PUMP: Pyramidal and Uniqueness Matching Priors for Unsupervised Learning of Local Descriptors

Jérome Revaud   Vincent Leroy   Philippe Weinzaepfel   Boris Chidlovskii
NAVER LABS Europe
firstname.lastname@naverlabs.com

Abstract

Existing approaches for learning local image descriptors have shown remarkable achievements in a wide range of geometric tasks. However, most of them require per-pixel correspondence-level supervision, which is difficult to acquire at scale and in high quality. In this paper, we propose to explicitly integrate two matching priors in a single loss in order to learn local descriptors without supervision. Given two images depicting the same scene, we extract pixel descriptors and build a correlation volume. The first prior enforces the local consistency of matches in this volume via a pyramidal structure iteratively constructed using a non-parametric module. The second prior exploits the fact that each descriptor should match with at most one descriptor from the other image. We combine our unsupervised loss with a standard self-supervised loss trained from synthetic image augmentations. Feature descriptors learned by the proposed approach outperform their fully- and self-supervised counterparts on various geometric benchmarks such as visual localization and image matching, achieving state-of-the-art performance. Project webpage: https://europe.naverlabs.com/research/3d-vision/pump.

1. Introduction

Local image descriptors, usually extracted sparsely as keypoints, are at the core of numerous computer vision tasks such as large-scale visual localization [60], pose estimation [25], Structure-from-Motion (SfM) [56, 70], dense 3D reconstruction [63] and SLAM [6]. Nowadays, learning-based approaches [1, 18, 26, 43, 48, 62, 68, 69, 76] significantly outperform the standard handcrafted keypoints such as SIFT [34] or ORB [53]. They are often trained assuming that numerous ground-truth pixel correspondences between pairs of images are available. These correspondences are most of the time obtained by considering a large collection of images for a given landmark and building a Structure-from-Motion (SfM) reconstruction, as done for instance for the MegaDepth dataset [31]. This SfM pipeline nevertheless fails in many cases, yielding an unfathomable bottleneck to the kind of ground-truth data that can be generated. The question we try to answer in this work is the following: *is it possible to exploit the sleeping potential of unsupervised image pairs,* i.e. *image pairs without any ground-truth pixel correspondences?*

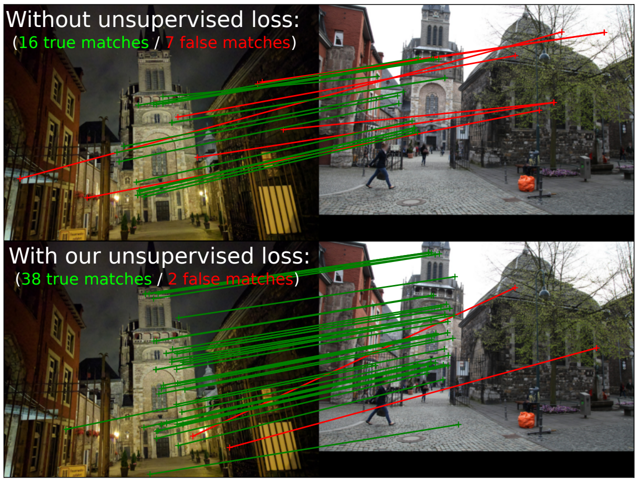
Without unsupervised loss:
(16 true matches / 7 false matches)
With our unsupervised loss:
(38 true matches / 2 false matches)

Figure 1. Qualitative impact of PUMP, our novel unsupervised loss, on a challenging image pair with illumination changes. We match keypoints extracted with models trained without (top) and with it (bottom), showing only matches that pass the geometric verification. Our unsupervised model finds more than twice as many true matches compared to a model trained without PUMP.

In the remainder of the paper, we follow Truong *et al.* [67] and adopt a practical definition of unsupervised learning w.r.t. the feature learning task. We denote a learning formulation ‘unsupervised’ if it does not require any supervision other than pairs of images depicting the same visual content. Inspired by the success of self-supervised learning for representation learning [8, 17, 24, 72], depth regression [15] and point cloud registration [2, 3], pure self-supervised learning approaches for local descriptors have provided partial answers to this question. They are trained on synthetically generated image pairs, where the second image is obtained by applying known transformations to the first image, such as a random homography, color jittering or even style transfer [41]. However, homographies and the like cannot model the full range of possible transformations between real image pairs. In parallel, weakly-supervised methods have been proposed and demonstrate the ability to train from *e.g.* known camera poses [71]. Yet, this is only achievable through the use of complex acquisition setups that require the deployment of sensors based on different modalities (IMU or GPS), or again, resorting to SfM reconstructions. Recently, unsupervised learning of local descriptors has been introduced in the form of cycle consistency constraints across multiple images [67, 79], either requiring more images to extract features for training, or at the cost of iterative training of descriptors and expensive model fitting [74].

In this paper, we introduce a novel method to learn local descriptors without supervision. It is based on jointly enforcing two key matching priors: *local consistency* and *uniqueness* of the matching. The former simply amounts to state that two neighboring pixels of one image will likely match with two pixels forming a similar neighboring pair in the other image, up to a small deformation. We assume that this holds in general at any scale, hence this prior can be efficiently enforced through a pyramidal structure. Inspired by DeepMatching [49], we employ a pyramidal non-parametric module that extracts higher-level correspondences enforcing the local consistency matching prior by design. The uniqueness prior, for its part, simply means that one pixel from the first image can correspond to at most one pixel in the second image. We enforce this property on high-level correspondences output by the DeepMatching module, which gracefully back-propagates along the pyramid to low-level pixel correspondences, enabling an effective training of local descriptors without supervision.

We coin our proposed approach PUMP for *Pyramidal and Uniqueness Matching Priors*. It is trained in conjunction with a self-supervised loss applied on synthetic image pairs. We experiment with both sparse and dense matching, either relying on external sparse keypoint detectors, or, in the case of dense matching, leveraging DeepMatching once again at test time to further enforce the two matching priors dynamically. We show that our unsupervised loss results in a significant increase of performance compared to a model trained solely using self-supervision and significantly outperforms the state of the art on several tasks and benchmarks. In short, we make the following contributions:

- We revisit the key notion of matching prior for descriptor learning, and show their unreasonable effectiveness at training and test time.
- We introduce a novel unsupervised loss derived from these priors, termed PUMP, for training deep descriptors at the pixel level.
- We present experimental evidence that our approach significantly outperforms state-of-the-art methods on both dense and sparse matching tasks in spite of requiring less supervision and training data.

