_{i=1}^n w_i \cdot p_i(x) \in [0, 1].$$

Note that the weight vector is constant, *i.e*. does not depend on the input image.

To learn the constant weight vector, we employed an evolution-based optimisation[1], which turned to be more accurate than gradient descent methods. The static ensemble was faster to train, as we were able to produce a decent detector in less time than it takes to train one epoch of the full model.

Image Detector. Secondly, we implemented a monomodal detector from Fig. 2b that relies exclusively on image information. Specifically, a classification head is applied to image embeddings extracted by the CLIP visual encoder. No pretrained detectors are involved in the model.

Formally, given an input image $x \in \mathbb{R}^{H\times W\times 3}$, the CLIP visual encoder extracts a feature vector $f_{\text{CLIP}}(x) \in \mathbb{R}^d$. This feature is passed through a classification head h_{MLP}, producing a scalar prediction:

$$\hat{y}(x) = h_{\text{MLP}}(f_{\text{CLIP}}(x)) \in [0, 1].$$

4. Experiments

4.1. Datasets

For our experiments, we collected a dataset of human face images. The dataset contains both manipulated images (*e.g*., face swap, face reenactment) and fully synthetic face images produced by various generators. To generate the images, we provided a detailed prompt for the sake of diversity, *e.g*. "close-up on face of a woman with gray hair, alabaster skin, and amber eyes, having zen expression set against a countryside background, with studio lighting." Real images were sampled from several datasets.

All images were subjected to the same preprocessing pipeline: face detection, cropping, and resizing to 224×224 px.

Training and validation set. The composition of our training set is given in Tab. 1. We have two variants; the Large set and the Small set. Intermediate and baseline models were trained on the small set, the final model was trained on the Large set. Fake images were mostly

[1]We used the `differential_evolution` implementation from `scipy.optimize`.

Table 1. **Training dataset composition.** The table shows the training dataset for the final ensemble model and smaller experimental variants. The validation set corresponds to a proportion of the original data.

Source / Config.	Large	Small	Label
FF++	13,557	2,500	Fake
RV	7,690	2,500	Fake
Flux	50	–	Fake
DF-IF	50	–	Fake
WFIR	10	–	Fake
LAION	18,878	5,000	Real
Total	40,235	10,000	–
Validation set	5%	10 %	–
Augmentations	45%	45 %	–

taken from Face Forensics++ [24] and generated by Realistic Vision (RV) model [5]. Real images for training were sampled from LAION dataset [35].

To prevent overfitting of the model and improve robustness to image degradations, we augmented 45% of the images in each epoch using techniques such as random quality level up to 60% JPEG compression, rotation, grayscale conversion, sharpening, Gaussian blur, affine transformations, color jittering, horizontal and vertical flips, and random cropping. The validation set is non-overlaping 5% and 10% for Large and Small sets, respectively. The rate of augmentation is the same on the validation set.

Test dataset. The composition of the test set is given in Tab. 2. It is on purpose different from the training set in both fake and real parts to test the model generalization to unseen content. For each generator and the real subset, we collected 1000 images. The images were either generated using public models (*e.g.*, via Hugging Face) or sourced from benchmark datasets such as DeepFakeBench [24] and DF40 [33]. Real images were sampled from publicly available datasets, including IMDB-WIKI [25], CelebA [15], FFHQ [13]. Note that the real images are taken from a different dataset than the training set, which is a challenging situation, as the detector might overfit to possible biases in the real data.

4.2. Detectors

The pre-trained deepfake detectors used in our ensemble models are listed in Tab. 3. Multiple publicly available detectors of various approaches that span years 2020–2026 were used.

We tested several of our models: `Ens-Small` – the model of architecture in Fig. 1, and two ablated baselines `Ens-Probs` – of architecture in Fig. 2a, and `MonoCLIP` – see Fig. 2b. They were trained on the small training set detailed in Tab. 1, and using transformer `ViT-B/32` of the `OpenCLIP` image encoder [11]. The final ensemble model `Ens-Final` was trained on the large training set and with a larger tranformer backbone `ViT-L/14`.

Table 2. **Test dataset.** The table lists all generators and real image sources used for evaluation, along with their respective origins and the reference names used throughout this work.

Generator	Source	Ref. Name
StyleGAN3	DF40 dataset [33]	`StyleG3`
FaceForensics++	FaceForensics++ [24]	`FF++`
PixArt-alpha	Generated [3]	`PixArt`
Realistic Vision v6	Generated [5]	`RV`
Stable Diffusion XL	Generated [21]	`SD-XL`
Kandinsky 3.0	Generated [1]	`Kndsky`
FLUX-dev	Generated [2]	`Flux`
DeepFloyd IF	Generated [8]	`DF-IF`
WhichFaceIsReal	DF40 dataset [33]	`WFIR`
MidJourney v6	DF40 dataset [33]	`MidJ`
Real images	IMDB + Celeba + FFHQ	`Real`

Table 3. **Pre-trained detectors.** The table lists all pre-trained detectors that were used in our ensemble models.

Detector Name	Year	Publication
CNND, CNND2 [30]	2020	CVPR
FreqD [9]	2020	ICML
GramNet [16]	2020	CVPR
Fusing [12]	2022	ICIP
DIMD-latent, DIMD-gan [6]	2023	ICASSP
UnivFakeD [18]	2023	CVPR
RPTC [36]	2023	ArXiv
DeFake [26]	2023	SIGSAC
NPR [28]	2024	CVPR
ClipBased, Corvi2023 [7]	2024	CVPRw
DeepfakeD [34]	2026	WACV

All models were trained to minimize binary cross-entropy loss. The final model was extensively trained for approximately five days using NVIDIA A100-SXM4 GPUs with 40GB of memory. Training epochs varied in duration, typically requiring between 45 minutes to two hours depending on computational resource utilisation and the complexity of the training scenarios.

4.3. Results

We evaluated all pretrained detectors and our models on subsets of our test set. Testing was performed under two scenarios: original images and degraded images produced by random cropping (up to 5/8), resizing (down to 0.5), and JPEG compression (down to 60%). The latter setting simulates typical downstream processing on social-media platforms. The results are shown in Tab. 4 and Tab. 5, respectively.

The tables report classification accuracy. All test subsets are balanced, containing 1000 real and 1000 fake images, so the chance level is 0.5. We gray out entries where the generators in the training and test sets overlap. We refer to these as "in-distribution" cases, and their accuracy values are excluded from the row-average calculations. While other detectors may encounter in-distribution data during evaluation and benefit from this overlap, our

Table 4. **Accuracy of detectors across generative models (original images).** Accuracy is evaluated per generator using balanced sets of fake and real images. The table shows detectors (rows) evaluated against various generative models (columns). Greyed-out numbers correspond to generators included in their training set; these values are excluded from the row-average calculations.

Detector	DF-IF	MidJ	RV	SD-XL	Flux	Kndsky	PixArt	FF++	StyleG3	WFIR	Row Avg.
CNND	0.53	0.76	0.47	0.47	0.52	0.47	0.56	0.47	0.59	0.91	0.57
CNND2	0.50	0.54	0.50	0.50	0.50	0.50	0.51	0.50	0.50	0.56	0.51
Corvi2023	0.45	**0.94**	0.95	**0.92**	0.56	0.86	**0.95**	0.45	0.62	0.45	0.72
DIMD-gan	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.96	**1.00**	0.60
DIMD-latent	0.45	0.92	0.95	0.91	0.58	0.81	0.95	0.45	0.45	0.45	0.69
DeFake	0.58	0.55	0.70	0.72	0.63	0.67	0.70	0.52	0.49	0.70	0.63
FreqD	0.46	0.32	0.64	0.62	0.40	0.50	0.58	0.50	0.72	0.78	0.55
Fusing	**0.84**	0.83	0.48	0.47	0.54	0.48	0.61	0.47	0.88	0.73	0.63
GramNet	0.78	0.73	0.57	0.29	0.74	0.78	0.54	0.35	0.79	0.79	0.64
NPR	0.65	0.88	0.53	0.51	**0.86**	0.82	0.78	0.41	0.88	0.53	0.69
RPTC	0.61	0.57	0.70	0.57	0.60	0.78	0.72	0.33	0.80	0.80	0.65
UnivFakeD	0.49	0.49	0.49	0.49	0.49	0.52	0.49	0.65	0.81	0.57	0.55
ClipBased	0.66	0.40	0.87	0.57	0.70	0.87	0.58	0.86	0.88	0.57	0.70
DeepfakeD	0.52	0.58	0.50	0.54	0.51	0.62	0.65	**0.93**	**0.98**	0.51	0.63
Ens-Small	0.64	0.62	0.96	0.63	0.70	**0.95**	0.84	0.89	0.96	0.79	0.77
Ens-Probs	0.76	0.88	0.93	0.61	0.82	0.93	0.89	0.47	0.93	0.92	**0.84**
MonoCLIP	0.69	0.87	0.88	0.83	0.72	0.80	0.84	0.85	0.73	0.68	0.76
Ens-Final	0.76	0.71	0.92	0.91	0.83	0.91	0.91	0.93	0.94	0.91	0.88

Table 5. **Accuracy of detectors on transformed images simulating social media degradation.** Images undergo random cropping, resizing, and JPEG compression to assess robustness under realistic post-processing conditions. The table shows detectors (rows) evaluated against various generative models (columns). Greyed-out numbers correspond to generators included in their training set; these values are excluded from the row-average calculations.

Detector	DF-IF	MidJ	RV	SD-XL	Flux	Kndsky	PixArt	FF++	StyleG3	WFIR	Row Avg.
CNND	0.51	0.52	0.50	0.50	0.50	0.50	0.51	0.50	0.52	0.51	0.51
CNND2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Corvi2023	0.57	**0.67**	0.61	0.64	0.58	0.70	0.56	0.59	0.80	0.66	0.64
DIMD-gan	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.50	0.50
DIMD-latent	0.51	0.63	0.64	0.59	0.53	0.62	0.53	0.51	0.58	0.58	0.57
DeFake	0.59	0.63	0.65	0.67	0.48	0.73	**0.67**	0.52	0.41	**0.61**	0.60
FreqD	0.50	0.61	0.46	0.55	0.49	0.59	0.62	0.46	0.46	0.54	0.53
Fusing	0.54	0.54	0.50	0.52	0.53	0.50	0.53	0.50	0.62	0.52	0.53
GramNet	0.53	0.51	0.52	0.52	0.51	0.52	0.52	0.47	0.47	0.51	0.51
NPR	0.50	0.50	0.50	0.50	0.50	0.51	0.50	0.50	0.50	0.50	0.50
RPTC	0.60	0.58	0.42	0.58	**0.61**	0.47	0.49	0.55	0.54	0.58	0.54
UnivFakeD	0.48	0.48	0.48	0.48	0.48	0.52	0.49	0.61	0.62	0.49	0.51
ClipBased	0.58	0.54	0.58	0.58	0.56	0.59	0.58	0.59	0.59	0.54	0.57
DeepfakeD	0.54	0.59	0.47	0.55	0.52	0.64	0.64	**0.89**	**0.93**	0.47	0.62
Ens-Small	**0.61**	0.64	0.75	**0.68**	0.58	**0.79**	0.66	0.86	0.88	0.55	**0.67**
Ens-Probs	0.57	0.57	0.60	0.57	0.55	0.59	0.56	0.50	0.51	0.56	0.56
MonoCLIP	0.54	0.55	0.81	0.57	0.55	0.58	0.55	0.79	0.73	0.51	0.57
Ens-Final	0.68	0.66	0.89	0.76	0.69	0.87	0.85	0.98	0.95	0.54	0.82

models are strictly tested on unseen images, ensuring a fair assessment of generalisation.

The proposed `Ens-Final` outperforms all competing methods by a large margin on average accuracy. The margin is grater for the the degraded images. The model is performing the best for many individual subsets.

For the smaller ensemble models, it is seen on Tab. 4 that our baseline ensemble model `Ens-Probs` exhibits the best average accuracy, while it does not achieve the top accuracy for any single generator. The result highlights that even without image features, the ensemble technique allows model to generalise effectively across unseen generators. However, when subjected to the degraded images, its accuracy drops slightly above random chance, reveal-

ing a significant lack of robustness against post-processing transformations, see Tab. 5.

On the other hand, as indicated in Table 5, the Ens-Small model maintains strong performance under image degradations, achieving the highest row-average accuracy among the models trained on the smaller datasets (Ens-Probs, MonoCLIP). The image only model MonoCLIP has a comparable average accuracy with Ens-Small on the original, but much lower on the degraded images. This result demonstrates the effectiveness of combining multiple pretrained detectors when models accesses both image features and detector outputs.

Interestingly, the architectures of our MonoCLIP and the ClipBased detector [7] are very similar, but there are still differences in accuracy. Since both models leverage CLIP embedding, we attribute the difference to a different training datasets.

Among other detectors, DeepfakeD and DeFake also demonstrate stable performance across both original and transformed images. Notably, DeepfakeD average accuracy remains nearly unchanged (0.63 to 0.62) under augmentations, and it remains the best-performing detector for StyleG3 and FF++. In contrast, several models CNND2, CNND, UnivFakeD, and FreqD operate close to random guessing levels and consistently achieve row averages between 0.50 and 0.55. Other models DIMD-gan, DIMD-latent, Fusing, GramNet, NPR, and RPTC experience significant performance degradation under transformations, approaching random chance levels.

4.4. Contribution of individual detectors

To understand how the final ensemble model Ens-Final allocates decision responsibility across detectors, we analysed the per-image weights assigned by the MLP head conditioned on CLIP embeddings. Since the model computes a different set of weights for each image, we evaluated 100 samples per generator and calculated the mean and variance of the assigned weights. Additionally, detectors with weights lower than 0.01 were omitted from the figures for clarity, as their contribution to the final decision is negligible.

The dynamic weights predicted by Ens-Final model are shown in Fig. 3, again for two scenarios – orginal and images distorted by common transofrmations. The figure shows the weights of the static ensemble Ens-Prob for comparison.

It is seen that the dynamic weights differ from the static weights, so the model indeed learned to use the image features. Moreover, the dynamic weights differ for generators and real images for both original and degraded images. On the other hand, we can observe certain similarities between similar models, *e.g*. Realistic Vision and SD XL. Interestingly, detector DeFake seemed to be used for many generators. For FF++ images, the most important was CLIPBased, which was the most useful for StyleGAN3 and used more or less in all tested subsets. To recognize real images, the most important detector was DIMID-gan.

Note that the weights are sparse. Weights of many detectors are always zero. Those detectors can be dropped to speed the evaluation up.

5. Conclusions

In this work, we conducted a comprehensive evaluation of fourteen publicly available deepfake detectors on a curated dataset of facial images comprising both real samples from multiple sources and synthetic samples generated by modern generative models. Building on these findings, we proposed an ensemble architecture that integrates the outputs of pretrained detectors and assigns them adaptive importance weights predicted directly from image features. Our experiments demonstrate that this approach consistently outperforms both individual detectors and simpler ensemble baselines, particularly in challenging scenarios where input images undergo degradations such as compression.

An analysis of the learned dynamic weights further revealed that, for certain generator families, only a subset of detectors contributes meaningfully to the final decision, while others receive negligible or effectively zero weight. Although this insight highlights a potential path toward computational efficiency, the current method requires running all detectors during inference, resulting in latency on the order of tens of seconds, since all the detectors do not fit into a GPU memory. Parallel computation on multiple GPUs will reduce latency. Future work will therefore focus on leveraging the discovered weight patterns to prune consistently uninformative detectors and reduce computational cost. Moreover, the observed generator-dependent weighting behavior opens promising opportunities for model attribution, which we plan to explore in subsequent research.

Acknowledgements

The research was supported by the National Recovery Plan project CEDMO 2.0 NPO (MPO 60273/24/21300/21000), the EC Digital Europe Programme project CEDMO 2.0 no. 101158609, and the CTU Student Grant SGS23/173/OHK3/3T/13.

References

[1] Vladimir Arkhipkin, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Anton Bukashkin, Konstantin Kulikov, et al. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In *Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2024. https://huggingface.co/kandinsky-community/kandinsky-3. 5

[2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1

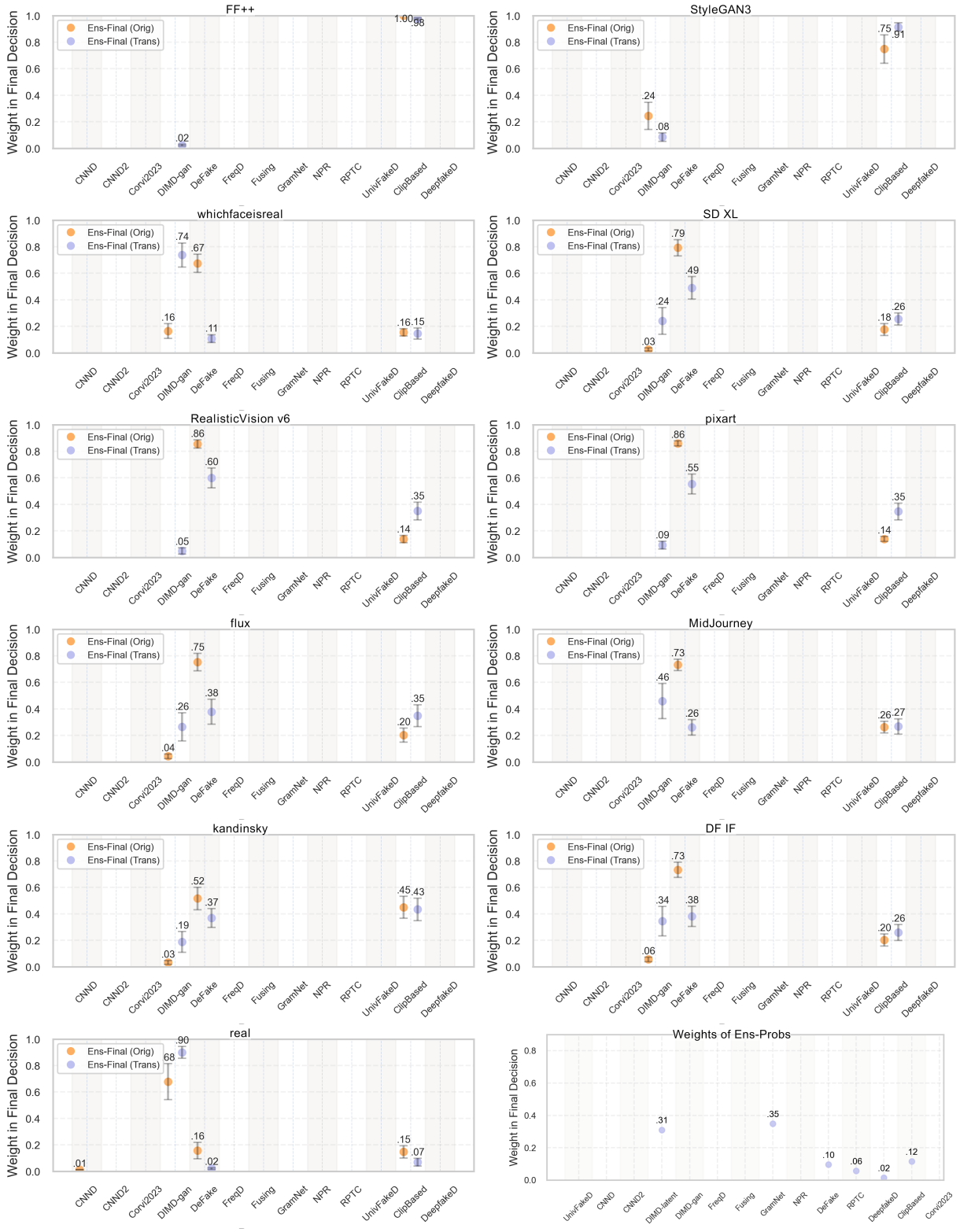


Figure 3. Importance weights of the full `Ens-Final` model calculated for both original (Orig) and degraded/transformed (Trans) images across individual generators and real test subsets. Means and standard deviations are computed over 100 random images. Static weights of `Ens-Probs` are shown for comparison.

kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. https://huggingface.co/black-forest-labs/FLUX.1-dev. 5

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. https://huggingface.co/PixArt-alpha/PixArt-XL-2-1024-MS. 5

[4] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. SimSwap: An efficient framework for high fidelity face swapping. In *Proc. ACM International Conference on Multimedia*. ACM, 2020. 2

[5] Civit AI. Realistic Vision v6, 2025. https://huggingface.co/SG161222/Realistic_Vision_V6.0_B1_noVAE. 5

[6] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023. 5

[7] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2, 5, 7

[8] Deep Floyd Lab. Deepfloyd, 2025. DeepFloyd/IF-I-XL-v1.0+DeepFloyd/IF-II-L-v1.0. 5

[9] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proc. International conference on machine learning (ICML)*. PMLR, 2020. 5

[10] Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, and Ben Y Zhao. Organic or diffused: Can we distinguish human art from ai-generated images? In *Proc. ACM SIGSAC Conference on Computer and Communications Security*, 2024. 2

[11] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Open CLIP, 2021. https://github.com/mlfoundations/open_clip/tree/v0.1. 5

[12] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized AI-synthesized image detection. In *Proc. International Conference on Image Processing (ICIP)*. IEEE, 2022. 5

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[14] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. DiffFace: diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163, 2025. 2

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision (ICCV)*, 2015. 5

[16] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proc. the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[17] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223, 2022. 2

[18] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[19] Gyana Panigrah, Prabira Sethy, Surya Prasada Borra, Nalini Barpanda, and Santi Behera. Deep ensemble learning for fake digital image detection: A convolutional neural network-based approach. *Revue d'Intelligence Artificielle*, 37, 2023. 3

[20] Nela Petrzelkova and Jan Cech. Detection of synthetic face images: Accuracy, robustness, generalization. In *Proc. DAGM German Conference on Pattern Recognition*, 2025. 2

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning (ICML)*. PmLR, 2021. 2, 3

[23] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022. 2

[24] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 5

[25] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. 5

[26] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proc. ACM SIGSAC conference on computer and communications security*, 2023. 5

[27] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[28] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[29] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[30] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[31] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. OpenSDI: Spotting diffusion-generated images in the open world. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[32] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[33] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37, 2024. 5

[34] Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2026. 2, 5

[35] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[36] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. 5

[37] Emre Şafak. Detection of fake face images using lightweight convolutional neural networks with stacking ensemble learning method. *PeerJ Computer Science*, 10: e2103, 2024. 3

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hraced, Czech Republic, February 9–12, 2026

Pi-GS: Sparse-View Gaussian Splatting with Dense π^3 Initialization

Manuel Hofer Markus Steinberger Thomas Köhler
Graz University of Technology
Austria
manuel.hofer@student.tugraz.at, steinberger@tugraz.at, t.koehler@tugraz.at

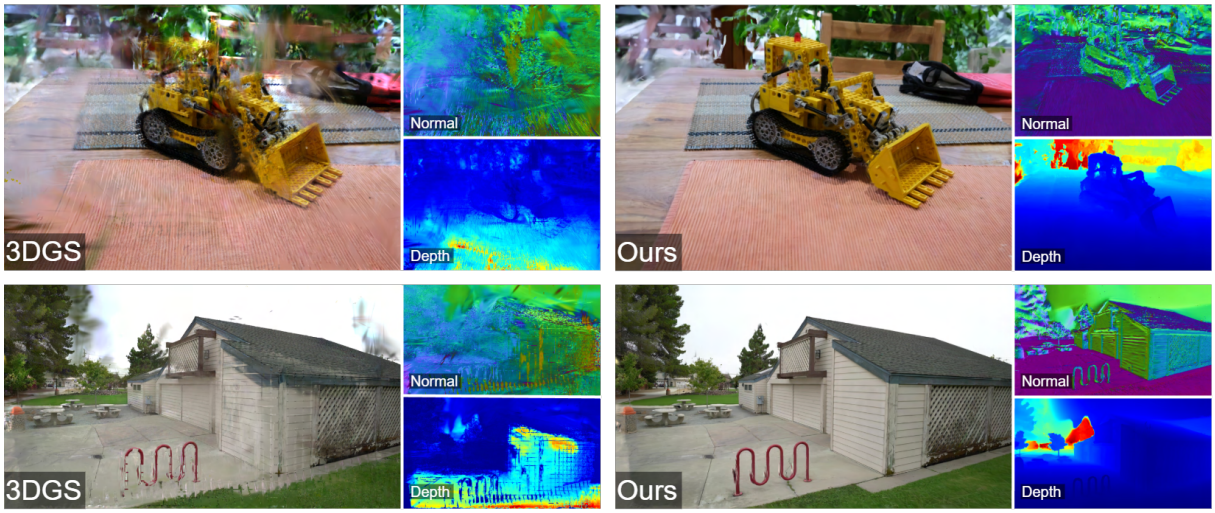


Figure 1. 3DGS exhibits floaters and view inconsistencies under sparse-view constraints. These artifacts are mostly caused by depth ambiguities and poor Gaussian alignment with the underlying geometry, as shown in the depth and normal maps. By incorporating depth supervision, normal supervision, and additional pseudo views, our method significantly reduces these artifacts and produces more view-consistent novel views with improved Gaussian alignment under sparse-view constraints.

Abstract

Novel view synthesis has evolved rapidly, advancing from Neural Radiance Fields to 3D Gaussian Splatting (3DGS), which offers real-time rendering and rapid training without compromising visual fidelity. However, 3DGS relies heavily on accurate camera poses and high-quality point cloud initialization, which are difficult to obtain in sparse-view scenarios. While traditional Structure from Motion (SfM) pipelines often fail in these settings, existing learning-based point estimation alternatives typically require reliable reference views and remain sensitive to pose or depth errors. In this work, we propose a robust method utilizing π^3, a reference-free point cloud estimation network. We integrate dense initialization from π^3 with a regularization scheme designed to mitigate geometric inaccuracies. Specifically, we employ uncertainty-guided depth supervision, normal consistency loss, and depth warping. Experimental results demonstrate that our approach achieves state-of-the-art performance on the Tanks and Temples, LLFF, DTU, and MipNeRF360 datasets.

