Dark Side Augmentation: Generating Diverse Night Examples for Metric Learning

Albert Mohwald Tomas Jenicek Ondřej Chum
VRG, Faculty of Electrical Engineering, Czech Technical University in Prague
mohwaalb@fel.cvut.cz jenicto2@fel.cvut.cz chum@cmp.felk.cvut.cz

Abstract

Image retrieval methods based on CNN descriptors rely on metric learning from a large number of diverse examples of positive and negative image pairs. Domains, such as night-time images, with limited availability and variability of training data suffer from poor retrieval performance even with methods performing well on standard benchmarks. We propose to train a GAN-based synthetic-image generator, translating available day-time image examples into night images. Such a generator is used in metric learning as a form of augmentation, supplying training data to the scarce domain. Various types of generators are evaluated and analyzed. We contribute with a novel light-weight GAN architecture that enforces the consistency between the original and translated image through edge consistency. The proposed architecture also allows a simultaneous training of an edge detector that operates on both night and day images. To further increase the variability in the training examples and to maximize the generalization of the trained model, we propose a novel method of diverse anchor mining.

The proposed method improves over the state-of-the-art results on a standard Tokyo 24/7 day-night retrieval benchmark while preserving the performance on Oxford and Paris datasets. This is achieved without the need of training image pairs of matching day and night images. The source code is available at https://github.com/mohwald/gandtr.

1. Introduction

Large-scale instance-level image retrieval is commonly used *e.g.* as a first step in visual place recognition and visual localization. As other computer vision problems, image retrieval is dominated by methods based on deep learning models. Fast and memory efficient approaches learn global image descriptors via metric learning. A large number and variety of corresponding image pairs is required to train well-performing global image descriptors.

One of the recent challenges in retrieval and visual localization is insensitivity to severe illumination changes, such as day and night [33, 34]. Methods trained mostly on the day domain perform well on that domain, while having relatively poor results when night images are observed. The goal is an approach that performs well on all domains; im-

Figure 1. Examples of day-to-night translations with various generators. Each row consists of (left to right) the source image, and images translated by: CycleGAN, CyEDA, and the proposed RCF^{N}GAN and HED^{N}GAN. All models are trained on the *SfM* dataset, except for CyEDA where a model pre-trained on BDD100k is used.

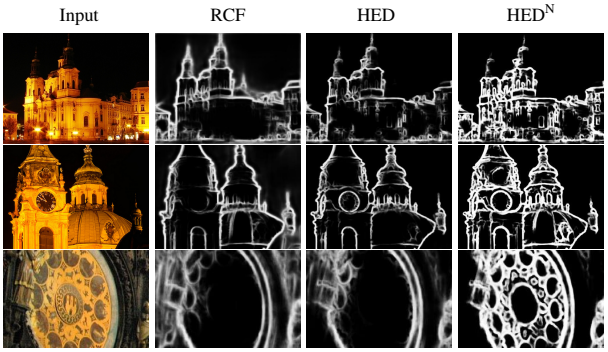

Figure 2. Comparison of edges extracted from real night images by RCF [19], HED [45], and proposed HED^N, which is trained jointly with the edge-consistency based HED^{N}GAN generator. RCF and HED were trained mainly on day images and do not detect some edges in the night.

proving results when night images are involved, while at the same time, the performance on the day domain should be preserved. The requirement of unharmed performance on the original (source) domain is, however, often neglected in the domain adaptation methods (*e.g*. [16], as shown by our experiments).

To achieve illumination invariant image retrieval, alignment of the day and night domains by metric learning was proposed by [13]. To achieve this, a large number of matching night to day image pairs is required. The acquisition of such pairs is a non trivial task, and suffers form significantly lower variability compared to day-to-day corresponding image pairs. Photo-sharing sites are a popular source of landmark image training data. A number of training and testing datasets were crawled from such sites, for example revisited Oxford and Paris [26], Google landmarks v1 [23], *SfM* [28] and *SfM N/D* [13]. For day images, these datasets exhibit sufficient visual variability to train image to descriptor mappings that generalize well to unseen scenes. In contrary, not only there are significantly less images taken during the night (*e.g*. in the Aachen Day-Night dataset [33], there are 30 times more day images than night images), but their variability is also lower, as only parts of the scenes are visually interesting (*e.g*. lit) during the night time. Therefore, only a small fraction of the scene reconstructed from the daytime images is photographed during night. This has been shown by day and night 3D reconstructions in [27].

As our *main contribution*, we propose to replace night training examples by synthetic images derived from day images by a generative adversarial network (GAN). In particular, a standard *SfM* dataset [30] with high variability of homogeneous (mostly day to day) matching image pairs is used and one of the matching images is transformed by a GAN into a night image, see Figure 1. This alleviates the necessity of obtaining night to day matching image pairs, and also significantly increases the variability of the training pairs. Even though night images are required to train the GAN, (i) a much lower number of night images is needed compared to performing the metric learning, (ii) these do not have to be paired with matching day counterpart. We compare various existing image translation methods that do not require pixel-aligned or visually related training data, in particular CycleGAN [49], DRIT [15], CUT [24], and CyEDA [3].

Inspired by the relative success of edge-based approaches to illumination invariance, as a *second contribution*, we propose a novel consistency enforcement through the edge consistency. Specifically, differentiable edge detector HED [45] is used to extract edges from the original and the translated image and their dissimilarity is penalized. The proposed method has a number of advantages: (i) it is an order of magnitude faster to train than CycleGAN while providing similar retrieval results, (ii) it provides insight into the importance and sufficiency of edges in night vision, (iii) it allows for simultaneous training of an edge detector (HED^{N}) that detects edges well in both day and night images. In this setup, HED [45] is compared to the more recent edge detector RCF [19].

Training data from automated 3D reconstructions are popular, as they are very clean and available without any human annotation. On the downside, the data distribution has strong modes that correspond to canonical views of popular landmarks. As a *third contribution*, we propose to further increase the variability in the training examples, by a novel method of diverse anchor mining. Instead of a random selection of training examples for each epoch, pseudo-random importance sampling is used, preventing over-using training data from the modes of the training distribution (avoiding using multiple similar examples in the training).