2. Related work

Our main contribution is an explicit integration of unsupervised priors for image matching. We thereby review related works on the different priors used in the literature and the type of supervision signal they require.

Local neighborhood consistency is arguably one of the most common prior for image matching. In fact, the use of image patches to detect features, and to describe and match pixels' appearances, stems from the assumption that local neighborhood is consistent across views. This prior can be traced back to the works of [16, 44] and is ubiquitous in image description works. Initially used for handcrafted methods [34, 53], the recent success of deep learning has motivated researchers to move to supervised CNN-based approaches for interest point detection [26, 43, 68, 76] description [1, 12, 18, 62] or both detection and description [4, 38, 48, 69]. A few works tried to improve the capability of CNNs to describe non-planar regions by introducing robust [33] or dynamic convolutional kernels [37], yet in a supervised training scenario. Similar in spirit, our dense matching procedure is able to dynamically adapt to local image deformations at test time in a hierarchical manner. A noticeable shift in supervision paradigm is emerging with the use of self-supervised learning strategies. Such approaches for training local descriptors rely on synthetically generated image pairs using known transformations, *e.g.* homographies [27, 48] with color jittering or stylization [41]. Several works [27, 42, 50] consider only the augmented input image pair to sample positive and negative local descriptors. Making a careful choice of hard negatives and coupling it with color augmentation and photo-realistic image stylization, Melekhov *et al.* [41] achieve superior performance to supervised methods using only synthetic homographies. DeTone *et al.* [10] have achieved some success by mining positives and negatives from homography-related image pairs, or by combining negative examples mining with homography [41]. However, homographies or other types of synthetic transformations are limited by nature and will fail to model complex appearance changes between real image pairs. Strong supervision at the pixel correspondence level therefore remains obligatory on challenging matching tasks such as visual localization [12, 48]. Our work yet shows that leveraging unsupervised image pairs is possible and effective for improving visual localization.

Local consistency for matching is an extension of the previous idea: since local neighborhoods are consistent, so are the matches between them. This idea has been explicitly formulated as a strong prior to remove false associations during the sparse matching step in [5, 32, 39]. Focusing on the dense matching problem alleviates the need for a detector. It has been tackled via hierarchical pyramidal matching, either handcrafted [49], optimized at test time [19], or learned in a fully supervised setting [21, 42, 66]. In this paper, we propose to leverage the non-parametric pyramidal structure of DeepMatching [49] to learn descriptors without supervision. Note that this is not to be confused with Deep Matching Prior [19] which, despite sharing the same name, is a fundamentally different approach relying on a test-time optimization procedure for each image pair.

In a similar direction, learning to predict dense matches from 4D cost volumes also received some attention lately with NCNet [52], that was first to estimate neighborhood consensus in the 4D space of correspondences. Later, multiple variants were proposed to overcome the large memory consumption, slow inference time and poorly localised correspondences. Rocco *et al*. [51] sparsify the correlation tensor containing tentative matches, and its subsequent processing with a 4D CNN using submanifold sparse convolutions. Li *et al*. [28] introduce non-isotropic 4D filtering to better deal with scale variation. DualRC-Net [29] avoids calculating the expensive full 4D correlation tensor by extracting first coarse resolution feature maps. The coarse maps are then used to produce a full but coarse 4D correlation volume, which is then refined by a learnable neighborhood consensus module. Reliability of matches can also be predicted using the correlation volume and used to improve matching in a self-supervised manner [65]. We also use a 4D correlation volume, but we process it using an efficient pyramidal structure that inherently encodes a strong matching prior. This allows us to learn local descriptors without supervision, and guide the matching at test-time.

Image Context preservation is another kind of prior often used in the literature. Local features are very accurate but prone to fail in ambiguous cases, *e.g*. repetitive structures, challenging lighting conditions or even seasonal changes. To circumvent this limitation, several previous works introduce the use of global context of the scene, either in the form of coarse image descriptors [9, 35], or graph operators that reason at the structure level [7, 54]. To increase the receptive field during feature extraction, LoFTR [59] proposes detector-free local features matching with transformers. Similarly, COTR [22] predicts matches with attention mechanisms in an asymmetric manner similar to [12]. Such methods however require strong supervision and do not enforce consistency of predicted matches.

Cycle Consistency is often used in complement to a photometric loss in the optical flow literature, and has been recently used in supervised [22] and unsupervised [40, 58] settings. Unsupervised learning of local descriptors can also rely on cycle consistency across multiple images [79], at the cost of requiring feature extraction on more images for training. Similarly, Truong *et al*. [67] regress dense correspondences in an unsupervised setting. Unfortunately, cycle consistency is difficult to optimize as it requires to minimize a differentiable flow. Different from these works, our unsupervised loss rather exploits uniqueness, *i.e*., the key property that a pixel in one image can correspond to at most one pixel in the other image.

Multi-View Geometry constraints can finally also be leveraged to improve the matching performance. For example, it is possible to use them for training data selection [36] or directly as a training loss [13, 14]. Usually, methods designed around this prior rely on supervision signals from epipolar geometry [9, 73, 75, 78] or relative camera poses [4, 13, 71]. Yang *et al*. [74] propose a self-supervised approach that alternates between two tasks, namely estimating camera poses and learning local descriptors, each task being supervised by the other. The main drawback of such approaches is that they require a complex and computationally heavy acquisition and training setup, requiring to process entire SfM datasets, building SfM map, knowing or computing camera intrinsics, *etc*. In comparison, our approach can in theory be trained from a set of image pairs obtained by different means, including baseline image retrieval coupled with an off-the-shelf geometric verification.