1. Introduction

3D scene reconstruction and novel view synthesis (NVS) are rapidly advancing, with many applications across different domains [32]. These methods can be applied in fields such as Virtual Reality (VR) for creating immersive worlds, cinematography to create visually appealing assets efficiently, or robot vision to help robots understand their physical environment [23]. The foundation of 3D scene reconstruction was laid by traditional Structure from Motion pipelines. More recently, significant advances in NVS were achieved by representing the scene as Neural Radiance Fields (NeRF) [16]. These methods achieve state-of-the-art results but suffer from slow training speeds and are unsuitable for real-time rendering due to high latency. Newer methods such as 3D Gaussian Splatting (3DGS) [11] enable high-quality NVS even for real-time rendering. Additionally, training speed is significantly reduced.

A major limitation of these novel view synthesis methods is the need for dense views, which often is not feasible for real-world applications. In sparse-view settings,

these methods tend to struggle with bad initialization, depth ambiguities and overfitting to training views. To improve the performance in these settings and counteract the depth ambiguities, certain priors are introduced to better generalize and escape minima throughout the optimization process. Methods such as *DNGaussian* [14] and *Few-shot Novel View Synthesis using Depth* [13] leverage monocular depth estimators to regularize the model with the help of the inferred depth. The depth regularization helps significantly to improve the depth ambiguities and increase the generalization capability of the models. A challenge for these models is correct depth scaling, proper point initialization, and accurate camera poses. The initial points and camera poses are traditionally generated using Structure from Motion (SfM) pipelines. However, these pipelines often struggle with sparse input views and limited overlap between views. Recent advancements for sparse-view settings were achieved by leveraging dense initialization with the help of point cloud estimation networks [27, 28, 31]. They replace the traditional SfM pipeline with models such as *MASt3R* [7] or *DUSt3R* [24] for the point cloud estimation and camera pose estimation. The resulting models achieve high-fidelity results but require good initial reference views for accurate predictions. In addition, a time-consuming iterative camera alignment process is required, which can take several minutes. Inaccurate camera poses may further reduce reconstruction quality.

We make the following contributions:

- We discuss a method for leveraging a Permutation-Equivariant point cloud estimation network for dense initialization without relying on traditional SfM.
- We introduce confidence aware pearson depth loss, to counteract uncertain depth estimations.
- We explore the use of PGSR in sparse-view settings for improved geometry alignment and reduced overfitting.

Our method achieves state-of-the-art results in sparse-view settings and significantly improves Gaussian surface alignment, while reducing floaters. Our code is publicly available at https://github.com/Mango0000/Pi-GS.

2. Related Work

This section reviews prior work on 3D reconstruction, covering classical geometry-based pipelines, neural radiance fields, and Gaussian splatting approaches, with a focus on sparse-view and pose-free scenarios.

2.1. Traditional 3D Reconstruction

Classical 3D reconstruction pipelines typically rely on Structure-from-Motion (SfM) to achieve camera pose estimation and to generate a point cloud from a given set of images taken from various viewpoints. Afterward, Multi-View Stereo (MVS) and surface reconstruction techniques such as Poisson reconstruction are used [10, 20, 34]. These methods perform well in textured and opaque scenes but struggle with transparent materials and sparse or low-overlap views. Moreover, they are highly sensitive to SfM failures, which can lead to unstable surface reconstruction.

2.2. Neural Radiance Fields

Neural Radiance Fields (NeRF) [16] represent scenes by continuous volumetric functions. This makes them capable of producing photorealistic novel views and handling view-dependent effects more accurately. However, a downside is that NeRFs are quite demanding in terms of computation. As a result, we see more efficient variants like *Instant-NGP* [17], *PlenOctree* [30] and *Efficient-NeRF* [9] that drastically shorten the training and rendering time by incorporating optimized data structures and improving the architecture.

2.3. 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [11] has emerged as a new method that improves on training and rendering speed by replacing the implicit radiance field of NeRF-based methods with an explicit representation. Its core idea is to use 3D Gaussians both for optimization and for rendering via rasterization, therefore achieving real-time rendering without losing either fine details or transparency. Advanced 3DGS methods, such as *PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction* (PGSR) [4], improve Gaussian surface alignment with the help of planar Gaussians and multi-view consistency losses. However, these methods generally rely on SfM for initialization and are optimized for dense and overlapping views.

2.4. Sparse-View Gaussian Splatting

Reconstruction from sparse views remains a major challenge for 3DGS. Several augmentations exist that address sparse-view reconstruction by introducing additional constraints and regularization terms. Depth-based supervision is explored in *Depth-Regularized 3D Gaussian Splatting* [6], *Few-shot NVS with Depth-Aware 3D Gaussian Splatting* [13], and *DNGaussian* [14]. This type of supervision results in fast convergence and reduces depth ambiguities. Meanwhile, *DropGaussian* [18] and *DropoutGS* [29] deactivate Gaussians at random in order to counteract overfitting. There are also more advanced methods like FSGS: Real-Time Few-Shot View Synthesis using Gaussian Splatting [33] which introduces a pooling strategy and fine-tunes the splitting strategy to improve sparse view reconstruction across different datasets. While these methods achieve very robust results in sparse-view scenarios, they typically rely on accurate camera poses from SfM.

2.5. SfM-Free Methods

Methods such as *COLMAP-Free 3D Gaussian Splatting* [8] and *InstantSplat* [31], eliminate the need for SfM by jointly optimizing the 3D Gaussians as well as the camera poses and using depth estimations for point cloud initialization. These methods are able to handle sparse-view

situations more robustly and recover from inaccurate camera poses.

2.6. Diffusion-Based Priors

More recent works incorporate diffusion priors not only to stabilize the reconstruction, but also to generate additional views from the limited number of input views. *GenFusion* [26], *SparseGS* [28], *Gaussian Scenes* [19], and *Intern-GS* [27] are some of the methods where these advantages can also be observed. While these methods achieve impressive results, they often struggle with high-frequency textures and view inconsistencies due to depth ambiguities and inaccurate Gaussian alignment.

Our method differs fundamentally from diffusion-based and optimization-heavy approaches. Instead of synthesizing novel views using generative priors, we improve reconstruction quality through dense geometric initialization and strong generalizability across datasets. We leverage depth and normal supervision from estimated depth maps and explicitly model depth uncertainty through confidence-aware constraints, allowing deviations from noisy estimates. Camera poses and point representations are predicted by a feed-forward network, reducing reliance on iterative optimization and increasing robustness in sparse-view settings. Consequently, our approach focuses on geometric consistency and generalization without relying on view hallucination or diffusion-based priors.

3. Method

We begin by outlining preliminaries on Gaussian Splatting and planar depth rendering. Section 3.2 details modifications to PGSR for sparse settings, followed by our dense initialization strategy in Section 3.3. We then present our uncertainty-aware Pearson loss in Section 3.4 and artifact-free normal supervision in Section 3.5. Finally, Section 3.6 describes our depth warping approach for improving view consistency.

3.1. Preliminaries

Gaussian Splatting. *3D Gaussian Splatting* (3DGS) introduced by Kerbl et al. [11] achieves great novel view synthesis results with high efficiency by leveraging a Gaussian scene representation. Another improvement of this scene representation over NeRF is the real-time rendering speed, as well as much faster training times. Our approach also builds upon 3DGS. The scene representation is defined by a set of 3D Gaussians. Each Gaussian can be defined by a 3D covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$ and the 3D center point $\mu \in \mathbb{R}^3$ in world space,

$$G(x) = e^{-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)}. \tag{1}$$

To project this 3D Gaussian onto the 2D image plane for rendering, the covariance matrix Σ' in clip space is defined as the following:

$$\Sigma' = JW\Sigma W^T J^T, \tag{2}$$

where J is the Jacobian of the affine approximation for this projection transformation and W is the view transformation matrix.

For the covariance matrix to be physically meaningful, it needs to be positive semi-definite. To ensure this throughout the training process, Σ is defined as the following:

$$\Sigma = RSS^T R^T, \tag{3}$$

where $S \in \mathbb{R}^{3\times3}$ is the scaling matrix, and $R \in \mathbb{R}^{3\times3}$ is the rotation matrix. This allows separate optimization of rotation and scaling and ensures that Σ is positive semi-definite. For increased memory efficiency, the rotation matrix is stored as a quaternion, and scaling as 3D vector.

Furthermore, for rendering the color C, we blend the colors of each Gaussian along the ray, as follows:

$$C = \sum_{i=1}^{N} T_i \alpha_i c_i, \tag{4}$$

where N is the number of Gaussians along a ray, c_i is the color of the i-th Gaussian represented by spherical harmonics (SH) to account for view dependent effects, α_i is the weighted opacity of the i-th Gaussian and T_i is the transmittance of the i-th Gaussian [11].

Transmittance T_i is defined as:

$$T_i = \prod_{j=1}^{i-1}(1-\alpha_j). \tag{5}$$

By calculating the color for each ray from the camera, we can render an image. The training of this Gaussian representation is done by back propagation with the following loss function:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{D-SSIM}, \tag{6}$$

where $\mathcal{L}_1$ is a simple l_1 loss between the rendered and ground-truth image and $\mathcal{L}_{D-SSIM}$ is an image similarity measure between rendered and ground-truth image [2, 11]. 3DGS relies on camera poses and points obtained from structure from motion (SfM). However, in sparse-view settings, the resulting point cloud can be highly sparse, and the overlap between the images may be insufficient to extract reliable structures or accurate camera poses. This leads to a challenging starting point for 3DGS optimization.

Depth and Normal Rendering. We use *Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction* [4] (PGSR) for normal and depth rendering. PGSR builds upon 3DGS, enabling the rendering and backpropagation of both the depth and normals. A naive

approach of computing the depth D of a pixel would be to use depth accumulation defined as:

$$D = \sum_{i=1}^{N} T_i \alpha_i z_i, \tag{7}$$

where T_i is the same as in Eq. (5), α_i is the weighted opacity of the i-th Gaussian and z_i is its distance from the camera [5]. PGSR on the other hand compresses the 3D Gaussians to get flat 2D planes, from which unbiased depth and normal maps can be rendered [4].

To get the 2D planes, PGSR flattens the 3D Gaussians by minimizing the minimum scale and therefore defining the scale loss $\mathcal{L}_s$ as following:

$$\mathcal{L}_s = ||\min(s_1, s_2, s_3)||_1, \tag{8}$$

where s_i is the i-th scale component of each Gaussian.

The direction of the minimum scale factor corresponds to the normal n_i. Therefore, the normals per ray, $\mathcal{N}$, can be rendered as following:

$$\mathcal{N} = \sum_{i=1}^{N} R_c^T n_i \alpha_i T_i, \tag{9}$$

where R_c is the rotation from the camera to the global world.

The distance d_i from the Gaussian plane to the camera center is defined as:

$$d_i = (R_c^T(\mu_i - T_c))R_c^T n_i^T, \tag{10}$$

where T_c is the camera center in the world and μ_i is the center of the i-th Gaussian.

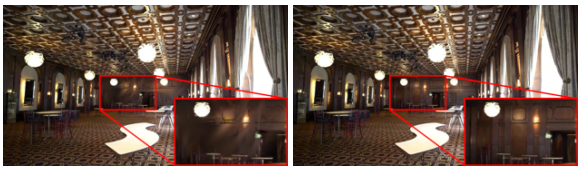
The distance D along a ray can now be defined as:

$$D = \sum_{i=1}^{N} d_i \alpha_i T_i. \tag{11}$$

PGSR extends 3DGS by introducing an Image Edge-Aware Single-View Loss $\mathcal{L}_{svgeo}$, which optimizes the Gaussian Scene with the Local Plane Assumption. This assumption states that two neighbouring pixels can be considered as an approximate local plane, but only if these pixels do not belong to an edge. The loss helps to improve the local depth and normal consistency. They also propose a Multi-View Geometric Consistency Loss, $\mathcal{L}_{mvgeom}$, which enhances geometric smoothness by projecting the depth and normals from one frame to another. Finally, they employ a Multi-View Photometric Consistency Loss, $\mathcal{L}_{mvrgb}$, which projects the grayscale image from one camera to another camera through depth warping [4].

3.2. PGSR Sparse-View

Default PGSR does not work well for the sparse-view setting out-of-the-box because of the multi-view observer trim, which assures that each point is observed by multiple cameras and this is not guaranteed in sparse-view settings. Therefore, we deactivate this trimming for our method. Another parameter that requires adjustment is the opacity reset interval. When opacity reset happens, fine details in the background will be lost and artifacts appear, as can be seen in Fig. 2. The details in Fig. 2a at the back wall are completely lost and artifacts in the window frame become visible. By continuing the training process even further, the artifacts' strength increases, and they become even more prominent. When deactivating opacity reset, the background details are retained and the artifacts vanish without sacrificing the overall quality. This can also be seen in Fig. 2b. The improvement is also reflected in the PSNR (Peak Signal-to-Noise Ratio), which increases from *22.76* to *23.73*. With these few settings, it is already possible to run the PGSR framework with acceptable results. For improved performance, we deactivate the splitting strategy as it is not needed for our dense point cloud initialization. The point cloud is already very detailed and this setting does not improve the final results (*cf.* Tab. 1).



(a) Ballroom scene with opacity reset.

(b) Ballroom scene without opacity reset.

Figure 2. Comparison of the Ballroom scene from Tanks and Temples with and without opacity reset [12]. Background details are lost when opacity reset is executed, and image quality further degrades over the training process.

3.3. Dense Initialization

Sparse-view settings pose a fundamental challenge for standard SfM frameworks like COLMAP [21, 22], where limited image overlap can lead registration to fail. Furthermore, the resulting sparse point clouds serve as a poor initialization for 3DGS, complicating the optimization of Gaussian primitives and compromising geometric fidelity. To mitigate this, we leverage a pre-trained feed-forward network to predict both depth and camera parameters. This strategy provides the dense geometric initialization and accurate poses required for high-quality sparse-view reconstruction. Figure 3a illustrates the point cloud generated by the feed-forward model π^3 [25], while Fig. 3b depicts the result from COLMAP [21]. Both methods use the same 24 input views from the "bicycle" scene of the MipNeRF360 dataset [3], rendered here from an identical viewpoint. The difference in density is significant: The COLMAP reconstruction contains only 1,028 points, whereas π^3 yields 1,013,106 points. Note that the π^3 output was filtered using the default confidence threshold of 20%.

(a) Point cloud inferred with π^3. (b) Point cloud created with COLMAP.

Figure 3. Comparison between π^3 point cloud and COLMAP point cloud, of the bike scene from MipNeRF360 Dataset with 24 training images [3, 21, 25].

3.4. Depth Supervision

From π^3, we obtain the per view point clouds which can be used as a depth map. For depth regularization, we evaluated different losses.

Standard L1 and L2 losses often cause the model to overfit to the limited fidelity of the inferred depth maps. We also evaluated the Global-Local Depth Normalization from *DNGaussian* [14] but found it unnecessary given the inherent scale consistency of our predictions. Instead, we utilize a Pearson correlation loss, which has demonstrated superior performance. This approach enforces structural consistency while enabling the recovery of high-frequency details that are missing from the initial depth estimation.

In addition to the default Pearson correlation loss, we also integrated the confidence given by π^3. As a result, the final depth can be modeled even more accurately by assigning low weights to uncertain regions. Our newly created confidence-aware depth loss, $\mathcal{L}_{pearson}$, is defined as:

$$\mu_p = \frac{\sum_{i=1}^{N} C_i D_i^p}{\sum_{i=1}^{N} C_i}, \quad \mu_t = \frac{\sum_{i=1}^{N} C_i D_i^t}{\sum_{i=1}^{N} C_i}, \tag{12}$$

$$\bar{D}_p = D_p - \mu_p, \quad \bar{D}_t = D_t - \mu_t, \tag{13}$$

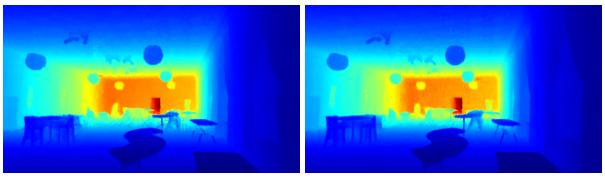
$$P_{conf} = \frac{\sum_{i=1}^{N} C_i \bar{D}_i^p \bar{D}_i^t}{\sqrt{\left(\sum_{i=1}^{N} C_i \left(\bar{D}_i^p\right)^2\right)\left(\sum_{i=1}^{N} C_i \left(\bar{D}_i^t\right)^2\right)}}, \tag{14}$$

$$\mathcal{L}_{pearson} = 1 - P_{conf}, \tag{15}$$

N is the number of pixels, D_i^p is the predicted depth of the i-th pixel, C_i the confidence of the i-th pixel and D_i^t is the ground truth of the i-th pixel, which is the depth estimated by π^3, and P_{conf} is the confidence-aware Pearson correlation. The resulting rendered depth after 7,000 iterations with the help of confidence-aware Pearson correlation can be seen in Fig. 4.

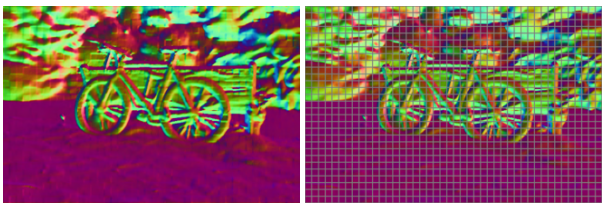
3.5. Normal Supervision

Surface Normals can be computed with the help of depth maps by calculating the pixel-wise partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$, where x and y are the pixel coordinates and z is the depth value, either rendered or estimated by π^3.



(a) Confidence-aware Peason loss. (b) Pearson loss.

Figure 4. Depth rendering of the Ballroom scene from the Tanks and Temples dataset, comparing the confidence-aware Pearson loss with the standard pearson loss [12]. The confidence-aware loss leverages uncertainty estimates to enhance detail, particularly in the background, and also improves performance with low-resolution depth estimates.



(a) Default normal map with artifacts. (b) Masked normal map.

Figure 5. The Normal map generated from the depth map using partial derivatives, which introduces grid artifacts, and the masked normal map, which removes grid artifacts introduced by π^3 architecture [25].

Because π^3 processes each image in patches of 14×14 pixels, the gradient is not continuous between adjacent patches, leading to grid-like artifacts, as can be seen in Fig. 5a. To alleviate this problem, we add a mask to ignore these discontinuous regions during loss computation. The mask is computed by creating a grid with 14×14 pixel cells, masking the 1-pixel-wide inner border of each cell. Therefore, the Gaussians are not regularized in these border regions, and the grid artifacts do not appear in the scene representation. The masked normal map can be seen in Fig. 5b. As supervision, we simply use the L1 loss between the rendered and ground-truth normal map defined as:

$$\mathcal{L}_{normal} = \frac{1}{N} \sum_{i=1}^{N} ||N_i^t - N_i^p||_1, \tag{16}$$

where N is the number of pixels, N_i^t is the ground-truth normal at pixel i and N_i^p is the predicted normal at pixel i.

3.6. Depth Warping

To improve generalization of our model further, we include pseudo-views which are generated with the help of depth warping. This is achieved by projecting the image pixels from one camera into 3D space, and then reprojecting the 3D points into the 2D image plane of a target camera. For accurate results, we only project pixels with high confidence and mask out the rest, including unseen regions. To generate high-quality pseudo-cameras, we use circle interpolation with the camera parameters as input. A circle can be defined by three points, so we use

Figure 6. The two Figures show two reprojection examples with the applied mask for the Barn scene of the Tanks and Temples dataset [12].

the two nearest cameras to the target camera for pseudo-view generation. The positions of the three cameras define our circle. Now then interpolate by a certain amount between each pair of neighbouring views, which results in two additional views per camera. We can generate an arbitrary number of pseudo-views by adjusting the interpolation step size. However, in our experiments, two pseudo-views between each pair yielded the best results. The nearest cameras are already computed by PGSR, therefore we can reuse them. A few examples of these generated pseudo-views can be seen in Fig. 6. These pseudo-views are then used throughout training for additional supervision with the help of SSIM and L1 loss, but with a weight set to *0.1*.

4. Evaluation

For testing strategy, we adhere to previous state-of-the-art models to ensure comparability. The datasets used for the evaluation are Tanks and Temples [12], MipNeRF360 [3], LLFF [15] and DTU [1].

Implementation Details. The Tanks and Temples dataset covers real-world indoor and outdoor scenes, but we only use a subset of 8 scenes, as done by other sparse-view models like *Intern-GS* and *InstantSplat*. We focus on the 3-view setting and therefore use the same train/test split. This means the testing set includes 12 images uniformly sampled without the first and last frame and the remaining set is the training set where we again uniformly sampled the 3 views [31]. For Tanks and Temples, no downsampling is applied.

The MipNeRF360 dataset contains real-world 360° indoor and outdoor scenes. For this dataset, two different approaches are used. One for the 3-view setting as defined by *Gaussian Scenes* [19] and one for the 12-view setting as defined by *SparseGS* [28]. For both settings, the 4x downsampled images are used, to adhere to the evaluation strategies of state-of-the-art models. For the 3-view setting, we use every 8th image as testing set and uniformly sample the 3 training views. For the 12-view setting, we use the split dataset provided by *SparseGS* [28]. The 12-view setting uses only 6 of the 9 scenes contained in the MipNeRF360 dataset, whereas the 3-view setting uses all 9 scenes.

The LLFF dataset contains real-world forward-facing images. For this dataset, we used the same evaluation strategy as defined by *DNGaussian*. A downsampling rate of 8 is used, and we adhere to the train/test split of the 3-view setting of *DNGaussian* [14].

Lastly, we also evaluated on the DTU dataset, which contains highly calibrated lab captures of object centric scenes. This dataset also provides bit masks to separate the background and real camera poses. We used our own inferred camera poses. We again used the testing strategy defined by *DNGaussian*. This time we used 4x downsampled images and the same train/test split of the 3-view setting of *DNGaussian* [14]. Similar to DNGaussian and other comparable methods, we applied the provided separation masks for the evaluation.

We use the exact same settings for all evaluations. π^3 [25] automatically downsamples the images to a certain pixel size, therefore we counteract the downsampling by rescaling the cameras to the full size. To make a fair comparison, we only project the training views to 3D space. The testing views are only used to get initial camera positions. We train for 7000 iterations, with depth loss, normal loss as well as pseudo views. The pseudo views are generated with a confidence threshold of 20%. This means that we mask out the projected pixel with confidence under 20%. Splitting of Gaussians is deactivated. We evaluate our model in terms of PSNR, SSIM and LPIPS.

4.1. Quantitative Evaluation

Tables 3 and 6 show the comparison between *Intern-GS* [27], *InstantSplat* [31], *SparseGS* [28], *DNGaussian* [14], *FSGS* [33], 3DGS [11] and Our method. On DTU and Tanks and Temples, our model can reconstruct the scene accurately, with good Gaussian surface alignment and without smoothing out high-frequency textures. On LLFF our model achieves slightly lower scores, because of missing information in unseen regions, as our model optimizes only on seen regions and known information. An example of this unseen region is illustrated in Fig. 7.

Tab. 4 shows the comparison between Gaussian Scenes, MASt3R Initialization, FSGS and Our method in the 3-view setting on MipNeRF360 [19, 33]. Our model achieves the lowest LPIPS score and second highest PSNR and SSIM. Compared to FSGS our model does not rely on accurate camera poses from traditional SfM.

Tab. 5 shows the comparison between 3DGS, *DNGaussian*, *SparseGS* and Our method in 12-view setting on MipNeRF360 [3]. Our model achieves the highest results with very coherent and view-consistent final scene, as our model improves the Gaussian surface alignment significantly. A comparison can be seen in Fig. 8.
To validate the accuracy of our camera pose estimates, we evaluate the Absolute Trajectory Error (ATE) on the Tank and Temples dataset. Our pose estimator, π^3, achieves a mean ATE of 0.0293 and a root mean squared error (RMSE) of 0.0325, demonstrating that it produces accurate camera poses suitable for fair comparison of photo-

(a) Ground Truth

(b) Intern-GS

(c) Ours

Figure 7. The Figures show a comparison between Intern-GS [27], Ours and the Ground Truth. Our model has very accurate reflections and fewer artifacts, nevertheless our model can not correctly reconstruct the unseen region at the ceiling.

(a) SparseGS

(b) Ours

Figure 8. The Figures show a comparison between SparseGS [28] and our method. Our model reconstructs the background and ground more accurately, and additionally decreases artifacts.

metric metrics in 3D Gaussian splatting.

4.2. Ablation

We evaluate the impact of each individual optimization on our final result. The evaluation is conducted using the Barn scene from the Tanks and Temples dataset. It is evident that all of our optimizations improve the result even further. Dense point cloud initialization with the help of π^3 significantly improves the result by also reducing the time required for SfM. Our custom depth loss improves the score by allowing low confidence depth regions to optimize more freely. Normal regularization encourages the Gaussians' normals to match the ground-truth geometry. Depth warping improves the results by adding more views, which helps the model generalize better and avoid overfitting to the training views. Our full model achieves a PSNR of *22.15* on the Barn scene. We also evaluated the effect of enabling splitting of Gaussians in our model. This setting results in a slight decrease in performance and was therefore deactivated. These results can be seen in Tab. 1.

Method	PSNR
Original 3DGS	17.53
PGSR	18.05
π^3 (dense) initialization	19.66
+ Depth Regularization	20.72
+ Normal Regularization	21.56
+ Depth Warping (Full Model)	22.15
+ Splitting Densification	21.97

Table 1. Ablation study of the regularization techniques introduced in our model on the Barn scene of Tanks and Temples. Additionally, we evaluated the impact of splitting Gaussians during densification.

In addition, we evaluate the impact of using PGSR compared to standard 3DGS for our sparse view setting (3-views). Table 2 shows that the planar depth created by PGSR helps significantly to place the Gaussians more accurately. Additionally, the losses introduced by PGSR help to improve the rendering results further. Our model remains stable even after increased training iterations and continues to show improved novel view synthesis results. A visual comparison between 3DGS and PGSR with different number of iterations can be seen in Fig. 9.