We explore the idea of using diverse synthetic data for metric learning, compare different generators, including a newly proposed one, and study the contribution of individual aspects of the proposed method on global descriptors. We evaluate the performance of our models on image retrieval and visual localization datasets. The contribution is applicable to other methods as well, which we demonstrate by applying the proposed method to HOW [41], a model that uses local descriptors for retrieval.

2. Related Work

In this section, we first review relevant approaches to day-night image retrieval and discuss their relation to our work. Then, we summarize the image-to-image translation and how it is utilized in the data augmentation task, and finally we outline how other works are tackling data augmentation for visual recognition and image retrieval through day–night domain adaptation.

Day-time image retrieval. In GeM [30], a CNN backbone produces global descriptors which can be compared by L2 norm to measure similarity between images. An alternative approach to use the CNN backbone to produce a set of local features was proposed in DELF [23], HOW [41], and FIRe [43]. In HOW, the last feature map is treated as a set of local features, from which the strongest features are used for image retrieval via ASMK [40]. FIRe follows the same pipeline, but adds a transformer-based head on top of the convolutional backbone.

The two paradigms can be combined, such as in DELG [5] where both global and local descriptors are utilized, each produced by a separate head. In DOLG [46], a slightly different approach is adopted, using the interactions between global and local descriptors to produce a stronger global descriptor. This principle is applied to vision transformers [7] in ViTGaL [25] where a cross-attention between the CLS token and spatial token embeddings is performed

at the end of the network. An alternative approach using a transformer-based backbone is taken by DToP [37], where local and global representations are produced by separate branches and the final representation is a concatenation of the two. The application of the proposed method to any of these methods is straightforward; we demonstrate its effectiveness on GeM [30] and HOW [41].

Real day-night training data. The closest method of CNN-based illumination-invariant image retrieval is the work of [13]. The alignment of the day and night domains consists of two steps – a photometric normalization of input images, e.g. using CLAHE [1], and the exploitation of *matching pairs* of day and night images from a 3D reconstruction [28] for training. In our work, we remove the requirement of obtaining matching night and day images. Further, we show that generating synthetic night images increases the variability in the training data which is reflected in a better retrieval performance.

Edges and color invariants. In EdgeMAC [29], metric learning is performed on edge-detector responses (edgemaps), without the need of night training images. It was shown that edges are preserved in the presence of a significant change in illumination such as day and night images, and even for images where colors and textures are corrupted. However, EdgeMAC was experimentally shown [29, 13] to perform poorly on standard datasets, as too much relevant information is lost in the process of turning images to edgemaps.

For edge detection, EdgeMAC exploits Dollár [6] which utilizes random decision forests. HED [45] proposed a CNN-based edge detector where detections from multiple intermediate feature maps are fused together, combining detections with different receptive fields. RCF [19] follows the same architecture, but proposes to exploit detections from every intermediate feature map, which yields better results at the cost of an increased inference time.

Recently, a zero-shot day-night domain adaptation was proposed by [16]. A layer with trainable parameters performing color-invariant edge extraction is preceded to the backbone network. The method achieves interesting results without any night images during training. However, the retrieval results are significantly lower than results of the method proposed in this paper. Further, our experiments show that while improving the retrieval results in day-night settings, application of the method [16] substantially harms retrieval in the original day domain.

Image-to-image Translation. In image-to-image translation, popularized by the pix2pix [12] and CycleGAN [49] models, an image in one visual domain is transformed into another domain, preserving the image content but modifying the image style. In CycleGAN [49], the training images from the two domains do not need to be paired. There are two generators trained there, each translating in one direction between the two domains. This enables to constrain the input image and output image after two consecutive translations in the opposing directions to be identical.

In a more recent method DRIT [15], a similar architecture is presented, but with the focus on generating diverse output images. This is achieved by using a single latent space for all encoders and decoders of generators and splitting this latent space into a content and style space while introducing a new cross-cycle consistency loss.

An uni-directional image translation, training a generator for a single direction between the domains, is proposed in CUT [24]. The consistency is enforced by patch-wise contrastive loss between corresponding patches from the original and target domain (positive pair) and other patches in the original domain (negative pairs).

An edge-like consistency for CycleGAN has been proposed in CyEDA [3]. Instead of comparing images in pixel-space, consistence of gradients (Sobel responses) is enforced. In our proposed method, sparse edge detections are used instead, and the edge detector is improved during the GAN training. Additionally, CyEDA generator introduces a blending mask mixing the input and output images as a form of skip-connection, which often fails in our setup, see Figure 1 top row for an example.

Another line of work focuses on architectures specifically for the task of object detection. In [18], cycle structure-consistency is enforced by measuring the difference between corresponding segmentation maps. Another approach is taken by [35] and consequent work of [4], where the translation is performed directly for individual objects, implicitly ensuring a structure consistency in the context of object detection. Despite the impressive results in case of object-rich images, such an approach is not suitable for landmark recognition.

Data augmentation. To increase the number of training examples, various types of augmentations [8] were introduced in computer vision problems. Image-to-image translation can be also used as a form of data augmentation for training, which is exercised by the proposed method. Recently, [21] demonstrated that synthetic data helps in feature matching, visual localization, and image retrieval.