3. Unsupervised learning of local descriptors

We aim to train a neural network f_θ with parameters θ that, given an image I of dimension $H \times W$, extracts a highly discriminative yet robust local descriptor for each pixel of I. Mathematically, we have $f_\theta : I \rightarrow F_I$ where $F_I \in \mathbb{R}^{H\times W\times d}$ is a d-dimensional feature map that can be seen as a collection of dense ℓ_2-normalized local descriptors. As many recent approaches [28, 29, 51, 52], our method builds upon the 4D correlation volume $C(F_1, F_2)$ computed as a dot-product between descriptors F_1, F_2 from the images I_1 and I_2. To ease readability, we simply denote $C(F_1, F_2)$ as C. Furthermore, we denote the correlation between two pixels $\boldsymbol{p} = (x_{\boldsymbol{p}}, y_{\boldsymbol{p}})$ in image I_1 and $\boldsymbol{q} = (x_{\boldsymbol{q}}, y_{\boldsymbol{q}})$ in image I_2 simply as $C_{\boldsymbol{p},\boldsymbol{q}}$. We now present our method to train f_θ given image pairs without any pixel-level supervision. As summarized in Figure 2, we first build a global correlation volume, which is aggregated and max-pooled over iterations in a pyramidal fashion using a non-parametric DeepMatching module (Section 3.1). The output consists of high-level correspondences, each spanning a large receptive field, as they result from the iterative aggregation of lower-level correspondences. We then apply a loss that encourages the uniqueness of these high-level, and thus reinforced, correspondences (Section 3.2).

3.1. Pyramidal local consistency matching prior

We first propose to integrate *local consistency*, a key property of matching stating that a pair of neighbor pixels in image I_1 will likely match a pair of pixels that are also

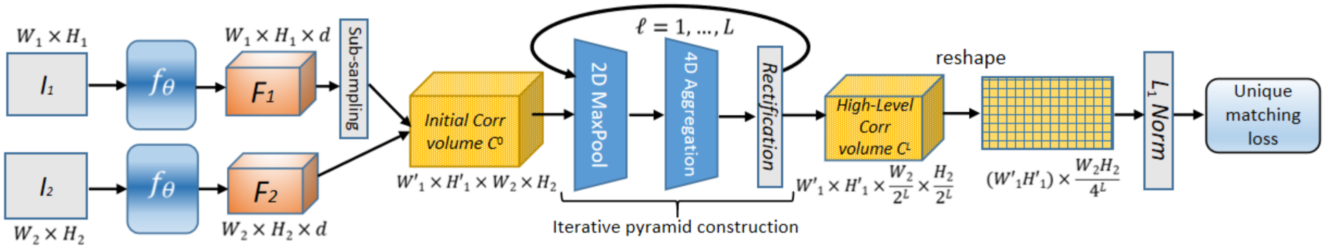
$W_1 \times H_1$
I_1
f_θ
$W_1 \times H_1 \times d$
F_1
Sub-sampling
I_2
f_θ
F_2
$W_2 \times H_2$
$W_2 \times H_2 \times d$
Initial Corr volume C^0
$W'_1 \times H'_1 \times W_2 \times H_2$
$\ell = 1, \ldots, L$
2D MaxPool
4D Aggregation
Rectification
Iterative pyramid construction
High-Level Corr volume C^L
$W'_1 \times H'_1 \times \frac{W_2}{2^L} \times \frac{H_2}{2^L}$
reshape
$(W'_1 H'_1) \times \frac{W_2 H_2}{4^L}$
L_1 Norm
Unique matching loss

Figure 2. Overview of our unsupervised framework. The initial, low-level 4D correlation volume C^0 is iteratively aggregated by integrating local neighborhood information until reducing image I_2 to $\frac{W_2}{2^L} \times \frac{H_2}{2^L}$ size. The consolidated, high-level correlation volume C^L is then unfolded in a 2D matrix from which the unique matching loss $\mathcal{L}_\mathcal{U}$ is applied.

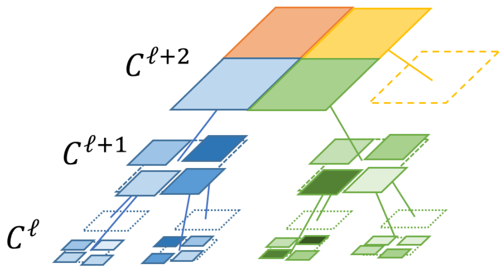

Figure 3. Illustration of the deformable pyramid. A correspondence in the parent-level correlation volume $C^{\ell+1}$ aggregates 4 correspondences in the child-level C^{ℓ} with a tolerance for small deformations, and so on for all levels. For the sake of clarity, we show only a subset of the parent-children patch relations.

neighbors in image I_2, with the same spatial offset up to a small deformation. To this end, we leverage a pyramidal aggregation technique similar to the non-parametric DeepMatching algorithm [49]; we remind the principal idea and procedure below.

DeepMatching proceeds asymmetrically by decomposing the first image I_1 in a regular grid of 4×4 patches. In our case, we simply subsample the feature map F_1 by a factor 4 in both spatial dimensions. We thus obtain the initial correlation volume C^0 of size $H'_1 \times W'_1 \times H_2 \times W_2$ where $H'_1 = H/4$ and $W'_1 = W/4$. It undergoes an iterative aggregation procedure that calculates the correlation maps of larger patches in subsequent pyramid levels. The key intuition of the aggregation is that the correlation $C^1_{\boldsymbol{p},\boldsymbol{q}}$ for a 8×8 parent patch at level $\ell = 1$ can be computed as the average correlation of its 4 children patches in C^0. $C^1(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{4}(C^0(\boldsymbol{p}-v^0, \boldsymbol{q}-v^0) + \ldots + C^0(\boldsymbol{p}+v^0, \boldsymbol{q}+v^0))$, where $v^\ell = (\pm 2^\ell, \pm 2^\ell)$ represents the offset between the parent patch center and its 4 children. This aggregation can be implemented using 4D convolutions with a fixed sparse kernel, where non-zero values encode the parent-children relations in the pyramid. The formulation we presented so far is able to handle only purely rigid transformations. To allow for local deformations, a 2D max-pooling stage with a 3×3 kernel and stride 2 along the second image dimensions is inserted before each 4D convolution. Likewise, a power rectification $x \rightarrow \max(0, x^\gamma)$ completes the aggregation at each level to further strengthen consistent correspondences and discard spurious ones.

Mathematically, the receptive field of a parent patch doubles along the x- and y- dimensions at each pyramid level, hence rapidly reaching the size of the full image, at which point the aggregation process stops. Figure 3 illustrates the deformable pyramidal structure enforced as a prior by this algorithm. The output is a consolidated correlation volume C^L, whose dimension along image I_2 is reduced by a factor 2^L due to the max-pooling stride, where L is the number of pyramid levels. In practice, we run the process for 5 levels with 224×224 input feature maps during training, yielding the final correlation volume of size 7×7 along the second image dimensions. We show in the supplementary material the correlation volume at different levels of the pyramid to give a better intuition of the procedure.

