PixOOD: Pixel-Level Out-of-Distribution Detection

Tomáš Vojíř, Jan Šochman, Jiří Matas
Czech Technical University in Prague, Faculty of Electrical Engineering
Department of Cybernetics, Visual Recognition Group

***Abstract*—We propose a pixel-level out-of-distribution detection algorithm, called PixOOD, which does not require training on samples of anomalous data and is not designed for a specific application which avoids traditional training biases. The PixOOD consists of two main parts - in-distribution data model and decision strategy estimator. In order to model the complex intra-class variability of the in-distribution data at the pixel-level, we propose an online data condensation algorithm which is more robust than standard K-means and is easily trainable through (stochastic) gradient descent techniques. Furthermore, we propose two models for estimating decision strategy, per-class and unified calibration models, each suitable for different applications. We evaluate PixOOD on a wide range of problems. It achieved state-of-the-art results on four out of seven datasets, while being competitive on the rest. The source code is available at https://github.com/vojirt/PixOOD.**

***Index Terms*—Out-of-distribution, Anomaly, Data condensation, Expectation maximization**

I. Introduction

State-of-the-art methods for many computer vision tasks rely on machine learning and therefore on the properties of the training data; consider segmentation, recognition, detection, optical flow, tracking, monocular depth estimation *etc*.. In the laboratory setting, when the training and test sets are obtained by a random split of the available corpus, the standard machine learning assumption of identical data distribution is satisfied by construction. However, in real-world deployment of computer vision systems, encountering domain shifts and out-of-distribution (OOD) data is unavoidable, as when a new product appears (recognition), new material is introduced (segmentation) or a unique 3D structure built (monocular depth estimation).

To be robust, computer vision methods need to address the problem of recognising OOD inputs, *i.e*. data not represented in the training set. Otherwise, for such data, their output is at best not guaranteed, at worst arbitrary, potentially disastrous. Besides the potential for reliable performance, methods that perform OOD, may benefit from human-in-the-loop intervention on novel data, often integrating collection of OOD or rare data for annotation.

In this paper, we present PixOOD, a novel method for out-of-distribution data detection at the level of pixels. The approach is general, it does not require training on samples of anomalous data, nor on synthetically generated outlier data, which requires rather specific knowledge of OOD data properties. These might be known in some case, but we aim at the scenario where minimal assumptions are made about the data not presented during training – such data cannot be synthesized. The proposed method does not exploit constraints that hold in specific applications like industrial inspection or road anomaly, or a particular problem setting like open set semantic segmentation. Fig. 1 shows examples from three diverse ODD benchmarks PixOOD is evaluated on in Sec. IV: the MVTec AD [1], a batch of industrial anomaly detection problems; the SMIYC [2] which contains three sub-tracks: a road anomaly detection (general anomaly segmentation in full street scenes), obstacle detection (obstacle segmentation with the road as region of interest), and LostAndFound; and finally the LaRS benchmark [3], a maritime obstacle segmentation problem with examples of obstacles provided.

Fig. 1. Examples of PixOOD results for road, maritime and industrial anomaly detection tasks. PixOOD is able to identify anomalous data not even considered, *i.e*. not labelled, in standard benchmarks, *e.g*. power cables (not in Cityscapes training classes) or spilled-out content or scratches in insulation.

The inspiration for our method is the recently published GROOD [4] approach for image-level classification and OOD detection. We particularly appreciate its simplicity and calibrated OOD score that is a result of solving the Neyman-Pearson task [5], [6] in a low dimensional projection space. However, several shortcomings of the GROOD method limit its applicability to diverse OOD problems. Specifically, two major issues are (i) pixel-level generalisation (*i.e*. decision making for each pixel instead of the whole image) which poses several engineering challenges and (ii) limited capability of complex intra-class variability modelling, which naturally arises in the pixel-level domain.

We address these weaknesses in PixOOD and demonstrate its efficacy on various downstream tasks. We show that in the case of pixel-level OOD decision problems with large intra-class variations the GROOD single mean representation is not sufficient. We propose a novel data condensation algorithm for rich class appearance modelling. The condensation method is derived from the K-means algorithm. We show it is related to a complete data log likelihood optimisation through the EM algorithm. The proposed method is more robust than the standard K-means and it is easily trained on large volumes of data through stochastic gradient descent. The condensation method is general and of interest on its own. We also introduce several

architectural improvements to accommodate the requirements of the pixel-level decision tasks. The contributions of the paper are summarised as follows:

1) A novel pixel-level OOD detection method called PixOOD. The approach is general (*i.e*. not designed for specific task/benchmark) and does not require any OOD training samples neither real nor synthetic (Sec. III).
2) A novel data condensation algorithm formulated as a stochastic optimisation with a novel loss function and re-initialisation mechanism (Sec. III-A).
3) We theoretically show the relation of the condensation loss function to a lower bound of the complete data log-likelihood optimisation (Sec. III-A2).
4) We demonstrate the applicability of the proposed method through applying it on three diverse benchmarks which are typically solved independently by specialised methods. The proposed method performs competitively on all (seven) datasets achieving state-of-the-art results on four (Sec. IV).

II. Related Work

On the very basic level we divide the methods to those which use real-world or synthetic OOD data [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] and those which do not [19], [20], [7], [21], [22], [23], [24], [25], [15], [26]. In our view, using a proxy for OOD data (*e.g*. objects from COCO in road anomaly [15], [9], [18] or synthetic textures in industrial inspection [16]) is the same as making an assumption about the form of allowed anomalies, which goes against the definition of the word 'anomaly' itself. We aim at a method applicable to the widest possible spectrum of problems, thus avoid introducing task specific biases through auxiliary (synthetic) OOD data.

Among the methods we compare with, one can observe several trends. In the road anomaly detection community, the reconstruction methods [20], [25], [21], [22] are trained, to reconstruct "normal" pixels. The assumption is that during inference an anomalous object would be poorly reconstructed and could be detected. These pixel-level reconstruction-based methods were slowly overtaken by energy-based models [23], [12], [14]. They introduce a regularization of the classification objective loss to enforce particular predictions for anomalous pixels, *e.g*. uniform distribution of posterior probability over in-distribution (ID) classes. A related approach is to add an extra class for OOD [18], but it requires at least synthetic OOD data to be generated. Recently, region-based methods [9], [10], [15], [27], [18] have gained popularity. This makes the ID/OOD reasoning more robust to noise and makes sense for instance in driving scenarios since most anomalous objects have strong boundary separation from its surrounding. The region-based methods, however, are not common in other domains such as the industrial anomaly detection, where the anomalies have different characteristics, *e.g*. deformations or missing parts.

In industrial anomaly benchmarks, recent successful methods are still based on reconstruction [16], [28], [29], [30], likelihood modelling [31], [32] or nearest-neighbour classification [33]. Interestingly, only the latest methods start to depart from a single model for each class and start to focus on a general models [28], [29], [30]. A possible reason is that the benchmarks are rather small and images are taken under well controlled conditions, so the generalisation is not the prime objective.

III. Method

In this section we describe the proposed PixOOD method for OOD (anomaly) detection at the pixel level. The method solves the original pixel-level classification problem into C classes (*e.g*. nineteen semantic classes in the Cityscapes benchmark), but also produces an extra score for deciding the pixels into in- and out-of-distribution with respect to the observed (in-distribution) training data. The PixOOD framework consists of three components (illustrated in Fig. 2): (i) extraction of the pixel/patch feature representation, (ii) building a two-dimensional projection space, and (iii) finding the optimal and calibrated ID/OOD decision strategy. This framework is inspired by GROOD method [4], where it was applied to image-level OOD detection. The components used in [4] are however not applicable to pixel-level tasks. In the following text, we cover the prerequisite knowledge by discussing each of the components from [4], but focus in detail on the new pixel-level components of the proposed PixOOD method. We validate the proposed changes experimentally in Sec. IV-A where applicable.

Representation. Every patch $\mathbf{p}$ in the input image (the patch size, in our case, is determined by the ViT [34] architecture) is first transformed into a feature vector $\mathbf{x} \in \mathbb{R}^D$ by a fixed pre-trained model. This representation needs to be rich enough to allow modelling ID data in a given pixel-level vision tasks. A good example of such model is the self-supervised trained DINOv2 encoder [35] that we employ in this work. Similarly, methods such as [4], [36], [37] utilized pre-trained CLIP [38] image encoder for image-level OOD detection tasks.

2D Projection Space. The 2D projection space serves two purposes: First, it makes it possible to estimate the ID densities with potentially limited data (*e.g*. ∼200 samples per class for MVTecAD dataset), second, it allows to make reasonable assumptions about the unknown OOD distribution, which would be difficult/impossible in the encoder's high-dimensional space.