In [2], augmentation by image-to-image translation is applied to a car detection problem. Their training data contain car annotations in day images only, so they propose to translate day images into the night domain, exploiting the day annotations with the generated night data. This is similar to the work of [17], with the difference that they design a custom GAN architecture in order to preserve objects in the images, aiding the consequent augmentation for a vehicle detection. The same idea of translating day images with

[1] Contrast Limited Adaptive Histogram Equalization, see [39] or [13] for a detailed description

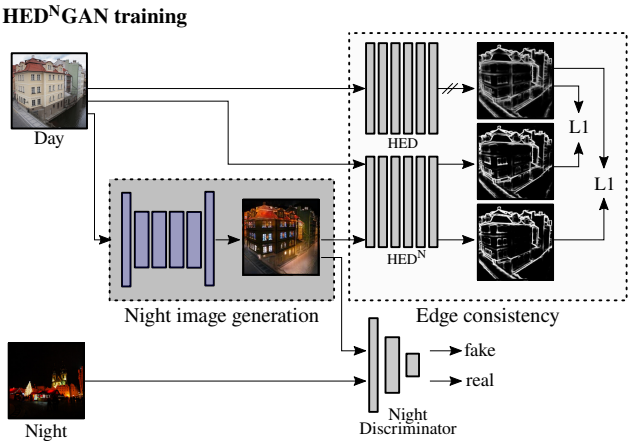

Figure 3. One training step with unpaired day and night images (left block) of our HED^{N}GAN architecture. The day → night generator translates the input day image (top left) into a fake night image (center), enforcing the edge consistency by L1 loss between HED and HED^{N}outputs (top right). The night discriminator predicts whether the generated night image (center) and the input night image (bottom left) are real or fake. HED^{N}edge detector (student) is trained by HED edge detector (teacher, not trained) to output night image edgemaps while preserving day image edgemaps.

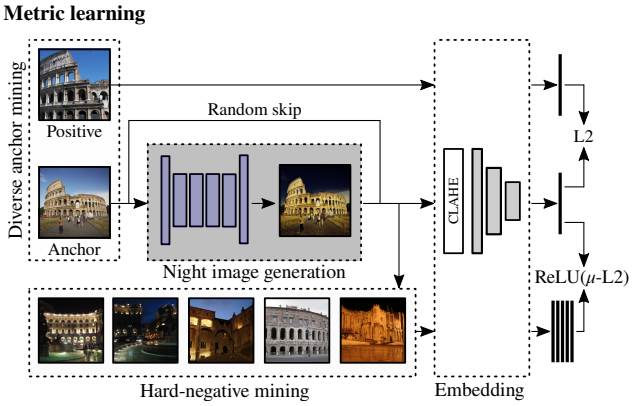

Figure 4. Training data generation and photometric normalization during embedding network fine-tuning. A mined diverse anchor (center left day image) is randomly translated into a night image (gray block, trained generator from Figure 3). The randomly translated image is used to mine a set of five negative images. The contrastive loss is applied on the global descriptors of the positive pair (L2) and of the negative pairs (ReLU(μ - L2)).

annotations into the night domain is used by [38] for image segmentation of vehicle-mounted camera images.

Image translation at the test time. Visual localization is approached by translating night images into day images during the inference in ToDayGAN [1]. Images from a camera mounted on a car are used, hence the variability in the images is relatively low. It is not clear whether such an approach would be applicable to general scenes[2]. We argue that generating a synthetic image commits to one of possible appearances, which, even if photo-realistic, is not guaranteed to be similar to the reality. Therefore, the applicability of image translation at the test time is limited. Instead, we attempt to learn an embedding that deals with all possible appearances, by using the image translation during the training. This is supported by [38] for the task of image segmentation with vehicle-mounted camera, where translating night images into day during inference yields significantly worse results compared to translating day images into night as a training data augmentation.

3. Method

In the proposed method, a GAN generator is first trained on unpaired day-night images. The trained generator is exploited in the consequent metric learning to generate day-night training examples from labelled day pairs. For metric learning, a standard global descriptor constrastive learning framework is followed [13] with three introduced changes: (a) diverse anchor images are mined, (b) anchor day images are translated to a night domain, and (c) negative mining is preformed after the optional translation step.

3.1. HED^{N}GAN training

It was shown previously [28, 13] that edges provide information that survives illumination changes between day and night. We propose a simple uni-directional image translator, named HED^{N}GAN, that attempts to generate images from the target domain and for which edgemaps are similar to edgemaps of the corresponding source images. For this purpose, a differentiable edge detector HED [45] is utilized.

The method is trained on examples from the day and night domain, where two unpaired images are randomly sampled in each iteration, one from each domain. The architecture consists of three models – the generator, discriminator, and edge detector, as depicted in Figure 3. In each iteration, a day image is translated via the generator into the night domain, resulting in a fake night image that is aligned with the real day image. An edge detection is performed on both the real day image (generator input) and the fake night image (generator output), and the resulting two edgemaps are compared pixel-wise, constraining the edges to be consistent between the two images. The discriminator is applied on the fake night image, training the generator adversarially, and ensuring that the images outputted are indistinguishable from the true night images. At the same time, the discriminator is trained on both sampled images – the fake night image and a randomly sampled real night image.

HED detector [45] is trained mainly on day images which causes it to miss edges in night images, negatively in-

[2]The official GitHub implementation of [1] proclaims "sensitivity to intrinsic camera characteristics".

fluencing the generator performance in our setup. Therefore HED$^{\mathrm{N}}$ detector is trained jointly with the generator training, so that HED$^{\mathrm{N}}$ (trained student) and HED (frozen teacher) have similar responses on both real day and generated night images. HED with HED$^{\mathrm{N}}$ is also used to measure the similarity of the generator input and output, enforcing the edge cosistency between them. We also test a variant where the HED edge detector is not trained, named HEDGAN. The alternative edge detector RCF [19] is evaluated in RCF$^{\mathrm{N}}$GAN method.

Learning details. In all experiments, each input image is first randomly downscaled by a scale from 0.8 to 1.0 and then random-cropped to the final size of 256x256 px. In a single iteration, two unpaired images from the two domains, day and night, are processed. Each training epoch consists of 10000 iterations.

In the HED$^{\mathrm{N}}$GAN architecture, ResNet generator [49], patchGAN discriminator [12], and HED edge detector [45][3] are exploited. All three networks are trained simultaneously with batch size of 10. The training step of the generator and discriminator is the same as in CycleGAN [49]. The generator and the discriminator use batch normalization and network weights are initalized following [9], both of which led to an increased stability of the generator during training in our case. The two HED edge detectors, student and teacher, are initialized with the weights from [22]; the teacher weights are not updated during training while the student weights are optimized with Adam optimizer [14] with learning rate 10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.0002. The edgemaps of the student and the teacher are compared pixel-wise via L1 distance – in the optimization step of the edge detector, the L1 loss is applied on output values before the sigmoid function, while in the optimization step of the generator, it is after the sigmoid function.

