2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Self-Supervised Learning of Domain-Invariant Local Features for Robust Visual Localization under Challenging Conditions

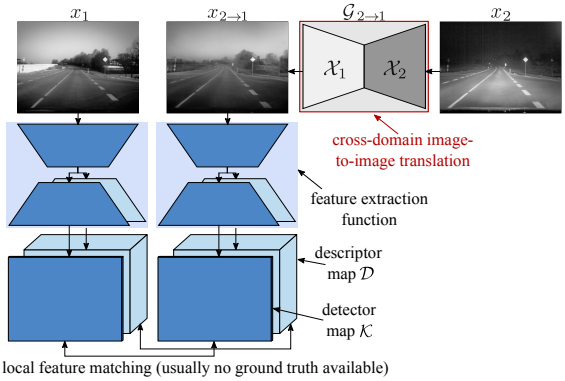
Moritz Venator[1,2], Yassine El Himer[3], Selcuk Aklanoglu[1], Erich Bruns[1], Andreas Maier[2]

***Abstract*— Visual localization provides the basis for many robotics applications such as autonomous navigation or augmented reality. Especially in outdoor scenes, robust localization requires local features which can be reliably extracted and matched under changing conditions. Previous approaches have applied generative image-to-image translation models to align images in a single domain before correspondence search. In this paper, we invert the concept and elaborate why it is more promising to use image domain adaptation for training of robust local features. Integrating this idea into a self-supervised training framework, we show in various experiments covering image matching, visual localization, and scene reconstruction that our Domain-Invariant SuperPoint (DISP) outperforms existing self-supervised methods in terms of repeatability, generalization, and robustness. In contrast to competitive supervised local features, our modular and fully self-supervised approach can be easily adapted to different domains and localization tasks as it does not require ground truth correspondences for training.**

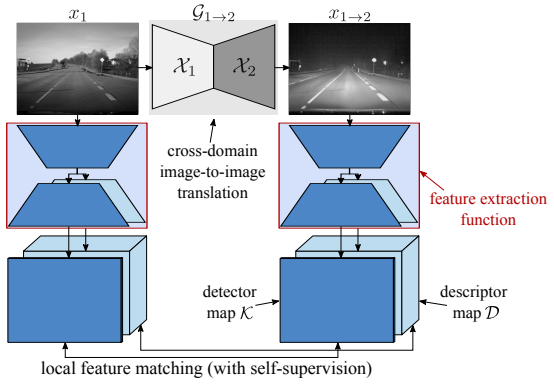
I. Introduction

Visual localization and reconstruction tasks such as Simultaneous Localization and Mapping (SLAM) [1], [2], Structure from Motion (SfM) [3], [4], [5], and long-term visual localization [6], [7] attempt to build a geometric model of an unknown environment and to localize agents such as smartphone cameras, robots, or self-driving cars inside this map. Being based on multi-view geometry [8], the successful application of those methods requires a reliable detection of local features that mark characteristic keypoints and can be robustly matched between images captured under varying viewing conditions [9], [10], [11], [7]. Especially in outdoor scenes, the appearance of keypoints can heavily change due to seasons, daytime, or weather, posing a challenge to correspondence search [9], [10], [11], [7], [12].

End-to-end trainable pipelines which simultaneously detect and describe keypoints [14], [15] have recently outperformed hand-crafted local features in this task [16], [17], [18]. However, they usually require ground truth information, i.e., pre-determined feature correspondences, for supervised training, which is hard to acquire for new domains and challenging environments. SuperPoint [19] constitutes an exception by using self-supervised learning based on pseudo-labeling and homographic adaptation. However, training does not account for feature repeatability under changing viewing conditions such as weather or daytime, which makes it inferior to the aforementioned supervised concepts.



(a) Image-to-image translation used as preprocessing for reaching domain invariance in image space before inference of feature extraction [10], [13].



(b) Our concept: self-supervised learning of domain-invariant local feature representations using domain adaptation during training.

Fig. 1. Comparison of concepts employing image-to-image translation for cross-domain local feature matching. The red rectangles indicate which networks are optimized for reaching domain invariance. Detailed explanations can be found in Section III.

Generative Adversarial Networks (GANs) [20] have shown impressive results in translating content taken from one image into a different representation or condition, so-called domains [21], [22]. Previous approaches combining domain adaptation with localization tasks have attempted to recover features in the pixel space by translating images captured under different viewing conditions into similar ones before extracting local or global features [11], [13]. However, this requires the GAN to reconstruct images in such a realistic way that features can be extracted and matched with real images of the target domain, which is especially challenging for non-urban environments with far-distance features and poor viewing conditions, e.g., night

Corresponding author: Moritz Venator (e-mail: moritz.venator@fau.de).
[1]Moritz Venator, Selcuk Aklanoglu, and Erich Bruns are with Volkswagen Car.SW Org., Germany.
[2]Moritz Venator and Andreas Maier are with the Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nürnberg, Germany.
[3]Yassine El Himer is with Technical University of Munich, Germany.

images [13]. It further implies that the model has already intrinsically learned meaningful feature representations as shown by recent work [23], [24], which employs adversarial losses for self-supervised learning of domain-invariant content representations for visual place recognition. However, these concepts do not enable learning of local feature representations as needed for precise six degrees of freedom (6-DoF) camera pose estimation [3], [7].