3.2. Unique matching prior

The consolidated correlation volume C^L represents the correlations between large deformable patches spanning the entire images; we call them *high-level patches*. Ideally, a high-level patch centered at pixel $\boldsymbol{p}$ has a unique match in image I_2, *i.e.*, there should exist only one $\boldsymbol{q}$ such that $C^L_{\boldsymbol{p},\boldsymbol{q}}$ has a high value, while all other correlations $C^L_{\boldsymbol{p},\boldsymbol{q'}}$ for $\boldsymbol{q'} \neq \boldsymbol{q}$ will be close to 0. While this constraint is not realistic for pixel-level descriptors due to repetitive/plain patterns or severe appearance changes, it appears as a natural property for high-level patches in C^L. In fact, the larger is a patch, the easier it is to resolve ambiguities and hard correspondences thanks to a larger context.

We thus propose a loss that encourages the key property of *unique match* for every high-level patch. Since each high-level patch is dynamically built at test-time upon a deformable subset of pixel-level correlations, this loss is automatically back-propagated to optimal pixel correlations fitting the pyramidal prior, and hence to pixel level descriptors. Formally, we reshape C^L as a two-dimensional tensor of size $(W'_1 H'_1) \times \frac{W_2 H_2}{4^L}$. We first normalize the correlation volume such that every row sums to 1:

$$\bar{C^L}_{\boldsymbol{p},\boldsymbol{q}} = \frac{C^L_{\boldsymbol{p},\boldsymbol{q}}}{\sum_{\boldsymbol{q}} C^L_{\boldsymbol{p},\boldsymbol{q}} + \epsilon}, \tag{1}$$

where ϵ serves as a regularization term that deals with occluded areas, for which all correlations would be close to zero. The uniqueness loss is then simply expressed as

$$\mathcal{L}_{\mathcal{U}}(F_1, F_2) = -\frac{1}{H_1' \times W_1'} \left\| \bar{C^L} \right\|_2^2. \tag{2}$$

Given that $\bar{C}^L$ is ℓ_1-normalized, the uniqueness loss concretely encourages all values in $\bar{C}^L$ to be close to 0 except one per row (*i.e.*, one per high-level patch) that will be close to 1, see supplementary for a proof. Note that this formulation is closely related to the sparse Lasso regularization [55].

3.3. Implementation details

Training. As our loss is asymmetric, we average its value for each pair and its reverse: $\mathcal{L}_{\mathcal{U}}' = \mathcal{L}_{\mathcal{U}}(F_1, F_2) + \mathcal{L}_{\mathcal{U}}(F_2, F_1)$. To reduce the computational cost and memory footprint, we subsample feature maps F_1 and F_2 by a factor 2 before passing them to the loss. We train our model with batches of 16 pairs, where one half of the pairs are trained with our proposed unsupervised loss, and the other half is generated synthetically via standard data augmentation of single images in a self-supervised manner. For these latter pairs, all ground-truth correspondences are obtained from the augmentation, and we use the same pixel-wise ranking loss $\mathcal{L}_{\mathcal{AP}}$ as in R2D2 [48]. We find important to use this auxiliary self-supervised loss to obtain good results. The final loss is calculated as a weighted sum $\mathcal{L} = \mathcal{L}_{\mathcal{AP}} + 0.3\mathcal{L}_{\mathcal{U}}'$. We fix $\epsilon = 0.03$ in Eq. 1 and set $\gamma = 1.5$ for power rectification as in the original DeepMatching [49]. We implement our approach in PyTorch [45]. We perform 50000 training iterations, that suffice for models to converge, with the Adam optimizer [23], a fixed learning rate of 10^{-4} and a weight decay of 5.10^{-4}.

Training data. We use 150,000 pairs from the SfM-120k dataset [46], which contains images from famous landmarks around the world. These training pairs are obtained using the overlap of observations in a SfM model built with COLMAP and provided with the dataset. Note that this process is done with SIFT, which requires no supervision, and only serves to verify whether two images depict the same scene, but does not guide the pyramidal matching at all. In terms of data augmentation, we only perform random crops of size 256×256 on these pairs. To generate the synthetic pairs, we randomly sample images from this dataset and apply standard data augmentation techniques. In particular, we use random pixel and color jittering, random rescaling, rotations and homographies.

Backbone architecture. Our network f_θ is built upon the recent ConvMixer architecture [64]. In detail, it first computes a 5×5 convolution with stride 1 and 128 output channels. It then embeds 4×4 non-overlapping gradient patches into 512-dimensional features using a convolution of kernel 4×4 with stride 4. Then, a series of pointwise and depthwise convolutions are applied. We use 7 such blocks, with depthwise convolutions using 9×9 kernels. We finally apply a last pointwise convolution and PixelShuffle operation to obtain the feature map F with $d = 128$ dimensions.

4. Experiments

After presenting datasets (Section 4.1), we evaluate our approach densely (Section 4.2) and sparsely using various keypoint detectors (Section 4.3). We finally provide an ablative study in Section 4.4.

4.1. Datasets and benchmarks

Hpatches [1] consists of 116 image sequences with varying photometric and viewpoint changes. Each sequence contains a reference image and 5 source images related by a homography to the source image taken under different viewpoint or illumination.

ETH3D [57] contains indoor and outdoor sequences captured using a hand-held camera and registered with SfM. Image pairs are generated by sampling frames with a fixed interval. We use it to evaluate the robustness to viewpoint changes as the baseline widens for increasing intervals.

Aachen Day-Night v1.1 [77] is a large-scale outdoor visual localization benchmark. We specifically consider the Day-Night split to measure the generalization performance of our approach, as it features large viewpoint changes and severe illumination changes due to the day/night duality. For this task, we use the Kapture [20] pipeline: in a first step, a global SfM map is built from the database images, and in a second step, query images are localized w.r.t. this mapping. The computational complexity of a complete matching is handled via the use of image retrieval with AP-GeM-LM18 [47] global descriptors. We reduce the number of image pairs to the top-20 nearest neighbors, during both the mapping and the query phases. We extract 20,000 local descriptors for each of these retrieved images, and match them to estimate first the global map and then the camera poses.