Framework	Iteration	Tanks and Temples			MipNeRF360		
		PSNR$^\uparrow$	SSIM$^\uparrow$	LPIPS$^\downarrow$	PSNR$^\uparrow$	SSIM$^\uparrow$	LPIPS$^\downarrow$
PGSR [4]	7000	19.99	0.503	0.355	23.36	0.791	0.156
3DGS [11]	7000	18.00	0.426	0.449	23.07	0.773	0.172
PGSR [4]	15000	20.19	0.517	0.343	23.41	0.795	0.169
3DGS [11]	15000	17.04	0.391	0.465	20.94	0.719	0.244

Table 2. Ablation study on the use of PGSR as base framework compared to 3DGS. The additional multi-view and single-view losses introduced by PGSR are activated after iteration 7000. This comparison shows that the PGSR depth rendering captures the underlying surface geometry more accurately by also reducing floaters significantly. With the help of PGSR we achieve view-consistent surfaces and reduce overfitting significantly.

Method	PSNR$^\uparrow$	SSIM$^\uparrow$	LPIPS$^\downarrow$
3DGS [11]	15.36	0.572	0.379
DNGaussian [14]	20.69	0.721	0.277
SparseGS [28]	21.20	0.717	0.231
InstantSplat [31]	22.20	0.743	0.199
FSGS [33]	22.31	0.693	0.197
Intern-GS [27]	22.67	0.736	0.191
Ours	22.87	0.764	0.189

Table 3. Evaluation on Tanks and Temples dataset with 3-view setting. Our model does not oversmooth high-frequency textures and accurately aligns the Gaussians with the underlying surface geometry.

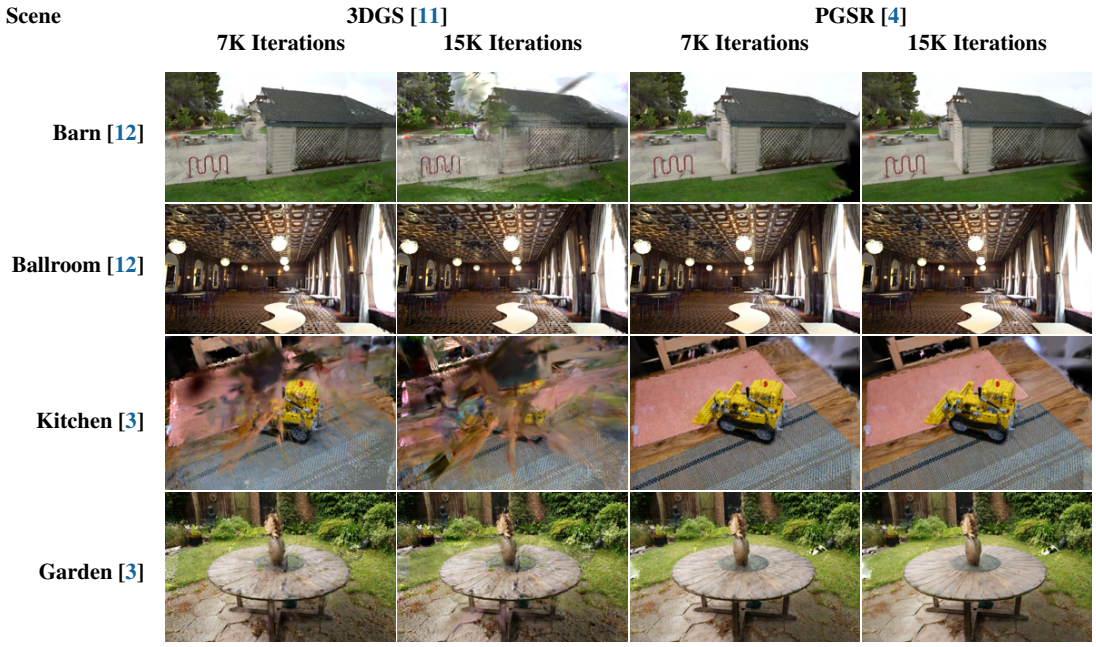


Figure 9. Visual comparison between 3DGS and PGSR with different Iterations. The planar depth from PGSR helps significantly to remove floaters and align the Gaussians accurately to the ground-truth geometry.

Method	PSNR↑	SSIM↑	LPIPS↓
MASt3R Initialization [19]	12.59	0.231	0.593
Gaussian Scenes [19]	13.81	0.265	0.547
FSGS [33]	14.17	0.318	0.578
Ours	14.14	0.310	0.523

Table 4. Evaluation on MipNeRF360 dataset with 3-view setting. Our model reconstructs seen regions accurately, but can not introduce geometry in unseen regions.

Method	PSNR↑	SSIM↑	LPIPS↓
3DGS [11]	17.49	0.490	0.431
DNGaussian [14]	16.28	0.432	0.549
SparseGS [28]	19.37	0.577	0.398
Ours	19.54	0.492	0.362

Table 5. Evaluation on MipNeRF360 dataset with 12-view setting. Our model can reconstruct the scenes with highly accurate surface alignment. The ground is view-consistent, and fewer floating artifacts compared to SparseGS [28].

5. Conclusion and Limitations

Our model shows strong performance under sparse-view constraints, specifically when handling between 3 and 12 views. The model demonstrates the importance of accurate dense point cloud initialization. We introduce a modified depth loss that enables correct scene generalization by reducing depth ambiguities without introducing artifacts in low confidence regions. In addition, we introduce normal and depth warping loss terms that improve alignment with the ground-truth surface geometry. Finally, we relax certain assumptions from PGSR to allow robust optimization in sparse-view settings.

Method	LLFF			DTU		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS [11]	15.52	0.408	0.405	10.99	0.585	0.313
DNGaussian [14]	19.12	0.591	0.294	18.91	0.790	0.176
SparseGS [28]	19.86	0.668	0.322	18.89	0.834	0.178
InstantSplat [31]	17.67	0.603	0.379	17.55	0.634	0.212
FSGS [33]	20.31	0.652	0.288	19.54	0.732	0.199
Intern-GS [27]	20.49	0.693	0.212	20.34	0.851	0.163
Ours	19.92	0.664	0.254	23.52	0.815	0.145

Table 6. Evaluation on LLFF and DTU dataset with 3-view setting. Following previous work, for evaluation on DTU the background masks are applied. Our model is able to reconstruct fine-grained textures accurately, but it underperforms in unobserved regions compared to methods that generate content for unseen regions.

Our model faces limitations when dealing with large datasets, as processing many input views with π^3 consumes a large amount of GPU memory, which is infeasible on consumer hardware. Additional limitations come from inaccurate depth estimations in specific scenes, such as the leaves scene from the LLFF dataset [15]. Future improvements could include the joint optimization of the camera poses and the Gaussian scene, which would result in improved reconstruction quality. Furthermore, the integration of generative priors could enhance the model's ability to maintain photometric and geometric consistency across occluded or sparse areas.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-Scale Data for Multiple-View Stereopsis. *IJCV*, pages 1–16, 2016. 6

[2] Allison H. Baker, Alexander Pinard, and Dorit M. Hammerling. On a Structural Similarity Index Approach for Floating-Point Data. *IEEE TVCG*, 30(9):6261–6274, 2024. 3

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Proc. CVPR*, pages 5470–5479, 2022. 4, 5, 6, 8

[4] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-Based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *IEEE TVCG*, 2024. 2, 3, 4, 7, 8

[5] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. GaussianPro: 3D Gaussian Splatting with Progressive Propagation. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024. 4

[6] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-Regularized Optimization for 3D Gaussian Splatting in Few-Shot Images. In *Proc. CVPRW*, pages 811–820, 2024. 2

[7] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion. In *Proc. 3DV*, 2025. 2

[8] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. COLMAP-Free 3D Gaussian Splatting. In *Proc. CVPR*, pages 20796–20805, 2024. 2

[9] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. EfficientNeRF: Efficient Neural Radiance Fields. In *Proc. CVPR*, pages 12902–12911, 2022. 2

[10] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proc. SGP*, 2006. 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM TOG*, 2023. 1, 2, 3, 6, 7, 8

[12] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM TOG*, 36(4), 2017. 4, 5, 6, 8

[13] Raja Kumar and Vanshika Vats. Few-Shot Novel View Synthesis Using Depth Aware 3D Gaussian Splatting. In *ECCV 2024 Workshops*, pages 1–13, Cham, 2025. Springer Nature Switzerland. 2

[14] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. DNGaussian: Optimizing Sparse-View 3D Gaussian Radiance Fields with Global-Local Depth Normalization. In *Proc. CVPR*, pages 20775–20785, 2024. 2, 5, 6, 7, 8

[15] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *ACM TOG*, 2019. 6, 8

[16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1): 99–106, 2021. 1, 2

[17] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM TOG*, 41(4): 102:1–102:15, 2022. 2

[18] Hyunwoo Park, Gun Ryu, and Wonjun Kim. DropGaussian: Structural Regularization for Sparse-view Gaussian Splatting. In *Proc. CVPR*, pages 21600–21609, 2025. 2

[19] Soumava Paul, Prakhar Kaushik, and Alan Yuille. Gaussian Scenes: Pose-Free Sparse-View Scene Reconstruction using Depth-Enhanced Diffusion Priors. *arXiv preprint arXiv:2411.15966*, 2024. 3, 6, 8

[20] Fabio Remondino, Ali Karami, Ziyang Yan, Gabriele Mazzacca, Simone Rigon, and Rongjun Qin. A Critical Analysis of NeRF-Based 3D Reconstruction. *Remote Sensing*, 15(14), 2023. 2

[21] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proc. CVPR*, 2016. 4, 5

[22] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Proc. ECCV*, 2016. 4

[23] Hamid Taheri and Zhao Chun Xia. SLAM; definition and evolution. *Engineering Applications of Artificial Intelligence*, 97:104032, 2021. 1

[24] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *Proc. CVPR*, 2024. 2

[25] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*, 2025. 4, 5, 6

[26] Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *Proc. CVPR*, pages 6078–6088, 2025. 3

[27] Sun Xiangyu, Chen Runnan, Gong Mingming, Xu Dong, and Liu Tongliang. Intern-GS: Vision Model Guided Sparse-View 3D Gaussian Splatting. *arXiv preprint arXiv:2505.20729*, 2025. 2, 3, 6, 7, 8

[28] Haolin Xiong, Sairisheek Muttukuru, Hanyuan Xiao, Rishi Upadhyay, Pradyumna Chari, Yajie Zhao, and Achuta Kadambi. SparseGS: Sparse View Synthesis Using 3D Gaussian Splatting. In *Proc. 3DV*, pages 1032–1041, 2025. 2, 3, 6, 7, 8

[29] Yexing Xu, Longguang Wang, Minglin Chen, Sheng Ao, Li Li, and Yulan Guo. DropoutGS: Dropping Out Gaussians for Better Sparse-view Rendering. In *Proc. CVPR*, pages 701–710, 2025. 2

[30] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for Real-Time Rendering of Neural Radiance Fields. In *Proc. ICCV*, pages 5752–5761, 2021. 2

[31] Fan Zhiwen, Wen Kairun, Cong Wenyan, Wang Kevin, Zhang Jian, Ding Xinghao, Xu Danfei, Ivanovic Boris,

Pavone Marco, Pavlakos Georgios, Wang Zhangyang, and Wang Yue. InstantSplat: Sparse-view Gaussian Splatting in Seconds. *arXiv preprint arXiv:2403.20309*, 2024. 2, 6, 7, 8

[32] Yiming Zhou, Zixuan Zeng, Andi Chen, Xiaofan Zhou, Haowei Ni, Shiyao Zhang, Panfeng Li, Liangxi Liu, Mengyao Zheng, and Xupeng Chen. Evaluating Modern Approaches in 3D Scene Reconstruction: NeRF vs Gaussian-Based Methods. In *Proc. DOCS*, pages 926–931, 2024. 1

[33] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *Proc. ECCV*, page 145–163, Berlin, Heidelberg, 2024. Springer-Verlag. 2, 6, 7, 8

[34] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A Survey of Structure from Motion. *Acta Numerica*, 26:305–364, 2017. 2

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hraced, Czech Republic, February 9–12, 2026

Dense Spatiotemporal Reconstruction of Sea Surface Temperature with Conditional Flow Matching

Grega Rovščček[1] **Matjaž Ličer**[2] **Matej Kristan**[1]
[1]Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[2]Slovenian Environment Agency, Ljubljana, Slovenia
grega.rovscek@fri.uni-lj.si

Abstract

Sea surface temperature (SST) reconstruction of missing values in cloud-covered regions is crucial for many downstream geophysics tasks. It is fundamentally ambiguous, yet most state-of-the-art methods remain deterministic, producing a single overly-smoothed reconstruction, often contaminated by artifacts, or offering a limited uncertainty estimation. We present DIRECT, a conditional generative model that reconstructs dense SST fields by learning the full conditional distribution of plausible solutions. DIRECT proposes a rectified flow-matching formulation, with the network conditioned on temporal context and day-of-year seasonality, and presents an observation-guided rectification that anchors the generative trajectory to measured pixels at every integration step. DIRECT produces an ensemble of physically consistent reconstructions, yielding both an accurate mean estimate and spatially resolved uncertainty that is well-calibrated via a lightweight post-hoc variance correction. DIRECT sets a new state-of-the-art on three Level-3 SST datasets (Mediterranean, Adriatic, Atlantic), delivering 6–17% improved reconstruction compared to the strongest published method and better preserving the mesoscale structure.

1. Introduction

Sea surface temperature (SST) is a key variable at the ocean–atmosphere interface, regulating exchanges of heat, moisture, and momentum and influencing atmospheric convection, storm development, and large-scale climate variability such as ENSO and monsoon systems [12, 18, 26, 27]. This designates SST as an essential climate variable [5], widely used in climate monitoring, numerical weather prediction, ocean reanalysis, and marine ecosystem studies [8, 28].

Global SST observations are primarily obtained from satellite-based infrared (IR) and microwave radiometers on low-Earth-orbit and geostationary platforms [24]. While IR sensors provide high spatial resolution, they are severely affected by cloud cover, which obscures large and spatially coherent portions of the ocean surface at any given time. This results in structured gaps that persist across space and time, posing a fundamental challenge for downstream applications that require spatially complete SST fields (Figure 1, top-left).

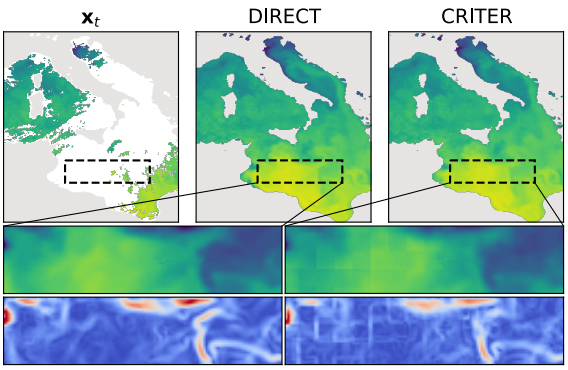


Figure 1. DIRECT recovers spatially coherent temperature fields from partially observed inputs and preserves smooth mesoscale structure in reconstructed regions. Compared to CRITER [34] (current state-of-the-art), DIRECT delivers more faithful patterns, which is further demonstrated by the spatial gradient magnitudes – notice that the blocky high-frequency structures emerging in CRITER are not present in DIRECT estimates.

Classical SST reconstruction methods based on statistical interpolation or low-rank decompositions (e.g., Optimal Interpolation, EOFs [1, 31]) recover large-scale variability but are limited by linearity and stationarity assumptions. Recent learning-based models, which range from convolutional autoencoders such as DINCAE/DINCAE2 [2, 3] to transformer-based methods like MAESSTRO [14] and CRITER [34], have substantially improved accuracy, but remain fundamentally *deterministic*, producing a single reconstruction that struggles to capture ambiguity under heavy cloud cover, complicates uncertainty estimation, and often smooths fine-scale structure or introduces long-range artifacts (Figure 1, top-right).

Acknowledgements. This work was supported by the ARIS program P2-0214 and projects L2-3169, J2-60054, and by the Slovenia supercomputing network SLING (ARNES, EuroHPC Vega - IZUM).

Diffusion-based generative models [7, 17, 33] have achieved state-of-the-art performance in image restoration by learning full conditional distributions rather than point estimates. In geophysical reconstruction they have been recently explored for satellite gap-filling and data assimilation [4, 32]. However, existing approaches typically operate on *single snapshots* without explicit spatiotemporal context [4], rely on *coarse or implicit conditioning* rather than structured temporal representations [32], or lack explicit mechanisms for handling *physical masks (e.g., land) and uncertainty calibration* [4]. Consequently, a principled generative formulation tailored to dense, spatiotemporal SST reconstruction remains an open challenge.

We propose DIRECT, a diffusion-inspired generative model for dense SST reconstruction from partially observed satellite measurements. DIRECT formulates SST gap-filling as a *conditional flow-matching* problem, learning a deterministic probability flow that maps noise to physically plausible SST fields conditioned on sparse observations, temporal context, and seasonal information. The model integrates (i) a unified U-Net backbone with SST-specific global conditioning, (ii) a spatiotemporal context representation that explicitly encodes the temporal origin of auxiliary information, and (iii) a task-specific rectification mechanism that enforces hard consistency with observed pixels throughout the generative process. By sampling multiple reconstructions, DIRECT produces both an accurate mean estimate and spatially resolved uncertainty, which is further automatically calibrated to avoid ensemble collapse. Our contributions are:

- We formulate dense SST reconstruction as a *conditional flow-matching* problem, integrating spatiotemporal context, seasonal conditioning, and hard observation consistency within a single end-to-end generative model.
- We introduce an explicit *temporal offset encoding* for auxiliary context frames, together with a multi-pass reconstruction strategy that iteratively refines the spatiotemporal context and the resulting reconstructions.
- We propose an effective *post-hoc uncertainty calibration* scheme that corrects ensemble under-dispersion and yields well-calibrated per-pixel uncertainty estimates without modifying the training objective.

DIRECT sets a new state-of-the-art on three multi-sensor SST benchmarks, reducing reconstruction error in occluded regions by 6–17% relative to the strongest existing baseline (CRITER [34]) and by up to 63% compared to other recent methods, while better preserving mesoscale spatial structure and producing well-calibrated uncertainty estimates.

2. Related work

Early SST reconstruction methods based on statistical interpolation and low-rank decompositions (e.g., Optimal Interpolation, EOFs [1, 31]) recover large-scale variability but are limited by linearity and stationarity assumptions. Learning-based models substantially improve reconstruction quality: convolutional autoencoders such as DINCAE/DINCAE2 [2, 3] remain overly smooth despite providing uncertainty surrogates, while transformer-based methods like MAESSTRO [14] and CRITER [34] better capture long-range structure but remain fundamentally *deterministic*, producing a single reconstruction that is limited in ambiguity representation and may introduce structured artifacts. In contrast, DIRECT models the full conditional distribution and produces ensembles of plausible reconstructions.

Inpainting by diffusion models has ben extensively explored in computer vision. The methods achieve state-of-the-art image restoration by learning full conditional distributions [7, 17, 33]. Pixel re-injection as in RePaint [23] and subsequent refinements [21] has been successfully exploited to enforce observation consistency during sampling for inpainting. However, these methods typically operate on single images and do not address spatiotemporally structured gaps, physical masks (e.g., land), or calibrated uncertainty in geophysical fields. DIRECT extends the generative restoration paradigm to spatiotemporal SST with explicit temporal conditioning and physical constraints. As alternative to diffusion models, flow matching and rectified-flow models learn deterministic probability flows that transport noise to data [22]. DIRECT builds on this framework for *conditional* reconstruction, combining flow-based sampling with hard observation consistency and domain-specific spatiotemporal conditioning.

A growing body of work applies diffusion-style generative models to Earth-science reconstruction tasks, motivated by their ability to generate ensembles and quantify uncertainty. Barth et al. [4] demonstrate diffusion-based gap-filling for satellite ocean color (chlorophyll-a), producing ensembles and evaluating uncertainty reliability. For SST specifically, CARE-SST [6] applies a DDPM-style approach with historical context to reconstruct cloud-contaminated SST. Related efforts also extend diffusion-based reconstruction to broader ocean-temperature settings using observation-guided sampling and simulation pretraining [29]. In atmospheric data assimilation, physics-guided diffusion frameworks incorporate physical regularization to improve coherence under sparse observations [32]. These works establish the promise of generative modeling for geosciences, but commonly differ from the SST gap-filling setting considered here in one or more aspects: operating on single snapshots or coarse temporal context, relying on guidance schemes without explicit temporal-offset encoding, and/or not targeting calibrated per-pixel uncertainty for dense spatiotemporal reconstruction. DIRECT addresses these gaps with an explicit temporal context formulation, hard observation consistency throughout sampling, and uncertainty calibration.

3. DIRECT

Let $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ denote a sequence of SST measurements $\mathbf{x}_t \in \mathbb{R}^{W \times H}$, and $\mathbf{M} = \{\mathbf{m}_t\}_{t=1}^{T}$ a sequence of

corresponding binary masks $\mathbf{m}_t \in \{0,1\}^{W\times H}$, with ones indicating valid observations, and let $\mathbf{m}_l$ be a constant binary mask, with zeros indicating land. Our task is to recover a sequence of dense reconstructed fields $\{\boldsymbol{\mu}_t\}_{t=1}^T$ along with per-pixel uncertainty estimates $\Sigma = \{\boldsymbol{\sigma}_t\}_{t=1}^T$.

The SST reconstruction is inherently a spatio-temporal inference problem, as the ocean surface temperature evolves smoothly over time under physical constraints. As a result, observations from adjacent days provide valuable information for reconstructing the field at time t, while the utility of temporally more distant observations diminishes. Previous studies [3, 34] have explored this temporal dependency and report that a window spanning one day before and after the target time captures most of the relevant information, meaning days at times $t-1$, t, and $t+1$ are considered in reconstruction of the missing values at day t.

Because adjacent context frames may themselves be partially observed, we densify the temporal context by opportunistically replacing missing pixels with observations from nearby days. Specifically, when a pixel is missing at time $t \pm 1$ it is filled using the closest available observation from days before/after up to a maximum temporal offset of $\Delta_{\text{filled}} = 3$ days. We denote these filled context frames as $\mathbf{x}'_{t-1}$ and $\mathbf{x}'_{t+1}$. To preserve the temporal origin of this information, we accompany each context frame with a one-hot encoded mask $\mathbf{M}_{t\pm1} \in \{0,1\}^{3\times W\times H}$, encoding specific offsets (1, 2, or 3) or denoting the pixel as invalid (0). For the central frame, a two-channel mask $\mathbf{M}_t \in \{0,1\}^{2\times W\times H}$ indicates the observed versus missing (or land) pixels. We denote the temporal context fields as $\mathbf{C}_{1t} = [\mathbf{x}'_{t-1}, \mathbf{x}_t, \mathbf{x}'_{t+1}]$ and the corresponding masks as $\mathbf{C}_{2t} = [\mathbf{M}_{t-1}, \mathbf{M}_t, \mathbf{M}_{t+1}]$. Finally, also following good practices of prior works [3, 34], we incorporate the day-of-year (DoY) d_t to provide the context of seasonal information.

3.1. A flow-matching architecture

We frame the SST reconstruction as a conditional flow matching problem [19], which specifies the reconstruction process as a deterministic probability flow from a simple noise distribution into the target data distribution. The process involves simulating a stochastic differential equation, for which the Euler integration at iteration step $k \in [0,1]$ is

$$\hat{\mathbf{x}}_t^{(k)+\Delta} = \tilde{\mathbf{x}}_t^{(k)} + \Delta \cdot \mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t), \qquad (1)$$

where Δ is the integration step, $\mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t)$ is the rectified flow network, with $\mathbf{C}_t$ the provided context, d_t the day-of-the-year variable, and $\tilde{\mathbf{x}}_t^{(k)}$ the previous iteration reconstruction $\hat{\mathbf{x}}_t^{(k)}$ with observed values re-injected to constrain the reconstruction steps, i.e.,

$$\tilde{\mathbf{x}}_t^{(k)} = \text{FUSE}(\hat{\mathbf{x}}_t^{(k)}) = (\hat{\mathbf{x}}_t^{(k)} \odot (1 - \mathbf{m}_t) + \mathbf{x}_t \odot \mathbf{m}_t) \odot \mathbf{m}_l. \qquad (2)$$

Figure 2 visualizes the $\mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t)$ architecture. The input $\mathbf{y}_t^{(k)} \in \mathbb{R}^{128\times W\times H}$ is formed by concatenating $\tilde{\mathbf{x}}_t^{(k)}$ with the context $\mathbf{C}_{1t}$ and mixed by a 1×1 convolution, which is summed with the $\mathbf{C}_{2t}$ also passed through 1×1 convolution to match the feature dimensions.

The core is a U-Net architecture following [20], where the encoder progressively increases the feature channels across four levels (128, 256, 384, 512). We employ 5 residual blocks at the first level and 3 at subsequent levels, with self-attention applied at the lowest resolution.

Global conditioning is injected via FiLM modulation [25]. We construct a conditioning token $\mathbf{e} \in \mathbb{R}^{512}$ by summing two embeddings: (i) the day of year d_t, encoded using its sine–cosine representation $[\sin(d_t \frac{2\pi}{365.25}), \cos(d_t \frac{2\pi}{365.25})]$ and mapped through a two-layer MLP, to provide seasonal context; and (ii) the flow iteration variable k, encoded via sinusoidal embeddings followed by a two-layer MLP. The resulting conditioning token modulates feature maps in every residual block.

3.2. A multi-pass formulation

Note that missing measurements in days before and after the reconstructed day $\mathbf{x}_t$ (i.e., its temporal context) are naiively filled with past/future observations before passing to $\mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t)$. Improved reconstruction of $\mathbf{x}_t$ may be expected if the missing values in its temporal context days were replaced by their own reconstructions. We thus propose to run the reconstruction of the entire time-series in multiple (N_p) passes, in each pass building the temporal context from the previous-pass reconstructions and setting the corresponding values in the context $\mathbf{C}_{2t}$ to ± 1 days, respectively.