In contrast with the image-level OOD detection task, the pixel/patch-level representations tend to produce more complex distributions in the (1024-dim in our case) embedding space over the image-level semantic classes. We observed that a simple projection (such as linear probing or distance to the mean representation of *in-distribution* data) is not able to model the rich intra-class appearance variations. For instance, pixels (patches) corresponding to the car label may be visually very different depending on their placement within the car region (*e.g*. wheels vs. windshield), thus causing their representations to differ as well. We show that this is a real problem in the ablation study in Sec. IV. To address this problem, we replace the first projection, the linear probe, with a multi-layer perceptron (Sec. III-B) and instead of the nearest class mean projection we propose to use our novel data

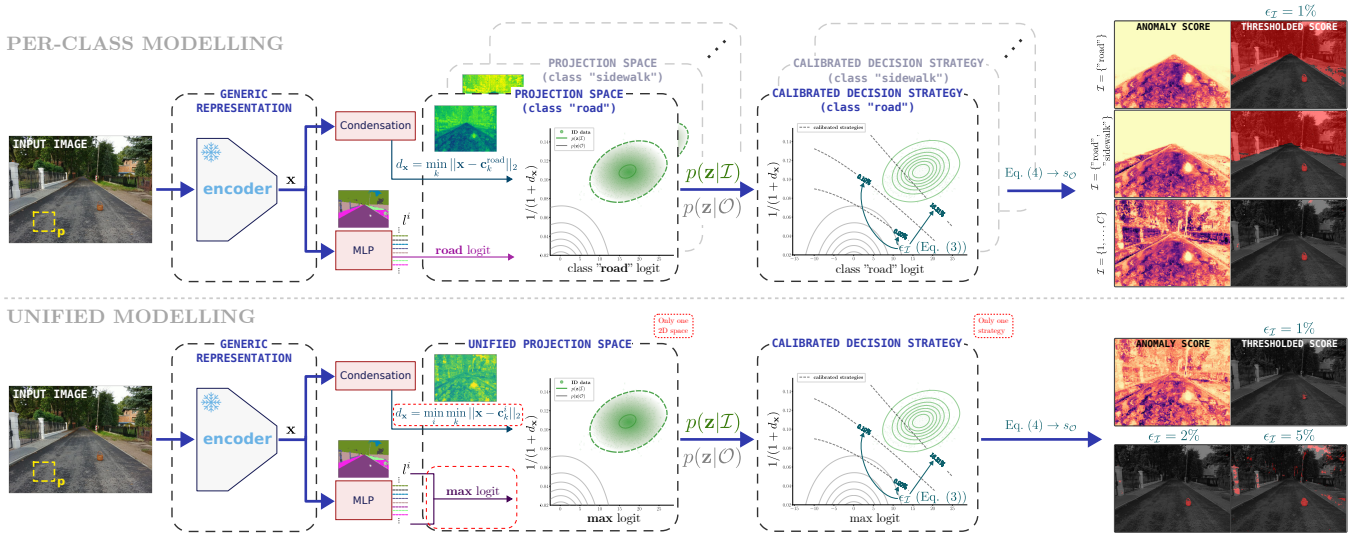


Fig. 2. PixOOD method overview. The Condensation and MLP blocks are train using backpropagation after which they are fixed. The ouputs of the Condensation and MLP for each pixel/patch p of training data, together with the corresponding ground-truth label, are then used to estimate the data distribution in the projection space and to compute the decision strategy. Note that the condensation is run independently for each class during training while MLP is trained jointly for all classes. In case of per-class modelling, the projection space and consequently the decision strategy is estimated for each class separately. In the unified approach, single projection space and decision strategy is found for all training data jointly. During inference, the per-class approach allow to select dymanically a subset of classes $\mathcal{I} \subseteq \{1, \dots, C\}$ which are considered in-distribution, *e.g.* road (and sidewalk) is in-distribution and everything else is OOD, or use all training classes as in-distribution *(top-right part of the figure)*. In unified model, only one OOD score is computed that takes into account all training classes as in-distribution. By changing allowed false negative error of in-distribution data ($\epsilon_\mathcal{I}$) we found different decision strategies that control the trade-off between true and false positive OOD detections *(illustrated at the bottom-right for the unified model)*. The main differences between per-class and unified models are highlighted in red dash boxes.

condensation algorithm (Sec. III-A). The condensation method is general and may also find its uses outside of the PixOOD method.

Given the two projections, there are two ways to construct the projection space: (i) build an independent space for each class, and (ii) build a single unified space[1]. Each construction has its own advantages given the task at hand, as demonstrated in Sec. IV. In the per-class construction, the projection space is formed as $\mathcal{Z} = \mathcal{Z}_{\text{MLP}^i} \times \mathcal{Z}_{\text{NN}^i}$ for each class $i \in \{1, \dots, C\}$ independently, where $\mathcal{Z}_{\text{MLP}^i}$ is a 1-D space of MLP class i logit scores l^i and $\mathcal{Z}_{\text{NN}^i}$ is a 1-D space of distances

$$d^i_{\mathbf{x}} = \min_k ||\mathbf{x} - \mathbf{c}^i_k||_2,$$

to the nearest etalon found by the condensation algorithm for the particular class i.

The unified model jointly models all classes in a single 2D projection space $\mathcal{Z} = \mathcal{Z}_{\text{MLP}} \times \mathcal{Z}_{\text{NN}}$, where $\mathcal{Z}_{\text{MLP}}$ corresponds to $\max_i l^i$ and $\mathcal{Z}_{\text{NN}}$ to

$$d_{\mathbf{x}} = \min_i \min_k ||\mathbf{x} - \mathbf{c}^i_k||_2$$

Decision Strategy. Given the projection space $\mathcal{Z}$ (per-class or unified), the next step is to find a strategy that classifies each $\mathbf{z} \in \mathcal{Z}$ as ID or OOD. Let $\mathcal{I}$ be a (meta-)class representing the in-distribution classes (single class or all classes depending on the projection space construction) and $\mathcal{O}$ a (meta-)class for the out-of-distribution. Both, $p(\mathbf{z}|\mathcal{I})$ and $p(\mathbf{z}|\mathcal{O})$ are modelled as multivariate normal distributions in $\mathcal{Z}$. The parameters of $p(\mathbf{z}|\mathcal{I})$ are estimated from the training data and $p(\mathbf{z}|\mathcal{O})$ is constructed with zero mean and a diagonal covariance matrix with large variances (see [4] for detailed reasoning about the OOD distribution and Fig. 2 for visual illustration).

The ID/OOD classification problem is formulated, same as in [4], as a Neyman-Pearson task [5], [40]: Find a strategy $q^*(z) : \mathcal{Z} \to \{\mathcal{I}, \mathcal{O}\}$ such that

$$\begin{aligned} q^* &= \arg\min_q \int_{\mathbf{z}:q(\mathbf{z})\neq\mathcal{O}} p(\mathbf{z}|\mathcal{O})\, d\mathbf{z} \\ \text{s.t.} \quad \epsilon_\mathcal{I} &= \int_{\mathbf{z}:q(\mathbf{z})\neq\mathcal{I}} p(\mathbf{z}|\mathcal{I})\, d\mathbf{z} \leq \epsilon \end{aligned} \tag{1}$$

This optimisations problem minimises the false positive rate (false acceptance of OOD data) and bounds the ID false negative rate by ϵ. It is known [40] that the optimal strategy for a given $\mathbf{z} \in \mathcal{Z}$ is constructed using the likelihood ratio:

$$q(\mathbf{z}) = \begin{cases} \mathcal{I} & \text{if} \quad r(\mathbf{z}) > \mu \\ \mathcal{O} & \text{if} \quad r(\mathbf{z}) \leq \mu \end{cases} \quad \text{where} \quad r(\mathbf{z}) = \frac{p(\mathbf{z}|\mathcal{I})}{p(\mathbf{z}|\mathcal{O})} \tag{2}$$

The optimal strategy q^* is obtained by selecting maximal threshold μ such that $\epsilon_\mathcal{I} \leq \epsilon$.

For the per-class modelling, the optimal decision strategy is found independently for each target class. This results in C decision strategies for the corresponding projection spaces. In the unified approach a single decision strategy is found.

ID/OOD Score. In practice, we would specify the acceptable false negative rate ϵ and obtain a single optimal strategy. However, most evaluation benchmarks require a score in the range $[0, 1]$ which is then used to compute the evaluation metrics by varying a threshold on this score. To produce such

[1]The per-class approach was used in both [4] and the conference version of this paper [39], the unified version is novel.

a score, we learn a mapping from the likelihood ratio to the false negative rate.

We cover the projection space $\mathcal{Z}$ by M samples in a uniform grid, and for each sample $\mathbf{z}_j$ we find the likelihood ratio $r(\mathbf{z}_j)$ and the corresponding error $\epsilon_{\mathcal{I},j}$ for the threshold set to $\mu = r(\mathbf{z}_j)$. This results in a set of sample pairs $R = \{(r(\mathbf{z}_j), \epsilon_{\mathcal{I},j})\}_{j=1}^{M}$. By linear interpolation, we build a 1D in-distribution scoring function $s_{\mathcal{I}} : \mathbb{R} \to [0, 1]$

$$s_{\mathcal{I}}\left(r(\mathbf{z}_j)\right) \approx \epsilon_{\mathcal{I},j} \ . \tag{3}$$

This score corresponds to the false negative error that we would make on the in-distribution data if we selected $\mu = r(\mathbf{z})$ in the optimal strategy q^* (Eq. (2)).

Finally, we define the OOD ("anomaly") score as

$$s_{\mathcal{O}}\left(r(\mathbf{z})\right) = 1 - s_{\mathcal{I}}\left(r(\mathbf{z})\right) \tag{4}$$

The score $s_{\mathcal{O}}$ is used in all experiments as the output of the PixOOD method. It is in the desired $[0, 1]$ range, is calibrated and explainable and with a clear meaning corresponding to the false negative rate on the ID data. Note that the usable range of this score is at the tail of the range, *e.g.* ≥ 0.95 which is equivalent to saying that the false negative error of the ID data is $\leq 5\%$. See $\epsilon_{\mathcal{I}}$, Eq. (3), and its corresponding decisions boundaries in Fig. 2 for illustration.

Given the OOD score function, there are several possible ways to incorporate it into the decision process during inference. The situation is straightforward for the unified approach, where only one per-pixel score is computed w.r.t. all known (training) classes. However, in the per-class approach there are several options. One may start with the classification by computing the most likely class $i^* = \arg\max_{i \in \{1,\dots,C\}} l^i$, and then test for its OOD score $s_{\mathcal{O}}^{i^*}$. This is a natural extension of the original classification task. Alternatively, we may compute the final OOD score as a minimum over all classes, $s_{\mathcal{O}}^* = \min_{\mathcal{I}} s_{\mathcal{O}}$, and do the classification independently. This setting corresponds more to the unified approach, but as shown in the ablation study, it does not perform as well. Finally, in some tasks, such as the road anomaly detection, there is a class of interest specified a priori (the road) and we are interested just in $s_{\mathcal{O}}^{\text{road}}$.

There are two main differences between the per-class and unified approaches. First, the OOD score for the case of per-class modelling is calibrated across all classes. Therefore, selecting the false negative error threshold (ϵ) will produce the desired error for each class. This is best suited for tasks where we are interested in one class OOD detection (*e.g.* anomalies on road only). In the unified model, the error is calibrated over all in-distribution data, thus implicitly taking into account class imbalance present in the training data. This dataset-wise calibration manifests in higher performance in the setting when all classes are considered.