In this work, we invert the idea of [13], [11]: Instead of using image-to-image translation at inference, we integrate it as an augmentation step into training of a feature extraction function (see Fig. 1). Extending the self-supervised learning concept of SuperPoint [19], we explicitly train the network to re-identify keypoints across different domains without requiring supervision. We evaluate the resulting Domain-Invariant SuperPoint (DISP) with different multi-domain datasets for outdoor scenes. Experiments targeting feature matching [25], visual localization [7], and collaborative road scene reconstruction [5], [12] show that the method can enhance repeatability, generalization, and distinctiveness of local features. Our approach demonstrates that using image-to-image translation for learning of domain-invariant representations in a convolutional feature space is more effective than attempting to reach domain invariance in the image space through preprocessing.

Our contributions are summarized as follows:

- Combining deep neural networks for feature extraction and image-to-image translation, we introduce a novel concept for self-supervised training of domain-invariant local features. Our modular approach can be easily finetuned for various datasets and domains without the need for ground truth correspondences.
- We theoretically motivate and practically demonstrate why our approach is superior to existing concepts using image-to-image translation as a preprocessing method for pixel domain alignment before feature extraction.
- Extensive experiments targeting keypoint matching, visual localization, and SfM show that the new training concept is able to boost feature matching between varying viewing conditions significantly.

II. Related Work

A. Local Feature Extraction

Local feature extraction aims to detect characteristic, distinctive keypoints in an image and to describe them by vectorial representations so that they can be recognized and matched between multiple images [8]. The quality of local features is usually evaluated based on their invariance under illumination and geometric changes [5], [25]. Despite the success of deep learning in many computer vision disciplines, hand-crafted features [16], [17], [18], [26] have remained competitive in certain tasks such as SLAM or SfM [2], [27].

On the other hand, learned local features have demonstrated superior repeatability under heavily changing viewing conditions as needed for long-term visual localization [19], [28]. Learned representations substitute detector [29], descriptor [30], [31], or whole extraction pipelines [32], [19], [14]. Usually, training of these methods is formulated as a supervised learning problem which requires pixel correspondences as ground truth [32], [30], [31], [14]. Using homographic adaptations to increase the viewpoint invariance, SuperPoint [19] proposes a self-supervised concept for learning local features. Our work goes beyond this approach by extending self-supervision to domain adaptation [33].

B. Structure from Motion and Visual Localization

Various problems in computer vision such as SfM [3], [4], [5] and visual localization tasks [9], [7] rely on finding 2D-2D correspondences between a query image x_i and a set of other images $S_N = \{x_j \mid j = 1..N\}$. Correspondence search in unordered collections often implies exhaustive matching requiring a high distinctiveness of local features to ensure completeness and accuracy of the resulting model [27]. A strategy often applied in place recognition [9], [6], 6-DoF localization [7], [34], [28], and SfM [5] is to reduce S_N to a smaller set S_{N_r} such that $N_r \ll N$, e.g., through k-nearest neighbor search with global descriptors [35], [23].

Local feature matching becomes especially challenging and ambiguous under strong appearance changes due to daytime, weather, or seasons [9], [6]. Therefore, Sattler et al. [7] have introduced benchmark datasets for long-term visual localization covering a wide set of challenging conditions. We will apply both exhaustive matching for SfM with COLMAP [5] and the hierarchical localization paradigm proposed by Sarlin et al. [28] to evaluate the performance of our learned local features in various settings.

C. Adversarial Training for Domain Adaptation

GANs [20] have shown impressive results in generative tasks such as image generation [36], [37] and image-to-image translation [22], [21], [38]. Cross-domain image-to-image translation translates an input image from a source domain to a target domain, where a domain may denote a class of images captured under certain viewing conditions [21] or other representations [22]. Models can either be trained with paired data employing the concept of conditional GANs [39], [22] or in an unsupervised fashion by enforcing cycle-consistency [21] and disentangled representations of image content and style [40], [38]. Recent work [23], [24] has proposed self-supervised concepts that learn to disentangle place and appearance features through adversarial losses. The domain-invariant features extracted by the content encoder can be used for place recognition, but do not allow local feature matching as needed for precise 6-DoF visual localization since there is no loss structuring the content feature map so that features are assigned to certain pixel locations.

Other approaches employ image-to-image translation to align the domains on a pixel level before conducting feature-based localization [11], [13], [41]. In their semi-supervised approach, Porav et al. [11] use an additional feature descriptor loss in order to recover local features lost due to changes in viewing conditions. Introducing a discriminator architecture which operates separately on different image transformations, ToDayGAN [13] enhances translations from

night to day images and subsequent place recognition. In contrast to these methods, we employ image-to-image translation as an augmentation technique for self-supervised learning of more robust local feature representations.

III. Domain-Invariant SuperPoint (DISP)

Local feature extraction comprises two tasks: keypoint detection and description. The detector function computes a map $\mathcal{K} \in \mathbb{R}^{H \times W}$ indicating keypoint locations in an image $x \in \mathbb{R}^{H \times W}$. In order to match keypoints found in different images by computing similarity metrics, the descriptor function outputs a tensor $\mathcal{D} \in \mathbb{R}^{H \times W \times D}$ with a fixed-length vector describing each pixel. In the following, we encapsulate the local feature representation, which is extracted by the function $f_p(x)$, in $\mathcal{P} = \{\mathcal{K}, \mathcal{D}\}$.