3.3. Probabilistic reconstruction

The time-series is reconstructed $N = 16$ times, starting with different random Gaussian samples initializing the missing regions, leading to an ensemble of reconstructions $\mathbf{S} = \{\hat{\mathbf{x}}_{t,n}\}_{n=1}^N$. Ideally, the ensemble per-pixel variance would capture the full posterior uncertainty, but our initial observations indicate that the variance is under-estimated at a small subset of pixels, whose locations vary. To address this, we consider the ensemble as a mixture of Gaussians, each centered at ensemble member with a small constant variance σ_0^2 across all pixels. The final reconstruction is thus obtained by moment-matching the mixture with a single Gaussian, i.e.,

$$\boldsymbol{\mu}_t = <\mathbf{S}>\,;\ \boldsymbol{\sigma}_t^2 = \text{var}(\mathbf{S}) + \sigma_0^2. \qquad (3)$$

The constant σ_0^2 is estimated on the validation set.

3.4. Training

DIRECT is trained by a modified self-supervised flow-matching objective to enable training with incomplete data. The training samples are created by sampling cloud masks $\mathbf{m}_m$ from other days and applying them to the central day, yielding $\mathbf{x}''_t = \mathbf{x}_t \odot \mathbf{m}_m$. Thus, a ground truth direction vector $\mathbf{u}_t = \mathbf{x}_t - \tilde{\mathbf{x}}_t^{\prime\prime(0)}$ is constructed from the simulated observed field with missing values initialized

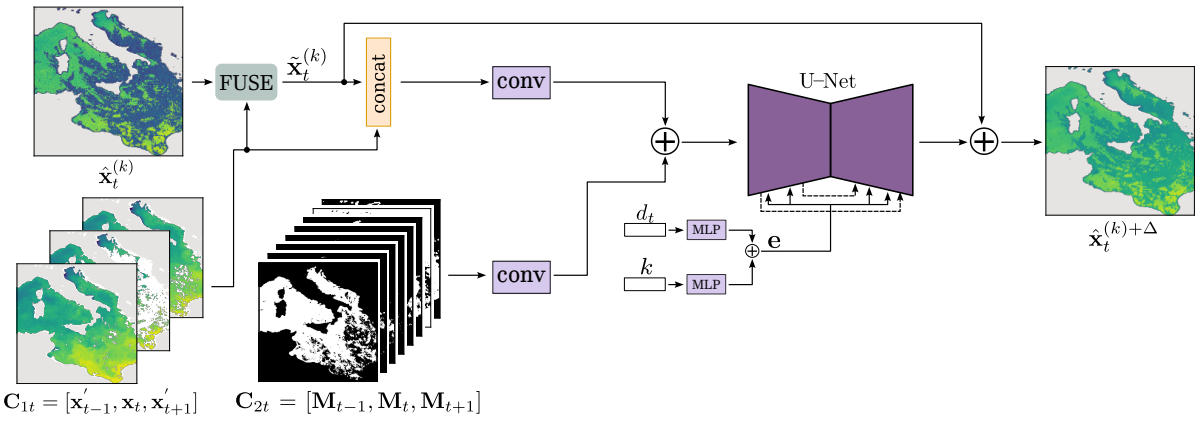


Figure 2. The DIRECT architecture overview. The SST field $\hat{\mathbf{x}}_t^{(k)}$ estimated at k/Δ-th iteration is accompanied by the temporal context $\mathbf{C}_{1t}$ and the time-stamp masks $\mathbf{C}_{2t}$, seasonal context d_t, and the current flow iteration value k. The FUSE rectified reconstruction $\tilde{\mathbf{x}}_t^{(k)}$ is concatenated with context frames $\mathbf{C}_{1t}$ and summed with the embedded origin masks $\mathbf{C}_{2t}$ to form the network input. Token $\mathbf{e}$ modulates the U-Net via FiLM [25]. The network output updated the initial SST estimate into $\hat{\mathbf{x}}_t^{(k)+\Delta}$ by Euler integration.

by noise ($\tilde{\mathbf{x}}_t''^{(0)}$) and the original observed field $\mathbf{x}_t$. The flow iteration variable k is sampled at uniform from interval $[0, 1]$, leading to the following loss at field indexed by time-step t:

$$\mathcal{L}_t = \frac{1}{N_{\text{obs},t}} \sum_{i=1}^{N_t} \mathbf{w}_{t,i} \left(\mathbf{v}_\theta(\tilde{\mathbf{x}}_t^{(k)} | \mathbf{C}_t, k, d_t)_{(i)} - \mathbf{u}_{t,i} \right)^2, \tag{4}$$

where N_t is the number of pixels in the SST field, $N_{\text{obs},t}$ is the number of observed pixels in $\mathbf{x}_t$, and $\mathbf{w}_t = \mathbf{m}_t \odot \mathbf{m}_l$ indicates observation presence at pixel i.

After training the model, the training set (or validation if available) can be reconstructed and the regularization variance σ_0^2 estimated as follows. Following [34] we define per-pixel scaled reconstruction error as

$$\boldsymbol{\epsilon}_{t,i} = (\mathbf{x}_{t,i} - \boldsymbol{\mu}_{t,i}) / \boldsymbol{\sigma}_{t,i}. \tag{5}$$

For a well-calibrated estimator, this error should follow a unit-variance, zero-centered Gaussian. The σ_0^2 can thus be estimated by minimizing the KL divergence [16] between the ideal Gaussian and the empirical Gaussian obtained from (5), which yields the following loss

$$\mathcal{L}_{\sigma_0^2} = (\mu_\epsilon^2) + (\sigma_\epsilon^2 - \log \sigma_\epsilon^2). \tag{6}$$

4. Results

Model parameters. During inference, each input produces an ensemble of $N = 16$ reconstructions, each obtained with $N_k = 25$ flow timesteps. The main results reported were obtained by using $N_p = 2$ passes.

Training details. The model is trained for 200 epochs using the AdamW optimizer ($\eta = 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.95$). We employ a hybrid learning-rate schedule consisting of a 20-epoch linear warm-up followed by cosine decay over the remaining 180 epochs, with a minimum learning rate of 1×10^{-8}.

Performance measures. We report the Root Mean Squared Error (RMSE) calculated over three specific regions: the region with observed data (RMSE_{obs}), the region in which we occluded the data (RMSE_{mis}), and the two regions combined for overall evaluation (RMSE_{all}). While RMSE_{obs} is included for completeness, it is not particularly informative: due to the FUSE rectification step (2), observed pixels are replaced with their measured values at every flow integration step, leading to near-zero error. Nevertheless, note that many competing methods do reconstruct also this part of the data, necessarily increasing the errors also in the *observed* regions. The metric of primary interest is therefore RMSE_{mis}, which reflects the quality of reconstruction in the occluded regions.

4.1. The SST datasets

We evaluate DIRECT on three level-3 (L3) multi-sensor SST datasets that represent diverse oceanographic regions with varying resolution and data sparsity: (i) The Mediterranean dataset [11] (from January 1, 2008, to December 31, 2021, remapped to a $0.0625° \times 0.0625°$ grid), (ii) the Adriatic dataset [9] (from August 25, 1981, to December 31, 2022, remapped to a $0.05° \times 0.05°$ grid), and (iii) the Atlantic dataset [10] (from January 1, 1982, to January 1, 2022, remapped to a $0.05° \times 0.05°$ grid). Following [34], we filter samples exceeding region-specific cloud coverage thresholds (100% for Mediterranean, 60% for Adriatic, 75% for Atlantic), resulting in 5114 valid samples for the Mediterranean, 7800 for the Adriatic, and 3454 for the Atlantic dataset, and employ chronological splits to prevent temporal leakage: 90% of the samples are used for training, 5% for validation, and 5% for testing.

4.2. Comparison with state-of-the-art

We compare DIRECT against three recent SST reconstruction methods: DINCAE2 [3], MAESSTRO [14], and CRITER [34]. We follow the evaluation procedure intro-

Table 1. Comparison of DIRECT to current state-of-the-art methods. All of the reported reconstruction errors are in °C, where the two numbers in parentheses correspond to the 10% and 90% percentiles.

Dataset	Model	RMSE_{all}	RMSE_{mis}	RMSE_{obs}
Mediterranean	MAESSTRO [14]	0.487 (0.320, 0.657)	0.607 (0.394, 0.856)	0.434 (0.299, 0.564)
	DINCAE2 [3]	0.209 (0.140, 0.300)	0.319 (0.226, 0.418)	0.148 (0.112, 0.184)
	CRITER [34]	0.127 (0.037, 0.235)	0.255 (0.168, 0.352)	0.017 (0.013, 0.021)
	DIRECT (ours)	**0.106 (0.026, 0.196)**	**0.220 (0.144, 0.302)**	**0.000 (0.000, 0.001)**
Adriatic	MAESSTRO [14]	0.456 (0.296, 0.635)	0.583 (0.362, 0.844)	0.392 (0.261, 0.539)
	DINCAE2 [3]	0.270 (0.111, 0.522)	0.433 (0.203, 0.769)	0.106 (0.087, 0.129)
	CRITER [34]	0.130 (0.045, 0.222)	0.243 (0.140, 0.358)	0.021 (0.014, 0.030)
	DIRECT (ours)	**0.109 (0.029, 0.193)**	**0.202 (0.113, 0.304)**	**0.001 (0.001, 0.002)**
Atlantic	MAESSTRO [14]	0.802 (0.508, 1.239)	0.832 (0.514, 1.301)	0.764 (0.479, 1.137)
	DINCAE2 [3]	0.444 (0.332, 0.581)	0.525 (0.396, 0.692)	0.302 (0.236, 0.364)
	CRITER [34]	0.391 (0.249, 0.542)	0.518 (0.386, 0.692)	0.036 (0.019, 0.046)
	DIRECT (ours)	**0.363 (0.234, 0.499)**	**0.489 (0.369, 0.640)**	**0.001 (0.000, 0.002)**

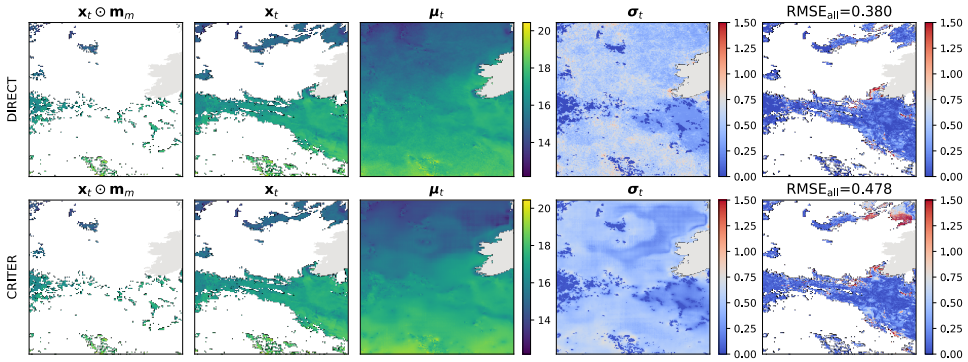


Figure 3. Comparison of DIRECT and CRITER reconstructions on the Atlantic region. The columns show (from left to right): the partially observed input ($\mathbf{x}_t \odot \mathbf{m}_m$), the ground truth $\mathbf{x}_t$, the reconstruction $\boldsymbol{\mu}_t$, the estimated uncertainty $\boldsymbol{\sigma}_t$, and the absolute reconstruction error (RMSE_{all}). All values are in °C.

duced by [34], in which each test sample is reconstructed 10 times, each time with a newly sampled cloud mask applied to the central frame. Final performance scores are obtained by averaging the reconstruction errors across these 10 passes over the test dataset. Because our evaluation replicates this protocol, we adopt the baseline results reported by [34] as reference values in this comparison.

Results in Table 1 show that across all three datasets, DIRECT achieves the lowest reconstruction error among all considered methods. Relative to DINCAE2, DIRECT reduces RMSE_{mis} by 53% on the Adriatic dataset, 31% on the Mediterranean, and 7% on the Atlantic dataset. Improvements over MAESSTRO are even greater, ranging from 41% on the Atlantic to more than 63% on both the Mediterranean and Adriatic datasets. Compared to CRITER, the strongest published baseline, DIRECT further reduces reconstruction error in occluded regions by 17% on the Adriatic, 14% on the Mediterranean, and 6% on the Atlantic. These gains also correspond to visibly more coherent reconstructions. An example from the Atlantic dataset is shown in Figure 3, where DIRECT recovers better spatially consistent SST fields while avoiding artifacts that persist in CRITER reconstructions (i.e., blocky structures). For subsequent experiments, we retain CRITER as the sole baseline, since it consistently outperforms DINCAE2 and MAESSTRO, and drop RMSE_{obs}, except when it offers some additional insight.

4.2.1. Uncertainty estimation and bias analysis

We evaluate the reliability of ensemble-based uncertainty estimates using the scaled error metric (5), which measures how well the predicted uncertainties explain the actual reconstruction error. Table 2 shows that the raw ensemble without the per-sample uncertainty σ_0 ($\text{DIRECT}_{\text{w.o.}\sigma_0}$), underestimates uncertainty ($\sigma_\epsilon > 1.4$). This reflects cases of ensemble collapse at sparse set of pixels, where reconstructions become overly similar, resulting in the predicted variance being too small. By using the dataset-specific learned regularization constant $\boldsymbol{\sigma}_0$ (0.114 for Mediterranean, 0.095 for Adriatic, 0.285 for Atlantic) in the total variance estimation (3), the underestimation is removed, reaching ($\sigma_\epsilon \approx 1.0$). DIRECT therefore achieves near-ideal estimation of standard deviation and bias on all three datasets. For further insights, we computed the ratio between the estimated constant $\boldsymbol{\sigma}_0$ and the average total per-dataset standard deviations. Overall,

Table 2. Uncertainty estimation analysis of DINCAE2, CRITER, and DIRECT. Mean standardized error (μ_ϵ), standard deviation (σ_ϵ), and bias are reported for each dataset.

Dataset	Model	μ_ϵ	σ_ϵ	Bias
Medite-	DINCAE2	-0.060	0.334	-0.060
rranean	CRITER	-0.022	1.116	-0.007
	$\text{DIRECT}_{\text{w.o.}\sigma_0}$	-0.036	1.728	**-0.003**
	DIRECT	**-0.018**	**1.062**	**-0.003**
Adriatic	DINCAE2	0.198	0.996	0.128
	CRITER	0.041	1.082	0.007
	$\text{DIRECT}_{\text{w.o.}\sigma_0}$	0.029	1.733	**0.003**
	DIRECT	**0.018**	**1.080**	**0.003**
Atlantic	DINCAE2	**-0.017**	0.801	**-0.006**
	CRITER	0.118	1.156	0.047
	$\text{DIRECT}_{\text{w.o.}\sigma_0}$	0.150	1.792	0.039
	DIRECT	0.099	**1.099**	0.039

the $\boldsymbol{\sigma}_0$ accounts for the 23%, 23%, and 33% of the total variance across the Mediterranean, Adriatic and Atlantic datasets, respectively.

4.2.2. Power spectrum density analysis

We perform a power spectral density (PSD) analysis [30] on a region of interest (ROI) in the central Mediterranean to evaluate how well DIRECT recovers spatial variability across scales compared to CRITER [34]. Time frames for which the ROI is fully observed are first identified, after which cloud masks are sampled, obscuring between 51% and 100% of the ROI. Both models reconstruct the full SST field from these masked inputs. For each reconstruction, we compute the 2D gradient magnitude $||\nabla \boldsymbol{\mu}_t^{\text{ROI}}||_2$ and the isotropic PSD using a FFT with a Blackman–Harris window [15].

The resuling graphs are shown in Figure 4. Both models successfully reproduce the large-scale spectral structure of the SST field and interestingly exhibit nearly identical spectral energy profiles at high wavenumbers. However, care is required in interpreting the density. Notably, while CRITER appears to recover the high-frequency parts equaly well as DIRECT, note that the gradient magnitudes (Fig. 4, row 3), reveal that this is partly driven by non-physical block artifacts, a byproduct of patch-based processing. In contrast, DIRECT's energy profile reflects coherent, albeit slightly smoothed, oceanic structures. Thus, DIRECT achieves a more physically faithful representation of the continuum, avoiding the spurious grid-like artefacts that inflate the PSD of the patch-based baselines.

4.2.3. Analysis of spatial scale correlation

To assess the preservation of spatial structure and mesoscale variability beyond pixel-wise error, we analyze the spatial correlation properties using empirical semivariograms [13]. Semivariance is computed as

$$\gamma(d) = \frac{1}{2N(d)} \sum_{i,j \in \mathcal{P}(d)} (p_i - p_j)^2, \tag{7}$$

where $\mathcal{P}(d)$ denotes the set of $N(d)$ pixel pairs seperated by distance d, and p_i is the observation value of $\mathbf{x}_t$ at i-th pixel. To ensure statistical robustness, we average the results across 10 independent mask realizations sampled from the dataset.

As shown in Figure 5, both models effectively capture local correlations (short lags). At intermediate to larger spatial lags, however, clearer differences appear. In the hidden region variograms, DIRECT consistently reproduces the growth of semivariance with distance more faithfully, while CRITER exhibits less stable long-range behavior, suggesting a struggle to maintain spatial coherence over large obscured gaps. When considering the full valid SST domain, the performance gap narrows significantly. In this setting, both DIRECT and CRITER follow the ground truth semivariance almost perfectly, with virtually no observable difference between the two models. This convergence is expected, as the calculation includes observed pixels. The results highlight that while both models are highly reliable when observations are present, DIRECT provides a more physically consistent spatial texture when reconstructing larger completely unobserved areas.

4.3. Influence of multi-pass reconstruction

We next evaluate the multipass inference strategy described in Section 3.2. The notation DIRECT_{N_p} is used to explicitly note the number of passes N_p used in DIRECT, with DIRECT_1 indicating a single pass, with initial fallback context used. Table 3 summarizes the results across three regions, with CRITER included for reference.

On the Mediterranean and Adriatic datasets, a two-pass version (DIRECT_2) provides modest but consistent improvements over the single pass, reducing RMSE_{mis} by 6% and 3%, respectively. This performance boost confirms that substituting the initial fallback context with the model's own high-fidelity estimates provides a cleaner conditioning signal. Compared to CRITER, DIRECT_2 achieves overall RMSE_{mis} reductions of 14% on the Mediterranean and 17% on the Adriatic. However, increasing to a three passes (DIRECT_3) yields no further benefits. On the Atlantic dataset, which covers a vast open-ocean area characterized by persistent, large-scale cloud cover, no measurable gain is observed even at $N_p = 2$. In such scenarios, both the initial fallback context and the first-pass reconstruction suffer from a lack of "anchor" observations. If the first-pass estimate contains significant uncertainty or reconstruction bias, using it as context for a second pass likely propagates these errors rather than reducing them – nevertheless the errors do not increase in this case, implying DIRECT's robustness.

4.4. Ablation study

Ablation studies were performed on the Mediterranean dataset with single-pass inference ($N_p = 1$) to expose the influences of individual parts.

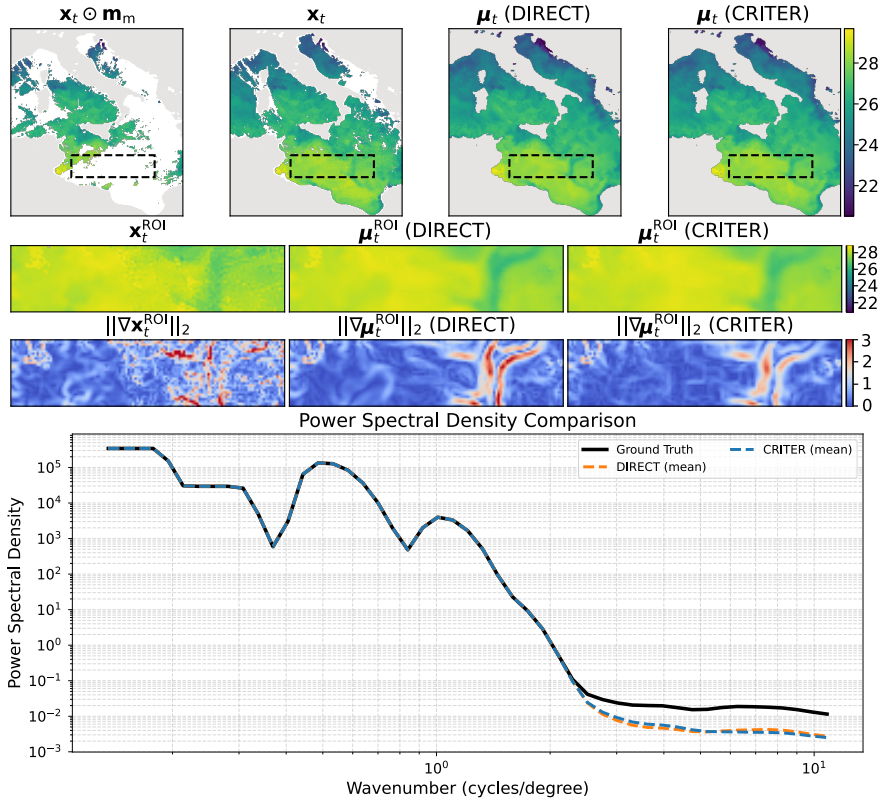


Figure 4. Masked SST input, ground truth field, and DIRECT/CRITER reconstructions (row 1). Structural SST field differences in selected region (row 2) are visualized by spatial gradient magnitudes (row 3), and by isotropic power-spectral density (row 4).

Table 3. Influence of the multi-pass strategy (DIRECT$_{N_p}$, with N_p passes), with CRITER results for reference. All reported reconstruction errors are in °C

Dataset	Model	RMSE$_{\text{all}}$	RMSE$_{\text{mis}}$
Medite-	DIRECT$_1$	0.113	0.234
rranean	DIRECT$_2$	**0.106**	**0.220**
	DIRECT$_3$	0.106	0.222
	CRITER	0.127	0.255
Adriatic	DIRECT$_1$	0.113	0.208
	DIRECT$_2$	**0.109**	**0.202**
	DIRECT$_3$	**0.109**	**0.202**
	CRITER	0.130	0.243
Atlantic	DIRECT$_1$	**0.363**	**0.489**
	DIRECT$_2$	**0.363**	**0.489**
	DIRECT$_3$	0.364	0.490
	CRITER	0.391	0.518

4.4.1. Analysis of input rectification (FUSE)

We evaluate the importance of the FUSE procedure (Equation (2)). To this end, we compare DIRECT against three ablated variants: one without injecting observed SST values (DIRECT$_{\overline{\text{OBS}}}$), one without zeroing land pixels (DIRECT$_{\overline{\text{ML}}}$), and one without any rectification (DIRECT$_{\overline{\text{FUSE}}}$). Results in Table 4 show that completely disabling FUSE causes a large drop in reconstruction accuracy, with RMSE$_{\text{mis}}$ increased by more than 20%. Restoring only the injection of SST values (DIRECT$_{\overline{\text{ML}}}$) or only the masking of land pixels (DIRECT$_{\overline{\text{OBS}}}$) recovers much of this loss (only a 2% increase for both), indicating that both corrections help anchor the generative process and prevent errors from propagating across integration steps. The best results are obtained when both corrections are applied, confirming that the full FUSE procedure provides complementary benefits and stabilizes the flow integration process.

Table 4. Performance of DIRECT variants. All of the reported reconstruction errors are in °C.

Variant	RMSE$_{\text{all}}$	RMSE$_{\text{mis}}$	RMSE$_{\text{obs}}$
DIRECT$_{\overline{\text{OBS}}}$	0.115	0.237	0.011
DIRECT$_{\overline{\text{ML}}}$	0.115	0.239	0.000
DIRECT$_{\overline{\text{FUSE}}}$	0.222	0.295	0.181
DIRECT	**0.113**	**0.234**	**0.000**

4.4.2. Influence of reconstruction ensemble size

Because DIRECT is a generative model, multiple reconstructions can be sampled for a single input by varying the initial noise seed. As per Section 3.3, we average these samples to obtain a final reconstruction. Table 5 shows

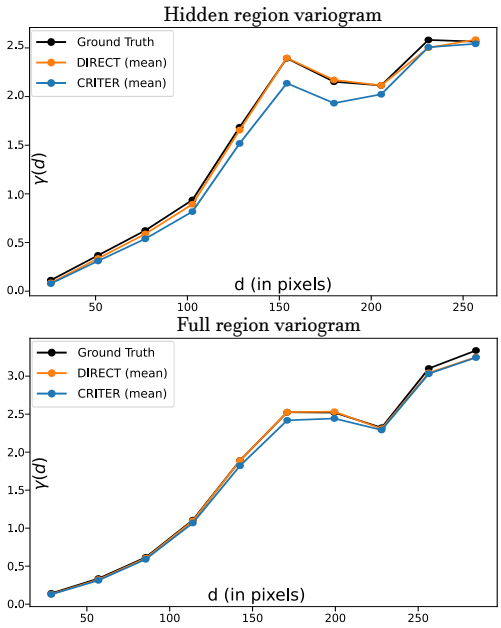


Figure 5. Visualization of the spatial scales and correlation properties analysis. Top plot: empirical semivariograms computed exclusively over hidden regions. Bottom plot: semivariograms computed over the full SST domain. Each curve represents the mean semivariance across 10 different mask realizations.

Table 5. Influence of ensemble size N on reconstruction accuracy. All of the reported reconstruction errors are in °C.

Ensemble size N	RMSE_{all}	RMSE_{mis}
$N = 1$	0.148	0.307
$N = 4$	0.125	0.251
$N = 8$	0.118	0.241
$N = 16$	**0.113**	**0.234**
$N = 32$	0.113	0.234
CRITER	0.127	0.255

that averaging just $N = 4$ reconstructions already reduces RMSE_{mis} by 18% compared to $N = 1$, and crucially, this already outperforms the deterministic state-of-the-art CRITER. Our default $N = 16$ yields a 23% improvement compared to $N = 1$. Although larger ensembles (e.g. $N = 32$) are comparable with our default, they also increase inference time, making 16 samples practical and effective.

Figure 6 visualizes the characteristics of this ensemble. While the mean field provides a stable and accurate reconstruction, the per-pixel standard deviation captures the model's uncertainty in obscured regions. The residuals $\boldsymbol{\Delta}_n = \boldsymbol{\mu}_t - \hat{\mathbf{x}}_{t,n}$ (Rows 2–4) highlight the diverse high-frequency fluctuations present in individual samples.