Secondly, the strategy for selecting the final OOD score is trivial for the unified model where only a single score per-pixel (patch) is produced. In case of per-class modelling, the user needs to define (based on application) which classes should be included in the computation of the final OOD score (see examples if top-right of Fig. 2), which in some cases may improve performance significantly. However, this also introduces another type of error in case of per-class modelling in application with multiple classes of interest. The aggregation mechanism (the most likely class selection or "min" over scores) may misclassify pixels (*e.g.* the pixels for which the aggregation operation did not select the correct label) and thus the OOD score will more likely be estimated higher for non-anomalous pixels, which will result in higher false positive OOD detections.

A. Condensation Algorithm

Using a single etalon for each class to model the ID data is insufficient in the pixel-level setting as discussed above. In this section, we present a novel soft-assignment condensation algorithm which for each (single class) dataset finds K etalons covering the data distribution, where K is significantly smaller than the dataset size. An important aspect of the algorithm is that it is compatible with stochastic gradient descent optimization, as the amount of data/patches is large. Next, we derive the condensation optimization criterion (Sec. III-A1) by generalizing the K-means criterion. The resulting criterion is closely related to the EM-algorithm for spherical Laplace distribution mixture model (Sec. III-A2), which provides more formal view of the condensation criterion. We also show how to alleviate the issue of local minima in the optimization process by re-initializing (Sec. III-A3) the unused etalons.

1) Condensation Optimization Criterion: Let $T = \{\mathbf{x}_i \in \mathcal{X} \equiv \mathbb{R}^D\}_{i=1}^{N}$ be a (potentially large) dataset of training samples from one class. The dimensionality of the samples depends on the representation used (later we use $D = 2$ and $D = 1024$). The **dataset condensation task** is to find up to K etalons (representatives) $\mathbf{c}_1, \dots, \mathbf{c}_K$ ($\mathbf{c}_k \in \mathcal{X}$) that cover the training data, *i.e.* all data points should be close to an etalon. The number of etalons is not fixed, but is upper-bounded by K (a "budget"). In addition, each etalon should represent a similar portion of the data.[2]

Let us denote the trainable parameters, the etalon coordinates $\mathbf{c} = (\mathbf{c}_1, \dots, \mathbf{c}_K)$ and let $d(\mathbf{x}, \mathbf{c}_k)$ be the L_2 distance between a data point $\mathbf{x}$ and an etalon $\mathbf{c}_k$. K-means optimises

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \frac{1}{N} \sum_{i=1}^{N} \min_k d(\mathbf{x}_i, \mathbf{c}_k)^2 \ . \tag{5}$$

This objective is not smooth due to the hard assignment to the closest etalon. Also, K-means tends to get stuck in local minima and is known to be sensitive to outliers in data due to the square of the distance. To make the criterion smooth, we use a smooth approximation [41] of the $\min$ operator:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} w(k, i; \tau) d(\mathbf{x}_i, \mathbf{c}_k)^2 \tag{6}$$

$$w(k, i; \tau) = \frac{\exp\left(-d(\mathbf{x}_i, \mathbf{c}_k)^2/\tau\right)}{\sum_{j=1}^{K} \exp\left(-d(\mathbf{x}_i, \mathbf{c}_j)^2/\tau\right)} \ , \tag{7}$$

[2]PixOOD used distances to nearest etalons as a proxy for density estimation. The distance will be smaller in dense areas and larger in sparser areas because of this requirement.

where Eq. (7) is the softmin function with temperature scaling τ. The temperature scaling allows a smooth transition from a soft assignment (easier to optimise, less susceptible to local minima) and a hard assignment (the goal of condensation). During the optimisation, the value of τ is scheduled to decrease from a large (soft assignment) to a low value (hard assignment).

The smooth objective in Eq. (6) is sensitive to outliers in data due to the square of the distance. When removing the square from Eq. (5), the solution to the problem is known as the K-medians algorithm which requires an iterative optimisation for finding the medians in each step. Instead, we remove the square in the smoothed version Eqs. (6) and (7) which is still easy to optimise and allows the soft-hard assignment transition through τ parameter:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} w(k,i;\tau) d(\mathbf{x}_i, \mathbf{c}_k) \tag{8}$$

$$w(k,i;\tau) = \frac{\exp\left(-d(\mathbf{x}_i, \mathbf{c}_k)/\tau\right)}{\sum_{j=1}^{K} \exp\left(-d(\mathbf{x}_i, \mathbf{c}_j)/\tau\right)} \tag{9}$$

The final requirement of the condensation task is that each etalon covers approximately the same number of data points. To that end, each etalon is assigned a trainable scale parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ to make it adaptive to the varying local data density. Instead of just dividing the distance by the parameter $(d(\mathbf{x}_i, \mathbf{c}_k)/\beta_k)$ in the formula, which would have a trivial minimiser for $\beta_k = \infty$, the distance is replaced by a logarithm of the Laplace distribution which regularises the β_k parameter to avoid this degenerated solution. This gives us the final loss optimised by the proposed algorithm[3]:

$$\begin{aligned} L(\mathbf{c}, \boldsymbol{\beta}; T, \tau) &= -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} w(k,i;\tau) \log\left(\frac{1}{\beta_k} e^{-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{\beta_k}} \right) \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} w(k,i;\tau) \left(\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{\beta_k} + \log \beta_k \right) , \end{aligned} \tag{10}$$

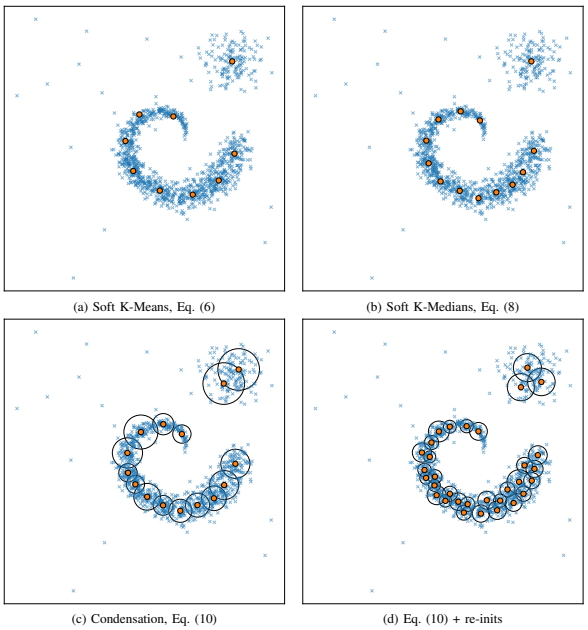
where $w(k,i;\tau)$ is defined in Eq. (9). The full iterative soft-hard condensation algorithm is summarised in Algorithm 1 (right) and the individual steps in its derivation are illustrated on toy example in Fig. 3 (left).

Interestingly, the objective of the loss defined in Eq. (10) has another interpretation. It is a lower bound on the complete data log-likelihood in EM algorithm for a mixture of spherical Laplace distributions with equal priors, the derivation is provided in the following section.

2) Relation to EM Algorithm.: Let us model the measurements $\mathbf{x}$ by a spherical Laplace distribution mixture model

$$\begin{aligned} p(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\beta}) &= \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k; \mathbf{c}_k, \beta_k) \\ &= \sum_{k=1}^{K} \pi_k \frac{1}{Z_k} \frac{1}{\beta_k^{D-1}} \exp\left(-\frac{d(\mathbf{x}, \mathbf{c}_k)}{\beta_k} \right) \end{aligned} \tag{11}$$

[3]The formula uses negative distance, hence the minus sign at the beginning.



(a) Soft K-Means, Eq. (6) (b) Soft K-Medians, Eq. (8) (c) Condensation, Eq. (10) (d) Eq. (10) + re-inits

Fig. 3. Methods derived in Sec. III-A applied to synthetic data (blue crosses) with outliers (isolated blue crosses). Only the "useful" etalons (from total of 50) with $s_t(k) >= \theta_r$ are displayed as orange circles. (a) K-means is sensitive to outliers and most etalons converge towards isolated data points (not shown) – only 9 useful etalons. (b) K-medians is more robust – 13 useful, (c) condensation adds the scale parameter β_k to each cluster (black circle) enabling adaptive region of influence – 15 useful, (d) re-inits with *(c)* preserve significantly more etalons by combining re-initialization strategy with adaptive scale – 32 useful.

Algorithm 1 Iterative soft-hard condensation algorithm.

Given: $T = \{x_i \in \mathbb{R}^D\}_{i=1}^N$
Initialise: $\mathbf{c}, \boldsymbol{\beta}, \tau, q, \lambda$
for epoch $<$ num. epochs **do**
  for $X = \{\mathbf{x}_i\}_{i=1}^n \subset T$ **do** ▷ iterate over batches
    Compute $d(\mathbf{x}_i, \mathbf{c}_k) \ \forall \mathbf{x}_i \in X, \forall \mathbf{c}_k \in \mathbf{c}$
    Compute $w(k,i;\tau)$ ▷ Eq. (9)
    Compute $L(\mathbf{c}, \boldsymbol{\beta}; T, \tau)$ ▷ Eq. (10)
    $\forall \mathbf{c}_k : \mathbf{c}_k \leftarrow \mathbf{c}_k - \lambda \frac{\partial L}{\partial \mathbf{c}_k}$
    $\forall \beta_k : \beta_k \leftarrow \beta_k - \lambda \frac{\partial L}{\partial \beta_k}$
    Update $s_t(k)$ ▷ Eq. (18)
  end for
  if cooldown $>$ epoch $>$ warmup and $s_t(k) < \theta_r$ **then**
    Re-initialise $\mathbf{c}_k$
  end if
  $\tau \leftarrow$ cos_scheduler(epoch)
end for

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the prior probabilities of each component, $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_K)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ characterise the component distributions as in Eq. (10) and Z_k's are the normalising factors of the spherical Laplace distribution independent of $\boldsymbol{\beta}$ and $\mathbf{c}$. The EM maximises the complete data log-likelihood

$$\begin{aligned} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\beta}; T) &= \sum_{i=1}^{N} \log p(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\beta}) \\ &= \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k; \mathbf{c}_k, \beta_k) \ . \end{aligned} \tag{12}$$

By introducing a variational distribution $w(k,i)$ (we show how to choose it below) a lower bound to $\mathcal{L}$ is derived using