A. Image-to-Image Translation for Domain-Invariant Feature Matching

Let $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ be images from two different domains, e.g., images captured under different viewing conditions. Image-to-image translation models try to approximate the conditional $p(x_2|x_1)$ by a function $\mathcal{G}_{1\to2}$ that models $p(x_{1\to2}|x_1)$. An assumption usually made in multimodal frameworks [38], [42] is that each image can be separated into a domain-invariant content c, which contains the geometric and semantic structure of the image scene, and a style s, which encodes illumination and texture and can be domain-specific.

If the scene content of the two images is the same, an ideal domain-invariant feature extraction function f_p^* should extract the same keypoints for both images:

$$\mathcal{P}_1 = f_p^*(x_1) = f_p^*(x_2) = \mathcal{P}_2 \,. \tag{1}$$

In multi-view geometry problems, the images x_1 and x_2 capture the same scene from different viewpoints. Still, an ideal domain-invariant feature extractor should find corresponding keypoints in both images and encode them in similar descriptors to minimize the matching loss $\mathcal{L}(\mathcal{P}_1, \mathcal{P}_2)$:

$$\mathcal{L}(f_p^*(x_1), f_p^*(x_2)) = \min_{\mathcal{P}_1, \mathcal{P}_2} \mathcal{L}(\mathcal{P}_1, \mathcal{P}_2) \,. \tag{2}$$

However, real local features are never perfectly domain-invariant. In fact, they are often severely affected by changes in viewing conditions [27], [7]. Previous approaches [11], [13] have therefore used image-to-image translation models to translate images into a common domain, e.g., $x_{2\to1} = \mathcal{G}_{2\to1}(x_2)$, and perform feature extraction afterwards. The domain with better viewing conditions is usually chosen as the target domain since more keypoints can be found. Unfortunately, image-to-image translation models often perform insufficiently in this direction of translation as they need to reconstruct objects which are barely visible in the input image. During GAN training, the discriminator forces the generator to fake content, which is rather harmful in terms of feature matching because it provokes outliers.

For optimization of cross-domain matching, we need to minimize the feature matching loss:

$$\min_{\mathcal{G}_{2\to1}} \mathcal{L}(\mathcal{P}_1, \mathcal{P}_{2\to1}) = \min_{\mathcal{G}_{2\to1}} \mathcal{L}(f_p(x_1), f_p(\mathcal{G}_{2\to1}(x_2))) \,. \tag{3}$$

Since the either hand-crafted or pre-trained feature extraction function f_p is fixed, $\mathcal{G}_{2\to1}$ needs to minimize the differences in viewing conditions in the image space between x_1 and $x_{2\to1}$. Due to the aforementioned problems in image-to-image translation models, this requirement can only be fulfilled in very constrained settings, e.g., if paired data for supervised training is available as in [11], limiting the applicability of the approach.

Instead, we invert the idea and propose to integrate image domain adaptation into a self-supervised learning process of a feature extraction function f_θ. Following the definition of a domain-invariant function f_p^* in (1), we translate an image into a different domain and train the network f_θ to find corresponding keypoints afterwards. Ideally, we get:

$$\mathcal{P}_1 = f_p^*(x_1) = f_p^*(\mathcal{G}_{1\to2}(x_1)) = f_p^*(x_{1\to2}) = \mathcal{P}_{1\to2} \,. \tag{4}$$

During training, we optimize the feature detector and descriptor instead of the domain translation model to minimize the matching loss:

$$\min_{f_\theta} \mathcal{L}(\mathcal{P}_1, \mathcal{P}_{1\to2}) = \min_{f_\theta} \mathcal{L}(f_\theta(x_1), f_\theta(\mathcal{G}_{1\to2}(x_1))) \,, \tag{5}$$

obtaining an improved feature extraction function $f_\theta \approx f_p^*$. This approach has a number of advantages:

- It explicitly trains feature extraction for matching images from different domains. Thus, the network automatically learns which keypoints are reliable and can be matched under changing conditions.
- Cross-domain matching is performed in deep convolutional layers instead of the image space and thus is able to incorporate more robust high-level features such as object and scene semantics.
- The modular solution design can incorporate different image-to-image translation models to make feature extraction more robust in the respective set of conditions.

In the following sections, we describe in detail how we integrate domain adaptation into a self-supervised training process for feature extraction. Extending [19], we call our model *Domain-Invariant SuperPoint (DISP)*.

B. Domain Adaptation

In SuperPoint [19], the authors employ homographic adaptation to make feature extraction more robust to viewpoint changes. They transform images with randomly sampled homographies which approximate how the appearance of keypoints changes when they are observed from different viewpoints. We add domain adaptation to this concept.

Let f_k be a given local feature detector extracting keypoint locations $\mathbf{x} = f_k(x)$ from an image x. Ideally, f_k is supposed to be covariant to the homography $\mathcal{H}$ so that the resulting keypoint locations follow the image transformation:

$$\mathcal{H}\mathbf{x} = f_k(\mathcal{H}(x)) \,. \tag{6}$$

Image-to-image translation functions model how the appearance of a scene changes under varying viewing conditions. Using an unsupervised model based on CycleGAN [21] and MUNIT [38], our approach can still be trained