4.2. Dense matching

We evaluate the performance of our PUMP descriptors in a dense or quasi-dense manner using DeepMatching [49] (DM). Similarly to the training phase, we replace the basic pixel descriptor of DeepMatching by our trained descriptor. The rest of the pipeline is left unchanged, except for the built-in cycle-consistency verification that we enhance to include nearest neighbors as well. We find this modification to be important as DeepMatching tends to produce many

	Method	AEPE↓	PCK@1↑	PCK@3↑	PCK@5↑
Dense flow	LiteFlowNet CVPR'18	118.85	13.91	-	31.64
	PWC-Net CVPR'18, TPAMI'19	96.14	13.14	-	37.14
	DGC-Net [42] WACV'19	33.26	12.00	-	58.06
	RAFT [61] ECCV'20	44.3	31.22	62.48	70.85
	GLU-Net CVPR'20	25.05	39.55	71.52	78.54
	GLU-Net+GOCor NeurIPS'20	20.16	41.55	-	81.43
	WarpC [67] ICCV'21	21.00	-	-	83.24
	COTR + Interp. [22] ICCV'21	7.98	33.08	77.09	86.33
	DMP [19] ICCV'21	5.21	-	-	90.89
	PUMP (S)+DM + Interp.	<u>4.19</u>	<u>76.36</u>	<u>90.11</u>	<u>92.29</u>
	PUMP (S+U)+DM + Interp.	**3.76**	**77.05**	**90.86**	**93.02**
Sparse	COTR ICCV'21	7.75	40.91	82.37	91.10
	PUMP (S)+DM	**2.87**	**74.72**	**96.05**	<u>97.14</u>
	PUMP (S+U)+DM	<u>2.97</u>	<u>74.01</u>	<u>95.86</u>	**97.27**

Table 1. Average End Point Error (AEPE) and Percent of Correct Keypoints (PCK) for different thresholds on HPatches. Sparse methods only return a subset of correspondences which they are confident of. The best and second best results are respectively in **bold** and <u>underlined</u>. DM stands for DeepMatching and 'Interp.' means Interpolation. We evaluate our approach with self-supervised pairs only (S) and with also unsupervised training pairs (S+U).

Method	AEPE↓						
	rate 3	rate 5	rate 7	rate 9	rate 11	rate 13	rate 15
LiteFlowNet CVPR'18	**1.66**	2.58	6.05	12.95	29.67	52.41	74.96
PWC-Net CVPR'18, TPAMI'19	1.75	2.10	5.59	14.35	27.49	43.41	
DGC-Net [42] WACV'19	2.49	3.28	4.18	5.35	6.78	9.02	12.23
GLU-Net CVPR'20	1.98	2.54	3.49	4.24	5.61	7.55	10.78
RAFT [61] ECCV'20	1.92	2.12	2.33	2.58	3.90	8.63	13.74
DMP [19] ICCV'21	1.78	2.07	2.52	3.07	4.72	6.14	7.47
COTR +Interp. [22] ICCV'21	1.71	<u>1.92</u>	<u>2.16</u>	2.47	<u>2.85</u>	**3.23**	3.76
PUMP (S)+DM +Interp.	1.77	2.81	2.39	<u>2.39</u>	3.56	3.87	4.57
PUMP (S+U)+DM +Interp.	<u>1.67</u>	**1.86**	**2.12**	**2.37**	**2.81**	<u>3.41</u>	**3.69**

Table 2. Average End Point Error (AEPE) for different rates on the ETH3D dataset. The best and second best results are respectively in **bold** and <u>underlined</u>.

isolated spurious correspondences that yet pass the built-in reciprocal verification. All in all, and as in our PUMP loss, DeepMatching enforces the local consistency and uniqueness priors *by design* in a global manner when computing the output set of correspondences. Our GPU implementation performs multi-scale matching on two 640-pixels images in about 3 seconds. For larger resolutions, we adopt a coarse-to-fine strategy as in COTR [22]. Note that the output of DeepMatching is not dense but quasi-dense, as it outputs one correspondence per atomic patch from the first image. We rely on a simple densification technique when dense warp fields are required. Namely, we follow COTR's scheme [22] and linearly interpolate matches using a Delaunay triangulation.

HPatches. We follow the evaluation protocol of [22,42,66] and evaluate on all image pairs from HPatches that feature viewpoint changes. We report results in Table 1 for both quasi-dense and fully-dense (*i.e*., interpolated) outputs. We evaluate two models: one trained solely from self-supervised pairs (S), *i.e*., obtained via data augmentation, and one including unsupervised pairs as well (S+U). Without interpolation, our self-supervised model (S) performs slightly better than the model trained with unsupervised pairs (S+U). This is not surprising given that it is trained exclusively from synthetic augmentations (homographies) fitting exactly the distribution of the test set. In fully-dense mode, our unsupervised model (S+U) outperforms the self-supervised model (S), indicating that the unsupervised loss allows to produce less outliers (as they strongly impair Delaunay interpolation) and is thus more robust. Overall, whether it is used with or without interpolation, both proposed models outperform all state-of-the-art approaches by a large margin. Note that we do not employ any explicit geometric constraints nor filtering, in contrast to RANSAC-Flow [58] for instance. PUMP also significantly outperforms the recently proposed unsupervised WarpC matching loss [67]. However, we hypothesize that the GLU-Net architecture of their model, required to train their unsupervised warp-consistency loss, is a bottleneck to their performance. Altogether, these results highlight the excellent (and expected) capacity of our pyramidal matching prior in the case of large planar areas without discontinuities.

ETH3D. Next we evaluate our model in a more challenging setting with real image pairs featuring viewpoint changes on complex 3D shapes and many discontinuities in the optical flow. Once again, we follow the evaluation protocol by [22, 42, 66]. Since the ground-truth is sparse and not necessarily aligned with the quasi-dense output, we only report results with the densely-interpolated variant for various frame intervals (*e.g*. rate) in Table 2. We observe that the model trained with unsupervised pairs significantly outperforms the self-supervised one by up to 25% (relative gain). This highlights the superior robustness against realistic noise of the model trained by injecting matching priors. Overall, it also outperforms all existing approaches, scoring the first or second AEPE for all rate intervals. Note that the self-supervised model still performs well, being ranked only after COTR [22], a recent approach trained under dense supervision using $\approx$ 50 times more data and a much larger network (18.5M *vs*. 3.5M parameters). This demonstrates the benefit of enforcing priors at test time in realistic conditions. Our approach is also significantly faster than other methods such as DPM [19] which requires multiple minutes of specific fine-tuning on each testing pair.