5. Conclusion

We presented DIRECT, a diffusion-inspired generative model for reconstructing dense SST fields from partially observed satellite measurements. By formulating SST gap-filling as a conditional flow-matching task, DIRECT departs from deterministic reconstruction and instead produces an ensemble of physically plausible realizations, enabling both accurate reconstructions and spatially resolved uncertainty estimation. The model integrates temporal context, seasonal conditioning, and observation-guided rectification within a single end-to-end framework. Experimental results across Mediterranean, Adriatic, and Atlantic datasets show that DIRECT consistently outperforms current state-of-the-art methods in both reconstruction accuracy and structural fidelity.

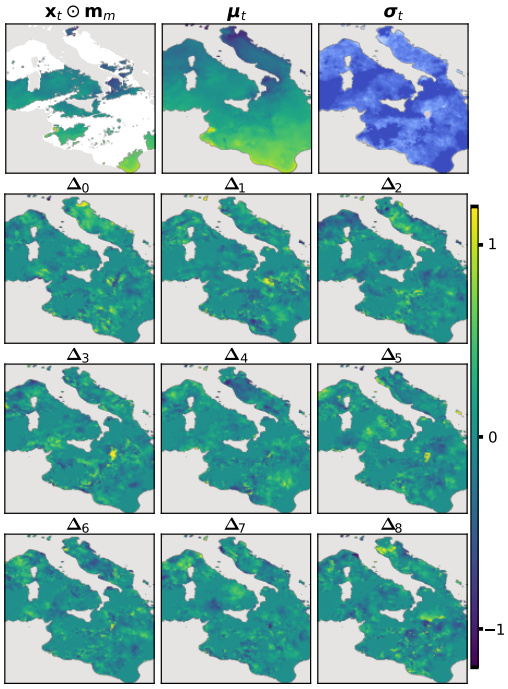


Figure 6. A visualization of a reconstruction ensemble. The first row shows the masked input, the ensemble mean ($\boldsymbol{\mu}_t$) and the per-pixel uncertainty ($\boldsymbol{\sigma}_t$). Rows 2–4 visualize the residual fields $\boldsymbol{\mu}_t - \hat{\mathbf{x}}_{t,n}$ for nine independent samples, illustrating the high-frequency structural diversity.

Limitations and Future Work. While the ensemble mean provides a stable and accurate estimate, the averaging process inherently acts as a low-pass filter, attenuating some of the high-frequency details present in individual generative samples. Future work could focus on improving the fidelity of single-sample reconstructions. Furthermore, while our post-hoc uncertainty calibration effectively corrects for under-dispersion, incorporating this calibration objective directly into the training phase via a proper scoring rule loss may lead to even sharper uncertainty estimates. Finally, extending the DIRECT architecture to multivariate oceanographic data represents a promising path toward a more holistic, physically-constrained generative model.

References

[1] A. Alvera-Azcárate, A. Barth, M. Rixen, and J.M. Beckers. Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the adriatic sea surface temperature. *Ocean Modelling*, 9(4): 325–346, 2005. 1, 2

[2] Alexander Barth, Aida Alvera-Azcárate, Matjaz Licer, and Jean-Marie Beckers. Dincae 1.0: A convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations. *Geoscientific Model Development*, 13(3):1609–1622, 2020. 1, 2

[3] A. Barth, A. Alvera-Azcárate, C. Troupin, and J.-M. Beckers. Dincae 2.0: multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations. *Geoscientific Model Development*, 15(5):2183–2196, 2022. 1, 2, 3, 4, 5

[4] A. Barth, J. Brajard, A. Alvera-Azcárate, B. Mohamed, C. Troupin, and J.-M. Beckers. Ensemble reconstruction of missing satellite data using a denoising diffusion model: application to chlorophyll a concentration in the black sea. *Ocean Science*, 20(6):1567–1584, 2024. 2

[5] Stephan Bojinski, Michel Verstraete, Thomas C. Peterson, Carolin Richter, Adrian Simmons, and Michael Zemp. The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, 95(9):1431 – 1443, 2014. 1

[6] Minki Choo, Sihun Jung, Jungho Im, and Daehyeon Han. Care-sst: context-aware reconstruction diffusion model for sea surface temperature. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220:454–472, 2025. 2

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 2

[8] C. Donlon, I. Robinson, K. S. Casey, J. Vazquez-Cuervo, E. Armstrong, O. Arino, C. Gentemann, D. May, P. LeBorgne, J. Piollé, I. Barton, H. Beggs, D. J. S. Poulter, C. J. Merchant, A. Bingham, S. Heinz, A. Harris, G. Wick, B. Emery, P. Minnett, R. Evans, D. Llewellyn-Jones, C. Mutlow, R. W. Reynolds, H. Kawamura, and N. Rayner. The global ocean data assimilation experiment high-resolution sea surface temperature pilot project. *Bulletin of the American Meteorological Society*, 88(8):1197 – 1214, 2007. 1

[9] E.U. Copernicus Marine Service Information. Mediterranean sea - high resolution l3s sea surface temperature reprocessed, 2024. Available from the Copernicus Marine Data Store (MDS), accessed 23-11-2024. 4

[10] E.U. Copernicus Marine Service Information. European north west shelf/iberia biscay irish seas – high resolution odyssea sea surface temperature multi-sensor l3 observations reprocessed, 2024. Available from the Copernicus Marine Data Store (MDS), accessed 23-11-2024. 4

[11] E.U. Copernicus Marine Service Information. Mediterranean sea - high resolution and ultra high resolution l3s sea surface temperature, 2024. Available from the Copernicus Marine Data Store (MDS), accessed 23-11-2024. 4

[12] Carlos Garcia-Soto, Lijing Cheng, Levke Caesar, Sunke Schmidtko, Elizabeth B Jewett, Alicia Cheripka, Ignatius Rigor, Ainhoa Caballero, Sanae Chiba, Jose Carlos Báez, et al. An overview of ocean climate change indicators: Sea surface temperature, ocean heat content, ocean ph, dissolved oxygen concentration, arctic sea ice extent, thickness and volume, sea level and strength of the amoc (atlantic meridional overturning circulation). *Frontiers in Marine Science*, 8:642372, 2021. 1

[13] Matheron Georges. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963. 6

[14] E. Goh, A. Yepremyan, J. Wang, and B. Wilson. Maesstro: Masked autoencoders for sea surface temperature reconstruction under occlusion. *Ocean Science*, 20(5):1309–1323, 2024. 1, 2, 4, 5

[15] F.J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978. 6

[16] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 4

[17] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement – a comprehensive survey, 2023. 2

[18] Zukun Li, Daoming Wei, Xuefeng Zhang, Yaoting Gao, and Dianjun Zhang. A daily high-resolution sea surface temperature reconstruction using an i-dincae and dnn model based on fy-3c thermal infrared data. *Remote Sensing*, 16(10), 2024. 1

[19] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[20] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. 3

[21] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8038–8047, 2024. 2

[22] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 2

[23] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022. 2

[24] Anne G. O'Carroll, Edward M. Armstrong, Helen M. Beggs, Marouan Bouali, Kenneth S. Casey, Gary K. Corlett, Prasanjit Dash, Craig J. Donlon, Chelle L. Gentemann, Jacob L. Høyer, Alexander Ignatov, Kamila Kabobah, Misako Kachi, Yukio Kurihara, Ioanna Karagali, Eileen Maturi, Christopher J. Merchant, Salvatore Marullo, Peter J. Minnett, Matthew Pennybacker, Balaji Ramakrishnan, RAAJ Ramsankaran, Rosalia Santoleri, Swathy Sunder, Stéphane Saux Picart, Jorge Vázquez-Cuervo, and Werenfrid Wimmer. Observational needs of sea surface temperature. *Frontiers in Marine Science*, Volume 6 - 2019, 2019. 1

[25] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with

a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. 3, 4

[26] Raheema Rahman and Hasibur Rahaman. Evaluation of sea surface temperature from ocean reanalysis products over the north indian ocean. *Frontiers in Marine Science*, Volume 11 - 2024, 2024. 1

[27] Antonio Ricchi, Lorenzo Sangelantoni, Gianluca Redaelli, Vincenzo Mazzarella, Mario Montopoli, Mario Marcello Miglietta, Alessandro Tiesi, Simone Mazzà, Richard Rotunno, and Rossella Ferretti. Impact of the sst and topography on the development of a large-hail storm event, on the adriatic sea. *Atmospheric Research*, 296:107078, 2023. 1

[28] A. Senatore, L. Furnari, and G. Mendicino. Impact of high-resolution sea surface temperature representation on the forecast of small mediterranean catchments' hydrological responses to heavy precipitation. *Hydrology and Earth System Sciences*, 24(1):269–291, 2020. 1

[29] Yuanyi Song, Pumeng Lyu, Ben Fei, Fenghua Ling, Wanli Ouyang, and Lei Bai. Reconmost: Multi-layer sea temperature reconstruction with observations-guided diffusion, 2025. 2

[30] Petre Stoica, Randolph L Moses, et al. *Spectral analysis of signals*. Pearson Prentice Hall Upper Saddle River, NJ, 2005. 6

[31] G. Taburet, A. Sanchez-Roman, M. Ballarotta, M.-I. Pujol, J.-F. Legeais, F. Fournier, Y. Faugere, and G. Dibarboure. Duacs dt2018: 25 years of reprocessed sea level altimetry products. *Ocean Science*, 15(5):1207–1224, 2019. 1, 2

[32] Hao Wang, Jindong Han, Wei Fan, Weijia Zhang, and Hao Liu. Phyda: Physics-guided diffusion models for data assimilation in atmospheric systems, 2025. 2

[33] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2): 1–42, 2024. 2

[34] M. Zupančič Muc, V. Zavrtanik, A. Barth, A. Alvera-Azcarate, M. Ličer, and M. Kristan. Criter 1.0: a coarse reconstruction with iterative refinement network for sparse spatio-temporal satellite data. *Geoscientific Model Development*, 18(17):5549–5573, 2025. 1, 2, 3, 4, 5, 6

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hradec, Czech Republic, February 9–12, 2026

Grading Handwritten Engineering Exams with Multimodal Large Language Models

Janez Perš, Jon Muhovič, Andrej Košir, Boštjan Murovec
University of Ljubljana, Faculty of Electrical Engineering
{janez.pers},{jon.muhovic},{andrej.kosir},{bostjan.murovec}@fe.uni-lj.si

Abstract

Handwritten STEM exams capture open-ended reasoning and diagrams, but manual grading is slow and difficult to scale. We present an end-to-end workflow for grading scanned handwritten engineering quizzes with multimodal large language models (LLMs) that preserves the standard exam process (A4 paper, unconstrained student handwriting). The lecturer provides only a handwritten reference solution (100%) and a short set of grading rules; the reference is converted into a text-only summary that conditions grading without exposing the reference scan. Reliability is achieved through a multi-stage design with a format/presence check to prevent grading blank answers, an ensemble of independent graders, supervisor aggregation, and rigid templates with deterministic validation to produce auditable, machine-parseable reports. We evaluate the frozen pipeline in a clean-room protocol on a held-out real course quiz in Slovenian, including hand-drawn circuit schematics. With state-of-the-art backends (GPT-5.2 and Gemini-3 Pro), the full pipeline achieves ≈8-point mean absolute difference to lecturer grades with low bias and an estimated manual-review trigger rate under 20% at $D_{\max} = 40$. Ablations show that trivial prompting and removing the reference solution substantially degrade accuracy and introduce systematic over-grading, confirming that structured prompting and reference grounding are essential.

1. Introduction

Handwritten, paper-based exams remain common in STEM education (science, technology, engineering, and mathematics), because they naturally elicit open-ended reasoning, intermediate work, and sketches (e.g., circuit diagrams) that are difficult to capture with purely digital assessments. Figure 1 illustrates the challenges. Yet manual grading of such exams is time-consuming and hard to scale. Learnosity reports that, in an online survey of 258 U.S. teachers, respondents spent an average of 9.9 hours per week grading and marking [14]. At the university level, faculty emotions research similarly suggests that grading can be experienced as comparatively unpleasant: in a large U.S. faculty sample, grading was associated with less positive and more negative emotions than research and teaching [24]. Beyond the immediate time cost, practitioner-facing syntheses argue that extensive grading can consume time and energy that would otherwise support course planning and instructional improvement [28].

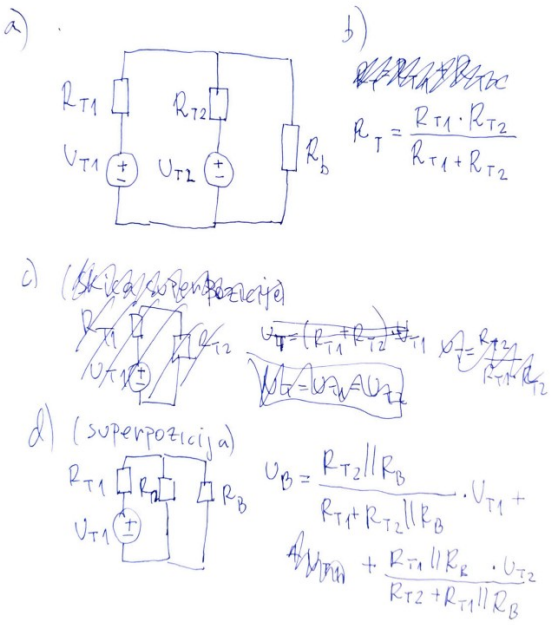


Figure 1. Sample handwritten answer with diagrams (e.g. circuits) that our system is designed to handle. Sample provided by class lecturer (not student).

The problem is amplified in STEM courses, where written exams and technical assignments require evaluating multi-step solutions and graphical artifacts under time pressure. A STEM-focused review notes that, as enrollments and workloads grow, instructors can struggle to provide timely feedback on labor-intensive assessments and may assign fewer practice opportunities despite pedagogical benefits [27]. More broadly, automated grading systems are often motivated by reducing educator workload and shortening feedback turnaround [27]. Workflow sys-

tems such as Gradescope show that digitizing and organizing scanned handwritten submissions can reduce grading overhead and support more consistent rubric application [25], but these platforms largely preserve a fundamental bottleneck: humans still read and score each response.

Recent progress in large language models (LLMs) opens a path toward more automated grading. On text-only short-answer grading, experiments on K–12 (primary and secondary school) short-answer data report that GPT-4 can achieve agreement close to human rater agreement under relatively simple prompting [10]. Controlled ASAG evaluations further show that grading performance is sensitive to prompt context—including whether a reference answer is provided or withheld—and can change substantially across datasets and setups [11]. For paper-based STEM exams, multimodal LLMs can be evaluated directly on images of handwritten solutions; recent results report improved alignment when prompts include official solutions and a grading rubric, while also noting that accuracy can remain insufficient for unsupervised, real-world deployment [5]. Practical limitations also remain in hybrid workflows: when handwriting is transcribed to text/LaTeX before grading, transcription errors can affect downstream scoring, and recent work treats LLM-produced grades as subject to subsequent human verification and explores confidence estimation via repeated sampling [15]. Separately, applied deployments on longer, multi-part responses highlight handwriting-to-text conversion as a recurring bottleneck and report challenges in applying fine-grained rubrics to long solutions and diagram/graph-heavy work [12]. Together, these findings suggest that reliable exam grading requires not only strong multimodal models, but also systems-level design: structured prompting, verification/aggregation, and deterministic post-processing to enforce consistent outputs.

In this work, we target the setting of end-to-end grading for scanned handwritten STEM exams that combine free-form text with hand-drawn diagrams. We present a multi-stage, multi-prompt grading workflow with deterministic post-processing to ensure reliable, auditable outputs, and evaluate it on undergraduate open-question engineering exams requiring handwritten textual answers and electrical schematics.

2. Related Work

Research into automated exam grading has a long history that predates modern LLMs. Before LLMs, many practical systems either (i) constrained answers into computer-readable formats (enabling direct scoring) or (ii) relied on handwriting recognition/OCR to extract text from scanned work before applying text-based scoring methods. A recurring challenge in this area is data availability: authentic exam scripts are frequently private and difficult to share, and publicly distributable datasets often only partially reflect real exam conditions.

2.1. Exam grading before the advent of LLMs

Prior to multimodal LLMs, automatic assessment largely decomposed into (i) text-based automated short-answer grading (ASAG) and (ii) document pipelines that first transcribed handwriting. Surveys of ASAG describe early systems built around engineered lexical/syntactic features and semantic similarity to reference answers, often using supervised models trained on scored responses and evaluated with standard agreement/correlation metrics [4]. Representative feature-integration approaches combine multiple linguistic feature families in discriminative scoring models [23], while vector-based approaches use distributed representations and similarity scoring for grading [16]. More recent surveys emphasize the shift toward deep learning and pretrained language models (including transformer-based approaches) for text ASAG [9].

For handwritten work, a common strategy was handwriting recognition/optical character recognition (OCR), followed by text-based scoring. Early examples integrated handwriting recognition with automated essay scoring and underscored the dependence of end-to-end scoring quality on transcription quality [26]. For short handwritten answers, Rowtula et al. propose a word-spotting-based approach that avoids full transcription and instead relies on retrieval-style signals for downstream evaluation [22]. In parallel, system-level tools such as Gradescope scaled the *workflow* of grading scanned submissions through dynamic rubrics and answer grouping, while still relying on humans to assign points [25]. For structured STEM problems, clustering-based methods have been proposed to group similar solution structures so that an instructor can label clusters and propagate (partial) credit [13]. Overall, pre-LLM grading systems either assumed clean text or relied on transcription as an error-prone front-end, and most did not robustly support unconstrained handwriting, diagrams, and multi-page exam context end-to-end.

2.2. Exam grading in the LLM era

Since 2023, large language models have been evaluated as general-purpose graders, first on text-only responses and increasingly on images of handwritten work. On K–12 short-answer data, Henkel et al. report that GPT-4 with relatively simple prompting can achieve agreement close to human rater agreement [10]. Complementary ASAG studies show that results depend on prompt context and experimental setup, including whether reference answers are provided or withheld [11]. In advanced mathematics settings, Gandolfi reports that GPT-4 can produce useful solutions and grading rationales, while also documenting reliability issues such as occasional loss of coherence and hallucinations that motivate explicit verification and guardrails [6].

A major shift is direct multimodal grading of scanned handwritten solutions. For university-level math exams, Caraeni et al. evaluate GPT-4o grading directly from handwriting images and report improved alignment when prompts include official solutions and a grading rubric,

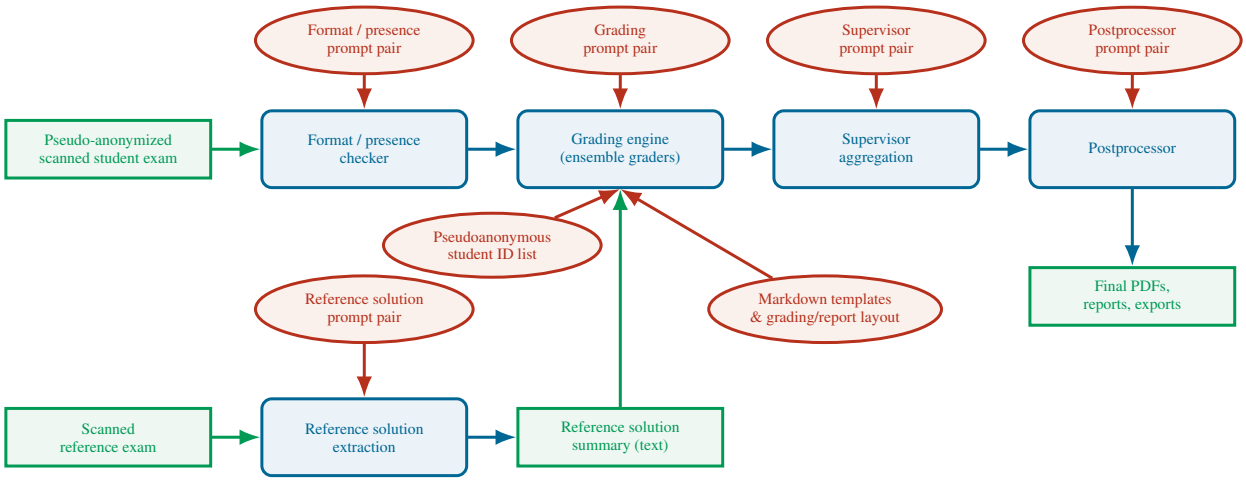


Figure 2. System overview. Green boxes denote data artifacts, red ellipses denote prompt pairs and templates, and blue boxes denote processing stages. All blue processing stages invoke a multimodal LLM backend (e.g., GPT-4o, GPT-5.x, Gemini, or Mistral), while reference solution scans are converted into a text-only summary that is injected into grading prompts. Only format checker, ensemble graders and reference solution extractor receive images, all other LLM interaction is text-only.

while noting that overall accuracy remains a limiting factor for real-world use [5]. Hybrid pipelines remain relevant: Liu et al. study AI-assisted grading of handwritten university mathematics exams using an OCR/LaTeX transcription stage and emphasize both transcription sensitivity and the role of human verification, while also exploring confidence estimation by sampling multiple grading runs [15]. At larger scale (e.g., hundreds of scripts in physics), workflow studies continue to highlight handwriting-to-text conversion as a practical bottleneck, challenges in applying fine-grained rubrics to long solutions, and persistent difficulty with diagrams/graphs [12]. For non-textual outputs such as hand-drawn graphs, recent comparative work evaluates meta-learning approaches alongside vision-language models on graph grading tasks [21]. Alternative multimodal scoring paradigms use CLIP-style embeddings (optionally combined with OCR) to incorporate visual information into scoring [2], and new benchmarks (e.g., DrawEduMath) systematically probe VLM interpretation of students' hand-drawn math images and document remaining weaknesses [3].

2.3. Datasets

Public datasets that enable *end-to-end* evaluation of handwritten exam grading (images in, numeric scores out) are rare, largely because authentic exam scripts are privacy-sensitive and typically collected under institutional consent/institutional review board (IRB), limiting public release. Consequently, recent multimodal grading studies evaluate on private course-exam collections with scanned pages, rubrics, and human scores [5, 12], while graph-focused handwritten corpora are likewise institutional [21]. The few open resources with handwriting either target short answers [7] or serve as VLM understanding benchmarks rather than points-based grading datasets [3]; large educational corpora with image responses are often not redistributable [1]. Therefore, to evaluate an end-to-end pipeline under realistic STEM exam conditions (multi-part solutions and diagrams), we collected and graded our own exam data.

3. Methods

The pipeline was designed to minimize interference with the exam process. Beyond using the standard A4 paper, no constraints are imposed on students or on how the lecturer administers the exam. The only additional requirement is that the lecturer provides a handwritten reference solution representing a perfect (100%) answer. These requirements directly shape the architecture shown in Fig. 2.

3.1. System architecture

The pipeline in Fig. 2 couples deterministic document handling with a small number of LLM-invoking stages. While the system is implemented as a substantial amount of orchestration code, the primary methodological contribution lies in the *structure of the pipeline* and, crucially, in the *prompt pairs and rigid templates* that make the overall behavior stable and machine-parseable. Due to space constraints, we do not reproduce the full prompts and templates.

Reference conditioning without exposing the reference scan. The lecturer provides a handwritten reference solution representing a perfect (100%) answer. A dedicated reference-extraction stage converts the scanned reference into a text-only summary that is injected into grading prompts; the reference image itself is not used during grading. This stage is run with a highly capable multimodal backend (GPT-5.2-pro in our experiments) to robustly interpret unconstrained handwriting and sketches.

Answer presence guardrail. Before any scoring, a format/presence checker predicts which tasks contain an actual student answer. This safeguard was introduced after observing rare cases where a model would hallucinate content when an entire answer region was left blank. Although such events are not reflected in aggregate metrics, they are operationally unacceptable, and the presence list prevents the grader from assigning points to missing answers.

Ensemble grading and supervisor aggregation. For each task, grading is performed by an ensemble of $K = 3$ independent, stateless model calls that produce structured drafts according to a fixed template. A supervisor model then merges the drafts into a single exam-level output, enforces template compliance, applies the presence decisions (unanswered tasks receive 0%), and flags inconsistencies for optional human resolution. This ensemble-plus-supervisor design reduces variance and improves robustness to occasional model glitches.

Postprocessing and exports. Finally, a postprocessor produces presentation-ready artifacts (e.g., report assembly and optional translation) while preserving all numeric fields and the template structure, enabling deterministic parsing, auditing, and export.

Privacy. Students are instructed *not to write their name and surname anywhere on the exam.* Instead, they are required to write their registration number, so the data is pseudo-anonymized before invoking LLM. The LLM gets sanitized *student roster* with all registration numbers for students participating in the exam and only needs to find which one of the finite set of registration number is written at the top of the first page of the exam. All de-anonymization is done locally after grading. Mandatory markdown templates ensure that the model outputs student's pseudo-identity in a specific place in the output, where it can be parsed. If parsing fails, human is required to read the number from the scan.

3.2. Prompts and templates

Each LLM-invoking stage is implemented as a *prompt pair* (system + user) backed by rigid Markdown templates. The system prompt fixes role, constraints, and prohibited behavior, while the user prompt injects instance-specific context (task labels, student scan, reference summary, and optional instructor rules). The templates strictly define section hierarchy and numeric fields, converting otherwise probabilistic outputs into artifacts that can be validated and parsed by deterministic post-processing; in contrast, directly asking an LLM to "grade the exam" is unreliable due to format drift and inconsistent application of rubrics.

Instructor-facing configuration. The only course-specific inputs are (i) the scanned handwritten reference solution and (ii) a short list of grading rules. All other pipeline stages, prompts, and templates are intended to remain unchanged across courses. Rules are numbered (`[R1]`, `[R2]`, ...), appended verbatim to grading-related prompts, and graders are instructed to cite applicable rule IDs in their explanations, improving auditability and facilitating human review. In our experiments, most rules transferred across STEM quizzes, with only minor course-specific adjustments (e.g., evaluating circuit sketches by topology rather than drawing orientation).