For the other three tested architectures, CycleGAN [49], DRIT [15], and CUT [24], their original implementations are used, in which the networks are trained for 100, 300, and 50 epochs, respectively. In all three architectures, the learning rate is linearly decayed to zero over the second half of epochs. For CyEDA [3], its pre-trained models on GTA [31] and BDD100k [47] are evaluated as well as a variant trained on *SfM*120k [30] using its original training implementation.

3.2. Metric learning

The learning of global image descriptors is cast as metric learning via Siamese network, the architecture is visualized in Figure 4. We follow the procedure used in [13] including the same hyper-parameter settings. First, to bring the appearance of images from different domains close together, a non-linear photometric CLAHE [39] normalization. CLAHE is performed on a grid 8x8 with clip limit of 1. The training is initialized with ImageNet pre-trained network [32], followed by fine-tuning on the *SfM* dataset [28, 13]. For the embedding network architecture, VGG-16 [36] or ResNet-101 [10] backbone is used, followed by GeM pooling and L2 normalization, as described in [30]. The network is trained for 40 epochs with 2000 iterations each epoch.

To show wide applicability of the proposed method, we also train a retrieval method based on aggregated local features – HoW [41]. The network is trained in a metric learning framework with a contrastive loss on global descriptors; the procedure of [41] is followed for both training and inference.

Night image generation. A trained day→night generator is used to generate day-night examples from day-only pairs. Before each epoch, 25% of anchors are translated from day to night domain, while the corresponding positive and negative images are left unchanged. In the Retrieval-SfM N/D dataset [13], the same ratio of night anchors is used. The generator weights are not updated during training.

Hard negative mining. In each iteration of the fine-tuning, the embedding network takes 7-tuple of images – one image is the *anchor*, one image is *positive*, and the remaining five images are *negative* examples, following [28, 13]. For each anchor, negative examples are mined from different 3D reconstructions, so that the distance between their descriptors and the anchor image descriptor is minimal. The negative mining takes place after the eventual anchor translation into the night domain.

Diverse anchor mining. In prior approaches of metric learning on the *SfM* dataset [28, 13], positive pairs for each epoch are selected from the set of all positive pairs in the dataset at random. Such a choice may lead to a repeated selection of similar anchor images (the same scene with a near-by view point) within an epoch. To vary the training examples, we propose to iteratively select diverse anchor images from a random pool of anchor images. The first anchor is selected at random. Before the next anchor is selected, the remainder of the pool is ordered by the minimal distance to already selected anchor images. The distance is measured as a Euclidean distance of image descriptors extracted by the network in its current state. New anchor image is selected at random from images between the 20th and 80th percentile of the pool ordering. Dropping 20% of the closest images encourages the diversity in anchors, dropping 20% of the most distant images prevents selecting images with outlying descriptors. In the training, 2000 anchor images are selected from a pool of 10000 anchors.

[3] Re-implementation https://github.com/sniklaus/pytorch-hed is used.

4. Experiments

The proposed method is experimentally evaluated on various standard datasets, including day-night datasets as well as mostly homogeneous day datasets. The ablation study shows contributions of individual steps for different settings and implementation choices. The impact of the choice of the training dataset and the amount of night images used in training is discussed.

4.1. Datasets

For training, we use the Retrieval-SfM dataset [30]. Three standard datasets are used for evaluation of image retrieval: Tokyo 24/7 [42, 13], revisited Oxford and Paris [26]. The night-time performance is also assessed on visual localization on two datasets: Aachen Day-Night v1.1 [48] and RobotCar Seasons [33, 20].

Retrieval-SfM [30] (*SfM*) contains 98045 images from reconstructed 3D models. This dataset was used in the prior work of [30, 13] to finetune a CNN for image retrieval. In this work, we also use the *SfM* dataset for the metric learning and day→night generator training. For the generator training, the generated day-night annotations from [27] were used and images with dimensions less than 512 px were removed, resulting in 86385 day and 10039 night images.

Retrieval-SfM N/D [13] (*SfM-N/D*) was introduced by [13] aiming to construct positive pairs with different lighting conditions. The *SfM* dataset was enriched by additional day-night positive pairs by using the information in the original 3D models. Note that this dataset is included for comparison only, it is *not* required to achieve results claimed as contribution of this paper (marked **!** in methods).

Tokyo 24/7 [42] (*Tokyo*) is a collection of 1125 smartphone-camera pictures capturing each of 375 scenes from 125 distinct locations in day, night and sunset light conditions. In this work, we use *Tokyo* to evaluate retrieval performance with the same evaluation protocol as proposed in [13] – each image is used as a query, images from the same scene but different lighting conditions are counted as positive, while images from different locations are considered as negative.

Oxford and Paris [26] ($\mathcal{R}$Oxf and $\mathcal{R}$Par) are standard image retrieval datasets (in their revisited version) and are used to evaluate retrieval performance on mostly homogeneous dataset with day-time images.

Aachen Day-Night v1.1 [34, 33, 48] (*Aachen*) contains images of the old inner city of Aachen in Germany. The database consists of 6697 daytime images taken by handheld cameras and the query set contains 824 day-time and 191 night-time query images taken by three mobile phones. Performance is reported for the night-time queries only.

RobotCar Seasons [20, 33] (*RobotCar*) consists of images captured from 3 vehicle-mounted cameras: 26121 database images and 11934 query images taken under different conditions. Performance is reported for the night-time evaluation protocol of visuallocalization.net benchmark [33] which contains images taken under *night* (1314 images) and *night-rain* (1320 images) conditions.