Figure 4 presents qualitative results on pairs from the 'lakeside' sequence with challenging viewpoint changes, complex 3D shapes, occlusions, lighting artefacts and illumination changes. Our method is able to accurately reconstruct the second frame (except, of course, in occluded areas) under challenging conditions. It can also match small regions, *e.g*. the white plate in the first column or the right side of the bench in the second column (see zoomed insets).

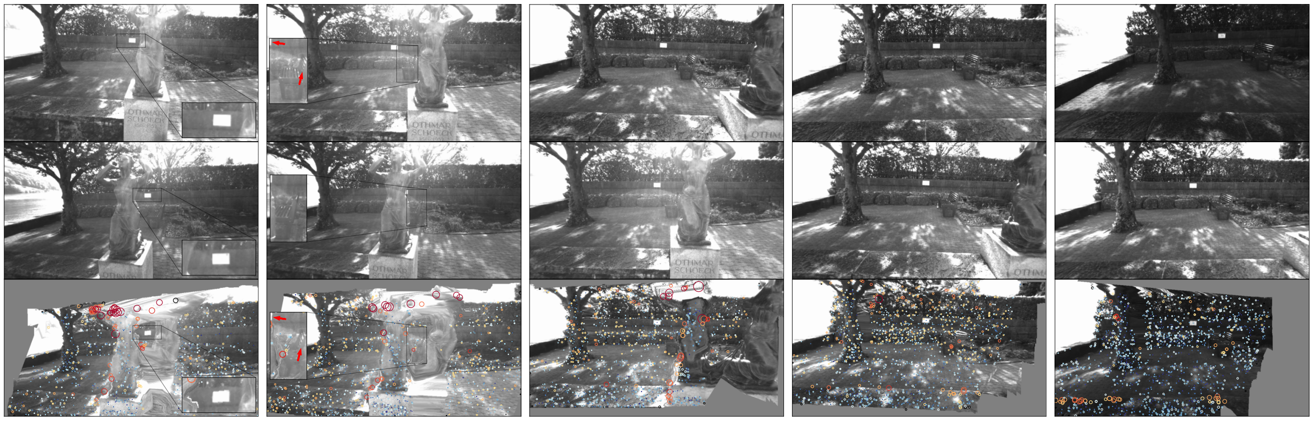

Figure 4. Wide baseline matching on the most challenging 'lakeside' sequence from ETH3D. The first two rows show pairs of images to match. The third row shows the first image warped to the second one according to the dense matching predicted by our model. Errors on the ground-truth control points are represented as circles whose area is proportional to the error, using the KITTI error color-code. We observe that large errors mostly appear around motion boundaries. More examples are shown in the supplementary video.

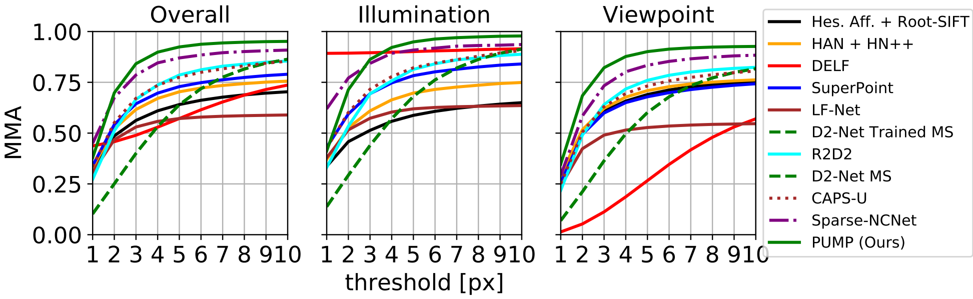

Figure 5. Sparse matching results on the HPatches dataset in term of Mean Matching Accuracy (MMA) for various error thresholds.

Det- ector	Desc- riptor	HPatches MMA@1↑	 MMA@3↑	 MMA@5↑	Localization on Aachen Day-Night 0.25m, 2°	 0.5m, 5°	 5m, 10°
SIFT	SIFT	29.79	43.89	46.74	45.55	53.40	63.87
	PUMP (S)	33.81	55.35	63.26	67.02	76.96	90.58
	PUMP (S+U)	**34.94**	**58.02**	**67.37**	**73.30**	**86.91**	**97.91**
	Abs. gain	↑ **+1.1**	↑ **+2.7**	↑ **+4.1**	↑ **+6.2**	↑ **+10.0**	↑ **+7.3**
R2D2	R2D2	33.17	75.53	83.84	72.25	85.86	97.91
	PUMP (S)	37.46	83.38	91.46	69.63	84.82	96.86
	PUMP (S+U)	**37.83**	**84.16**	**92.42**	**73.30**	**86.91**	**98.43**
	Abs. gain	↑ **+0.4**	↑ **+0.8**	↑ **+1.0**	↑ **+6.2**	↑ **+2.1**	↑ **+1.6**
SuperPoint	SuperPoint	27.03	65.22	75.54	70.16	86.91	97.91
	PUMP (S)	32.48	71.44	78.81	67.54	81.68	93.19
	PUMP (S+U)	**33.36**	**73.41**	**81.4**	**74.35**	**87.96**	**98.43**
	Abs. gain	↑ **+0.9**	↑ **+2.0**	↑ **+2.6**	↑ **+6.2**	↑ **+6.3**	↑ **+5.2**

Table 3. Mean Matching Accuracy (MMA) on HPatches and percentage of localized queries on Aachen-Night within three error thresholds, with different sparse keypoint detectors. Absolute gain shows the performance increase when training with unsupervised pairs (S+U) compared to self-supervised pairs only (S).