Template constraints. Two templates are used: a per-grader template and a supervisor template. Both prohibit adding or removing sections and enforce a fixed scoring line with a deterministic numeric pattern (achievement, weight, contribution). Per task, the template separates the question text, a plain-text summary of the student answer, a correctness explanation with required short meta-tags (including rule citations), and a single scoring line. The final exam total must equal the sum of per-task contributions (no normalization), which enables automatic consistency checks.

Language. Unless stated otherwise, the prompts, rules, and templates used in our experiments are written in Slovenian. The pipeline itself is language-agnostic: adapting to other languages requires only translating these text artifacts, without changing the processing stages.

4. Experiments

Automated grading quality has important qualitative aspects that are not fully captured by aggregate metrics alone (e.g., the coherence and pedagogical usefulness of the generated feedback). Therefore, in addition to the quantitative evaluation reported below, we provide *supplementary material*, comprised of the scanned reference solution of the "Class B" exam, and a corresponding mock solution, which was actually graded by the pipeline using GPT-5.2 with thinking set to "high". We provide these materials to illustrate the structure and depth of the produced feedback. The link to the supplementary material is provided at the end of the manuscript.

4.1. Experimental protocol

Our grading pipeline is not trainable and therefore operates in a fully zero-shot setting. To limit potential bias from iterative prompt and rubric engineering, we use a clean-room protocol with two parallel courses: **Class A** (development) and **Class B** (held-out evaluation).

During the semester, the system was deployed weekly on Class A quizzes and its outputs were reviewed by the lecturer and students, providing formative feedback on system behavior. After week 9, the pipeline and all prompts were frozen. The frozen system was then applied to one scanned quiz from Class B *without any modifications to the pipeline, prompts, or templates*. The

only change relative to Class A was an instructor-facing grading-rule adjustment reflecting standard practice for circuit sketches: circuit answers were evaluated by electrical topology rather than drawing orientation. All quantitative and qualitative results reported in this paper are based exclusively on this held-out Class B evaluation and are compared to grades assigned by the Class B lecturer.

To maintain separation, Class B quizzes were graded only by the lecturer during the semester and scanned for archival purposes. Apart from limited pilot runs in week 2 (not used for analysis), the Class B materials were not accessed or inspected by the system developers prior to the held-out evaluation.

4.2. Dataset

Class A is an undergraduate course on communication technologies (approximately 30 enrolled students, weekly exams were attended by 20-25 students), while Class B is an industrial electronics course (approximately 15 enrolled students). Both courses used short weekly written quizzes administered for 20 minutes at the beginning of each lecture, covering material from the previous week. The quizzes consist of open-ended questions requiring handwritten text and, where appropriate, hand-drawn diagrams or schematics. Dataset for testing was obtained by collecting 15 exams from the third week of lectures, one from each students.

Weekly quizzes were not strictly mandatory, but passing a subset was required to qualify for the final exam; strong weekly performance could optionally substitute the final exam grade. This provided meaningful incentive while keeping overall pressure moderate. For privacy reasons, we do not release student submissions, scans, or grades; we report only aggregate performance metrics and the exact text of the held-out evaluation questions.

Language. The quizzes and student answers were in a non-English language (Slovenian). Accordingly, all prompt pairs, grading rules, and report templates used in the main evaluation were written in Slovenian. The pipeline itself is language-agnostic: language-specific content is confined to prompts, rules, and templates, which can be translated to other languages (including with the assistance of modern LLM-based tools) without changing the pipeline stages.

Held-out exam content. Table 1 lists the three questions (with weights) from the held-out Class B quiz used for evaluation.

Weight	Question text
25%	When using voltage dividers, we encounter a trade-off: for certain reasons we want to construct the divider using resistors with as small resistance values as possible, while on the other hand we want the resistances to be as large as possible. Explain this contradiction and the reasons behind it.
25%	What condition must be satisfied when connecting a load to a voltage divider consisting of resistors R_1 and R_2 in order for the load to be current-driven?
50%	Two batteries with Thevenin voltages U_{t1} and U_{t2} and Thevenin internal resistances R_{t1} and R_{t2} are connected in parallel and then connected to a load R_b. • Sketch the corresponding circuit. • Write the expression for the Thevenin resistance of the combined source. • Write the expression for the Thevenin voltage of the combined source. • Write the expression for the voltage across the load.

Table 1. Exam questions used in the held-out evaluation. Only question text and weights are disclosed; no student data are shared. Questions are translated from Slovenian; translation is provided for readability.

4.3. Evaluation Metrics

We evaluate the grading system using exam-level metrics that quantify agreement with human grading. As ground truth, we use the exam grade assigned by the lecturer of Class B, which is the only available reference.

Unless stated otherwise, all metrics are computed at the level of the full exam (three tasks), thereby evaluating the system in a true end-to-end setting. Importantly, we distinguish two sources of variability: (i) an internal ensemble used by the grading pipeline as part of its fixed design, and (ii) independent repetitions of the full evaluation used only to estimate experimental variability.

Pipeline parameter: ensemble size. Within a single pipeline execution, each task is graded by an ensemble of $K = 3$ independent, stateless model calls, and the resulting drafts are merged by a supervisor model into a single exam-level grade. We denote this final, supervisor-aggregated output for student i as g_i^{LLM}. The ensemble size K is a fixed pipeline parameter (not varied in the experiments), chosen early as a compromise between inference cost and robustness to occasional model failures.

Let N denote the number of students, g_i^{LLM} the exam grade assigned by the system for student i, and g_i^{H} the corresponding grade assigned by the human lecturer. We define the signed grading difference as $\Delta_i = g_i^{\mathrm{LLM}} - g_i^{\mathrm{H}}$.

Mean Absolute Difference (μ). Overall grading accuracy is measured using the mean absolute difference between automated and human-assigned grades:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} |\Delta_i| \,. \tag{1}$$

Standard Deviation of Absolute Differences. To capture the variability of grading errors across students, we

compute the standard deviation of absolute differences:

$$\sigma_{|\Delta|} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(|\Delta_i| - \mu)^2}. \tag{2}$$

Grading Bias. Systematic over- or under-grading is quantified by the mean signed difference:

$$b = \frac{1}{N}\sum_{i=1}^{N}\Delta_i. \tag{3}$$

Manual Review Trigger Rate. In addition to agreement with human grades, we quantify the expected amount of manual consolidation required when multiple automated graders disagree. This metric does not rely on human reference grades and is computed solely from the ensemble outputs.

Let $s_{i,k}$ denote the exam-level score assigned to student i by grader k within the ensemble, with $k = 1, \ldots, K$. For a given disagreement threshold $D_{\max}$, a manual review is triggered for student i if the maximum pairwise disagreement between graders exceeds the threshold, i.e.,

$$\max_k s_{i,k} - \min_k s_{i,k} \geq D_{\max}. \tag{4}$$

We define the corresponding trigger indicator as

$$T_i(D_{\max}) = \begin{cases} 1, & \text{if a review is triggered for student } i, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The manual review trigger rate at threshold $D_{\max}$ is then given by

$$\mathrm{TR}(D_{\max}) = \frac{1}{N}\sum_{i=1}^{N} T_i(D_{\max}). \tag{6}$$

Here $D_{\max}$ is measured in absolute points on the 0–100 exam scale. In this work, $\mathrm{TR}(D_{\max})$ is estimated at the *exam level*, using final exam scores produced by each grader. In practical deployments, the same criterion could be applied at a finer granularity, such as the level of individual questions or answers, to further localize and reduce the required amount of human intervention.

Experimental parameter: evaluation repetitions. Separately from the pipeline ensemble, we repeat the *entire* evaluation $R = 3$ times to measure run-to-run variability due to stochastic model outputs. Each repetition corresponds to a full rerun of the grading pipeline (including all model calls and supervisor aggregation) with the system configuration unchanged. For each metric, we report the three per-run values, together with their mean and standard deviation across the R repetitions. The choice $R = 3$ reflects a practical compromise between robustness and the computational cost of repeated multimodal inference.

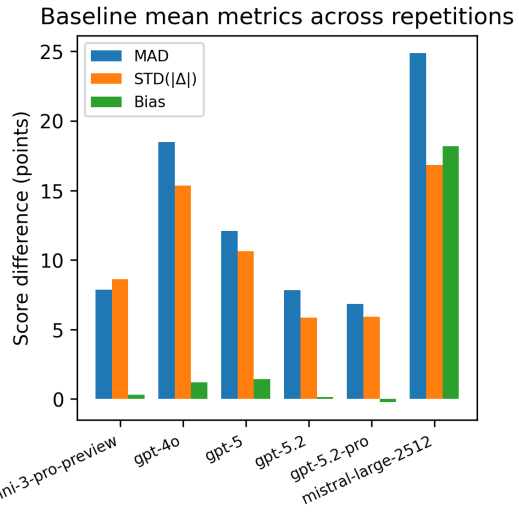


Figure 3. Baseline exam-level performance across backends (mean over $R = 3$ repetitions where available; GPT-5.2-pro: single run). **Legend note:** In all figures, "MAD" denotes the mean absolute difference (μ) and "Bias" denotes the mean signed difference (b) as defined in Sec. 4.3.

5. Results

All results are reported on the held-out Class B exam (Sec. 4). We use the exam-level metrics defined in Sec. 4.3 (μ, $\sigma_{|\Delta|}$, b) and, where relevant, the manual review trigger rate based on grader disagreement.

Legend note. In all figures, MAD denotes the mean absolute difference (μ), STD($|\Delta|$) the standard deviation of absolute differences ($\sigma_{|\Delta|}$), and Bias the mean signed difference (b), all measured in points on the 0–100 exam scale.

Unless noted otherwise, each backend was evaluated with the full pipeline configuration and repeated $R = 3$ times to estimate experimental variability; GPT-5.2-pro is reported from a single run due to cost. OpenAI models were accessed via the official OpenAI API [17], while other models were accessed via OpenRouter [20].

Model viability screening. Figure 3 compares backends under the full pipeline. The strongest backends (GPT-5.2, GPT-5.2-pro, and Gemini-3 Pro) achieve single-digit μ with low bias, indicating close agreement with the lecturer's exam grades. In contrast, GPT-4o and Mistral 3 exhibit substantially larger deviations; Mistral 3 also shows a pronounced positive bias, consistent with systematic over-grading on this exam.

Ablation on pipeline guidance. To isolate the effect of prompt engineering and reference conditioning, we evaluate two strong backends under a trivial prompting regime (Fig. 4). In this setting, the prompts only enforce the

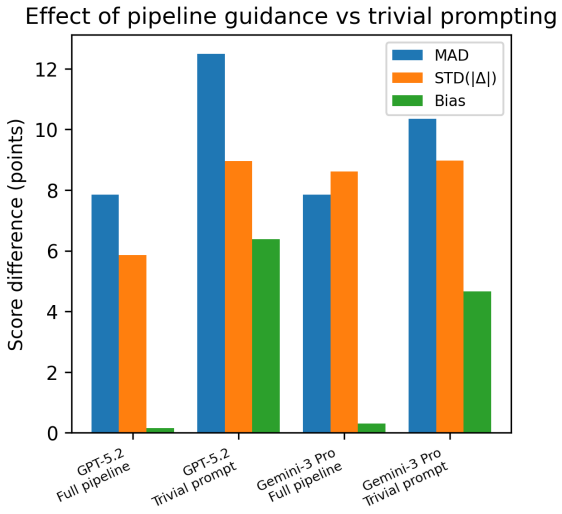


Figure 4. Full pipeline vs. trivial prompting for two strong backends (trivial: no rules, no reference, no supervisor).

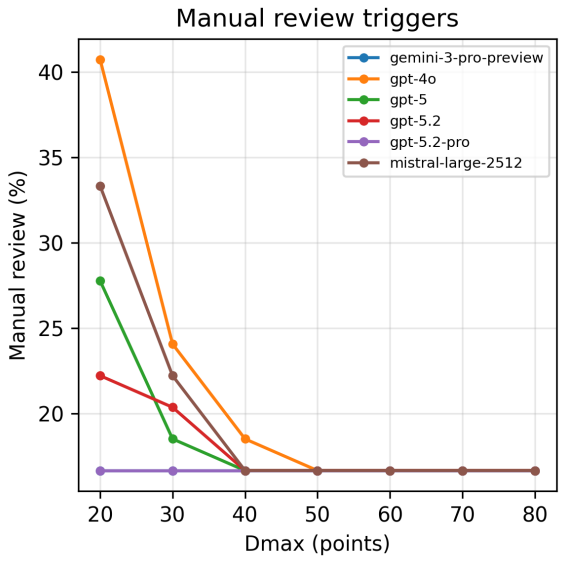


Figure 5. Estimated manual review trigger rate vs. disagreement threshold $D_{\max}$ (mean over repetitions where available). Note that the rate never falls to zero – this means there are samples where reviewers always disagree.

Markdown output format, but omit grading rules, omit the reference-solution summary, and disable supervisor aggregation. Because aggregation is disabled, we treat the three per-grader outputs as independent grading attempts and report their mean. For both GPT-5.2 and Gemini-3 Pro, trivial prompting increases μ and introduces a strong positive bias, confirming that structured prompting and reference grounding are necessary to obtain reliable grading.

Estimated manual review workload. Figure 5 reports the fraction of submissions that would require manual consolidation as a function of the disagreement threshold $D_{\max}$, computed at the exam level from the ensemble grader outputs. At strict thresholds (e.g., $D_{\max} = 20$–30 points), weaker backends yield substantially higher review rates, reflecting less stable grading. For larger thresholds the trigger rate drops for all models, indicating that only a small subset of submissions exhibit severe grader disagreement. This analysis complements accuracy metrics by quantifying the expected human effort required to safely deploy the system.

Ablation on reference conditioning. Table 2 ablates reference usage for the two best backends (GPT-5.2, Gemini-3 Pro) under three regimes: *Full pipeline* (reference extracted to text and injected into prompts), *No reference*, and *Image reference* (reference image only, no text extraction). Each cell reports run1/run2/run3 and mean±std; best results per model and metric are boldfaced. We report μ, $\sigma_{|\Delta|}$, b, and the exam-level manual review trigger rate $TR(D_{\max} = 40)$ from per-grader scores. For GPT-5.2, removing the reference increases μ and introduces a strong positive bias; image-only reference partly recovers μ but remains biased, highlighting the role of text extraction for calibrated scoring. For Gemini-3 Pro, both ablations primarily increase positive bias with limited gains in $\mu/\sigma_{|\Delta|}$, consistent with pipeline development being carried out mainly on GPT-5 models, yielding a larger benefit for GPT-5.2 than for Gemini.

Student feedback (Class A). We collected preliminary student feedback in Class A after real use of the system on 8 weekly AI-graded quizzes. A total of 14 students completed an anonymous questionnaire; at the time of the survey, students had already received detailed PDF feedback and had access to a complaint process. Attitudes toward AI-first grading were mostly positive: 64% reported a positive stance or preference, 21% negative, and 14% indifferent. When asked whether they benefit from the system, 71% answered yes (29% no). Reported benefits (multiple-choice) were primarily detailed explanations (41%), perceived fairness/impartiality (34%), and fast turnaround (21%). The main concerns were missed answers (35%), more mistakes than professors (26%), and changed exam difficulty (22% less demanding, 17% more demanding). Overall, 43% judged that advantages outweigh disadvantages, 29% the opposite, and 21% reported no difference.

6. Discussion

To the best of our knowledge, the literature does not yet describe a comparable end-to-end framework that grades *scanned, multi-page, handwritten STEM exams with diagrams* using multimodal LLMs while producing deterministically parseable outputs with explicit guardrails, ag-

Model	Regime	μ	$\sigma_{\lvert\Delta\rvert}$	b	$\mathbf{TR}(D_{\max} = 40)$ [%]
GPT-5.2	Full pipeline	**7.7/8.4/7.4** (7.8±0.4)	**4.2/6.5/6.9** (5.9±1.2)	**0.6/-0.4/0.3** (0.2±0.4)	**16.7/16.7/16.7%** (16.7±0.0)%
	No reference	9.8/11.2/9.5 (10.2±0.8)	8.4/6.4/6.0 (6.9±1.1)	6.5/6.4/6.1 (6.4±0.2)	22.2/16.7/16.7% (18.5±2.6)%
	Image reference	7.7/8.2/8.2 (8.1±0.2)	5.9/4.9/7.1 (6.0±0.9)	3.9/3.4/4.8 (4.0±0.5)	**16.7/16.7/16.7%** (16.7±0.0)%
Gemini-3 Pro	Full pipeline	**7.1/9.8/6.7** (7.9±1.4)	5.0/15.1/5.7 (8.6±4.6)	**3.0/-2.7/0.7** (0.3±2.3)	**16.7/16.7/16.7%** (16.7±0.0)%
	No reference	7.4/8.9/9.6 (8.6±0.9)	6.8/8.2/8.6 (7.9±0.8)	6.6/8.4/9.0 (8.0±1.0)	**16.7/16.7/16.7%** (16.7±0.0)%
	Image reference	7.9/8.5/8.4 (8.3±0.3)	**6.9/6.2/7.4** (6.9±0.5)	7.1/7.8/7.7 (7.5±0.3)	**16.7/16.7/16.7%** (16.7±0.0)%

Table 2. Ablation study on the two best-performing backends under three reference regimes: *Full pipeline* (reference extracted into text and injected into prompts), *No reference*, and *Image reference* (reference image only, no text extraction). Each cell reports run1/run2/run3 and mean±std across runs. Metrics are exam-level μ, $\sigma_{|\Delta|}$, b, and the manual review trigger rate $TR(D_{\max} = 40)$ computed from per-grader exam scores. Best results per model and metric are highlighted in bold (lowest mean; for b, smallest absolute mean).

gregation, and auditable reports. Our results show that, with such workflow design, modern multimodal backends can grade short engineering quizzes with agreement close enough to enable practical use with limited manual intervention. In informal discussions with instructors, we repeatedly encountered skepticism that this would be achievable for unconstrained handwriting and sketches at the level of accuracy reported here; these experiments provide evidence that the capability is now real when the system is engineered around model failure modes rather than idealized prompts. A key contextual point is timing: in our experience, this type of end-to-end approach only became practically viable with the late-2025 generation of multimodal *reasoning*-capable models released by major providers (e.g., GPT-5/GPT-5.2 and Gemini 3) [8, 18, 19]. Earlier backends in our screening exhibit substantially larger deviations and stronger bias (Fig. 3), reinforcing the need for both capable models and systems-level safeguards.

The presented pipeline was built first for real instructional use, not as a benchmark-optimized research prototype. Accordingly, our evaluation is preliminary: results are reported on one held-out quiz with a single human grader as reference, and the underlying exam data cannot be released in its raw form due to privacy constraints. We therefore plan to expand validation to a larger and more diverse collection and, following privacy review and institutional approval, release the code, prompts, and an accompanying dataset suitable for standardized evaluation. The absence of widely usable end-to-end datasets for authentic handwritten exam grading remains a practical barrier for the field; we view this work as an initial step toward making such evaluation feasible and repeatable.

7. Conclusion

We presented an end-to-end workflow for grading scanned handwritten engineering exams with multimodal LLMs. The core contribution is a robust system design that couples prompt pairs and rigid templates with deterministic validation, ensemble grading, and supervisor aggregation to turn probabilistic model behavior into auditable grading artifacts. On a held-out real course quiz, state-of-the-art multimodal backends achieve close agreement with lecturer grades and manageable estimated manual-review rates, indicating that deployment is plausible for short formative assessments. We release this as preliminary evidence that automated grading of realistic handwritten STEM work is now achievable under careful workflow constraints, and we plan broader evaluation and open-sourcing (with an accompanying dataset) after further validation and privacy review.

Ethical considerations

The ethical consideration of this research research aims to protect both study participants and future users of the proposed technology. Key ethical concerns include the handling of sensitive data (e.g., grades), the risk that participation in the experiment could affect student performance, and the power imbalance between students and instructors.

Our approach is guided by the three principles of the Belmont Report [29]:

1. **Respect for persons**: safeguarding autonomy through dignity, agency, and informed consent;
2. **Beneficence**: maximizing potential benefits while minimizing risks and harms; and
3. **Justice**: ensuring equitable treatment and a fair distribution of benefits and burdens.

In this context, a central consideration is whether the anticipated benefits justify the burdens placed on the af-

fected population. Potential benefits to students include (i) more objective grading and (ii) higher-quality feedback from their instructors. Individualized feedback is especially valuable because it is often infeasible for instructors to provide at scale. The resulting feedback may also support improvements to course design and teaching practices. Importantly, the automated grading system did not affect students' official course or exam grades; all assessments were graded manually as they would have been without the study.

A full-scale deployment would require formal ethical review, including evaluation of the experimental design and data-protection measures. Nonetheless, we conclude that the anticipated benefits to students outweigh the associated burdens.

Disclosure of AI use

The whole orchestration pipeline and the experimental code (approximately 13.000 lines of python code) has been written with the help of GPT5-codex tool and GPT5-pro models by OpenAI. Search for related work was done using OpenAI's AI agent (DeepResearch), and manuscript text was written by dictating the contents to the GPT5-pro. The manuscript has been throroughly checked by the authors and revised where necessary.

Supplementary material

We provide the scanned reference solution of the "Class B" exam, and a corresponding mock solution, which was actually graded by the pipeline using GPT-5.2 with thinking set to "high". It can be accessed via the following link: https://lmi.fe.uni-lj.si/en/janez-pers-2/supplementary-material/

Acknowledgements

We acknowledge the support of the EC/EuroHPC JU and the Slovenian Ministry of HESI via the project SLAIF (grant number 101254461). This research was also supported by project P2-0246 ICT4QoL - Information and Communications Technologies for Quality of Life and ARIS research program P2-0095.

References

[1] Sami Baral, Anthony F. Botelho, John A. Erickson, Priyanka Benachamardi, and Neil T. Heffernan. Improving automated scoring of student open responses in mathematics. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, pages 130–138, Paris, France, 2021. International Educational Data Mining Society. Conference dates: June 29–July 2, 2021. 3

[2] Sami Baral, Anthony Botelho, Abhishek Santhanam, Ashish Gurung, Li Cheng, and Neil Heffernan. Auto-scoring student responses with images in mathematics. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 362–369, Bengaluru, India, 2023. International Educational Data Mining Society. 3

[3] Sami Baral, Lucy Li, Robert Knight, Alice Ng, Luca Soldaini, Neil Heffernan, and Kyle Lo. DrawEduMath: Evaluating vision language models with expert-annotated students' hand-drawn math images. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7110–7132. Association for Computational Linguistics, 2025. arXiv version: https://arxiv.org/abs/2501.14877. 3

[4] Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25 (1):60–117, 2015. 2

[5] Adriana Caraeni, Alexander Scarlatos, and Andrew S. Lan. Evaluating GPT-4 at grading handwritten solutions in math exams, 2024. 2, 3

[6] Alberto Gandolfi. GPT-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *International Journal of Artificial Intelligence in Education*, 35(1):367–397, 2024. 2

[7] Christian Gold and Torsten Zesch. Handwritten ASAP short answer scoring. Zenodo dataset, 2020. Version 1.0; accessed 2025-12-14. 3

[8] Google. A new era of intelligence with Gemini 3. https://blog.google/products/gemini/gemini-3/, 2025. Accessed: 2025-12-23. 8

[9] Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. Survey on automated short answer grading with deep learning: from word embeddings to transformers, 2022. 2

[10] Owen Henkel, Adam Boxer, Libby Hills, Bill Roberts, and Zachary Levonian. Can large language models make the grade? an empirical study evaluating LLMs' ability to mark short answer questions in K-12 education. In *Proceedings of the 11th ACM Conference on Learning @ Scale (L@S)*, pages 300–304, 2024. arXiv version: https://arxiv.org/abs/2405.02985. 2

[11] Gerd Kortemeyer. Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, 4(1):47, 2024. Open PDF also available via ETH Research Collection (see Springer page for links). 2

[12] Gerd Kortemeyer, Julian Nöhl, and Daria Onishchuk. Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. *Physical Review Physics Education Research*, 20(2): 020144, 2024. 2, 3

[13] Andrew S. Lan, Divyanshu Vats, Andrew E. Waters, and Richard G. Baraniuk. Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In *Proceedings of the Second ACM Conference on Learning @ Scale (L@S)*, pages 167–176, 2015. 2

[14] Learnosity. A third of US teachers considered leaving education in last 12 months due to grading workload. Learnosity EdTech Blog (Press Release), 2025. Published 26 Mar 2025. 1

[15] Tianyi Liu, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. AI-assisted automated short answer grading of handwritten university level mathematics exams, 2024. HTML:

https://arxiv.org/html/2408.11728v1. 2, 3

[16] Ahmed E. Magooda, Mohamed A. Zahran, Mohsen A. Rashwan, Hazem M. Raafat, and Magda B. Fayek. Vector based techniques for short answer grading. In *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 238–243, 2016. 2

[17] OpenAI. OpenAI API Reference. Online documentation, 2025. Accessed 2025-12-23. 6

[18] OpenAI. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2025-12-23. 8

[19] OpenAI. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025. Accessed: 2025-12-23. 8

[20] OpenRouter. OpenRouter API Documentation. Online documentation, 2025. Accessed 2025-12-23. 6

[21] Behnam Parsaeifard, Martin Hlosta, and Per Bergamin. Automated grading of students' handwritten graphs: A comparison of Meta-Learning and Vision-Large language models, 2025. 3

[22] Vijay Rowtula, Subba Reddy Oota, and C. V. Jawahar. Towards automated evaluation of handwritten assessments. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 19–24, 2019. 2

[23] Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1049–1054, Denver, Colorado, 2015. Association for Computational Linguistics. 2

[24] Carolin Schwab, Anne C. Frenzel, Jordan Jaeger, Allison BrckaLorenz, and Robert H. Stupnisky. How do university faculty feel about grading? insights from a control-value theory perspective. *Studies in Higher Education*, 49(8): 1486–1503, 2024. Published online 15 Oct 2023. 1

[25] Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter Abbeel. Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work. In *Proceedings of the Fourth ACM Conference on Learning @ Scale (L@S)*, pages 81–88, 2017. 2

[26] Sargur N. Srihari, Rohini Srihari, Pavithra Babu, and Harish Srinivasan. On the automatic scoring of handwritten essays. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2880–2884, 2007. 2

[27] Le Ying Tan, Shiyu Hu, Darren J. Yeo, and Kang Hao Cheong. A comprehensive review on automated grading systems in STEM using AI techniques. *Mathematics*, 13 (17):2828, 2025. 1

[28] Youki Terada and Stephen Merrill. Why teachers should grade less frequently. Edutopia, 2024. Published 08 Nov 2024. 1

[29] The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The belmont report: Ethical principles and guidelines for the protection of human subjects of research, 1979. Published April 18, 1979. U.S. Department of Health, Education, and Welfare. Hosted by the U.S. HHS Office for Human Research Protections (OHRP). 8

29^{th} *Computer Vision Winter Workshop*
Jindřichův Hradec, Czech Republic, February 9–12, 2026

Exploring Multimodal Large Language Models for Morphing Attack Detection

Nikola Marić
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
nikola.maric@ijs.si

Marija Ivanovska, Vitomir Štruc
Faculty of Electrical Engineering, University of Ljubljana
Tržaška c. 25, 1000 Ljubljana
{marija.ivanovska, vitomir.struc}@fe.uni-lj.si

Abstract

Existing single-image morphing attack detection (S-MAD) systems often suffer from poor cross-dataset generalization and operate as opaque "black boxes," which is particularly problematic in high-stakes border control scenarios. This paper investigates the adoption of open-source Multimodal Large Language Models (MLLMs) for S-MAD under strict cross-dataset evaluation through two different approaches. First, we assess selected MLLMs in zero-shot settings using a structured forensic prompting framework, which elicits multi-step semantic analysis with human-readable regional attributions. Second, leveraging the lightweight and parameter-efficient LoRA approach and a synthetic training dataset of morphs, we adapt the best-performing MLLM to the morphing attack detection (MAD) task in an efficient, generalizable, and privacy-preserving manner, enhancing the model's sensitivity to diverse morphing artifacts. Our experimental results show that the proposed prompting strategy significantly improves overall attack detection accuracy compared to naive prompting. Moreover, our LoRA-adapted MLLM, Gemma-3 12B, achieves an average equal error rate (EER) of 14.81% across various morphing attack benchmarks, outperforming widely used MAD models.