VGG-16 backbone

Method	Avg	Tokyo	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GeM [30]	69.9	79.4	**60.9**	69.3
GeM N/D [13] **!**	71.1	83.5	60.0	69.8
CIConv [16]	-	83.3	-	-
CLAHE $\mathcal{C}$ [13]	71.6	84.1	60.8	69.8
CLAHE N/D $\mathcal{C}$ [13] **!**	72.4	87.0	60.2	70.0
HED$^{\text{N}}$GAN $\mathcal{CD}$ (ours)	**73.4**	**88.9**	**61.1**	**70.3**
CycleGAN $\mathcal{CD}$ (ours)	**74.0**	**90.2**	60.7	**71.0**

ResNet-101 backbone

Method	Avg	Tokyo	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GeM [30]	75.7	85.0	65.3	**76.7**
CIConv [16]	75.0	88.3	62.0	74.7
HED$^{\text{N}}$GAN $\mathcal{CD}$ (ours)	**78.4**	**92.2**	**66.3**	**76.6**
CycleGAN $\mathcal{CD}$ (ours)	**78.4**	**92.0**	**66.8**	76.4

HOW ResNet-18 backbone

Method	Avg	Tokyo	$\mathcal{R}$Oxf	$\mathcal{R}$Par
HOW [41]	80.8	87.8	**75.1**	79.4
HOW N/D **!**	**82.0**	89.2	**75.5**	**81.4**
HED$^{\text{N}}$GAN $\mathcal{CD}$ (ours)	**82.0**	**91.6**	74.6	79.7
CycleGAN $\mathcal{CD}$ (ours)	**82.4**	**92.9**	74.6	**79.8**

Table 1. Comparison in terms of mAP on Tokyo 24/7, $\mathcal{R}$Oxf Medium and $\mathcal{R}$Par Medium datasets and their average. Methods marked by **!** use paired day-night training data. The best score for each backbone architecture (in separate tables) is emphasized by red bold, second best by bold.

Method	Avg	Tokyo	$\mathcal{R}$Oxf	$\mathcal{R}$Par
DOLG [46]	-	-	81.5	91.0
DOLG	82.6	75.4	82.4	91.0
ViTGaL [25]	-	-	82.4	91.4
ViTGaL	83.6	79.8	79.6	91.4

Table 2. Performance of the SoTA methods with publicly available code. DOLG [46] uses the convolutional backbone ResNet-101, ViTGaL [25] uses transformer backbone XCiT-S24. All methods are trained on GLDv2 dataset [44] which overlaps with test datasets $\mathcal{R}$Oxf and $\mathcal{R}$Par. For each method, we provide the results as reported in their paper (marked by the paper reference) and as reproduced by their publicly available code. Note the poor perfomance on the Tokyo 24/7 dataset.

4.2. Results

We provide the results for embedding networks with the VGG-16 as well as ResNet-101 backbone. GAN is

trained on the *SfM* dataset (using unpaired day and night images), the embedding network is fine-tuned also on the *SfM* dataset (using matching day images pairs), and the final retrieval performance is evaluated on *Tokyo*, $\mathcal{R}$Oxf and $\mathcal{R}$Par datasets. Our method is trained on the GAN-augmented *SfM* dataset, contains CLAHE normalization step, diverse anchor mining, and night examples are generated from day anchors with probability of 25%. The full version of tables can be found in the Supplementary Material.

We compare with the baselines GeM [30], CLAHE [13], CLAHE N/D [13] (day-night training pairs used) and CI-Conv [16]. The baseline methods in Table 1 are referred to by their original name and their reference. For GeM [30], the results of the publicly available github pytorch models[4] are reported. Methods proposed in this work are referred to as a combination of the retrieval training data (CycleGAN, HED$^{\text{N}}$GAN), whether CLAHE photometric normalization [13] was used (marked $\mathcal{C}$), and whether diversity mining (contribution of this paper) was used (marked $\mathcal{D}$). For comparison, results by recent state-of-the-art (day time) retrieval methods are shown in Table 2.

We trained all methods for 40 epochs, starting from ImageNet-pretrained backbones. This differs from [13], where the ImageNet-pretrained embedding network was trained for 20 epochs – pre-fine-tuned for 10 epochs and then fine-tuned for 10 epochs in the final configuration[5]. Otherwise the setup from [13] was followed precisely for all methods trained on *SfM* and *SfM-N/D* datasets (methods starting with GeM or CLAHE and GeM N/D **!** or CLAHE N/D $\mathcal{C}$ **!** respectively).

Beyond global descriptors. The proposed method of generating night data was also applied to HoW [41], a retrieval method based on aggregated local features. The results are summerized in the bottom of Table 1, showing improvement over the baseline and similar or better results compared to a version of the HOW network trained on *SfM-N/D*.

Beyond image retrieval. Visual localization is exploited to measure the night-time performance on a related task. We utilize the Kapture pipeline [11], in particular its benchmark for image retrieval in the context of visual localization. The percentage of images localized within each of the three thresholds for three paradigms is reported, following visuallocalization.net benchmark [33]. Three localization paradigms from Kapture pipeline [11] are followed - Paradigm 1: pose approximation by returning the pose of the top-1 retrieved image. Paradigm 2a & 2b: pose estimation from top 20 retrieved images without (2a) and with (2b) a global map respectively. Three methods from Table 1 were evaluated: baseline GeM [30], and our CycleGAN and HED$^{\text{N}}$GAN models. The results in Table 3 show a consistent performance improvement over the baseline across all setups. It should be noted that the results are not comparable to SotA methods trained for visual localization.

[4] https://github.com/filipradenovic/cnnimageretrieval-pytorch

[5] Please note that this difference, apart from providing a fair comparison, has preserved or slightly increased the performance of the re-trained baseline methods.