4.3. Sparse keypoint-based matching

We evaluate the impact of the matching priors leveraged during training in a sparse matching setting by comparing again the performance achieved by the PUMP (S) and **PUMP (S+U)** models. Since our method produces dense descriptor maps, we need to resort to an external keypoint detector to select repeatable locations in the image scale-space. To make the evaluation as comprehensive as possible, we measure the performance for 3 standard detectors: SIFT [34], R2D2 [48] and SuperPoint [10]. Note that for each detector, we extract descriptors for each method at the exact same locations and scales, making the evaluation fair and strictly centered on the descriptors.

We perform a comprehensive study of the overall descriptor quality by evaluating jointly on two complementary tasks, namely in terms of keypoint matching on HPatches and localization accuracy on Aachen-Night. For HPatches, we follow the experimental protocol by [11] and measure the Mean Matching Accuracy (MMA). The MMA corresponds to the average percentage of correct matches for all image pairs w.r.t. a specified error threshold in pixels. Visual localization performance is measured as the percentage of queries successfully localized w.r.t. specified thresholds on the camera position and orientation. Table 3 reports results for each keypoint detector and each descriptor on both benchmarks. We first note that our models, including the self-supervised model (S), significantly outperform their respective keypoint baselines on HPatches. Interestingly, this does not translate into localization accuracy: in fact the self-supervised model constantly yields an inferior localization

Train	+ Style	Loss	HPatches			Aachen Day-Night		
data	Transfer?		MMA@1↑	MMA@3↑	MMA@5↑	0.25m, 2°	0.5m, 5°	5m, 10°
SfM120k		S	37.46	83.38	91.46	69.63	84.82	96.86
SfM120k		S+U	37.83	84.16	92.42	**73.30**	86.91	**98.43**
SfM120k	✓	S	**37.97**	84.77	**92.67**	72.77	86.91	**98.43**
SfM120k	✓	S+U	37.64	**84.96**	**92.97**	**73.30**	87.43	**98.43**
Aachen	✓	S+F	36.38	83.77	91.49	72.77	**89.01**	**98.43**

Table 4. Ablative study on HPatches and Aachen Day-Night on the training set and supervision level. We evaluate the impact of training on SfM120k with or without image pairs generated using style-transfer, or using the R2D2 Aachen training set with full supervision. Self-, Un- and Fully- supervised losses are respectively denoted as S, U and F. The best and second best results are resp. in **bold** and underlined.

accuracy compared to baseline keypoints. This discrepancy is explained by the fact that self-supervision covers well simple transformation like homographies, but fails to model more realistic changes. In contrast, the model trained with unsupervised pairs (S+U) largely outperforms the self-supervised model by 6 points on average and all baseline keypoints as well, despite being trained *without pixel-level supervision*. This clearly demonstrates that injecting a powerful yet unsupervised prior during training helps the model to establish hard, realistic correspondences, and makes an important difference on challenging tasks.

Figure 5 compares our performance with the state of the art on HPatches (we use R2D2's keypoint detector in this case). Our approach significantly outperforms all state-of-the-art methods, including the recent Sparse NCNet [51] and the self-supervised method CAPS-U [41].

4.4. Advanced augmentations and full supervision

Most state-of-the-art approaches for descriptor learning are currently trained with full-supervision at the correspondence level, either thanks to external supervision or from self-supervision with advanced data augmentation techniques [41, 48]. In order to evaluate the individual impact of these components w.r.t. our method, we perform a joint study on the HPatches and Aachen Day-Night benchmarks. Specifically, we consider automated style-transfer [30] as an advanced data augmentation technique, since it has been shown effective to learn robust descriptors. We append style-transfer pairs downloaded from the R2D2 [48] official training set that specifically target Day-Night illumination changes on Aachen images. We also consider the full R2D2 training set, mostly composed of Aachen images (75%), which also comprises fully-supervised pairs pre-computed using a complex flow estimation pipeline. To establish a fair comparison, each time we retrain the ConvMixer backbone model from scratch using the same hyper-parameters and loss functions (when applicable) for every training set. Results are reported in Table 4. The first and second rows correspond to the models used in all previous experiments. We observe that adding style-transfer pairs to the SfM120k training set results in a steady performance increase for both models. However, our unsupervised approach without style-transfer performs overall on par with a self-supervised approach augmented with style-transfer pairs. While it is extremely difficult to estimate the proportion of Day-Night pairs in SfM120k, such pairs certainly exist since photos are taken at different times of the day. This shows that our unsupervised approach can exploit these difficult pairs, resulting in a significant improvement of the matching robustness. Finally, we point out that our unsupervised approach trained with additional style-transfer pairs performs overall better than a fully-supervised approach specifically trained on a dedicated Aachen-centered dataset in identical conditions. Indeed, while performance of the two methods on Aachen Day-Night are on par, our method significantly outperforms the fully-supervised approach on HPatches.

4.5. Limitations

While our approach does not require any supervision given real image pairs, it still needs to receive pairs depicting the same scene or object. While these are theoretically easy to collect using *e.g.* image retrieval methods, this remains to be demonstrated. Furthermore, despite outperforming the state of the art, PUMP might still fail in classical challenging cases such as untextured areas or repetitive patterns, especially in the absence of matching priors at test time for sparse matching.

5. Conclusion

Learned pixel descriptors have become the gold standard for multiple vision tasks such as SfM and visual localization. Their training nevertheless typically requires large amounts of ground-truth annotations, which is bothersome and expensive to collect, *e.g.* using SfM techniques themselves relying on local image features. In this work, we showcased the feasibility of learning discriminative and robust local descriptors in an unsupervised setting. We foresee a great increase in the amount and diversity of potential new sources of training data where the SfM pipeline currently fails, thus expanding the range of possible applications.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 1, 2, 5

[2] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *ICCV*, 2021. 1

[3] Gabriele Moreno Berton, Carlo Masone, Valerio Paolicelli, and Barbara Caputo. Viewpoint invariant dense matching for visual geolocalization. In *ICCV*, 2021. 1

[4] Aritra Bhowmik, Stefan Gumhold, Carsten Rother, and Eric Brachmann. Reinforced feature points: Optimizing feature detection and description for a high-level task. In *CVPR*, 2020. 2, 3

[5] Jia-Wang Bian, Wen-Yan Lin, Yun Liu, Le Zhang, Sai-Kit Yeung, Ming-Ming Cheng, and Ian Reid. GMS: grid-based motion statistics for fast, ultra-robust feature correspondence. *IJCV*, 2020. 2