Jensen's inequality as

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\beta}; \mathbf{x}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} w(k,i) \frac{\pi_k p(\mathbf{x}|k; \mathbf{c}_k, \beta_k)}{w(k,i)} \quad (13)$$

$$\geq \sum_{i=1}^{N} \sum_{k=1}^{K} w(k,i) \log \frac{\pi_k p(\mathbf{x}|k; \mathbf{c}_k, \beta_k)}{w(k,i)} \quad (14)$$

which for the case of a mixture of spherical Laplace distributions looks like

$$\sum_{i=1}^{N} \sum_{k=1}^{K} w(k,i) \log \pi_k + w(k,i) \log \frac{1}{\beta_k^{D-1}} \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{\beta_k}\right) - w(k,i) \log w(k,i) \; . \quad (15)$$

Since the M-step maximises this lower bound only over $\mathbf{c}$ and $\boldsymbol{\beta}$, the last term becomes a constant and could be omitted from the optimisation. Further, if all priors π_k are assumed constant $\pi_k = 1/K$, also the first term disappears. What is left is maximisation of the negative of the loss (10). In the E-step any variational distribution $w(k,i)$ defines a lower bound for $\mathcal{L}$. It is easy to show that the optimal $w^*(k,i) = p(k|\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\beta})$. Using the Bayes formula

$$w^*(k,i) = \frac{p(\mathbf{x}_i; \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\beta})}{\sum_{j=1}^{K} p(\mathbf{x}_i; \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\beta})} = \frac{\frac{1}{\beta_k^{D-1}} \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{\beta_k}\right)}{\sum_{j=1}^{K} \frac{1}{\beta_j^{D-1}} \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{c}_j)}{\beta_j}\right)} \; , \quad (16)$$

where the constant a priori probabilities cancel out. Although optimal, when the embedding dimension D is high (*e.g.* 1024), this formula becomes numerically unstable. Replacing all β_k's by a constant value τ reduces the formula into Eq. (9). This in principle breaks the EM monotonicity convergence property, but practically avoids the numerical instability and allows for independent treatment of the etalon scale parameters and the soft-to-hard transition parameter τ. In practice, EM converges to local maxima, even with the optimal lower bound, thus we address both problems (monotonicity and local minima) by introducing etalon re-inits.

3) Re-Inits.: K-means, EM algorithm and also the proposed condensation algorithms converge, in general, to a local optimum. The soft-hard scheduling helps to alleviate this to some extent, but the result still depends on the initialisation. Although there exist heuristic initialisation methods like K-means++, they are computationally expensive. Hence, we introduce etalon re-inits inside the optimisation loop. For each etalon, we compute its data sample (batch) support as

$$s(k) = \sum_{i=1}^{n} w(k,i;\tau) \; . \quad (17)$$

In the EM interpretation, this corresponds to a posteriori probability of the mixture component given (batch) data (multiplied by n, the batch size). A small support is indicative of a misplaced etalon. In the stochastic optimisation, the mini-batch data samples are often small and biased resulting in noisy $s(k)$ estimate. Thus, we estimate the support using init-unbiased running exponential weighted average [42] over mini-batches. Given a decay rate q, at the iteration t it is estimated as

$$s_t(k) = (1 - q_t) s_{t-1}(k) + q_t s(k), \quad \text{where} \quad q_t = q \,/\, \big(1 - (1-q)^t\big). \quad (18)$$

An etalon is reset if, after a warm-up number of iterations, its $s_t(k) < \theta_r$. The θ_r is user-defined threshold and it controls the balance between number of used etalons and coverage of low density areas. We reset the etalon to a random data point from the mini-batch and add a little noise to avoid degenerate configurations.

B. Discriminative Classifier

The discriminative power of simple Linear Probing (LP) technique, used in [4], is inadequate for the more complex pixel-level classification task. The LP is generally used to assess zero-shot performance of large pre-trained model on image classification task and most often is implemented as a linear regression using off-the-shelf solver (*e.g.* L-BFGS [43]) that requires all data at the same time. However, these techniques are prohibitive in the pixel domain due to the sheer amount of data. Thus, we propose to replace the LP by a small multi-layer Perceptron (MLP) network consisting of two layers with GELU [44] non-linearity. That increases the representation power and is easy to use in the batch processing regime with an increased volume of data of the pixel domain.

C. Pixel-Level Adaptations

Pixel-level tasks introduce two additional technical challenges due to the use of the ViT architecture [34], which operates on patches rather than individual pixels.

First, as with any architecture that reduces spatial resolution, fine-grained details may be lost. In our case, this issue arises only at the inference time, but *not* during training, since the modelling is performed on patch representation. The common solutions include either increasing the input image resolution or upsampling the intermediate representations. We propose using simple linear upsampling of the backbone feature maps before OOD score estimation. We provide an ablation study (Sec. IV-A) that explores different interpolation magnitudes, ranging from no interpolation to 50% ($\times 7$ upscaling) of the original input image resolution. Other, often more complex, techniques that follow the same core idea of producing higher-resolution feature maps could also be used, *e.g.* FeatUp [45]. However, we did not explore such techniques, as we did not identify prediction upsampling as a limiting factor for our method. To obtain the final full-resolution OOD score at inference time, we first up-scale the feature maps to estimate OOD score map at higher-resolution followed by final upscaling of the scores to the original input image resolution.

Second, semantic segmentation labels are defined per pixel, whereas a patch may cover multiple semantic classes at class boundaries. For example, a patch on the boundary between a road and a sidewalk can contain labels from both classes. Standard practice is to ignore this mismatch by resizing labels using nearest-neighbour interpolation or by upscaling predictions before loss or metric computation. Since our method

explicitly relies on patch representations in embedding space for modelling, and we observed a negative impact of mixed-label patches on downstream performance, we instead filter out ambiguous ("uncertain") patches during training. Specifically, we keep only patches in which more than 90% of the labels belong to a single class. This results in a cleaner training signal for condensation, in-distribution data modelling in the projection space, and OOD score calibration. At test time, all patches are processed equally. Note that the training of the MLP is performed on the original input resolution by up-scaling the MLP output logits, which is a common approach, and thus the mixed patches have no effect.

IV. Experiments

In all experiments, the proposed PixOOD uses no anomaly or OOD data during training (such as auxiliary datasets or synthetic anomalies through augmentation) except for the LaRS dataset where the obstacle class is part of training data.

Implementation Details. We use frozen DINOv2 [35] ViT-L variant with 14 patch size as a backbone in all experiments. The AdamW optimizer is used for training with batch size 4 and input image resized such that the longer dimension is 1792 while keeping the aspect ratio. We train the condensation algorithm for 100 epochs with 0.0 weight decay, learning rate 0.1, cosine learning rate decay, and 'budget' K set to 1000 (unless stated otherwise). We set the hyper-parameters of the condensation algorithm as follows: warmup and cooldown epochs to 1 and 90 respective; $\theta_r = 0.5\frac{B}{K}$, where B is an average number of data in a batch for a given class computed as running average during training (this is due the fact that the condensation is trained for all classes at the same time and there is different proportion of data for each class in a given training batch). The softmin temperature scaling τ is initially set to 2.0 and end value 0.5 with cosine scheduling. The classification MLP is trained for 30 epochs with 0.0005 weight decay, and learning rate 10^{-4} without any learning rate scheduling. The final calibrated strategy is found on the training data for the respective tasks. For the 1D scoring function (Eq. (3)) we sample 2000×2000 uniform grid for the interpolation data points.

Datasets. While we are primarily focused on autonomous driving application, we also consider other domains to showcase the generic aspect of the proposed method. Commonly used benchmarks from recent literature for the respective domains are: (i) for the road anomaly detection, where the evaluation is limited to the road region only – Road Anomaly [21], FischyScapes LaF [46], [47], SMIYC [2] (Obstacle Track, LaF NoKnown), (ii) SMIYC (Anomaly Track) for semantic segmentation anomaly, (iii) MVTec AD [1] for industrial inspection, and (iv) LaRS [3] for maritime obstacle detection.

Evaluation metrics. We use metrics standard for the respective tasks. For road and semantic segmentation anomaly detection, commonly used metrics are Average Precision (AP) – area under precision-recall curve, False Positive Rate at 95% of True Positive Rate, denoted as FPR, and mean F1 score computed component-wise and averaged over different detection thresholds. In the industrial inspection domain, image and pixel level AUROC – area under ROCs curve are most often used together with AUPRO used in [1] – per-region overlap (PRO), to account for anomalies of different sizes within one image. The maritime obstacle detection (LaRS) uses standard mean Intersection over Union (mIoU) for the semantic classes and a domain specific metrics [48] such as: 1) wateredge accuracy – computed from boundary between water and static obstacles and (2) obstacle detection accuracy – precision, recall and F1 score where true positive detections are considered if predicted obstacle covers the ground-truth pixels from more than 70%. The final ranking in LaRS is determined by the Q measure calculated as F1 $\times$ mIoU. Qualitative results with typical failure cases are shown in Figs. 1 and 4.

TABLE I
Ablation study – baselines, novel condensation, "budget" K, and pixel adaptation techniques. All reported results use the best performing setup of the PixOOD pipeline except one change that is being ablated. Components of the best setup are highlighted in green. The evaluation use road as the ROI.

PixOOD [road/sidewalk] with	K	RA+FS+RO	
		mean AP ↑	mean FPR ↓
Baseline MSP	-	73.93	6.93
Baseline NN dist	1000	81.10	5.65
Single Mean	1	88.14	3.49
Soft K-Means – Eq. (6)	200	93.01	2.26
Soft K-Medians – Eq. (8)	200	93.54	1.99
Soft-to-Hard Cond. – Sec. III-A	200	94.11	1.82
	500	94.71	1.67
	1000	94.80	1.64
Linear Probe		92.36	2.38
Multi-Layer Perceptron		94.80	1.64
Patch label percentage $> 50\%$		94.22	1.77
$> 90\%$		94.80	1.64
Feature map up-sampling $\times$ 1		92.14	2.35
$\times$ 2		94.20	1.75
$\times$ 5		94.78	1.65
$\times$ 7		94.80	1.64

A. Ablation Study

We use road anomaly detection as the downstream task on which we demonstrate the effectiveness of the proposed condensation algorithm and ablate its core components such as the effect of different 'budget' size K, two baselines formed from outputs of the MLP classifier and nearest neighbour distance to condensed data, and other technical decisions. All hyperparameters and random seeds (set to 42) for different variants of the method were fixed to limit the nuisance factors as much as possible.