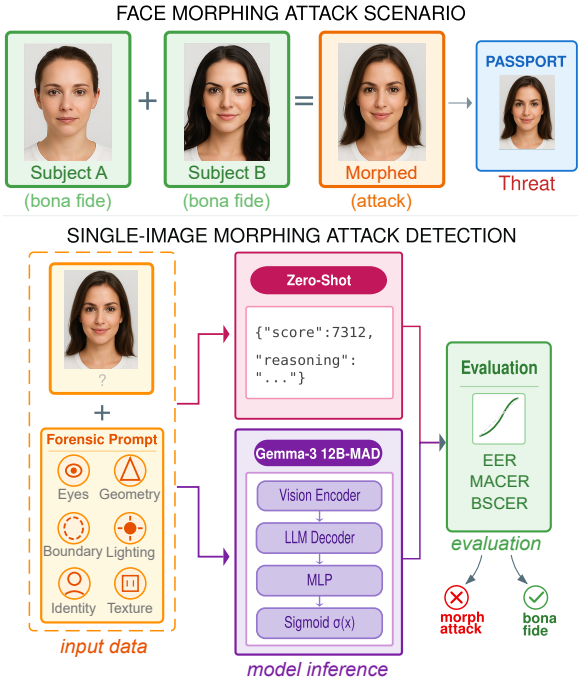


Figure 1. Face morphing attack threat and proposed detection pipeline: Blended facial identities compromise document security (top); our framework performs detection through zero-shot structured forensic analysis or fine-tuned classification (bottom).

1. Introduction

Face-morphing attacks pose a serious threat to the integrity of biometric security systems by blending facial images of two individuals into a single composite image that simultaneously represents both identities, as illustrated in Figure 1 [15, 19]. By embedding such a morphed image into an identity document, such as a passport, an attacker and an accomplice can jointly and repeatedly bypass automated face matching systems during identity verification [7]. This fundamental vulnerability has motivated the development of dedicated morphing attack detection (MAD) techniques aimed at reliably distinguishing bona fide facial images from morphed ones [7, 9, 14, 21].

MAD methodologies are generally categorized into differential and single-image approaches. Differential MAD methods assess the authenticity of a presented face image by directly comparing it to a trusted reference image acquired during enrollment [6, 33]. In contrast, single-image MAD operates solely on the probe image, making it particularly suitable for practical real-world border control scenarios where reliable reference images are often unavailable, outdated, or of insufficient quality [7]. In this paper, we limit our focus to the more challenging single-image morphing attack detection (S-MAD) task, which poses stricter constraints on available information.

While traditional S-MAD detectors rely on handcrafted features, modern approaches leverage deep learning supervised classifiers, which often suffer from severe generalization issues. Recent studies show cross-dataset equal error rates degrading to near-random performance when models are trained on one type of morphing attack techniques and tested on another [11, 14]. This ef-

fect is caused not just by the domain shift, but also due to the variability of morphing attack artifacts that characterize individual morphing techniques [22]. This issue has been tackled with the development of different unsupervised [14, 16, 20] and self-supervised MAD techniques [21, 22], but these methods either fail to learn a well-defined boundary between bona fides and morphs, or lack human-interpretable results of the decision [39].

Recent advances in foundation models, spanning vision-only architectures (e.g., ViT), vision–language models (e.g., CLIP), and Multimodal Large Language Models (MLLMs) capable of more complex reasoning, offer a promising route to simultaneously improve MAD generalization and interpretability. General-purpose vision–language models, such as CLIP, have previously shown competitive MAD performance when adapted to the downstream MAD task [7], but they lack explicit semantic reasoning, which is important for forensic analysis in security-critical applications. Conversely, proprietary MLLMs such as GPT-4 have demonstrated both impressive zero-shot detection capabilities and decision interpretability, as reported in preliminary MAD studies [1, 39]. However, their closed-source nature limits its reproducibility, transparency, and task-specific adaptation, motivating the exploration of open-source MLLMs for morphing attack detection. These observations leave two fundamental questions unanswered: *i)* whether open-source MLLMs can match or even exceed specialized CNN-based MAD systems under strict cross-dataset evaluation, and *ii)* whether parameter-efficient adaptation on privacy-friendly synthetic data can equip such models with robust and explainable morphing detection capabilities without overfitting to specific attack processes.

In this paper, we address these questions through a systematic investigation of widely used open-source MLLMs for single-image morphing attack detection, focusing on reproducible, locally deployable architectures rather than proprietary APIs. Our contributions are multifold:

- We conduct the first systematic zero-shot and cross-dataset benchmarking of open-source MLLMs across various, i.e. landmark-, GAN-, and diffusion-based morphing attacks, revealing pronounced differences in their latent forensic sensitivity to MAD.
- We introduce a structured multi-step forensic prompting protocol that leverages chain-of-thought (CoT) reasoning to substantially improve zero-shot morphing attack detection performance over naive prompting, while simultaneously providing interpretable, region-level semantic attributions of detected morphing artifacts.
- We propose a self-supervised, parameter-efficient MLLM adaptation strategy that leverages privacy-preserving synthetic data to achieve strong cross-dataset generalization and competitive performance against widely used state-of-the-art MAD methods.

The remainder of the paper is organized as follows. Section 2 reviews related work on MADs. Section 3 describes our proposed MAD approach. Sections 4 and 5 present experiments and results. Section 6 concludes the paper.

2. Related Work

Face morphing attacks have evolved from early landmark-based warping techniques to increasingly sophisticated generative approaches. Modern attacks span classical landmark-based morphs [13, 28], GAN-based synthesis [8, 38], and more recent diffusion-based morphs [3, 11]. This rapid progression has, in turn, driven the evolution of Morphing Attack Detection (MAD) methodologies, which have advanced from hand-crafted forensic features to supervised deep learning approaches, and more recently to generalized unsupervised frameworks and foundation model-based solutions.

Early Single-Image MAD. Early S-MAD approaches relied on hand-crafted texture descriptors and image forensics. Techniques employing Local Binary Patterns (LBP), Binarized Statistical Image Features (BSIF), and Photo Response Non-Uniformity (PRNU) analysis were effective at identifying blending artifacts or sensor noise inconsistencies [12, 27, 32]. In parallel, Image Quality Assessment (IQA) strategies, such as MagFace [23] and CNNIQA [16], leveraged the observation that morphing processes often degrade facial utility or natural image statistics. These methods established important baselines and remain relevant as quality-based detectors in our comparative evaluation. However, their reliance on low-level cues limits robustness against high-quality, seamless morphs produced by advanced generation techniques, resulting in poor generalization across datasets [31].

Supervised CNN-based MAD. The advent of deep learning shifted the focus toward supervised Convolutional Neural Networks (CNNs) as the dominant paradigm for morphing attack detection. Architectures such as MixFaceNet-MAD [4] and adaptations of Inception and ResNet [17, 35] demonstrated high intra-dataset detection accuracy under controlled laboratory conditions. To further enhance interpretability and spatial precision, Pixel-Wise MAD (PW-MAD) introduced explicit pixel-level supervision to localize morphing regions at fine granularity and provide more transparent decision cues [9]. Despite achieving strong detection performance on known attack types, supervised methods remain prone to severe overfitting to training distributions and dataset-specific artifacts. As a result, they often fail when confronted with previously unseen morphing techniques, such as diffusion-generated morphs, particularly when trained exclusively on landmark-based examples [14]. This lack of robustness highlights a fundamental limitation of CNN-based detectors, whose performance can degrade substantially when exposed to novel and rapidly evolving attack generation mechanisms encountered in real-world deployments.

Unsupervised and Self-Supervised MAD. To address the generalization gap, recent research has pivoted toward the unsupervised and self-supervised learning paradigms, where MADs are trained on bona fide data only, treating morphs as out-of-distribution samples. Self-Paced Learn-

ing MAD (SPL-MAD) [14] and MAD-DDPM [20], for example, train reconstruction models that, based on the reconstruction error during testing time flag, morphs as statistical outliers. To improve the estimation of the bona fide distribution, approaches such as OrthoMAD [25] and IDistill [5] optimize their models by simultaneously performing identity disentanglement. Some recent works, e.g., SelfMAD [21] and SelfMAD++ [22], also leverage self-supervised signals, that train the model in a binary manner, by utilizing synthetic morphs that represent typical morphing artifacts, created using augmentations of bona fide data. These methods have significantly reduced cross-dataset error rates by learning generic definitions of face authenticity rather than memorizing specific attack signatures. Our LoRA-based adaptation of MLLMs follows a similar path, as training is performed exclusively on bona fide images with online generation of synthetic artifacts to preserve generalization to unseen attacks.

Foundation Models for MAD. Most recently, the emergence of foundation models has opened a new frontier in MAD. *Caldeira et al.* proposed MADation [7], which fine-tunes the CLIP vision-language model using Low-Rank Adaptation (LoRA) [18], achieving state-of-the-art generalization by leveraging broad pre-trained visual knowledge. Concurrently, *Caldeira et al.* introduced MAD-Prompts [6], exploring multi-prompt aggregation for zero-shot MAD with proprietary MLLMs. However, their study remains limited to closed-source APIs and does not investigate parameter-efficient adaptation or open-weights models. Furthermore, zero-shot evaluations using Multimodal Large Language Models (MLLMs) like GPT-4 Vision have shown that these models possess inherent forensic capabilities, offering both detection and textual explanations without task-specific training [1, 39].

In contrast to MADation, which adapts only the CLIP vision-language model without explicit reasoning mechanisms, our work employs MLLMs that integrate vision and language for semantic analysis. We introduce a structured multi-step Chain-of-Thought forensic prompting protocol for zero-shot MAD. Moreover, we implement self-supervised LoRA fine-tuning of MLLMs applied to both the vision and language components of the models, for their adaptation to the downstream MAD task. Unlike existing MAD methods, our approach provides both interpretability through structured reasoning and cross-dataset performance across diverse morph types.

3. Methodology

In this section, we present two distinct options related to the adoption of open-source MLLMs for MAD. First, we propose a *zero-shot forensic prompting strategy* designed to elicit latent expert knowledge from off-the-shelf models without parameter updates. Second, we introduce a *synthetic-data-driven MLLM adaptation*, where we fine-tune an MLLM to a downstream task using on-the-fly generated synthetic artifacts. The latter approach aims to learn generalized representations of morphing attacks without relying on labeled datasets of specific morphing algorithms, thereby addressing the critical issue of overfitting in current MAD approaches.

3.1. Zero-Shot Forensic Prompting Strategy

Standard prompting strategies (e.g., asking "Is this face morphed?") fail to produce reliable results in forensic contexts, often leading to model hallucinations or refusals due to safety alignment [24]. To overcome this issue, we developed a structured prompting methodology grounded in Chain-of-Thought (CoT) reasoning [37], transforming the MLLM from a passive classifier into an active forensic analyst. We additionally condition the MLLM with a forensic analyst system prompt to reduce generic safety refusals and anchor the model in the MAD context.

Structured Analytical Protocol. Our approach moves beyond binary classification by implementing a six-step analytical protocol inspired by NISTIR 8584 [26] guidelines. This protocol explicitly guides the model's attention to anatomical face regions where morphing artifacts typically appear. These six steps are presented as numbered sub-questions in the prompt, and the model must provide a brief textual assessment for each before issuing a final decision. Our proposed prompt sequentially evaluates:

1. *High-Frequency Features:* Scrutinizing fine-grained details around the eyes and lips for ghosting, double edges, or unnatural sharpness discontinuities.
2. *Facial Geometry:* Detecting subtle asymmetries, spatial misalignments, or warping inconsistencies introduced by landmark-based blending operations.
3. *Skin Texture Analysis:* Identifying unnatural smoothing, loss of skin porosity, or texture homogenization commonly observed in attacks generated with deep learning-based methods or heavily retouched imagery.
4. *Boundary Consistency:* Checking for blending artifacts, color mismatch, or edge disruptions at common areas of interest such as hairline, jawline, face contour.
5. *Lighting Coherence:* Verifying consistent illumination direction, shadow placement, and reflectance properties across different facial regions in the image.
6. *Identity Integrity:* Performing a holistic assessment of overall identity coherence, ensuring that facial attributes remain semantically consistent and plausible.

Semantic Scoring and Output Constraints. A key challenge in zero-shot evaluation with MLLMs is the reliable extraction of calibrated, continuous confidence estimates for quantitative performance analysis. In preliminary experiments, we observed that coarse confidence scales (e.g., 0–100, where lower values indicate bona fide images and higher values indicate morphs) induce pronounced score quantization, with predictions collapsing onto a small set of discrete values. This behavior reduces score resolution and degrades the reliability of threshold-based evaluation metrics used to quantify detection error.

To mitigate this issue, we adopt a high-resolution confidence scale ranging from 0 to 10,000, coupled with an explicit semantic interpretation of score intervals. This

choice is not intended to increase numerical precision in a statistical sense, but to counteract the tendency of MLLMs to collapse predictions onto a small set of preferred numeric tokens when prompted with coarse ordinal scales. Low-resolution ranges (e.g., 0–100) encourage categorical reasoning and rounded outputs, whereas a larger numeric range supports finer-grained ordinal differentiation. To further stabilize score usage, semantic anchors are defined within the prompt. Scores above 9,000 thus indicate high certainty of a morph, while scores in the 1,000–3,000 range denote likely bona fide images. This guides the model to utilize a full dynamic range and yields smoother score distributions for threshold-based evaluation.

To support automated evaluation and reproducibility, we constrain the model output to a strict JSON schema of the form {`"step1_reasoning": "...", "step1_score": "...", ...`}. For each of the six forensic analysis steps introduced above, the model is required to produce textual reasoning and a score denoting whether the input image represents an attack. The final decision score is obtained by averaging the six step-wise confidence scores and is used for quantitative evaluation and threshold-based decision making. This structured output ensures machine parsability while enforcing a clear separation between reasoning and decision making, thereby improving interpretability and consistency across inference runs. The exact prompts used in our experiments are provided in the Appendix.

3.2. Synthetic-Data-Driven MLLM Adaptation

Zero-shot MLLMs rely solely on broad pretraining and prompt-based reasoning, without task-specific calibration to the subtle visual artifacts characteristic of morphing attacks. Consequently, their sensitivity to fine-grained, low-level inconsistencies, such as localized geometric distortions or frequency-domain artifacts, may be insufficient for reliable morph detection. These limitations motivate targeted adaptation of MLLMs to improve detection accuracy while preserving generalization. We adapt MLLMs using a binary training objective on synthetic data.

Generation of Synthetic Training Data. We adopt a training strategy that simulates typical morphing artifacts rather than using real morphs, similar to [21]. By defining the "attack" class through synthetic perturbations, we force the model to learn generic indicators of manipulation rather than the specific characteristics of actual morphing techniques (e.g., StyleGAN fingerprints). The pipeline for simulation of training data generates training pairs of bona fide and morphed images $(I_{\mathrm{BF}}, I_{\mathrm{M}})$ by processing bona fide inputs I through three separate stages:

- *Pixel-Space Artifact Simulation*: introduces artifacts that simulate irregularities created by landmark-based morphing techniques. Specifically, given an input bona fide image I, this stage first applies a set of randomly parametrized geometrical image transformations ζ:

$$I_{PA} = \zeta(I), \tag{1}$$

where ζ is randomly sampled from {`Translation`, `ElasticTransform`, `Scaling`}. The pixel-augmented image I_{PA} is blended with the source I using a binary blending mask M:

$$I'_{PA} = I_{PA} \odot a \cdot M + I \odot (1 - a) \cdot M, \tag{2}$$

where a is the blending factor, uniformly sampled from a predefined set $\{0.5, 0.5, 0.5, 0.375, 0.25, 0.125\}$.
- *Frequency-Space Artifact Simulation*: injects structured noise patterns into the Fourier spectrum of the blended image I'_{PA} to mimic the spectral inconsistencies introduced by deep learning morphing techniques, i.e., GANs and diffusion models. Specifically, this stage first creates a random structured geometrical pattern Φ, uniformly chosen to represent one of the following: a symmetrical grid, an asymmetrical grid, a square checkerboard, a circular checkerboard, randomly distributed squares, a set of random lines, or a set of random stripes. The magnitudes of its Fourier transform $F_\Phi = |\mathrm{FFT}(\Phi)|$ are then superimposed on the magnitudes of the Fourier transform of I'_{PA}, $F_{\mathrm{PA}} = |\mathrm{FFT}(I'_{\mathrm{PA}})|$, and transformed back to the image space, by applying the inverse Fourier Transformation FFT^{-1}:

$$I_{\mathrm{FA}} = \mathrm{FFT}^{-1}\big((1-k)F_{\mathrm{PA}} \oplus kF_\Phi\big), \tag{3}$$

where k is a constant that defines the contribution of Fourier spectra F_{PA} and F_Φ to the summation.
- *Visual Variability Simulation*: focuses on transforming the visual appearance of images to simulate subtle, global visual variations commonly encountered in real-world imagery. Specifically, given an input bona fide image I and a synthetic morph I_{FA}, this stage applies a set of randomly parametrized transformations ψ, to generate a bona fide image I_{BF} and morph I_M:

$$I_{BF} = \psi(I), \qquad I_M = \psi(I_{FA}) \tag{4}$$

where ψ is uniformly sampled from {`RGBShift`,`HueSaturationValue`, `RandomBrightnessContrast`, `RandomDownScale`, `Sharpen`} - a set comprising five basic (global) image transformations.

An example of a train pair (I_{BF}, I_M) is shown in Figure 2.

Parameter-Efficient MLLM Adaptation. Full fine-tuning of multi-billion-parameter models is computationally prohibitive and may lead to catastrophic forgetting of pre-trained knowledge. Therefore, we employ LoRA [18] to adapt our MLLM by injecting a small number of trainable parameters while keeping the pre-trained weights frozen. Specifically, for a frozen pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the weight of the adapted model is

$$W = W_0 + \Delta W,\ \Delta W = BA \tag{5}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices with rank $r \ll min(d, k)$.

(a) Bona Fide Image I_{BF}

(b) Simulated Morph I_M

Figure 2. Example of a synthetic training image pair used for the MLLM adaptation. (a) represents a bona fide image, (b) is generated via pixel-space and frequency-space artifact simulation.

(a) OpenCV

(b) StyleGAN2

(c) Greedy-DiM

Figure 3. Examples of a landmark-based (a), a GAN-based (b), and a diffusion-based (c) morph, all three generated using bona fide images from FRLL. The visual characteristics of morphs vary substantially depending on the morphing technique.

Importantly, during the adaptation of our MLLM, we apply LoRA adapters to query (q) and value (v) projections of the self-attention layers in *both* the Vision Encoder and the Language Decoder towers. This dual-tower strategy is essential for MAD, as adapting the vision tower allows the model to extract forensic visual cues (e.g., noise patterns) that are likely suppressed in standard pretraining, while adapting the language tower aligns the reasoning engine to the description of typical visual artifacts. Additionally, we append the final aggregated token output of the decoder with a lightweight MLP classification head, optimized using Binary Cross-Entropy (BCE) loss:

$$L_{\text{BCE}} = -\left[y \cdot \log(p) + (1-y) \cdot \log(1-p)\right] \quad (6)$$

where p is the model's probability to classify an image as a bona fide or a morph, while y is the ground truth label.

The classification head is trained jointly with the LoRA adapters in the vision and language towers, while original MLLM parameters are not updated as they are frozen.

4. Experiments

Experimental MLLMs. For our experiments, we select four different widely used open-source MLLMs: Gemma-3 (27B and 12B) [36], optimized for strong multimodal reasoning with efficient instruction tuning; Qwen2.5-VL 32B [2], known for robust visual understanding and multilingual reasoning; Llama-4-Scout 17B, designed for efficient deployment and fast multimodal inference; and Mistral Small 3.1 24B - a compact yet powerful MLLM emphasizing efficiency and strong language modeling performance. With this selection, we aim to cover various architectural innovations, including Mixture-of-Experts and varying parameter scales, ensuring our findings are robust across different model backbones. In our experiments all models are evaluated in zero-shot settings, while the parameter-efficient LoRA adaption is performed only with Gemma-3 12B, as the best-performing MLLM.

Testing Datasets. To ensure rigorous assessment of MLLMs' MAD performance, we evaluate models across different testing datasets spanning classical landmark-based morphs, GAN-, and diffusion-based attacks. Figure 3 illustrates how visual characteristics of morphs differ depending on the type of the underlying morphing technique. In our experiments, we utilize seven widely used benchmark datasets: FRLL-Morphs, FRGC-Morphs, FERET-Morphs [30], containing morphs generated with morphing algorithms AMSL, FaceMorpher (FM), OpenCV (OCV), StyleGAN2 (SG), and WebMorph (WM), and higher-quality sets MorGAN [8], MIPGAN-II [38], MorDIFF [11], and Greedy-DiM [3]. In addition to these datasets, we also use LMA-DRD [9], included to evaluate performance on printed and scanned images, to assess robustness against re-digitalization noise.

Training Data. During the synthetic-data-driven MLLM adaptation, we utilize the bona fide training subset of SMDD [10]. This subset consists of 25,000 synthetic images generated with StyleGAN2 utilized to generate simulated training morphs as described in Section 3.

Evaluation Metrics. In our evaluations, we follow the ISO/IEC 20059:2025 standard[1] by computing the Equal Error Rate (EER), where the Morphing Attack Classification Error Rate (MACER) equals the Bona Fide Sample Classification Error Rate (BSCER). MACER corresponds to the proportion of morphs incorrectly accepted as bona fide, whereas BSCER measures bona fide images falsely rejected as attacks. Beyond EER, in some experiments we also report BSCER at fixed MACER operating points of 1% and 5%. These stricter operating points more accurately reflect real-world identity verification tasks, where low attack-acceptance is essential for system security.

Implementation Details. Our experimental protocol explicitly differentiates between zero-shot evaluation of MLLMs and their adaptation to the MAD task.

During the *zero-shot evaluations*, all models were configured with a temperature of 0.1 to balance output determinism with the creative reasoning required for forensic analysis. Inference was performed using the vLLM engine on a cluster of four NVIDIA RTX 4090 GPUs. To fit memory constraints, Llama-4-Scout utilized 4-bit quantization, while other models used bfloat16. Images were preprocessed following the requirements of each MLLM.

During the lightweight adaptation of Gemma-3 12B,

[1]International Organization for Standardization (ISO). ISO/IEC 20059:2025 — Information technology — Biometric presentation attack detection — Testing and reporting. (This standard supersedes ISO/IEC 30107-3:2017.)

Table 1. Zero-shot MLLM evaluations vs. LoRA-adapted Gemma-3 performance in terms of EER(%). The LoRA-adapted Gemma-3 significantly outperforms all zero-shot baselines, including the classification head trained on top of the pretrained Gemma-3.

Dataset	Morph type	Zero-shot evaluations				Gemma-3 + classification head	
		Mistral Small 3.1	Llama-4-Scout	Qwen2.5-VL	Gemma-3	Pretrained backbone	LoRA adapted
FRGC-M	FM	49.20	—	42.41	32.19	24.48	**4.98**
	OCV	52.82	—	42.62	43.55	29.72	**9.23**
	SG	50.45	—	59.81	57.06	40.82	**17.32**
FERET-M	FM	53.07	—	34.52	18.20	13.80	**9.30**
	OCV	53.26	—	32.46	19.62	11.73	**7.40**
	SG	48.57	—	34.27	40.94	**27.40**	34.97
FRLL-M	AMSL	42.13	49.63	44.13	25.10	48.62	**12.74**
	FM	37.29	41.50	43.01	13.08	18.15	**0.49**
	OCV	40.24	37.63	39.51	13.33	6.38	**1.47**
	SG	52.06	47.59	26.86	27.39	16.72	**6.83**
	WM	40.78	39.22	41.06	12.88	21.47	**2.37**
LMA-DRD	D	49.01	—	45.40	47.05	31.56	**18.90**
	PS	51.50	—	47.43	43.88	39.74	**28.28**
MorGAN	GAN	55.60	—	48.63	52.58	46.69	**45.26**
	LMA	46.88	—	51.67	52.87	50.00	**23.06**
MIPGAN II	SG	50.24	46.58	20.75	35.56	31.67	**17.92**
Greedy-DiM	DiffAE	50.49	49.93	24.55	**6.15**	11.24	11.78
MorDIFF	DiffAE	47.77	—	45.91	36.13	24.88	**17.33**
Average		48.41	44.58	40.28	32.09	27.50	**14.98**

we leverage LoRA adapters with parameters $r = 16$, and $\alpha = 32$, injected into the query and value projections of all self-attention layers in the vision and the language tower of the MLLM. Weights were optimized for 30 epochs with an effective batch size of 32. To promote stable optimization, we employ a differential learning rate strategy across model components. Specifically, we use a higher learning rate for the randomly initialized classification head (1×10^{-4}), a moderate learning rate for the vision tower (6×10^{-6}), and a substantially lower learning rate for the language tower (3×10^{-7}). This design reflects the differing levels of sensitivity to parameter updates: the classification head requires rapid convergence from scratch, while the vision and language towers—adapted via LoRA—benefit from more conservative updates to preserve pre-trained representations and prevent destabilization of linguistic reasoning. The optimization was performed on two NVIDIA A100 (80GB) GPUs.