[6] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. Robotics*, 2016. 1

[7] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *ICCV*, 2021. 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[9] François Darmon, Mathieu Aubry, and Pascal Monasse. Learning to guide local feature matches. In *3DV*, 2020. 3

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, 2018. 2, 7

[11] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint description and detection of local features. In *CVPR*, 2019. 7

[12] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2DNet: Learning image features for accurate sparse-to-dense matching. In *ECCV*, 2020. 2, 3

[13] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Neural reprojection error: Merging feature learning and camera pose estimation. In *CVPR*, 2021. 3

[14] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Visual correspondence hallucination. In *ICLR*, 2022. 3

[15] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. PackNet-SfM: 3D packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 1

[16] Christopher G. Harris and Mike Stephens. A combined corner and edge detector. In *AVC*, 1988. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[18] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *CVPR*, 2018. 1, 2

[19] Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In *ICCV*, 2021. 3, 6

[20] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar de Souza, Vincent Leroy, and Gabriela Csurka. Robust image retrieval-based visual localization using kapture. *arXiv preprint arXiv:2007.13867*, 2020. 5

[21] Philip Torr James Thewlis, Shuai Zheng and Andrea Vedaldi. Fully-trainable deep matching. In *BMVC*, 2016. 3

[22] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *ICCV*, 2021. 3, 6

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

[24] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, 2019. 1

[25] Jogendra Nath Kundu, M. V. Rahul, Aditya Ganeshan, and R. Venkatesh Babu. Object pose estimation from monocular image using multi-view keypoint correspondence. In *ECCV Workshop*, 2018. 1

[26] Axel Barroso Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.net: Keypoint detection by handcrafted and learned CNN filters. In *ICCV*, 2019. 1, 2

[27] Zakaria Laskar, Iaroslav Melekhov, Hamed R. Tavakoli, and Juha Ylioinas. Geometric image correspondence verification by dense pixel matching. In *WACV*, 2020. 2

[28] Shuda Li, Kai Han, Theo W. Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *CVPR*, 2020. 3

[29] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *NeurIPS*, 2020. 3

[30] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *ECCV*, 2018. 8

[31] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1

[32] Wen-Yan Lin, Fan Wang, Ming-Ming Cheng, Sai-Kit Yeung, Philip H. S. Torr, Minh N. Do, and Jiangbo Lu. CODE: coherence based decision boundaries for feature correspondence. *IEEE Trans. PAMI*, 2018. 2

[33] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. GIFT: learning transformation-invariant dense visual descriptors via group cnns. In *NeurIPS*, 2019. 2

[34] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 1, 2, 7

[35] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *CVPR*, 2019. 3

[36] Zixin Luo, Tianwei Shen, Lei Zhou, Siyu Zhu, Runze Zhang, Yao Yao, Tian Fang, and Long Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *ECCV*, 2018. 3

[37] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *CVPR*, 2020. 2

[38] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *IJCV*, 2021. 2

[39] Jiayi Ma, Ji Zhao, Junjun Jiang, Huabing Zhou, and Xiaojie Guo. Locality preserving matching. *IJCV*, 2019. 2

[40] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018. 3

[41] Iaroslav Melekhov, Zakaria Laskar, Xiaotian Li, Shuzhe Wang, and Juho Kannala. Digging into self-supervised learning of feature descriptors. In *3DV*, 2021. 2, 8

[42] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. DGC-Net: Dense geometric correspondence network. In *WACV*, 2019. 2, 3, 6

[43] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *ECCV*, 2018. 1, 2

[44] Hans P. Moravec. Rover visual obstacle avoidance. In *IJCAI*, 1981. 2

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017. 5

[46] Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 5

[47] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019. 5

[48] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1, 2, 5, 7, 8

[49] Jérôme Revaud, Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. DeepMatching: Hierarchical deformable dense matching. *IJCV*, 2016. 2, 3, 4, 5

[50] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. *IEEE Trans. PAMI*, 2019. 2

[51] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *ECCV*, 2020. 3, 8

[52] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 3

[53] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: an efficient alternative to SIFT or SURF. In *ICCV*, 2011. 1, 2

[54] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 3

[55] Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 2017. 5

[56] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[57] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 5

[58] Xi Shen, François Darmon, Alexei A. Efros, and Mathieu Aubry. RANSAC-Flow: generic two-stage image alignment. In *ECCV*, 2020. 3, 6

[59] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 3

[60] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomás Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. *IEEE Trans. PAMI*, 2021. 1

[61] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 6

[62] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *CVPR*, 2019. 1, 2

[63] Engin Tola, Vincent Lepetit, and Pascal Fua. A fast local descriptor for dense matching. In *CVPR*, 2008. 1

[64] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 5

[65] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *CVPR*, 2021. 3

[66] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-Local Universal Network for Dense Flow and Correspondences. In *CVPR*, 2020. 3, 6

[67] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *ICCV*, 2021. 1, 2, 3, 6

[68] Yannick Verdie, Kwang Moo Yi, Pascal Fua, and Vincent Lepetit. TILDE: A temporally invariant learned detector. In *CVPR*, 2015. 1, 2

[69] Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu, Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, and Andrew Markham. P2-net: Joint description and detection of local features for pixel and point matching. In *ICCV*, 2021. 1, 2

[70] Jianyuan Wang, Yiran Zhong andYuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *CVPR*, 2021. 1

[71] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *ECCV*, 2020. 2, 3

[72] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 1

[73] Guandao Yang, Tomasz Malisiewicz, and Serge J. Belongie. Learning data-adaptive interest points through epipolar adaptation. In *CVPR Workshops*, 2019. 3

[74] Heng Yang, Wei Dong, Luca Carlone, and Vladlen Koltun. Self-supervised geometric perception. In *CVPR*, 2021. 2, 3

[75] Yuan Yao, Yasamin Jafarian, and Hyun Soo Park. MONET: multiview semi-supervised keypoint detection via epipolar divergence. In *ICCV*, 2019. 3

[76] Xu Zhang, Felix X. Yu, Svebor Karaman, and Shih-Fu Chang. Learning discriminative and transformation covariant local feature detectors. In *CVPR*, 2017. 1, 2

[77] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *IJCV*, 2021. 5

[78] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixé. Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*, 2021. 3

[79] Tinghui Zhou, Yong Jae Lee, Stella X. Yu, and Alexei A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *CVPR*, 2015. 2, 3