The results for these ablations are presented in Tab. I. First, only the MLP classifier (the same as in the full method) is used and we compute the maximum softmax probabilities as anomaly score proxy (commonly used technique known as MSP method [49]). Second, anomaly score is computed using a normalized distance to the nearest etalon obtained by the proposed condensation algorithm. These two baselines essentially test the signal that is used to established the 2D projection space from which the decision strategy of the proposed PixOOD method is computed. This comparison

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON ROAD ANOMALY AND FS LAF DATASETS. THE METHODS ARE GROUPED BY THE USAGE OF AUXILIARY OOD DATA. NOTE THAT METHODS DACUP AND JSR-NET (AND ALSO PIXOOD [ROAD, SIDEWALK] WITH SELECTED ID CLASSES) ARE DESIGNED TO PERFORM BINARY DECISION (*i.e.* ROAD OR NOT ROAD) AND THUS NOT SUITABLE FOR ROI: IMAGE EVALUATION PROTOCOL.

Method	OOD Data	ROI: Road						ROI: Image					
		Road Anomaly		FishyScapes LaF		Average		Road Anomaly		FishyScapes LaF		Average	
		AP ↑	FPR ↓	AP ↑	FPR ↓	AP ↑	FPR	AP ↑	FPR ↓	AP ↑	FPR ↓	AP ↑	FPR ↓
SynBoost [8]	✓	63.72	52.27	92.46	0.66	78.09	26.47	35.39	67.83	60.51	31.16	47.95	49.50
PEBAL [7]	✓	67.31	39.41	84.05	1.58	75.68	20.49	44.06	43.10	58.79	4.79	51.43	23.95
Maximized Entropy	✓	96.23	6.03	77.16	10.14	86.70	8.08	85.01	11.20	40.93	38.21	62.97	24.71
RbA [15]	✓	92.99	9.01	85.98	2.87	89.48	5.94	2 85.36	1 6.91	66.20	6.12	75.78	6.51
Mask2Anomaly [9]	✓	94.10	16.91	1 94.49	1 0.52	94.30	8.71	82.20	13.95	71.28	30.29	2 76.74	22.12
EAM [10]	✓	2 96.43	1 2.26	2 93.67	0.77	2 95.05	2 1.52	69.54	7.72	1 82.68	2 4.04	76.11	2 5.88
UNO [18]	✓	1 96.76	2 2.32	93.56	2 0.65	1 95.16	1 1.48	1 88.52	2 7.39	2 81.92	1 1.29	1 85.22	1 4.34
Image Resynthesis [21]	✗	76.39	48.08	66.75	3.10	71.57	25.59	35.22	63.56	4.65	61.45	19.94	62.50
EAM [10]	✗	88.96	17.89	64.86	16.73	76.91	17.31	69.42	13.37	53.54	28.28	61.48	20.82
RbA [15]	✗	89.76	16.54	74.70	15.13	82.23	15.84	78.65	11.82	61.84	18.68	70.24	15.25
JSR-Net [20]	✗	94.42	9.25	78.30	3.96	86.36	6.60	19.27	54.12	0.26	69.41	9.76	61.77
DaCUP [25]	✗	96.19	5.46	2 89.75	2 1.45	92.97	3.45	21.19	51.35	0.95	62.59	11.07	56.97
UNO [18]	✗	1 97.05	1 2.77	89.38	1.51	2 93.22	1 2.14	2 82.43	1 9.20	1 74.75	2 6.82	1 78.59	2 8.01
PixOOD [road,sw]	✗	2 96.39	2 4.30	1 93.55	1 0.54	1 94.97	2 2.42	42.01	55.85	9.63	59.42	25.82	57.64
PixOOD unified	✗	92.74	7.76	81.18	2.75	86.96	5.25	1 84.13	2 10.93	2 69.37	1 4.84	2 76.75	1 7.88

highlights the substantial improvement when incorporating the complementary information from MLP and the nearest neighbour distances to the condensed etalons and using the calibrated decision strategy.

The next part of the ablation shows that a single mean representation is insufficient for complex class modelling in pixel-level domain. Furthermore, the robustness and efficacy of the proposed condensation method is demonstrated by increased performance compared to K-{Means, Medians} algorithms. Performance can be further improved by allowing a larger 'budget' K, and it saturates for K around 1000, suggesting that the proposed method can utilise extra clusters, up to a saturation point, efficiently. Furthermore, we evaluate the proposed generalisations and technical improvements from Sec. III-C. The feature map up-sampling improves mean AP from 92.14 to 94.80 and reduces mean FPR from 2.35 to 1.64. The largest gain comes primarily from introducing feature-map upsampling (factor ≥ 2) and further gains saturates quickly (around $\times 5$). The improvement when filtering "uncertain" patches, *i.e.* with $\leq 90\%$, is modest, but we argue that this approach is better aligned with our modelling assumptions than assigning every patch a hard label by nearest-neighbour interpolation. Note that for the *Linear Probe* we used a single fully connected layer and trained it by SGD, since it can be processed in batches as discussed in Sec. III-B. The results are presented in a leave-one-out manner in the bottom three blocks of Tab. I. By modifying only one part and leaving the rest in the best configuration it clearly shows the benefit of every design choice.

Lastly, we evaluated several strategies to select the final score (Tab. III). For region of interest (ROI) limited to a particular image region, *e.g.* road region, using the corresponding class OOD score performs better than considering all scores or using a unified model. The results can be improved by including similar classes that are most often bordering with each other and the ROI. This consideration addresses false positives detections on boundaries of the ROI and most likely misclassifications. However, this class selection strategy can not be used when ROI is the whole image, where the unified model performs better than any selection strategy using the per-class modelling (see Fig. 5 for qualitative comparison).

TABLE III
ABLATION STUDY – OOD SCORE SELECTION. THE BEST RESULTS FOR EACH ROI ARE HIGHLIGHTED IN GREEN.

PixOOD w/ classes	Score	ROI: Road (RA+FS+RO)		ROI: Image (RA+FS)	
		mean AP ↑	mean FPR ↓	mean AP ↑	mean FPR ↓
$i \in$[road]	$\min_i s^i_{\mathcal{O}}$	90.77	2.14	25.79	60.79
$i \in$[road,sw]	$\min_i s^i_{\mathcal{O}}$	94.80	1.64	25.82	57.64
$i \in$[all]	$\min_i s^i_{\mathcal{O}}$	72.85	13.78	51.61	17.59
$i \in$[all]	$s^{i^*}_{\mathcal{O}}$	78.69	12.46	55.21	18.33
$i \in$[all] unified	$s_{\mathcal{O}}$	87.08	3.56	76.75	7.88

B. Road Anomaly Detection

We compare the proposed method with recent state-of-the-art methods on two standard datasets, Road Anomaly and FishyScapes LaF, and also on SMIYC benchmark designed for road anomaly segmentation task. Results on the standard datasets are presented and compared with most recent publications [15], [10], [9], [25] in Tab. II. The proposed method outperforms other state-of-the-art methods (even several of those that use an auxiliary dataset to model the OOD data) often by a large margin.

Evaluation results on the SMIYC benchmark are reported in Tab. IV. We report all three tracks, two for road anomaly detection (Obstacle Track and LaF NoKnown) and one for semantic segmentation anomaly detection (Anomaly Track). Compared to methods that do not use any auxiliary OOD data, our method performs best in Obstacle Track together with the UNO method (that however uses synthetic anomalies even in the version marked as *without OOD data*) and second best in LaF NoKnown track.

Anomaly Track is the only experiment where per-class PixOOD is "formally" under-performing; The typical observed failures are: (i) under-segmenting, (ii) detecting semantic shifts, such as mountains or forests which are not annotated

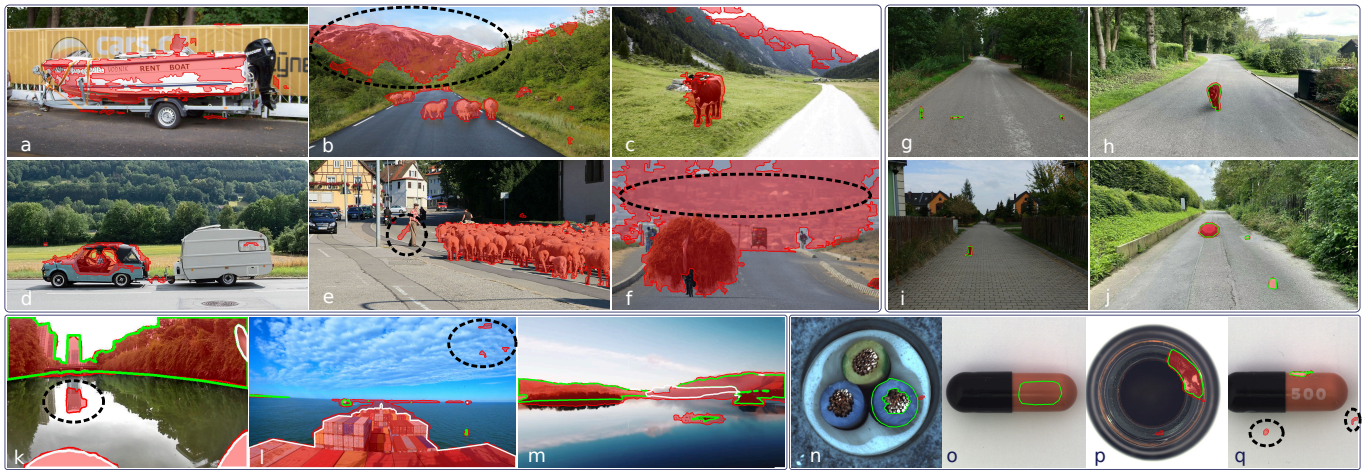


Fig. 4. Typical outputs with focus on "failure" cases. PixOOD anomalies are sometimes under-segmented (a) or over-segmented (b,c,m), but it often finds unexpected but reasonable anomalies: a stick (e), reflection in the water (k), birds (l), extra scratches (p), pill remains (q). It is also unable to detect logical anomalies like the switched cable in (n) or missing label in (o). Moreover, semantic/domain shifts are considered as anomaly: mountains (b,c), convertible roof with see-through (d) and city view (f) which are not present in the Cityscapes. Legend: red - detected anomaly, green - GT when available, white - ignore region; ellipses mark relevant regions.