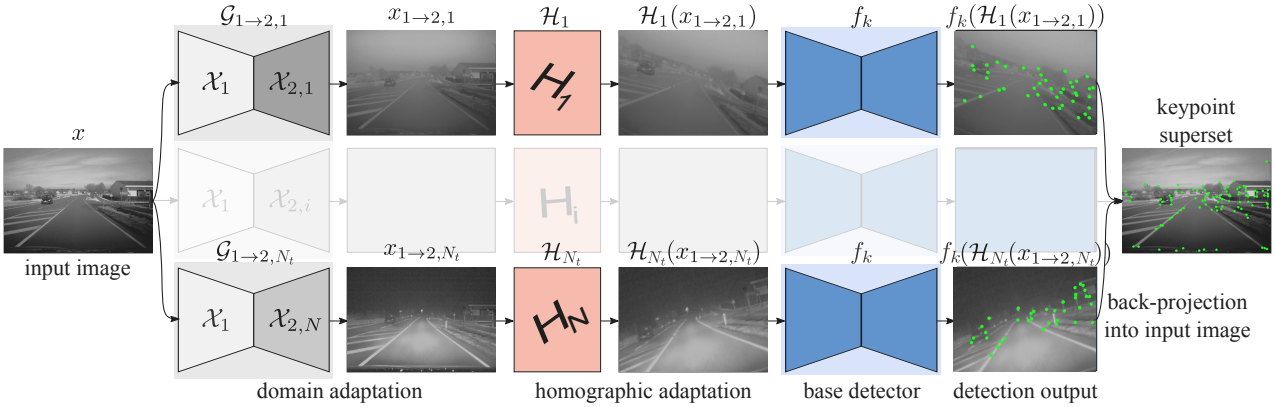


Fig. 2. Self-supervised labeling of keypoint superset with domain and homographic adaptation which is used for training step 3 (see Section III-C) and extends the SuperPoint concept [19]. For each of the N image copies, we sample a target domain (which can also be the source domain) and style.

with unlabeled data. The ideal extraction function shall be invariant to domain changes, i.e., the keypoint locations are not affected by domain adaptation:

$$\mathbf{x} = f_k(\mathcal{G}(x))\,, \tag{7}$$

where $\mathcal{G}$ represents an image-to-image translation model:

$$\mathcal{G}(x) = \mathcal{G}_{1\to2}(x) = G_2(c, s_2) = G_2(E_1^c(x), s_2)\,. \tag{8}$$

The decoder G_2 takes the content c of image $x \in \mathcal{X}_1$ and a random style s_2 sampled from $p(s_2) \sim \mathcal{N}(0, 1)$ to generate an image in the target domain $\mathcal{X}_2$. The content c is extracted by the encoder E_1^c of the source domain $\mathcal{X}_1$. Combining domain and homographic adaptation, we obtain:

$$\mathcal{H}\mathbf{x} = f_k(\mathcal{H}(\mathcal{G}(x)))\,, \tag{9}$$

so that keypoint reprojection into the input image results in:

$$\mathbf{x} = \mathcal{H}^{-1} f_k(\mathcal{H}(\mathcal{G}(x))) = \mathcal{H}^{-1} f_k(\mathcal{H}(G_2(E_1^c(x), s_2)))\,. \tag{10}$$

Since local features do usually not fulfill (6) and (7), output $\mathbf{x}$ differs between various transformations $\mathcal{H}$ and $\mathcal{G}$. By aggregating outputs of N_t different transformations as illustrated in Fig. 2, we can generate a self-labeled keypoint superset:

$$\hat{F}(x; f_k) = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathcal{H}_i^{-1} f_k(\mathcal{H}_i(G_{2,i}(E_1^c(x), s_{2,i})))\,. \tag{11}$$

C. Network Architectures

Our concept combines network architectures based on MUNIT [38] and SuperPoint [19], two state-of-the-art methods for multimodal image-to-image translation and local feature extraction, which we will briefly summarize here. More details on the exact implementation can be found in the respective papers. Due to our modular approach, the networks can be replaced by other architectures that can be trained in an unsupervised or self-supervised way.

Multimodal Image-to-Image Translation: The MUNIT [38] framework couples two generator-discriminator pairs as known from CycleGAN [21]. In domain $\mathcal{X}_1$, the generator consists of an encoder E_1 extracting content c_1 and style s_1 from an image x_1, and a decoder G_1 reconstructing the image from this latent representation. Combining the encoder and decoder of two domains, e.g., E_1 and G_2, allows to translate an image x_1 from one domain to the other $x_{1\to2}$.

Local Feature Representation: SuperPoint [19] uses a fully convolutional neural network with a single encoder and two decoder heads for local feature detection and description. The weights of the encoder, whose layer design is inspired by the VGG network [43], are shared by the two functions and thus learn features for both tasks simultaneously. The decoder heads are task-specific and consist of convolutional operations followed by explicit, non-learned upsampling layers which output a detection map $\mathcal{K} \in \mathbb{R}^{H\times W}$ and a corresponding descriptor map $\mathcal{D} \in \mathbb{R}^{H\times W\times D}$ (see Fig. 1).

D. Training of Domain-Invariant SuperPoint

Training of a DISP model comprises four steps in total. After the base detector and domain adaptation models have been trained in the first two steps, we can generate different variants of DISP by repeating steps 3–4 with various combinations of datasets and domain adaptations.

1) The base detector is trained on a *Synthetic Shapes* dataset as proposed in [19], where random images with geometric shapes are generated and labeled on-the-fly.
2) For each domain pair DISP shall be able to deal with, we need to train a multimodal image-to-image translation model $\mathcal{G}$ which we can employ for domain adaptation in the following steps.
3) The base detector is improved by re-training it on a dataset with real images. A keypoint superset created with randomly sampled transformations of domain adaptation $\mathcal{G}$ and homographic adaptation $\mathcal{H}$ (as described in Section III-B) serves as ground truth.
4) Finally, both the detector and descriptor part of the SuperPoint network are trained jointly (see Fig. 3). Here, we use two copies of the same input image with one of them being transformed by consecutive domain and homographic adaptation (with a more restrictive range of homographies). The output of the detector fine-tuned in step 3 is used for self-labeling of correspondences.