VGG-16 backbone

Method	top-1	w/o global map	with global map
GeM [30]	0/0/16.2	59.2/73.8/87.4	62.3/76.4/91.1
CycleGAN $\mathcal{CD}$	0/0/**18.8**	61.3/78.0/**90.6**	62.8/78.5/92.1
HED$^{\text{N}}$GAN $\mathcal{CD}$	0/0/18.3	**62.3**/**79.1**/90.1	62.8/78.5/92.1

ResNet-101 backbone

Method	top-1	w/o global map	with global map
GeM [30]	0/0.5/16.8	60.7/75.4/88.0	62.8/79.1/90.6
CycleGAN $\mathcal{CD}$	0/0.5/**20.4**	63.4/77.0/92.1	**66.0**/80.6/**95.8**
HED$^{\text{N}}$GAN $\mathcal{CD}$	0/0.5/19.4	**64.9**/**79.6**/**94.2**	65.4/80.6/94.2

Aachen v1.1 dataset - night

VGG-16 backbone

Method	top-1	w/o global map	with global map
GeM [30]	0/0.6/7.1	5.9/11.4/16.1	8.2/14.6/20.3
CycleGAN $\mathcal{CD}$	0.1/1.3/**17.4**	**12.2**/21.9/29.2	**14.5**/26.1/35.6
HED$^{\text{N}}$GAN $\mathcal{CD}$	0.1/1.3/16.3	12.1/**22.7**/**31.1**	14.0/**27.1**/**37.8**

ResNet-101 backbone

Method	top-1	w/o global map	with global map
GeM [30]	0.1/0.9/10.1	7.6/14.6/20.3	10.3/18.1/25.2
CycleGAN $\mathcal{CD}$	**0.5**/**2.2**/**22.6**	**14.1**/**26.2**/**36.0**	**16.3**/29.2/**40.9**
HED$^{\text{N}}$GAN $\mathcal{CD}$	0.3/1.7/21.4	13.4/24.7/34.3	15.9/**29.7**/40.0

RobotCar Seasons dataset - all night

Table 3. Visual localization evaluation on *Aachen* and *RobotCar* datasets – only results for night conditions are reported. Scores in each table cell correspond to the percentage of localized query images within their respective accuracy thresholds: (0.25m, 2°) / (0.5m, 5°) / (5m, 10°). The columns correspond to three paradigms from Kapture pipeline [11]: pose approximation from top-1 retrieved image, and pose estimation from top 20 images without global map and with global map. The best score for each dataset and backbone architecture (in separate tables) is emphasized by by bold.

Night edge detection. The qualitative comparison of RCF, HED, and HED$^{\text{N}}$ detectors is shown in Figure 2. To provide quantitative evaluation, we train EdgeMAC [29] edge embedding with HED, HED$^{\text{N}}$, and RCF$^{\text{N}}$detectors, increasing the image size to 362 and not binarizing the edgemaps to improve performance. The results are summarized in Table 4. The embeddings based on the proposed HED$^{\text{N}}$ outperform those based on both RCF$^{\text{N}}$and HED when used alone or in an ensemble with an image embedding (*e.g.* GeM). Similarly to [13], ensembles of edge and image embeddings obtain better performance at the cost of double dimensionality. The best performance is obtained for an en-

Method	Avg	Tokyo	$\mathcal{R}$Oxf	$\mathcal{R}$Par
EdgeMAC [29]	45.6	75.9	17.3	43.5
HEDMAC	56.8	79.5	38.3	52.5
HED$^{\text{N}}$MAC	**59.2**	81.9	**38.4**	**57.2**
RCF$^{\text{N}}$MAC	58.5	**88.9**	35.1	51.4
HEDMAC+GeM ‡	72.0	84.8	60.9	70.3
HED$^{\text{N}}$MAC+GeM ‡	72.6	85.7	**61.1**	70.9
HED$^{\text{N}}$MAC+$^{\text{N}}$GAN ‡	74.4	91.4	60.6	71.3
HED$^{\text{N}}$MAC+GAN ‡	**74.7**	**91.8**	60.4	**71.9**

Table 4. The effect of our trained HED$^{\text{N}}$ detector (from HED$^{\text{N}}$GAN) on the EdgeMAC [29] method. HEDMAC and HED$^{\text{N}}$MAC is a variant of EdgeMAC method with the HED [45] edge detector with either original or our weights, respectively. RCF$^{\text{N}}$MAC is a variant of EdgeMAC with the RCF [19] edge detector with our weights. In the bottom block, ensembles of EdgeMAC variants with chosen methods from Table 1 are reported. GeM is from [30], $^{\text{N}}$GAN corresponds to HED$^{\text{N}}$GAN $\mathcal{CD}$, and GAN to CycleGAN $\mathcal{CD}$, all from Table 1. Ensembles have double the dimensionality (1024) and are marked with ‡. The best score for each dataset in each block is in bold.

semble of our HED$^{\text{N}}$ edge embedding and HED$^{\text{N}}$GAN $\mathcal{CD}$ or CycleGAN $\mathcal{CD}$ image embedding.

Discussion. Results in Table 1 show that using GAN training data is superior even to using the paired day-night images from the *SfM-N/D* dataset (CLAHE N/D $\mathcal{C}$ [13]). For CIConv [16] with the ResNet-101 backbone, we have used the publicly available model trained by the authors and evaluated it on Oxford and Paris benchmarks.[6] We observe drop in both these benchmarks compared to GeM [28]. Our method outperforms [16] on both backbone architectures; in fact, our VGG-16 model "HED$^{\text{N}}$GAN $\mathcal{CD}$" outperforms CIConv [16] ResNet-101 model on the *Tokyo* dataset, despite the weaker backbone architecture.

Further, an alternative approach [1] of using the generator in inference rather than during training was evaluated. On *Tokyo* dataset, all images with the night class label (*i.e.* using oracle, as this label is not exploited in any other method) were translated to the day domain prior to retrieval. Two models, the one provided by [1] and our generator, were tested. However, in both cases the performance was substantially worse than the retrieval baseline, specifically 39 and 52 mAP for [1] and our generator respectively.