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART ON THE SMIYC [2] BENCHMARK. FOR CLARITY, ONLY TOP PERFORMING METHODS AND MAIN EVALUATION METRICS ARE SHOWN.

Method	OOD Data	Anomaly Track (ROI: Image)			Obstacle Track (ROI: Road)			LaF NoKnown (ROI: Road)		
		AP ↑	FPR ↓	mean F1 ↑	AP ↑	FPR ↓	mean F1 ↑	AP ↑	FPR ↓	mean F1 ↑
SynBoost [8]	✓	56.44	61.86	9.99	71.34	3.15	37.57	❶ 81.71	❷ 4.64	48.72
ATTA [11]	✓	67.04	31.57	20.64	76.46	2.81	36.57	–	–	–
DenseHybrid [12]	✓	77.96	9.81	31.08	87.08	0.24	50.72	❷ 78.67	❶ 2.12	❶ 52.33
RPL+CoroCL [13]	✓	83.49	11.68	30.16	85.93	0.58	56.69	–	–	–
Maximized Entropy [14]	✓	85.47	15.00	28.72	85.07	0.75	48.51	77.90	9.70	❷ 49.92
Mask2Anomaly [9]	✓	88.72	14.63	47.16	❷ 93.22	0.20	68.15	–	–	–
EAM [10]	✓	93.75	❷ 4.09	❷ 60.86	92.87	0.52	❷ 75.58	–	–	–
RbA [15]	✓	❷ 94.46	4.60	51.87	❶ 95.12	❶ 0.08	57.44	–	–	–
UNO [18]	✓	❶ 96.33	❶ 1.98	❶ 62.61	93.19	❷ 0.16	❶ 77.65	–	–	–
ODIN [19]	✗	33.06	71.68	5.15	22.12	15.28	9.37	52.93	30.04	34.53
JSRNet [20]	✗	33.64	43.85	13.66	28.09	28.86	11.02	74.17	6.59	35.97
Image Resynthesis [21]	✗	52.28	25.93	12.51	37.71	4.70	8.38	57.08	8.82	19.17
NFlowJS [23]	✗	56.92	34.71	14.89	85.55	❷ 0.41	50.36	❶ 89.28	❶ 0.65	❶ 61.75
ObsNet [24]	✗	75.44	26.69	45.08	–	–	–	–	–	–
DaCUP [25]	✗	–	–	–	81.50	1.13	46.01	81.37	7.36	❷ 51.14
RbA [15]	✗	86.13	15.94	42.04	87.85	3.33	50.42	–	–	–
CSL [27]	✗	80.08	❷ 7.16	❷ 50.39	87.10	0.67	❷ 51.02	–	–	–
cDNP [26]	✗	❷ 88.90	11.42	28.12	–	–	–	–	–	–
UNO [18]	✗	❶ 96.10	❶ 2.27	❶ 58.87	❶ 88.97	0.61	❶ 76.32	–	–	–
PixOOD [road,sw]	✗	–	–	–	❷ 88.90	❶ 0.30	50.82	❷ 85.07	❷ 4.46	44.41
PixOOD [all] $s_{\mathcal{O}}^{i^*}$	✗	72.98	53.44	25.89	74.70	0.97	35.74	51.77	20.09	25.37
PixOOD unified	✗	86.00	25.04	35.09	82.61	0.50	41.90	64.16	9.41	23.04

in the Cityscapes dataset, and (iii) detecting domain shift, *i.e.* same class as in Cityscapes but with very different appearance. The semantic/domain shifts may not be considered OOD in the Anomaly Track due to (vague) definition of anomaly, and it requires an algorithm to generalise semantic concepts (like a car). Instead, our method detects anomalies w.r.t. the given ID data. This allows detecting these semantic/domain shifts, which is important in many tasks. For qualitative examples see Fig. 4. The unified OOD model improves the Anomaly Track results, mainly because it incorporates inherent priors of driving scenes and thus makes fewer errors on the most represented classes, *i.e.* classes with high image areas, thus, greatly reducing false positive detections.

C. ISSU Road Anomaly Detection Benchmark

In this section we present a results on the recently introduced ISSU benchmark [50]. This road anomaly dataset contains data from Indian driving scenes [51], [52], [53] and allows controlled in-domain and cross-domain evaluation setting. The tested methods are trained either on images from Indian roads or from standard Cityscapes dataset and evaluated on the Indian driving data. This tests the generalisation of the methods to new environments. The ISSU benchmark,

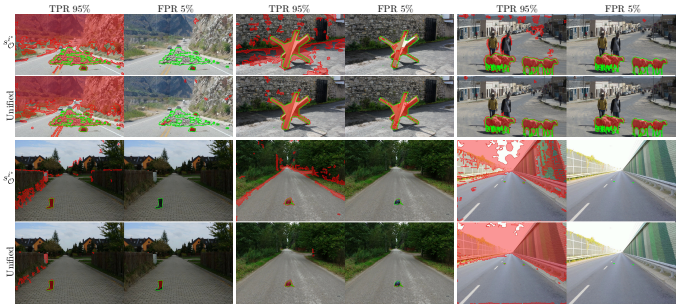


Fig. 5. Qualitative comparison of the unified approach *vs*. the per-class model using the logits to first select most likely class i^* ($s^{i^*}_{\mathcal{O}}$) for image ROI. The images were selected to showcase most common behaviour difference. The threshold for anomaly score for each image was set to operation point of TPR (FPR) equal to 95% (5%) w.r.t. ground-truth anomaly pixels, *i.e.* classify correctly 95% of anomalous pixels (false positive classifications equal to 5% of anomalous pixels). The per-class approach makes consistently errors on boundaries between classes. Both approaches struggle with overexposure and small objects (bottom right).

TABLE V
COMPARISON WITH THE STATE-OF-THE-ART ON THE ISSU [50] BENCHMARK.

Method	OOD Data	ROI: Road				ROI: Image			
		Static		Temporal		Static		Temporal	
		AP ↑	FPR↓	AP ↑	FPR↓	AP ↑	FPR↓	AP ↑	FPR↓
In-domain									
RbA [15]	✓	❶ 95.8	1.7	❷ 57.2	❷ 33.7	❶ 79.1	❷ 3.9	❷ 37.7	❷ 29.4
EAM [10]	✓	❷ 95.6	❷ 1.6	❶ 62.1	96.2	❷ 76.8	4.2	❶ 38.7	91.4
M2F Pebal [7]	✓	92.5	1.9	48.9	❶ 23.8	64.5	4.4	23.6	❶ 24.7
UNO [18]	✓	94.0	❶ 1.2	56.1	92.3	71.4	❶ 3.0	30.4	89.7
M2A [9]	✓	48.9	78.5	30.0	79.5	32.0	66.9	10.7	78.6
JSR-Net [20]	✗	85.7	8.4	52.1	❷ 26.5	4.2	56.1	2.3	58.7
DaCUP [25]	✗	85.5	100.0	56.8	100.0	5.4	100.0	2.9	100.0
RbA [15]	✗	92.7	77.5	53.4	98.9	❷ 75.7	73.4	❷ 36.5	94.9
EAM [10]	✗	❶ 94.5	❶ 2.2	❷ 70.0	98.1	❶ 77.1	❶ 5.9	❶ 45.2	92.9
M2F Pebal [7]	✗	92.3	❷ 3.4	54.2	95.6	69.9	❷ 9.2	32.4	92.6
PixOOD [road,sw]	✗	❷ 93.1	4.3	❶ 83.1	❶ 10.1	2.4	64.7	1.5	65.2
PixOOD [all] $s^{i^*}_{\mathcal{O}}$	✗	69.8	37.9	45.0	60.0	20.3	39.4	6.2	❷ 56.5
PixOOD unified	✗	79.7	13.1	43.1	29.3	50.5	28.6	21.5	❶ 44.2
Cross-domain									
UNO [18]	✓	66.3	90.8	49.1	90.5	❷ 55.5	92.9	❶ 37.2	92.4
RbA [15]	✓	❷ 76.1	68.9	37.9	87.9	❶ 56.4	80.7	❷ 24.6	91.6
RbA [15]	✗	62.4	99.1	32.5	99.3	43.3	97.3	15.7	98.5
PixOOD [road,sw]	✗	❶ 92.3	❶ 5.1	❶ 84.3	❶ 10.8	2.4	❷ 65.3	1.5	❷ 65.6
PixOOD [all] $s^{i^*}_{\mathcal{O}}$	✗	49.1	58.4	34.4	65.2	11.4	73.7	4.8	80.7
PixOOD unified	✗	74.2	❷ 19.9	❷ 51.8	❷ 26.4	26.6	❶ 50.0	11.5	❶ 59.5

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART ON MVTEC AD [1]. THE METHODS ARE GROUPED BY THE USE OF A PER-CLASS MODELS FOR EACH OBJECT CATEGORY VS UNIFIED MODEL. THE BOTTOM LINE SHOWS RESULTS OF COMBINING PIXOOD WITH SPECIALISED METHOD DRAEM [16].