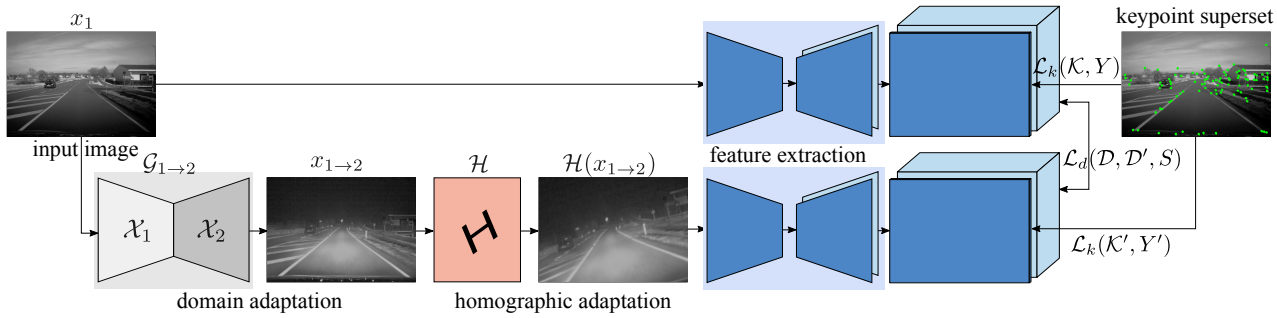


Fig. 3. Joint training of detector and descriptor. For each input image, we randomly sample a target domain (can also coincide with source domain) and style for image-to-image translation before applying homographic adaptation. Training aims to minimize detector $\mathcal{L}_k$ and descriptor $\mathcal{L}_d$ loss.

As in SuperPoint [19], steps 1 and 3 exclusively optimize the detector by minimizing the detection loss $\mathcal{L}_k(\mathcal{K}, Y)$ with $\mathcal{K}$ being the predicted local feature detection map and Y the keypoint superset. Step 4 combines a detection loss for each image with a descriptor loss $\mathcal{L}_d$ for the image pair:

$$\mathcal{L}(\mathcal{P}, \mathcal{P}') = \mathcal{L}_k(\mathcal{K}, Y) + \mathcal{L}_k(\mathcal{K}', Y') + \lambda \mathcal{L}_d(\mathcal{D}, \mathcal{D}', S)\,. \quad (12)$$

The losses are computed as follows:

$$\mathcal{L}_k(\mathcal{K}, Y) = \frac{1}{HW} \sum_{\substack{h=1 \\ w=1}}^{H,W} l_p(\mathbf{k}_{hw}, y_{hw})\,, \quad (13)$$

$$\mathcal{L}_d(\mathcal{D}, \mathcal{D}', S) = \frac{1}{H^2W^2} \sum_{\substack{h=1, \\ w=1}}^{H,W} \sum_{\substack{h'=1, \\ w'=1}}^{H,W} l_d(\mathbf{d}_{hw}, \mathbf{d}'_{h'w'}, s_{hwh'w'})\,, \quad (14)$$

where l_p is a cross-entropy loss, l_d a hinge loss, and $s_{hwh'w'}$ a binary function indicating the homographic correspondence between two descriptors $\mathbf{d}_{hw}$ and $\mathbf{d}'_{h'w'}$ as described in [19].

IV. Experimental Setup

A. Training

We train our models on three of the most diverse and prominent domains of viewing conditions in road scenes: sun, rain, and night images. In order to evaluate dataset bias, we compare models trained with two different datasets:

The **Oxford RobotCar Dataset (OX)** [10] was recorded during repetitive test drives on a route through the city of Oxford over a time period of one year. Among others, it includes video streams of multiple cameras. We only use RGB front camera images from sequences not included in the RobotCar Seasons benchmark (only rear camera images) to avoid any overlap between training and evaluation data.

Our own **Non-Urban Dataset (NU)** consists of grayscale images that were captured during repetitive test drives on a single test route, but cover a wide range of non-urban road scenes (highways, rural roads, villages). These environments are especially challenging for cross-domain feature matching, but are usually underrepresented in academic datasets.

For DISP, we first train two image-to-image translation models (see training step 2 in Section III-D): sun↔rain and sun↔night. For MUNIT [38], this takes up to three days on a single NVIDIA Titan X GPU. For steps 3 and 4, we contrast the effect of unidirectional domain adaptation from good to poor viewing conditions (e.g., sun→night) and bidirectional translations (e.g., sun↔night) as well as training with two and three domains (rain↔sun↔night). Requiring max. 600,000 iterations with a batch size of eight and a learning rate of 0.0001, training steps 3 and 4 take ca. 26 hours in total on a single GPU for an input resolution of 320×240 px (16% more than SuperPoint), while inference time stays the same (10.5 ms / image). Furthermore, we investigate the influence of domain adaptation in the last two training stages by omitting it in step 4 of some experiments. We call the resulting models DISP-DL (detector learning).