4.3. Diverse anchors

The effect of diverse anchor mining ($\mathcal{D}$) is ablated in Table 5. Results show consistent improvement of diverse anchor mining across both baselines and our method on day-night retrieval (*Tokyo* dataset), while having little impact on Oxford and Paris datasets. In Table 6, the results

[6]For VGG-16 backbone, there is no publicly available model, therefore only results on *Tokyo* dataset published in [16] are reported

Method	Avg	Tokyo	$\mathcal{R}$Oxf	$\mathcal{R}$Par
CLAHE $\mathcal{C}$	71.9	85.4	60.0	70.1
CLAHE $\mathcal{CD}$	72.2	85.9	60.3	70.5
CLAHE N/D $\mathcal{C}$!	72.5	87.5	59.9	70.1
CLAHE N/D $\mathcal{CD}$!	**73.0**	**87.7**	**60.8**	**70.7**
HED$^{\text{N}}$GAN	72.7	88.0	60.2	70.0
HED$^{\text{N}}$GAN $\mathcal{C}$	73.2	88.7	60.5	70.4
HED$^{\text{N}}$GAN $\mathcal{CD}$	**73.4**	**88.8**	**60.7**	**70.6**

Table 5. The effect of diverse anchors ($\mathcal{D}$). Methods CLAHE $\mathcal{C}$ [13] and CLAHE N/D $\mathcal{C}$ [13] from Table 1 are reported in the top block. Please note that we re-train the models for this ablation, so we obtain a slightly higher performance. In the bottom block, the effect of CLAHE ($\mathcal{C}$) and diverse anchors ($\mathcal{D}$) is reported on the proposed method HED$^{\text{N}}$GAN. The best score for each dataset in each block is in bold.

Method	{day, sunset}		{sunset, night}		{day, night}	
	D→S	S→D	S→N	N→S	D→N	N→D
CLAHE $\mathcal{C}$	97.7	98.2	80.1	81.3	70.9	76.1
CLAHE N/D $\mathcal{C}$!	97.5	98.2	80.3	86.2	73.0	81.3
HED$^{\text{N}}$GAN $\mathcal{C}$	97.1	98.2	84.5	86.9	77.1	80.3
HED$^{\text{N}}$GAN $\mathcal{CD}$	97.5	98.0	84.3	88.3	77.9	81.1

Table 6. Retrieval performance (mAP) on *Tokyo* for a combination of three different subsets of the dataset – day (D), sunset (S), and night (N). Images from the first class are always queries and from the second class are positives (query→positive); the last image of the scene from the unused class is excluded from the evaluation. Scores for selected methods from Tab. 5 are reported.

are broken down into combinations of query and result domains (for example, column D→S means querying with a Day image where the Sunset image of the same scene is the only positive). Experiments further show that the proposed HED$^{\text{N}}$GAN $\mathcal{CD}$ method substantially improves the performance when querying or retrieving night images. We interpret this observation as the embedding learned by the proposed method being better at discriminating night images. A similar trend with a smaller improvement can be seen when training with paired day-night data or when diverse anchors are used. We also observe consistent gains when the image photometric normalization CLAHE, proposed by [13], is used.

4.4. Impact of training data

In this part, we study how sensitive the retrieval performance is to the choice of the night image generator and the amount of night images used to train the embedding network.

Night image generator. Different night image generators were compared: CycleGAN, DRIT [15], CUT [24], CyEDA [3], and a proposed edge-consistency RCF$^{\text{N}}$GAN, HEDGAN, and HED$^{\text{N}}$GAN. In our experiments, the best-

performing generator (Top GAN) is CycleGAN, see Table 7. Despite different levels of a photo-realistic perception, ranging from DRIT to rather abstract HED^{N}GAN, the difference in performance is not so severe, all generators outperforming training with real night images *SfM-N/D*. Despite similar scores, the generator training times differ greatly, as illustrated in Table 8. The lightest model to train is HEDGAN, followed by HED^{N}GAN which both train a degree of magnitude faster than CycleGAN and DRIT.

To some observers, the translated night images (see Figure 1) may appear as intensity inverted images. Using images with inverted intensity channel in the LAB colour space improves the results (avg 71.2) over the baseline (avg 70.0), however, such a simple colour augmentation is far below the results achieved by the proposed trained generators or training on the *SfM-N/D* dataset.

Night data amount. We tested the sensitivity of using a different ratio of day and night training data in the embedding learning. We observe that using 25% or 50% of night images for learning the embedding does not make a significant difference, with a minor increase of scores on the *Tokyo* dataset. In all experiments, we report results for the variant with 25% of night images in order to stay consistent with the prior work of [13]. We also observe that adding true night images from *SfM-N/D* (GAN + *SfM-N/D*) has a minimal impact on the retrieval results.

The experiments show that photo-realistic appearance of the synthetic training images is *not* important for retrieval performance, and that using more powerful generator would not improve the performance, since even adding true night data does not. However, (local) similarity to real night images encouraged by the discriminator is important (which is not the case for a simple intensity inversion augmentation).

5. Conclusions

The training of a deep neural network that embeds images into a descriptor space suitable for image retrieval insensitive to severe day-night illumination changes was proposed. Synthetically generated night images are used so that the training does not require corresponding pairs of night and day images. The proposed method outperforms prior work, including the work of [13] which uses ground-truth day-night annotated image pairs.

We have shown that the proposed method is capable of generating diverse training examples, and that a larger diversity of synthesized training data outperforms smaller diversity of real training data.

Besides evaluating a number of existing generators, a light-weight generator exploiting edge consistency was proposed. In our HED^{N}GAN method, day/night edge detector HED^{N} is trained. Its superior performance to HED detector was shown both qualitatively and quantitatively.

Night Data	Avg	Tokyo	$\mathcal{R}$Oxf	$\mathcal{R}$Par
DRIT $\mathcal{CD}$	73.5	**90.2**	59.8	70.5
CUT $\mathcal{CD}$	73.0	87.7	60.3	**71.1**
CyEDA [3] BDD $\mathcal{CD}$	72.9	87.9	60.4	70.3
CyEDA [3] GTA $\mathcal{CD}$	72.9	87.9	60.3	70.4
CyEDA $\mathcal{CD}$	70.9	82.0	60.1	70.5
RCF^{N}GAN $\mathcal{CD}$	73.2	88.3	60.4	70.8
HEDGAN $\mathcal{CD}$	73.2	88.1	61.0	70.5
HED^{N}GAN $\mathcal{CD}$	73.4	88.9	**61.1**	70.3
C-GAN $\mathcal{CD}$	**74.0**	**90.2**	60.7	71.0
C-GAN+N/D $\mathcal{CD}$ **!**	73.5	88.6	60.8	**71.1**
C-GAN+N/D 50% $\mathcal{CD}$ **!**	**73.9**	**90.1**	**61.1**	70.6

Table 7. The impact of retrieval training data. In the top block, generator architectures DRIT [15], CUT [24], CyEDA [3] (pretrained models from [3] and trained by us on *SfM*), RCF^{N}GAN (trained RCF), HEDGAN (frozen HED), HED^{N}GAN (trained HED), and CycleGAN [49] are tested. In the bottom block, the best performing CycleGAN generator architecture is further combined with the *SfM-N/D* dataset with ratio 1:1 (CycleGAN+N/D); scores for 25% (default in experiments) and 50% of night images in the training data are reported. Methods marked by **!** use paired day-night training data. The best score for each dataset in each block is in bold.