Comparison With Existing MADs. We assess the MAD performance of evaluated MLLMs against various established MAD methods. Among supervised baselines, we consider MixFaceNet-MAD [4], Inception-MAD [29], and PW-MAD [9]. As the performance of the supervised methods and their generalization strongly depend on the training data, we train each method on three different datasets, i.e., SMDD, MorGAN, and LMA-DRD, following a protocol established in [14]. In addition to supervised MADs, we also include comparison with unsupervised MADs FIQA-MagFace [16], CNNIQA [16], SPL-MAD [14], and MAD-DDPM [20], conceptually similar self-supervised models SBI [34] and SelfMAD [21], and the foundation model-based method MADation [7].

5. Results

MLLM Evaluation in Zero-Shot Settings. Results obtained during the zero-shot evaluation of selected MLLMs are summarized in Table 1. Among the four selected MLLMs, Gemma-3 27B achieved the best average EER of 32.09%, outperforming the runner-up Qwen2.5-VL by 8.19%. Both Llama-4-Scout and Mistral Small 3.1 performed substantially worse, with an overall EER of 44.58% and 48.4%, respectively. These results demonstrate that MLLMs possess different zero-shot capabilities for detecting morphed faces. The sensitivity of the models to specific morphing techniques also varies considerably. However, MLLMs in general achieve lower attack detection error when tested on artifact-rich morphs, as opposed to the accuracy measured on higher-quality morphed images. Gemma-3, for example, relatively accurately detects blending artifacts in FRLL FaceMorpher, OpenCV, and WebMorph attacks, with an EER ranging between 12.88% and 13.33%. Nevertheless, detection errors are significantly higher on FRGC-StyleGAN morphs (57.1% EER), probably due to the seamless latent-space interpolation performed by the morphing technique StyleGAN, which produces very few perceptible artifacts. Qualitative evaluation examples with corresponding confidence scores and reasoning are shown in Figure 4.

Evaluation of the Adapted MLLM. To isolate the impact of LoRA adaptation, we evaluate Gemma-3 12B using a lightweight, probability-based classifier rather than prompt-derived numeric scores. Specifically, we first attach and train an MLP classification head on top of the frozen (unadapted) MLLM and use its sigmoid output as the morph probability. This bypasses the need to interpret free-form numeric confidence values generated by the language decoder, which are subject to token-level biases and scale-dependent discretization. We then evaluate the LoRA-adapted Gemma-3 12B in the same manner. Results are reported in Table 1. As can be seen, the unadapted MLLM with an added classification head achieves an average EER of 27.50%, outperforming zero-

Table 2. Comparison of the LoRA adapted Gemma-3 with supervised MAD models trained on different datasets in terms of EER(%). Gemma-3 outperforms competitors in terms of average detection accuracy, achieving stable performance across different morph types.

Dataset	Morph type	MixFaceNet-MAD [4]					PW-MAD [9]					Inception-MAD [29]					Gemma-3
		D	PS	LMA	GAN	SMDD	D	PS	LMA	GAN	SMDD	D	PS	LMA	GAN	SMDD	[LoRA]
FRGC-M	OCV	23.81	25.04	31.62	21.11	20.67	57.06	48.60	29.74	53.55	26.45	34.32	13.65	36.17	59.66	19.63	**9.23**
	FM	22.83	23.54	29.38	19.98	18.10	56.00	50.70	30.49	51.61	23.40	34.96	19.71	35.10	56.91	16.06	**4.98**
	SG	32.71	28.68	21.70	21.95	**11.62**	37.38	38.42	16.43	26.62	14.32	41.14	25.85	36.19	47.03	15.26	17.32
FERET-M	OCV	28.12	32.19	31.57	33.86	31.74	37.27	45.29	34.27	43.11	39.93	**6.39**	7.23	42.12	13.62	59.32	7.40
	FM	22.57	29.48	27.90	31.81	23.69	35.16	44.30	28.24	40.40	29.41	**5.17**	6.91	36.53	18.36	46.94	9.30
	SG	29.57	29.02	35.46	39.41	39.85	44.25	45.30	29.70	42.47	47.20	9.03	**7.12**	35.29	15.09	60.05	34.97
FRLL-M	OCV	8.82	13.22	8.91	17.66	4.39	17.33	15.69	13.96	45.59	2.42	13.72	10.76	6.86	55.89	5.38	**1.47**
	FM	7.80	10.97	7.34	15.65	3.87	13.88	15.14	10.92	44.57	2.20	16.62	15.81	6.32	66.14	3.17	**0.49**
	SG	20.07	15.29	13.41	23.51	8.89	29.97	27.64	18.11	48.53	16.64	37.24	19.58	20.56	55.03	11.37	**6.83**
	WM	25.97	29.04	20.61	30.39	12.35	33.78	28.51	35.75	52.43	16.65	57.38	58.32	30.88	77.42	9.86	**2.37**
	AMSL	24.53	27.59	19.24	30.03	15.18	36.25	32.95	34.38	48.52	15.18	49.02	61.44	**9.80**	86.49	10.79	12.74
LMA-DRD	D	15.68	18.03	17.06	25.01	19.42	20.80	25.10	22.34	40.21	17.06	7.64	17.06	15.68	50.77	**15.11**	18.90
	PS	21.77	18.44	27.05	27.05	23.72	26.48	23.72	29.41	44.11	20.39	**11.37**	12.75	22.34	38.42	19.01	28.28
MorGAN	LMA	39.42	**22.89**	10.61	46.42	30.12	34.20	34.14	9.71	34.37	27.31	38.55	31.73	8.43	40.16	28.51	23.06
	GAN	53.01	50.44	42.57	24.90	42.64	52.04	46.59	42.80	8.84	43.78	50.84	38.79	**27.41**	0.40	44.34	45.26
Greedy-DiM	DiffAE	45.10	41.67	40.69	48.04	39.71	17.16	33.82	17.16	15.20	42.16	31.86	51.96	25.98	29.90	56.86	**11.78**
MorDIFF	DiffAE	21.30	23.70	28.83	30.19	20.40	3.21	**0.98**	11.60	16.00	13.80	21.08	21.78	19.41	56.09	15.23	17.33
Average		26.71	26.30	25.21	28.88	21.55	33.21	33.32	25.33	40.46	23.43	28.67	25.48	25.42	47.94	25.70	**14.81**

* Train data: **D** (LMA-DRD - digital), **PS** (LMA-DRD - print&screen), **LMA** (MorGAN - landmark-based), **GAN** (MorGAN - GAN-based), **SMDD**

Table 3. Comparison of the LoRA adapted Gemma-3 with unsupervised MAD models in terms of EER(%) and BSCER at MACER 5% and 10%. Gemma-3 shows competitive average EER, while highlighting complementary strengths with state-of-the-art MADs.

Dataset	Morph type	FIQA-MagFace [16]			CNNIQA [16]			SPL-MAD [14]			MAD-DDPM [20]			SBI [34]			SelfMAD [21]			MADation ViT-B[7]			MADation ViT-L[7]			Gemma-3 [LoRA]		
		EER	5%	10%	EER	5%	10%	EER	5%	10%	EER	5%	10%	EER	5%	10%	EER	5%	10%	EER	5%	10%	EER	5%	10%	EER	5%	10%
FRGC-M	FM	33.82	73.79	62.84	42.84	75.94	66.86	16.91	25.39	21.47	25.62	95.12	90.15	16.68	38.07	26.14	5.59	6.43	2.80	–	–	–	–	–	–	**4.98**	**4.91**	**2.15**
	OCV	33.30	74.71	62.52	43.15	74.64	66.35	20.75	32.50	25.42	28.22	95.12	90.15	15.32	36.31	25.10	**2.59**	**1.14**	**0.41**	–	–	–	–	–	–	9.23	14.11	8.31
	SG	**14.21**	26.46	17.60	36.51	70.34	57.93	16.80	**26.13**	**21.09**	9.02	95.12	90.15	52.90	97.10	94.40	15.84	45.23	25.52	–	–	–	–	–	–	17.32	34.35	25.21
FERET-M	FM	25.14	61.22	44.44	13.23	35.17	19.32	20.42	40.85	27.09	27.98	95.27	90.17	26.47	60.87	52.36	**3.19**	**1.70**	**0.38**	–	–	–	–	–	–	9.30	19.50	8.04
	OCV	26.14	61.50	43.95	20.45	58.60	37.23	25.71	57.45	45.60	31.38	95.27	90.17	28.73	70.08	60.61	**1.13**	**0.57**	**0.38**	–	–	–	–	–	–	7.40	12.75	4.97
	SG	**12.67**	**24.63**	**15.71**	33.84	79.55	66.17	25.33	62.06	49.72	32.14	95.27	90.17	41.83	90.55	82.42	18.14	46.12	32.33	–	–	–	–	–	–	34.97	71.23	61.56
FRLL-M	AMSL	30.94	77.94	66.18	21.61	60.29	39.22	3.26	0.50	0.50	27.13	94.94	90.02	11.76	24.23	16.78	**0.99**	**0.05**	**0.05**	**3.85**	–	**2.89**	7.26	–	10.63	12.74	20.71	14.22
	FM	27.99	73.04	57.35	19.97	57.84	36.76	1.03	0.99	0.99	10.40	95.19	90.38	13.73	36.99	26.10	**0.00**	0.26	0.17	1.35	–	**0.00**	0.74	–	0.98	**0.49**	**0.00**	**0.00**
	OCV	24.73	66.18	53.43	7.53	11.76	4.41	1.88	0.50	0.50	13.76	95.17	90.01	12.25	27.85	18.84	**0.00**	**0.00**	**0.00**	2.97	–	0.49	**0.99**	–	**0.00**	1.47	0.49	**0.00**
	SG	7.53	8.82	5.39	35.92	75.49	68.14	14.65	32.18	24.75	14.32	95.17	90.18	44.61	94.68	90.92	10.34	24.22	12.52	17.21	–	26.69	24.96	–	49.03	**6.83**	**10.29**	**5.39**
	WM	27.19	68.14	55.39	21.54	46.57	33.33	6.39	11.39	3.47	30.30	95.09	90.34	39.22	89.93	83.37	3.45	1.64	0.41	3.42	–	**0.49**	4.07	–	1.47	**2.37**	**1.47**	**0.49**
MIPGAN II	SG	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	22.21	–	47.55	**9.06**	–	**5.39**	17.92	–	23.57
Greedy-DiM	DiffAE	47.00	94.61	85.78	49.40	96.08	93.14	37.72	80.69	71.78	36.10	95.20	89.70	33.82	90.60	81.60	**7.60**	37.60	27.80	–	–	–	–	–	–	11.78	**13.73**	**11.76**
MorDIFF	DiffAE	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**1.10**	–	**0.00**	20.40	–	37.25	17.33	–	26.04
Average (*)		25.89	59.25	47.55	28.83	61.86	49.07	15.90	30.89	24.36	23.86	95.16	90.13	28.11	63.11	54.89	**5.74**	**13.75**	**8.56**	–	–	–	–	–	–	9.91	16.96	11.84
Average (†)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**7.44**	–	11.16	9.64	–	14.96	8.45	–	**9.96**

* Test data: FRGC-M, FERET-M, FRLL-M, Greedy-DiM; † Test data: FRLL-M, MIPGAN II, MorDIFF

shot evaluations by 4.59 percentage points. However, the LoRA-adapted model significantly reduces the EER to 14.81% (a 46.2% relative improvement), suggesting that unadapted features are not as informative for MAD.

In our experiments, the adaptation was especially beneficial for landmark-based morphs. On FRLL-FaceMorpher, for example, the adapted Gemma-3 12B achieved an EER of 0.49%, a substantial gain over the previously reported zero-shot EER of 13.08%. Similar gains appear on FRLL-OpenCV and FRLL-WebMorph. After the adaptation, the detection accuracy was also significantly improved for some GAN-based morphs, such as FRLL-StyleGAN2 (improved from 27.39% EER to 6.83%). However, GAN-based MorGAN attacks remained challenging even after the MLLM adaptation. To better assess such failures, a further analysis on the impact of the training data used for the adaptation is needed.

Comparison Against Supervised MADs. The empirical comparison of the adapted Gemma-3 12B with existing supervised MADs is given in Table 2. As can be seen, our adapted MLLM outperforms all competitive models, achieving an average EER of 14.81%, a 6.74% improvement over the best supervised MAD baseline MixFaceNet-MAD, trained on SMDD. In addition, we note that supervised baselines exhibit severe overfitting. Inception-MAD trained on GAN morphs, for example, degrades to 55.03% on FRLL-StyleGAN2. In contrast, our Gemma-3 12B-MAD maintains consistent performance across attack types, showing better robustness and generalization across different test data.

Comparison Against Unsupervised MADs. The empirical comparison of the adapted Gemma-3 12B with existing unsupervised and self-supervised MADs is given in Table 3. The HRNet-W18-based SelfMAD achieves the best average EER of 5.74%, outperforming Gemma-3 12B-MAD by 4.17 percentage points. Unlike Gemma-3 12B-MAD, whose vision encoder represents a transformer, SelfMAD extracts visual features with a high-resolution CNN, which is especially good at detecting localized artifacts, important for the MAD task. We also note that while SelfMAD dominates landmark-based attacks, Gemma-3 12B-MAD performs better on certain generative attacks. On Greedy-DiM, Gemma-3 12B-MAD achieves 13.73% BPCER at 5% APCER versus 37.60% for SelfMAD. Moreover, Gemma-3 12B-MAD outperforms both SPL-MAD and MAD-DDPM, showing the strength and the underexplored potential of MLLMs in the context of MAD.

Comparison With Foundation Model-based MADs. We compare against MADation and proprietary MLLMs.

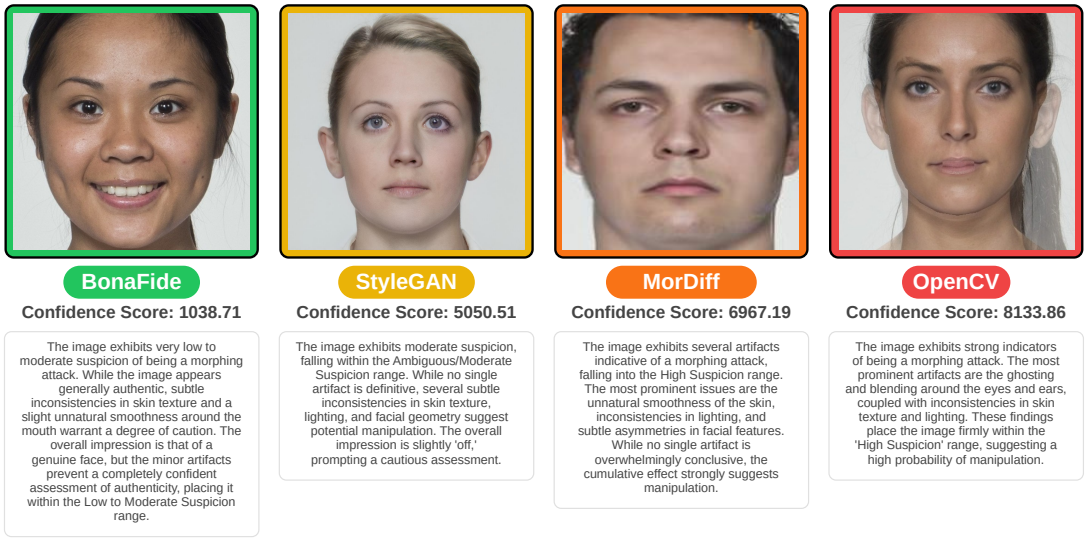


Figure 4. Qualitative zero-shot results generated with Gemma-3. Examples illustrate the step-wise forensic reasoning produced by the proposed six-step prompting protocol, together with the final aggregated confidence score for one bona fide and three morphing attacks.

Prior work [39] reports GPT-4 Turbo achieves 37.0% EER on MIPGAN-II in zero-shot evaluation. Our LoRA-tuned Gemma-3 12B achieves 17.92%, demonstrating that domain-specific adaptation outperforms larger proprietary models. Compared to MADation ViT-B (7.44% average EER on FRLL/MIPGAN-II/MorDIFF), Gemma-3 12B performs comparably (8.45%). At strict thresholds, Gemma-3 achieves 9.96% BPCER at 10% APCER versus 14.96% for MADation ViT-L (Table 3).

6. Conclusion

This paper explored the potential of open-source Multimodal Large Language Models (MLLMs) for single-image morphing attack detection (S-MAD) under strict cross-dataset evaluation, addressing both generalization and interpretability, two longstanding challenges in biometric security. We explored MLLMs through two paradigms: structured zero-shot forensic prompting and parameter-efficient, synthetic-data-driven adaptation.

First, we demonstrated that carefully designed multi-step forensic prompting, inspired by established forensic analysis guidelines, can effectively elicit latent morphing-related knowledge from pretrained MLLMs without task-specific training. The proposed Chain-of-Thought protocol significantly improves zero-shot detection performance compared to naive prompts, while simultaneously producing human-readable, region-level semantic explanations. Notably, zero-shot Gemma-3 exhibits competitive performance on diffusion-based morphs, outperforming specialized CNN-based detectors in certain cases, highlighting the complementary forensic sensitivity of MLLMs to emerging face morphing attack types.

Second, we showed that parameter-efficient LoRA adaptation, guided by privacy-preserving synthetic artifacts, substantially enhances MLLM detection accuracy and cross-dataset robustness. The adapted Gemma-3 12B-MAD model achieves a strong average EER across eight diverse benchmarks, outperforming widely used supervised and unsupervised MAD methods and approaching the performance of state-of-the-art self-supervised and foundation model–based detectors. This confirms that MLLMs can be effectively adapted to the MAD task without reliance on real-world datasets or proprietary models.

Despite these advances, important limitations remain. Zero-shot MLLM performance, while inherently interpretable and informative, is not yet sufficient for deployment in high-security operational settings with strict accuracy requirements. Conversely, LoRA-adapted models currently operate as binary classifiers and do not retain the rich, structured forensic explanations available in zero-shot inference. Additionally, performance degradation on low-resolution and re-digitized images highlights the need for improved robustness to adverse data acquisition.

These findings point to a promising future direction: conversational and multi-objective fine-tuning of MLLMs, enabling models to jointly deliver classifier-level accuracy and structured, step-by-step forensic reasoning. Such models could effectively bridge the gap between transparency and performance, positioning MLLM-based MAD systems as trustworthy, explainable, and operationally viable tools for biometric security applications.

Acknowledgements

The research presented in this work was supported in part by the ARIS Research Program P2-0250(B) “Metrology and Biometric Systems”, and the ARIS Research Project J2-50065 “DeepFake DAD”, as well as ARIS Research Program P2-0103 “Knowledge Technologies”, young researcher grant no. PR-13689, grant no. PR-12379, the EU project “LLMs4EU” (Contract No. 101198470) and the “ANOZ” project ON-2309.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint:2303.08774*, 2023. 2, 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint:2502.13923*, 2025. 5

[3] Zander W. Blasingame and Chen Liu. Greedy-DiM: Greedy Algorithms for Unreasonably Effective Face Morphs. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2024. 2, 5

[4] Fadi Boutros, Naser Damer, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. MixFaceNets: Extremely Efficient Face Recognition Networks. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2021. 2, 6, 7

[5] Eduarda Caldeira, Pedro C Neto, Tiago Gonçalves, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Unveiling the Two-Faced Truth: Disentangling Morphed Identities for Face Morphing Detection. In *31th European Signal Processing Conference (EUSIPCO)*, 2023. 3

[6] Eduarda Caldeira, Fadi Boutros, and Naser Damer. MADPromptS: Unlocking Zero-Shot Morphing Attack Detection with Multiple Prompt Aggregation. In *Proceedings of the 1st International Workshop & Challenge on Subtle Visual Computing*, 2025. 1, 3

[7] Eduarda Caldeira, Guray Ozgur, Tahar Chettaoui, Marija Ivanovska, Peter Peer, Fadi Boutros, Vitomir Struc, and Naser Damer. MADation: Face Morphing Attack Detection with Foundation Models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025. 1, 2, 3, 6, 7

[8] Naser Damer, Alexa Moseguí Saladié, Andreas Braun, and Arjan Kuijper. MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018. 2, 5

[9] Naser Damer, Noémie Spiller, Meiling Fang, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. PW-MAD: Pixel-Wise Supervision for Generalized Face Morphing Attack Detection. In *International Symposium on Visual Computing (ISVC)*, 2021. 1, 2, 5, 6, 7

[10] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly Synthetic Data for the Development of Face Morphing Attack Detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 5

[11] Naser Damer, Meiling Fang, Patrick Siebke, Jan Niklas Kolf, Marco Huber, and Fadi Boutros. MorDIFF: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Diffusion Autoencoders. In *11th International Workshop on Biometrics and Forensics (IWBF)*, 2023. 1, 2, 5

[12] Luca Debiasi, Christian Rathgeb, Ulrich Scherhag, Andreas Uhl, and Christoph Busch. PRNU Variance Analysis for Morphed Face Image Detection. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018. 2

[13] Lisa DeBruine and Benedict Jones. Face Research Lab London Set, 2017. 2

[14] Meiling Fang, Fadi Boutros, and Naser Damer. Unsupervised Face Morphing Attack Detection via Self-paced Anomaly Detection. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2022. 1, 2, 3, 6, 7

[15] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. *On the Effects of Image Alterations on Face Recognition Accuracy*. Springer International Publishing, 2016. 1

[16] Biying Fu and Naser Damer. Face morphing attacks and face image quality: The effect of morphing and the unsupervised attack detection by quality. *IET Biometrics*, 11 (5), 2022. 2, 6, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)*, 2022. 3, 4

[19] Marco Huber, Fadi Boutros, Anh Thi Luu, Kiran B. Raja, Raghavendra Ramachandra, Naser Damer, Pedro C. Neto, Tiago Gonçalves, Ana F. Sequeira, Jaime S. Cardoso, et al. SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data. In *International Joint Conference on Biometrics (IJCB)*, 2022. 1

[20] Marija Ivanovska and Vitomir Štruc. Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models. In *11th International Workshop on Biometrics and Forensics (IWBF)*, 2023. 2, 3, 6, 7

[21] Marija Ivanovska, Leon Todorov, Naser Damer, Deepak Kumar Jain, Peter Peer, and Vitomir Štruc. SelfMAD: Enhancing Generalization and Robustness in Morphing Attack Detection via Self-Supervised Learning. In *IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, 2025. 1, 2, 3, 4, 6, 7

[22] Marija Ivanovska, Leon Todorov, Peter Peer, and Vitomir Štruc. SelfMAD++: Self-supervised foundation model with local feature enhancement for generalized morphing attack detection. *Information Fusion*, 2026. 2, 3

[23] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[24] Kartik Narayan, VS Vibashan, and Vishal M. Patel. FaceXBench: Evaluating Multimodal LLMs on Face Understanding (T-BIOM). *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2026. 3

[25] Pedro C Neto, Tiago Gonçalves, Marco Huber, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. OrthoMAD: Morphing Attack Detection Through Orthogonal Identity Disentanglement. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2022. 3

[26] Mei Ngan and Patrick Grother. Face Analysis Technology Evaluation (FATE) MORPH Part 4B: Considerations for Implementing Morph Detection in Operations. NIST Interagency Report (NISTIR) 8584, National Institute of Standards and Technology, Gaithersburg, MD, 2025. 3

[27] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29 (1), 1996. 2

[28] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5), 1998. 2

[29] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. Detecting Face Morphing Attacks with Collaborative Representation of Steerable Features. In *3rd International Conference on Computer Vision and Image Processing (CVIP)*, 2020. 6, 7

[30] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel. Vulnerability Analysis of Face Morphing Attacks from Landmarks and Generative Adversarial Networks. *arXiv preprint:2012.05344*, 2020. 5

[31] Ulrich Scherhag, Luca Debiasi, Christian Rathgeb, Christoph Busch, and Andreas Uhl. Detection of Face Morphing Attacks Based on PRNU Analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 1(4), 2019. 2

[32] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face Recognition Systems Under Morphing Attacks: A Survey. *IEEE Access*, 7, 2019. 2

[33] Ria Shekhawat, Hailin Li, Raghavendra Ramachandra, and Sushma Venkatesh. Towards Zero-Shot Differential Morphing Attack Detection with Multimodal Large Language Models. In *IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, 2025. 1

[34] Kaede Shiohara and Toshihiko Yamasaki. Detecting Deepfakes with Self-Blended Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 7

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[36] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 Technical Report. *arXiv preprint:2503.19786*, 2025. 5

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *International Conference on Neural Information Processing Systems (NIPS)*, 35, 2022. 3

[38] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Bylappa Raja, Naser Damer, and Christoph Busch. MIPGAN: Generating strong and high quality morphing attacks using identity prior driven GAN. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 3(3), 2021. 2, 5

[39] Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. ChatGPT Encounters Morphing Attack Detection: Zero-Shot MAD With Multi-Modal Large Language Models and General Vision Models. *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, 2025. 2, 3, 8