As state-of-the-art baselines for self-supervised methods, we train SuperPoint (SP) models without domain adaptation as in the original paper [19] for image collections from different domain sets: sun/rain, sun/night, and all available domains. Furthermore, we evaluate if our approach is superior to results obtained after images have been aligned in one domain by image-to-image translation as proposed in [11], [13] and contrasted in Section III-A.

TABLE I

Detector Repeatability and Correctness of Homography Estimation on HPatches Dataset [25]

Model	Training data	Detector rep.		Homography	
		Illum.	Viewp.	Illum.	Viewp.
SP [19]	COCO	0.641	0.622	0.925	0.743
SP	NU: sun/rain	0.653	0.592	0.941	0.722
SP	NU: sun/night	0.649	0.586	0.942	0.765
SP	NU: all domains	0.661	0.677	0.946	0.728
DISP-DL	NU: sun→rain	0.648	0.591	0.964	0.770
DISP-DL	NU: sun→night	0.528	0.579	0.950	0.721
DISP-DL	NU: sun↔rain	0.635	0.576	0.945	0.709
DISP-DL	NU: sun↔night	0.636	0.573	0.972	0.719
DISP-DL	NU: rain↔sun↔night	0.656	0.662	0.948	0.751
DISP	NU: sun→rain	0.658	0.653	**0.997**	0.742
DISP	NU: sun→night	**0.671**	**0.686**	0.967	0.750
DISP	NU: sun↔rain	0.655	0.652	0.981	0.733
DISP	NU: sun↔night	0.647	0.684	0.979	**0.772**
DISP	NU: rain↔sun↔night	**0.671**	0.681	0.975	0.745

B. Evaluation Datasets

We evaluate the performance of our DISP models in terms of repeatability, generalization, and robustness of local fea-

TABLE II

RESULTS ON VISUAL LOCALIZATION BENCHMARKS [7] SHOWING RECALL RATES [%] FOR DIFFERENT DISTANCE AND ORIENTATION THRESHOLDS. NV+X DENOTES HIERARCHICAL LOCALIZATION [28] WITH NETVLAD [35] AS GLOBAL AND X AS LOCAL FEATURE.

Method	Training data (local features)	Aachen Day-Night		RobotCar Seasons	
		day	**night**	**day all**	**night all**
		0.25 m / 0.5 m / 5.0 m 2° / 5° / 10°	0.25 m / 0.5 m / 5.0 m 2° / 5° / 10°	0.25 m / 0.5 m / 5.0 m 2° / 5° / 10°	0.25 m / 0.5 m / 5.0 m 2° / 5° / 10°
AP-GeM+R2D2 [44], [15]	Aachen Day [45] i.a.*	**88.7** / **95.8** / **98.8**	81.6 / 88.8 / **96.9**	**55.1** / **82.1** / **97.3**	**28.8** / **58.8** / **89.4**
NV+D2 [14]	MegaDepth [46]*	84.8 / 92.6 / 97.5	**84.7** / **90.8** / **96.9**	52.2 / **80.1** / **95.9**	20.3 / 46.8 / 73.7^{+}
NV+SP [19], [28]	COCO [47]	80.5 / 87.4 / 94.2	68.4 / 77.6 / 88.8	**53.1** / 79.1 / 95.5	7.2 / 17.4 / 34.4
NV+SP	NU: sun/rain	73.2 / 81.4 / 89.7	36.7 / 52.0 / 74.5	34.0 / 55.2 / 66.0	1.0 / 7.5 / 14.2
NV+SP	NU: sun/night	75.5 / 84.6 / 90.7	48.0 / 65.3 / 80.6	37.8 / 61.1 / 74.2	0.8 / 5.6 / 10.6
NV+SP	NU: all domains	75.5 / 84.8 / 90.8	61.2 / 72.4 / 81.6	42.6 / 68.0 / 83.7	4.0 / 9.9 / 20.7
NV+DISP	NU: sun→rain	75.5 / 84.3 / 90.7	49.0 / 68.4 / 81.6	45.6 / 74.8 / 91.8	3.4 / 10.9 / 19.0
NV+DISP	NU: sun→night	77.7 / 85.8 / 91.5	61.2 / 73.5 / 80.6	45.1 / 73.9 / 90.7	3.6 / 12.8 / 21.8
NV+DISP	NU: rain↔sun↔night	79.2 / 87.4 / 93.9	62.2 / 72.4 / 81.6	48.0 / 77.1 / 94.5	4.9 / 12.0 / 24.3
NV+SP	OX: all domains	86.9 / 93.1 / 96.1	75.5 / 82.7 / 88.8	45.4 / 75.4 / 91.9	20.6 / 52.3 / 74.9
NV+DISP	OX: rain↔sun↔night	**88.8** / **95.1** / **98.2**	**82.7** / **90.8** / **96.9**	49.0 / 78.0 / 95.1	**28.2** / **67.7** / **94.5**

*Ground truth correspondences required color code: **best overall** – **second best** $^{+}$query images preprocessed with ToDayGAN [13]

tures by conducting different experiments demanding robust matching under changing viewing conditions.

HPatches: The HPatches dataset [25] includes 116 scenes where each scene contains five image pairs affected by large illumination or viewpoint changes. Given the ground-truth homographies, we compute scores for detector repeatability ($N = 300$, NMS $= 4$) and correctness of homography estimation ($\epsilon = 3$, $N \leq 1000$, NMS $= 8$), following the evaluation process in the original SuperPoint paper [19].