Training	Train (h)	Epoch (h)	Epochs	Params
DRIT	196	0:39	300	75.5
CycleGAN	102	1:01	100	51.0
CUT	35	0:42	50	26.0
CyEDA	28	0:24	69	39.6
RCF^{N}GAN	27	0:32	50	40.3
HED^{N}GAN	21	0:25	50	40.2
HEDGAN	16	0:19	50	25.5

Table 8. Generators training comparison. In the first two columns, training times in hours are reported (measured on NVIDIA Tesla P100 16GB). In the pre-last column, the total number of epochs necessary to converge to the top performance is reported. In the last column, a number of trainable parameters in millions is illustrated.

Finally, we have introduced a method of mining diverse anchor images, that further improves the diversity in the training data which is reflected in increased retrieval performance. Such an approach is applicable to other metric-learning and similar tasks where strong modes of the training-data distribution do not correspond to the distribution of test data.

6. Acknowledgements

This research was supported by the Research Center for Informatics (project CZ.02.1.01/0.0/0.0/16 019/0000765 funded by OP VVV) and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS23/173/OHK3/3T/13.

References

[1] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *ICRA*, 2019. 4, 8

[2] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *IJCNN*, 2019. 3

[3] Jing Chong Beh, Kam Woh Ng, Jie Long Kew, Che-Tsung Lin, Chee Seng Chan, Shang-Hong Lai, and Christopher Zach. Cyeda: Cycle-object edge consistency domain adaptation. In *ICIP*. IEEE, 2022. 2, 3, 5, 8, 9

[4] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *CVPR*, 2020. 3

[5] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*. Springer, 2020. 2

[6] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1841–1848, 2013. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[11] Martin Humenberger, Yohann Cabon, Noé Pion, Philippe Weinzaepfel, Donghwan Lee, Nicolas Guérin, Torsten Sattler, and Gabriela Csurka. Investigating the role of image retrieval for visual localization: An exhaustive benchmark. *IJCV*, 130(7):1811–1836, 2022. 7

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3, 5

[13] Tomas Jenicek and Ondrej Chum. No fear of the dark: Image retrieval under varying illumination conditions. In *ICCV*, 2019. 2, 3, 4, 5, 6, 7, 8, 9

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5

[15] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *IJCV*, 2020. 2, 3, 5, 8, 9

[16] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *ICCV*, 2021. 2, 3, 6, 7, 8

[17] Che-Tsung Lin, Sheng-Wei Huang, Yen-Yi Wu, and Shang-Hong Lai. Gan-based day-to-night image style transfer for nighttime vehicle detection. *IEEE T-ITS*, 2020. 3

[18] Che-Tsung Lin, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai. Multimodal structure-consistent image-to-image translation. In *AAAI*, 2020. 3

[19] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Jia-Wang Bian, Le Zhang, Xiang Bai, and Jinhui Tang. Richer convolutional features for edge detection. *IEEE TPAMI*, 2019. 1, 2, 3, 5, 8

[20] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000km: The oxford robotcar dataset. *The International Journal of Robotics Research (IJRR)*, 2017. 6

[21] Markus S Mueller, Torsten Sattler, Marc Pollefeys, and Boris Jutzi. Image-to-image translation for enhanced feature matching, image retrieval and visual localization. *ISPRS*, 2019. 3

[22] Simon Niklaus. A reimplementation of HED using PyTorch. https://github.com/sniklaus/pytorch-hed, 2018. 5

[23] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. 2

[24] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*. Springer, 2020. 2, 3, 5, 8, 9

[25] Lam Phan, Hiep Thi Hong Nguyen, Harikrishna Warrier, and Yogesh Gupta. Patch embedding as local features: Unifying deep local and global features via vision transformer for image retrieval. In *ACCV*, 2022. 2, 6

[26] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 2, 6

[27] Filip Radenovic, Johannes L Schonberger, Dinghuang Ji, Jan-Michael Frahm, Ondrej Chum, and Jiri Matas. From dusk till dawn: Modeling in the dark. In *CVPR*, 2016. 2, 6

[28] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016. 2, 3, 4, 5, 8

[29] Filip Radenovic, Giorgos Tolias, and Ondřej Chum. Deep shape matching. In *ECCV*, 2018. 3, 7, 8

[30] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE TPAMI*, 2018. 2, 3, 5, 6, 7, 8

[31] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *ICCV*, 2017. 5

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[33] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *CVPR*, 2018. 1, 2, 6, 7

[34] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *BMVC*, 2012. 1, 6

[35] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S. Huang. Towards instance-level image-to-image translation. In *CVPR*, 2019. 3

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 5

[37] Chull Hwan Song, Jooyoung Yoon, Shunghyun Choi, and Yannis Avrithis. Boosting vision transformers for image retrieval. In *WACV*, 2023. 3

[38] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*. International Society for Optics and Photonics, 2019. 4

[39] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer London, 2010. 3, 5

[40] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *ICCV*, 2013. 2

[41] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *ECCV*. Springer, 2020. 2, 3, 5, 6, 7

[42] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 6

[43] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning Super-Features for Image Retrieval. In *ICLR*, 2022. 2

[44] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 6

[45] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 1, 2, 3, 4, 5, 8

[46] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *ICCV*, 2021. 2, 6

[47] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 5

[48] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *IJCV*, 2021. 6

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3, 5, 9