Method	per-object model	OOD Data	Image-level	Pixel-level	
			AUROC ↑	AUPRO ↑	AUROC ↑
PaDiM [54]	✓	✗	97.9	92.1	97.5
DRAEM [16]	✓	✓	98.0	92.8	97.3
DAF [17]	✓	✓	97.6	93.0	❷ 98.1
CFLOW-AD [32]	✓	✗	98.3	❶ 98.6	94.6
PatchCore (25%) [33]	✓	✗	❷ 99.1	93.4	❷ 98.1
PNI [31]	✓	✗	❶ 99.6	❷ 96.1	❶ 99.0
UniAD [29]	✗	✗	96.5	–	96.8
HVQ-Trans [28]	✗	✗	❷ 98.0	–	❷ 97.3
MambaAD [30]	✗	✗	❶ 98.6	❶ 93.1	❶ 97.7
PixOOD	✗	✗	97.0	91.3	94.6
PixOOD unified	✗	✗	96.9	❷ 91.9	95.5
PixOOD + DRAEM	✓	✓	98.6	95.3	98.0
PixOOD unified+ DRAEM	✓	✓	98.6	95.4	98.2

TABLE VII
COMPARISON WITH THE STATE-OF-THE-ART ON LARS [3]. RESULTS FROM OFFICIAL LEADERBOARD AT THE DATE OF SUBMISSION LOJZEZUST.GITHUB.IO/LARS-DATASET (CONSIDERING LATEST MODEL SUBMISSIONS FROM INDIVIDUAL INSTITUTIONS; UNIL - UNIVERSITY OF LJUBLJANA, UNIC - UNIVERSITY OF CAGLIARY).

Model Name (MaCVi year)	Q	μ	Pr	Re	F1	mIoU
WaSR-T (UniL) '24	60.1	71.1	59.7	64.7	62.1	96.7
DeepLabv3 (UniL) '24	62.9	77.5	61.1	72.0	66.1	95.2
SegFormer (UniL) '24	67.8	78.6	63.8	77.5	70.0	96.8
KNet (UniL) '24	71.3	78.8	67.6	80.4	73.4	97.2
Mask2Former (HSU) '24	73.8	78.5	76.9	73.8	75.3	98.0
NFormer (Wisdom) '25	75.0	79.2	73.1	80.6	76.7	97.9
PSAM (UniL) '25	75.3	79.0	74.3	81.6	77.8	96.8
Mask2former (DLMU) '24	75.7	78.4	79.7	75.1	77.3	97.8
Mask2former (BUPT) '25	76.3	74.7	78.9	77.3	78.1	97.7
Swin-L (HKUST) '24	77.8	79.6	78.5	82.0	80.2	97.1
Snarciv3 (UniC) '24	78.1	79.7	76.9	83.0	79.9	97.8
Hk9084 (individual) '25	79.0	79.7	77.6	84.2	80.8	97.8
WaterScenes '25	80.8	80.6	82.9	79.8	81.3	98.1
WaterFormer (GIST) '25	83.2	80.5	81.6	88.0	84.7	98.3
PixOOD '24	73.9	74.6	70.7	81.6	75.8	97.5
PixOOD ft '25	76.9	74.5	75.3	83.2	79.0	97.3
PixOOD ft unified '25	77.4	74.4	76.0	82.7	79.2	97.7

similarly to previous experiments, evaluates the methods on ROI: {Road, Image}. The results are presented in Tab. V. Note that the results for the state-of-the-art methods were obtained from the official benchmark publication [50].

The results show that the Mask2Former-based methods struggle with some types of anomalies (e.g. thin structures) that are not detected at all (very high FPR). On the other hand, pixel-level (including the proposed PixOOD) can detect even thin structures since they classify each pixel individually. PixOOD produces smaller, but still high number of false positives, but this is probably due to the lack of spatial regularization. In general, the FPR is much higher in the ISSU benchmark for all methods because of the naturally more cluttered environment of the Indian driving scenes. The proposed PixOOD method demonstrated the highest level of generalisation in the cross-domain setting among the evaluated methods.

D. Industrial Anomaly Detection

The industrial anomaly detection problem is most often addressed by training a model for each individual class (product). Only recently, new methods started to address the problem in a holistic manner by training a single model for all objects categories. PixOOD inherently falls into this category of a single model for all products with the caveat that, for the per-class decision modelling, we use the corresponding product class OOD score at the test time. The results for the standard MVTec AD [1] benchmark are shown in Tab. VI and qualitative results in Fig. 6. PixOOD performs competitively to domain specialised methods that use unified model.

Furthermore, we can incorporate domain specific knowledge, *e.g.* from DRAEM [16], to increase the performance for this specific task. The resulting task-augmented PixOOD (denoted as "PixOOD + DRAEM") significantly improves the results, especially in the case of the AUPRO metric.

E. Maritime Obstacle Detection

This experiment demonstrates the versatility of the proposed method in different settings and domains. The maritime obstacle detection problem is posed as semantic segmentation. The task is to segment the image into three classes – *Water, Sky and Obstacle*. During training, examples of all classes are available. To fit PixOOD to the setting, we model all three classes as in-distribution data, and during inference we use the classification network to produce class labelling. For each pixel and predicted class we then decide based on the OOD score if the pixel is ID (the class label remains as predicted) or OOD for which we change the class label to *Obstacle*.

The results of our method are compared to the state-of-the-art methods (taking the latest version of the method submitted) from the official leaderboard and are shown in Tab. VII. The proposed method performs favourably placing sixth (with small margin between 4-6th places) with the main issue being low precision due to over-segmentation (see Fig. 4 and Fig. 7 for qualitative examples). Note that due to availability of training data for the obstacle class, most of the heavy lifting is done by the discriminative part of the method, *i.e*. MLP logits predictions, and the task can be addressed by semantic segmentation methods. This is the case of other state-of-the-art methods in the leaderboard, which are "standard" semantic segmentation approaches. To highlight this fact, we included version of the PixOOD that only differ in fine-tuning the backbone during training of the MLP (with learning rate 10^{-6} and cosine scheduling) to improve the semantic segmentation part, and thus, the overall performance improved.

F. Computational Costs

All the reported training times assumes a single NVIDIA A100-40GB graphics card. The training on the Cityscapes dataset (~ 2500 images of original resolution 2048×1024) takes around 18 hours for MLP and the runtime of the condensation algorithm for all classes was 5, 8 and 17 hours for K equal to 200, 500 and 1000 respectively.

The complete inference for one model on datasets used in Tab. I takes around 22 (7) minutes using the $7(1)\times$ feature map upsampling respectively. The inference for one image of resolution around 2000×1000 (1200×720) around 4.7 (2) second for the $7\times$ feature map upsampling and 1.2 (0.4) seconds for no feature upsampling.

V. Conclusions

In this paper, we proposed a novel pixel-level OOD detection method. The method is not designed for a specific task or benchmark and performs competitively on a range of pixel-level OOD problems. The method does not require OOD training samples, neither real nor synthetic. The method builds on a proposed data condensation algorithm which is theoretically linked to the optimization of a complete data log likelihood in the EM algorithm. We applied the proposed method to three very diverse pixel-level anomaly benchmarks and achieved state-of-the-art results on four out of the seven considered datasets (i.e. Road Anomaly, FS LaF, SMIYC-Obstacle, SMIYC-LAF) and performed competitively on others.

The presented experiments also reveal the limitations of the chosen problem formulation. First, PixOOD relies on patch-level features and scoring, without an explicit spatial prior. Consequently, its errors are predominantly of spatial nature: under(over)-segmented anomalous regions, "holes" in large anomalous objects, higher false positive rate in cluttered scenes.

Second, the method is calibrated to the support of the training distribution rather than to semantic class generalization. For example, if a specific "car" instance is present in the training data, PixOOD does not assume that all possible visual realizations of semantic class "car" should therefore be treated as in-distribution. This explains why unusual but semantically valid appearances and domain shifts are detected as anomalous. At the same time, this behaviour may be desirable in applications where the goal is to flag previously unseen appearances despite semantic similarity.

Third, the method is fundamentally appearance-based, and therefore misses logical anomalies whose local texture remains plausible while the global configuration is wrong, such as switched cables or missing pill labels. The logical anomalies go beyond the intended scope of the method and indicates an interesting future direction.

Acknowledgments

The work was supported by Toyota Motor Europe and Czech Science Foundation Grant No. 25-15993S.

References

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," in *CVPR*, June 2019.

[2] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann, "SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation," in *NeurIPS Datasets and Bench.*, 2021.

[3] L. Žust, J. Perš, and M. Kristan, "LaRS: A Diverse Panoptic Maritime Obstacle Detection Dataset and Benchmark," in *ICCV*, 2023.

[4] T. Vojíř, J. Šochman, R. Aljundi, and J. Matas, "Calibrated Out-of-Distribution Detection with a Generic Representation," in *ICCVW*, October 2023.

[5] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference," *Biometrika*, 1928.

[6] J. Neyman and E. S. Pearson, "IX. On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 1933.

[7] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes," in *ECCV*, 2021.

[8] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-Wise Anomaly Detection in Complex Driving Scenes," in *CVPR*, June 2021.

[9] S. N. Rai, F. Cermelli, D. Fontanel, C. Masone, and B. Caputo, "Unmasking Anomalies in Road-Scene Segmentation," in *ICCV*, October 2023.

[10] M. Grcić, J. Šarić, and S. Šegvić, "On Advantages of Mask-Level Recognition for Outlier-Aware Segmentation," in *CVPRW*, June 2023.

[11] Z. Gao, S. Yan, and X. He, "ATTA: Anomaly-aware Test-Time Adaptation for Out-of-Distribution Detection in Segmentation," in *NeurIPS*, 2023.

[12] M. Grcic, P. Bevandic, and S. Segvic, "DenseHybrid: Hybrid Anomaly Detection for Dense Open-Set Recognition," in *ECCV*, 2022.