Long-Term Visual Localization: We evaluate the models for the task of long-term visual localization based on two benchmarks introduced in [7]: Aachen Day-Night [45] and RobotCar Seasons [10]. Both contain database images that were captured in consistent viewing conditions (Aachen: day; RobotCar: overcast) and registered in a sparse SfM model. A set of query images from different viewing conditions must be localized in the reference model. For every local feature function, we initially have to compute image correspondences and re-triangulate 3D reference points using the ground truth camera poses of the provided database [28]. For all models, we employ a computation-efficient hierarchical localization concept as proposed in [34]: First, we find candidate images by matching query with database images using NetVLAD [35] as a global image descriptor. After that, we match local features between the query image and the retrieved candidates to calculate the 6-DoF camera pose [8]. As baselines, we report the results of state-of-the-art methods for local feature extraction [14], [15] from the benchmark website (www.visuallocalization.net/benchmark), which also provides more details about datasets and evaluation metrics.

Multi-Domain SfM: In order to evaluate distinctiveness and repeatability of local features in exhaustive matching settings, we consider the completeness and accuracy for collaborative road scene reconstructions, following the concept introduced in [12]. As input for the SfM pipeline based on COLMAP [5], we use unordered image sets from the Non-Urban Dataset which were recorded at ten test spots and excluded from training. We attempt to reconstruct each test spot for two domain pairs: sun–rain and sun–night. For each pair, we sample up to 100 images per domain and test spot. Matches are computed exhaustively, i.e., we search for correspondences between every image pair. We follow the evaluation protocol of [12]: After successful reconstruction, we align the final SfM model with the positions obtained by GPS interpolation (accuracy: < 3 m). The number of images localized within a certain GPS deviation indicates completeness and accuracy of the camera poses reconstructed.

(a) SuperPoint: all domains (b) DISP: rain↔sun↔night

Fig. 4. Correspondences found by SuperPoint [19] and DISP models in images taken from Aachen Day-Night [7] and our Non-Urban test dataset.

V. RESULTS

A. Image Matching

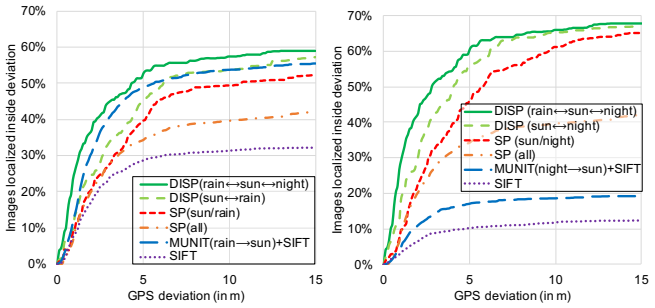
Table I shows the results obtained with different models of SuperPoint and DISP for the HPatches dataset. The scores reveal that the robustness of local features can be increased by adding domain adaptation to the training process. We obtain well-balanced results by training DISP with bidirectional translations between all three domains (see examples in Fig. 4). As expected, using domain adaptation only for self-supervised labeling of the keypoint superset (DISP-DL) results in inferior descriptor performance.

B. Long-Term Visual Localization

Table II depicts the results obtained for the visual localization benchmark datasets Aachen Day-Night and RobotCar Seasons, showing that integration of domain adaptation into training is able to improve visual localization accuracy

(a) SuperPoint: all domains (b) Domain alignment + SuperPoint [19] (c) DISP (ours)

Fig. 5. Examples of local feature matching for image pairs taken from RobotCar Seasons benchmark (top row) and Non-Urban test dataset (bottom row). SuperPoint was trained on images from all domains, DISP with rain↔sun↔night translations. Domain alignment translates one image into the conditions of the other before feature extraction, similar to [11], [13].



(a) Image domains: sun/rain (b) Image domains: sun/night

Fig. 6. Camera pose localization accuracy for Multi-Domain SfM. The diagrams show the overall ratio of camera poses localized within the respective distances to the corresponding GPS measurements using different local features (aggregated over all ten test spots of each collection).

significantly. In all training setups, DISP outperforms the corresponding SuperPoint models by achieving superior recall rates in both day and night domains. As in HPatches, multiple and bidirectional translations lead to the best results. SuperPoint and DISP models trained on the NU dataset show inferior generalization on RobotCar Seasons due to the large differences in image quality, especially for challenging illumination conditions (see Fig. 4 and 5).

Comparison with state-of-the-art baselines shows that DISP (trained with OX dataset) outperforms the original SuperPoint model [19], [28] while it is comparable (Aachen Day-Night) or even superior (RobotCar Seasons night images) to state-of-the-art supervised methods [14], [15] currently leading the benchmarks for local features (with classical matching pipelines). Note that we extracted features from downsampled images (width: 480 px) and did not train on the target cameras, which leaves further room for improvement. Comparison with results produced by D2-Net [14] and SuperPoint after preprocessing of night query images from RobotCar Seasons with ToDayGAN [13] in Table II and Fig. 5 reveals that our concept (domain adaptation during feature training) outperforms the contrary approach (pixel domain alignment before feature extraction).

C. Multi-Domain SfM

Fig. 6 visualizes road scene reconstruction results obtained for different local features. We compare reconstructions based on the previously shown SuperPoint and DISP as well as the hand-crafted SIFT [16]. The results demonstrate that the learned local features can outperform SIFT as they provide both high repeatability and distinctiveness. As in the experiments described before, we observe a rise in localization accuracy for the DISP variants compared to the SuperPoint baselines.

In addition, we depict the reconstruction results with SIFT after the images captured under poor viewing conditions (rain or night) have been translated to the sun domain. While this method can almost reach the localization accuracy of DISP for the sun/rain collections, it completely fails for sun/night. This supports our argumentation in Section III-A that domain invariance in the image space can only be reached via preprocessing if image-to-image translation can accurately reconstruct the scene content in the target domain.

VI. Conclusion

In this work, we contrasted two ways to improve cross-domain feature matching through GAN-based image-to-image translation. Reasoning why it is most effective to employ image-to-image translation already during the learning process, we proposed a self-supervised training pipeline that explicitly optimizes local feature extraction for domain invariance. In extensive experiments covering image matching, long-term visual localization, and SfM, we showed that DISP outperforms previous self-supervised methods and is on par with state-of-the-art supervised features. DISP complements recent approaches [23], [24] for domain-invariant place recognition features: In future work, both essential steps for visual localization can be trained self-supervised.

References

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo Tourism: Exploring Photo Collections in 3D," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 835–846, 2006.

[4] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a Day," in *IEEE International Conference on Computer Vision*, 2009.

[5] J. L. Schönberger and J.-M. M. Frahm, "Structure-from-Motion Revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware Visual Localization under Challenging Perceptual Conditions," in *IEEE International Conference on Robotics and Automation*, 2017.

[7] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[8] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, UK: Cambridge University Press, 2003.

[9] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free," in *Robotics: Science and Systems*, 2015.

[10] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[11] H. Porav, W. Maddern, and P. Newman, "Adversarial Training for Adverse Conditions: Robust Metric Localisation using Appearance Transfer," in *IEEE International Conference on Robotics and Automation*, 2018.

[12] M. Venator, E. Bruns, and A. Maier, "Robust Camera Pose Estimation for Unordered Road Scene Images in Varying Viewing Conditions," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 1, pp. 165–174, 2019.

[13] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. van Gool, "Night-to-Day Image Translation for Retrieval-based Localization," in *IEEE International Conference on Robotics and Automation*, 2019.

[14] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-Net: A Trainable CNN for Joint Description and Detection of Local Features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[15] J. Revaud, P. Weinzaepfel, C. De Souza, and M. Humenberger, "R2D2: Repeatable and Reliable Detector and Descriptor," in *Advances in Neural Information Processing Systems*, 2019.

[16] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool, "Speeded-Up Robust Features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *International Conference on Computer Vision*, 2011.

[19] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2014.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *IEEE International Conference on Computer Vision*, 2017.

[22] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, and B. A. Research, "Image-to-Image Translation with Conditional Adversarial Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] L. Tang, Y. Wang, Q. Luo, X. Ding, and R. Xiong, "Adversarial Feature Disentanglement for Place Recognition Across Changing Appearance," *IEEE International Conference on Robotics and Automation*, 2020.

[24] C. Qin, Y. Zhang, Y. Liu, S. Coleman, D. Kerr, and G. Lv, "Appearance-invariant place recognition by adversarially learning disentangled representation," *Robotics and Autonomous Systems*, vol. 131, 2020.

[25] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[26] J. Dong and S. Soatto, "Domain-Size Pooling in Local Descriptors: DSP-SIFT," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[27] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative Evaluation of Hand-Crafted and Learned Local Features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[28] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From Coarse to Fine: Robust Hierarchical Localization at Large Scale," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[29] N. Savinov, A. Seki, U. Ladický, T. Sattler, and M. Pollefeys, "Quad-networks: unsupervised learning to rank for interest point detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[30] A. Mishchuk, D. Mishkin, F. Radenović, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Advances in Neural Information Processing Systems*, 2017.

[31] Y. Tian, B. Fan, and F. Wu, "L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[32] K. M. Yi, E. Trulls, V. Lepetit, P. Fua, K. Moo Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned Invariant Feature Transform," in *European Conference on Computer Vision*, 2016.

[33] M. Wang and W. Deng, "Deep Visual Domain Adaptation: A Survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[34] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging Deep Visual Descriptors for Hierarchical Efficient Localization," in *Conference on Robot Learning*, 2018.

[35] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[36] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv:1511.06434*, 2015.

[37] X. Wang and A. Gupta, "Generative Image Modeling using Style and Structure Adversarial Networks," in *European Conference on Computer Vision*, 2016.

[38] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal Unsupervised Image-to-Image Translation," in *European Conference on Computer Vision*, 2018.

[39] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv:1411.1784*, 2014.

[40] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward Multimodal Image-to-Image Translation," in *Advances in Neural Information Processing Systems*, 2017.

[41] M. S. Mueller, T. Sattler, M. Pollefeys, and B. Jutzi, "Image-to-Image Translation for Enhanced Feature Matching, Image Retrieval and Visual Localization," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 2/W7, 2019.

[42] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse Image-to-Image Translation via Disentangled Representations," in *European Conference on Computer Vision*, 2018.

[43] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, 2015.

[44] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, J. Revaud, P. Rerole, N. Pion, C. de Souza, V. Leroy, and G. Csurka, "Robust Image Retrieval-based Visual Localization using Kapture," *arXiv:2007.13867*, 2020.

[45] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image Retrieval for Image-Based Localization Revisited," in *British Machine Vision Conference*, 2012.

[46] Z. Li and N. Snavely, "MegaDepth: Learning Single-View Depth Prediction from Internet Photos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[47] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014.