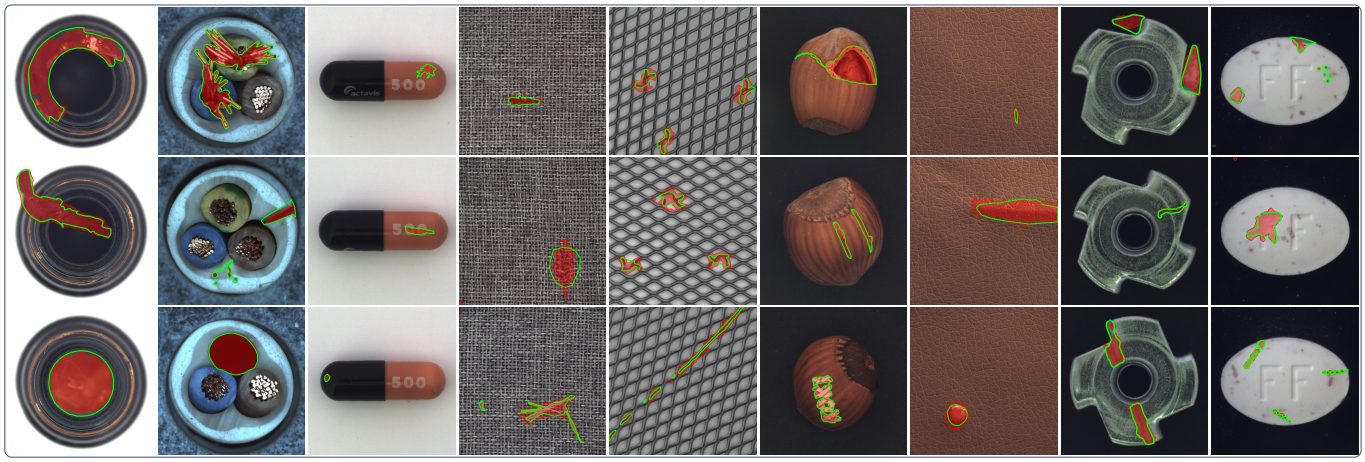


Fig. 6. Industrial anomaly detection examples from the MVTec AD dataset. Legend: red - detected anomaly, green - ground-truth.

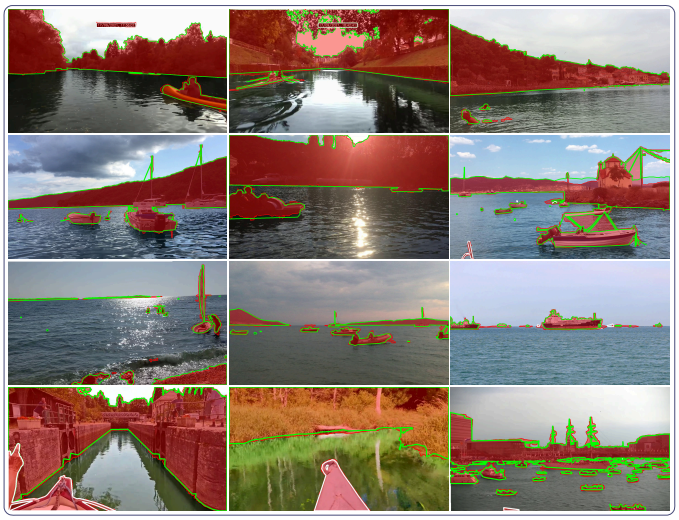


Fig. 7. Maritime Obstacle (LaRS) anomaly detection examples. Legend: red - detected anomaly, green - ground-truth, white - ignore region.

[13] Y. Liu, C. Ding, Y. Tian, G. Pang, V. Belagiannis, I. Reid, and G. Carneiro, "Residual Pattern Learning for Pixel-Wise Out-of-Distribution Detection in Semantic Segmentation," in *ICCV*, October 2023.

[14] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy Maximization and Meta Classification for Out-of-Distribution Detection in Semantic Segmentation," in *ICCV*, October 2021.

[15] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney, "RbA: Segmenting Unknown Regions Rejected by All," in *ICCV*, 2023.

[16] V. Zavrtanik, M. Kristan, and D. Skocaj, "DRAEM - A Discriminatively Trained Reconstruction Embedding for Surface Anomaly Detection," in *ICCV*, 2021.

[17] Y. Cai, D. Liang, D. Luo, X. He, X. Yang, and X. Bai, "A Discrepancy Aware Framework for Robust Anomaly Detection," *IEEE Tran. Indust. Info.*, 2023.

[18] A. Delić, M. Grcic, and S. Šegvić, "Outlier detection by ensembling uncertainty with negative objectness," in *British Machine Vision Conference (BMVC)*, 2024.

[19] S. Liang, Y. Li, and R. Srikant, "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks," in *ICLR*, 2018.

[20] T. Vojir, T. Šipka, R. Aljundi, N. Chumerin, D. O. Reino, and J. Matas, "Road Anomaly Detection by Partial Image Reconstruction With Segmentation Coupling," in *ICCV*, 2021.

[21] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the Unexpected via Image Resynthesis," in *ICCV*, October 2019.

[22] K. Lis, S. Honari, P. Fua, and M. Salzmann, "Detecting road obstacles by erasing them," in *arXiv 2012.13633*, 2020.

[23] M. Grcić, P. Bevandić, Z. Kalafatić, and S. Šegvić, "Dense Out-of-Distribution Detection by Robust Learning on Synthetic Negative Data," in *arXiv 2112.12833*, 2023.

[24] V. Besnier, A. Bursuc, D. Picard, and B. Alexandre, "Triggering Failures: Out-Of-Distribution detection by learning from local adversarial attacks in Semantic Segmentation," in *ICCV*, 2021.

[25] T. Vojíř and J. Matas, "Image-Consistent Detection of Road Anomalies As Unpredictable Patches," in *WACV*, 2023.

[26] S. Galesso, M. Argus, and T. Brox, "Far Away in the Deep Space: Dense Nearest-Neighbor-Based Out-of-Distribution Detection," in *ICCVW*, October 2023.

[27] H. Zhang, F. Li, L. Qi, M.-H. Yang, and N. Ahuja, "CSL: Class-Agnostic Structure-Constrained Learning for Segmentation Including the Unseen," *Proc. of the AAAI Conf. on Artif. Intell.*, vol. 38, no. 7, 2024.

[28] R. Lu, Y. Wu, L. Tian, D. Wang, B. Chen, X. Liu, and R. Hu, "Hierarchical Vector Quantized Transformer for Multi-class Unsupervised Anomaly Detection," in *NeurIPS*, 2023.

[29] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A Unified Model for Multi-class Anomaly Detection," in *NeurIPS*, 2022.

[30] H. He, Y. Bai, J. Zhang, Q. He, H. Chen, Z. Gan, C. Wang, X. Li, G. Tian, and L. Xie, "MambaAD: Exploring State Space Models for Multi-class Unsupervised Anomaly Detection," *NeurIPS*, 2024.

[31] J. Bae, J.-H. Lee, and S. Kim, "PNI : Industrial Anomaly Detection using Position and Neighborhood Information," in *ICCV*, October 2023, pp. 6373–6383.

[32] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "CFLOW-AD: Real-Time Unsupervised Anomaly Detection With Localization via Conditional Normalizing Flows," in *WACV*, January 2022, pp. 98–107.

[33] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards Total Recall in Industrial Anomaly Detection," in *CVPR*, 2022, pp. 14 318–14 328.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021.

[35] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

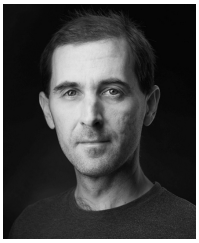
[36] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "AnomalyGPT: Detecting Industrial Anomalies using Large Vision-Language Models," *arXiv preprint arXiv:2308.15366*, 2023.

[37] H. Wang, Y. Li, H. Yao, and X. Li, "CLIPN for Zero-Shot OOD Detection: Teaching CLIP to Say No," in *ICCV*, October 2023.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[39] T. Vojíř, J. Šochman, and J. Matas, "PixOOD: Pixel-Level Out-of-Distribution Detection," in *ECCV*, 2024.
[40] M. I. Schlesinger and V. Hlavac, *Ten Lectures on Statistical and Structural Pattern Recognition*. Computational Imaging and Vision, 2002.
[41] M. Cho, K. Alizadeh-Vahid, S. Adya, and M. Rastegari, "DKM: Differentiable k-Means Clustering Layer for Neural Network Compression," in *ICLR*, 2022.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[43] P. L. R. H. Byrd and J. Nocedal, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Jour. on Scien. and Stat. Comp.*, 1995.
[44] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv 1606.08415*, 2023.
[45] S. Fu, M. Hamilton, L. E. Brandt, A. Feldmann, Z. Zhang, and W. T. Freeman, "FeatUp: A Model-Agnostic Framework for Features at Any Resolution," in *ICLR*, 2024.
[46] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, "Lost and found: detecting small road hazards for self-driving vehicles," in *IROS*, 2016.
[47] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation," *arXiv 1904.03215*, 2019.
[48] B. Bovcon, J. Muhovič, D. Vranac, D. Mozetič, J. Perš, and M. Kristan, "MODS–A USV-Oriented Object Detection and Obstacle Segmentation Benchmark," *ITS*, 2021.
[49] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in *ICLR*, 2017.
[50] Z. Laskar, T. Vojíř, M. Grcic, I. Melekhov, S. Gangisetty, J. Kannala, J. Matas, G. Tolias, and C. Jawahar, "A Dataset for Semantic Segmentation in the Presence of Unknowns," in *CVPR*, 2025.
[51] G. Varma, A. Subramanian, A. M. Namboodiri, M. Chandraker, and C. V. Jawahar, "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments." in *WACV*, 2019.
[52] F. A. Shaik, A. R. Malreddy, N. R. Billa, K. Chaudhary, S. Manchanda, and G. Varma, "IDD-AW: A benchmark for safe and robust segmentation of drive scenes in unstructured traffic and adverse weather." in *WACV*, 2024.
[53] C. Parikh, R. Saluja, C. V. Jawahar, and R. K. Sarvadevabhatla, "IDD-X: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic." in *IEEE Int. Conf. on Robotics and Automation*, 2024.
[54] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization," in *ICPRW*, 2020.

Tomáš Vojíř received the Ph.D. degree from the Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic, in 2018. He is currently working as research fellow with the Visual Recognition Group, FEE, Czech Technical University in Prague, Prague, Czech Republic. His research interests include out-of-distribution detection, uncertainty modelling and vision for autonomous driving, such as road anomaly detection.



Jan Šochman received the Ph.D. degree from the Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic, in 2009. He is currently working as research fellow with the Visual Recognition Group, FEE, Czech Technical University in Prague, Prague, Czech Republic. His research interests include out-of-distribution detection, video motion segmentation, and active feedback loop learning.

Jiří Matas received the Ph.D. degree from the University of Surrey, Guildford, U.K., in 1995. He is a Professor with the Visual Recognition Group, FEE, Czech Technical University in Prague, Prague, Czech Republic. His research interests include visual tracking, out-of-distribution detection, object recognition, image matching and retrieval, sequential pattern recognition, and RANSAC-type optimization methods.