This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Learning Affine Correspondences by Integrating Geometric Constraints

Pengju Sun[1, 2] Banglei Guan[1, 2(✉)] Zhenbao Yu[1, 2] Yang Shang[1, 2] Qifeng Yu[1, 2] Daniel Barath[3, 4]

[1]College of Aerospace Science and Engineering, National University of Defense Technology, China.
[2] Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, China.
[3]ETH Zurich, Switzerland. [4] HUN-REN SZTAKI, Hungary.

Abstract

Affine correspondences have received significant attention due to their benefits in tasks like image matching and pose estimation. Existing methods for extracting affine correspondences still have many limitations in terms of performance; thus, exploring a new paradigm is crucial. In this paper, we present a new pipeline designed for extracting accurate affine correspondences by integrating dense matching and geometric constraints. Specifically, a novel extraction framework is introduced, with the aid of dense matching and a novel keypoint scale and orientation estimator. For this purpose, we propose loss functions based on geometric constraints, which can effectively improve accuracy by supervising neural networks to learn feature geometry. The experimental show that the accuracy and robustness of our method outperform the existing ones in image matching tasks. To further demonstrate the effectiveness of the proposed method, we applied it to relative pose estimation. Affine correspondences extracted by our method lead to more accurate poses than the baselines on a range of real-world datasets. The code is available at https://github.com/stilcrad/LearningACs.

1. Introduction

In computer vision, image matching and geometric estimation stand as fundamental problems, playing crucial roles in domains ranging from autonomous driving to robotics [8, 24, 36]. Affine correspondences (ACs) have attracted significant attention in recent years, due to their ability to provide valuable insights into the underlying 3D geometry of the surrounding environment [3, 5, 22]. Notably, previous research has demonstrated the efficacy of affine correspondences in tasks such as homography, epipolar geometry, and focal length estimation [7, 30, 50, 65].

(a) LoFTR (b) DKM (c) Ours

Figure 1. Image matching with large viewpoint change. Correct matches are green, and incorrect ones red. Our method leads to more correct matches than the LoFTR [56] and DKM [16].

Affine correspondences offer distinct advantages in addressing fundamental challenges in visual perception, particularly in image-matching and relative pose estimation tasks, owing to its capacity to encode higher-order information about the scene geometry [8, 19, 24]. By locally approximating the image deformation caused by changes in camera pose through affine mapping, geometric information about corresponding local regions are obtained [26, 45]. These approximations enhance the robustness of affine covariants for matching and recognition tasks [41]. It is a valuable property for matching planar surfaces in the presence of extreme viewpoint changes, improving the reliability of wide baseline image matching [47]. Moreover, affine correspondences are used to estimate complex geometric relationships between images, such as essential matrices, outperforming the results when relying solely on point correspondences [5, 8]. Benefiting from the smaller sample sizes required, robust homography and relative pose estimation is significantly faster than when using the point-based solvers while leading to more accurate results [5, 8]. Through exploiting these informative affine correspondences, such algorithms can attain enhanced precision and reduced runtime [8]. The reduction in the number of matches needed to estimate a model, *e.g.*, homography, leads to advantages, like

✉: Corresponding author.
Email: {sunpengju23, guanbanglei12, shangyang1977, yuqifeng}@nudt.edu.cn zhenbaoyu@whu.edu.cn dbarath@ethz.ch

reduced computational complexity and better efficiency of the outlier removal process [23, 26, 27].

Obtaining high-quality affine correspondences in real-world scenarios remains an open problem [3, 39]. There are many limitations to existing methods, such as the limited quantity and accuracy. These drawbacks arise from many of these methods that use detector-based techniques and do not fully exploit geometric constraints. Existing extractors that use sparse detectors perform poorly with repetitive and weak textures [47, 60]. View-synthesis-based methods, such as ASIFT [48], rely not only on the detector, but also involve computationally expensive image transformations. These limitations severely affect the use of affine correspondences in tasks related to geometric estimation [12].

In recent years, dense matching with neural networks has been shown to effectively overcome the limitations imposed by traditional detection methods [16, 56], which rely on detected sparse keypoints [47, 48, 60]. Having a dense warp between the two images allows one to extract Aundant accurate keypoints. Due to the use of global context, such methods also excel in weakly textured regions and repetitive structures. At the same time, geometric constraints are also proving their effectiveness in matching tasks [61].

This paper presents a novel pipeline for robust affine correspondence estimation through synergistic integration of dense feature matching and geometric constraint optimization. This approach allows us to extract a large number of accurate corespondences even between images with large viewpoint changes, as shown in Fig.1. In summary, the main contributions are as follows.

- We propose a novel framework for estimating affine correspondences. By combining a dense matcher, geometric constraints, and a local affine transformation extractor using a soft scale and orientation estimator, The framework surpasses state-of-the-art in match number and accuracy.
- A novel affine transformation loss, represented by the Affine Sampson Distance, is introduced to further enhance the conformity of affine correspondences with the scene geometry. This approach enhances training supervision, improving affine correspondence quality.
- We show that the proposed model is applicable to a range of matching tasks, producing high-quality affine correspondences and achieving state-of-the-art results on both image-matching and relative pose estimation benchmarks, indicating the robustness, performance, and applicability of the method in practical scenarios.

2. Related Work

2.1. Image Matching

Classical image matching involves three key steps: keypoint detection, descriptor extraction, and correspondence estimation [31]. Keypoints are traditionally detected using scale pyramids with handcrafted response functions [35]. Feature matching relies on optimizing metrics like Sum of Squared Differences or correlation. [63]. Examples of such functions include the Hessian [9], Harris [28], Difference of Gaussians (DoG) [35], as well as learned ones such as SOSNet [57], Key.Net [32], SuperPoint [14], PoSFeat [33], and DKDNet [20]. SuperPoint adopts a detector-based architecture similar to handcrafted methods and proposes a self-supervised training approach through homography adaptation, yielding improved performance. Subsequently, several algorithms have been devised based on these paradigms [17, 51, 59]. DKDNet integrates a dynamic keypoint feature learning module and a guided heatmap activator to enhance the performance of keypoint detection [20]. These methods depend on the efficacy of the designed detectors, encountering limitations in scenarios with repetitive structures or weak textures.

Concurrent with detector-based matching, another line of works [14, 15, 51] focus on generating matches directly from raw images, where richer context can be leveraged and the keypoint detection step skipped. They execute global matching uniformly across the image grid at a coarse scale and extract matches via mutual-nearest neighbors or optimal transport [52, 56, 62, 66]. In contrast to detector-free methods, dense methods generate a dense warp. This warp is typically predicted by regression based on the global 4D-correlation volume [58]. DKM [16] introduces a dense kernelized matching approach that significantly improves two-view estimation. Building on this, RoMa [18] represents a significant advancement in dense feature matching by applying a Markov chain framework to analyze and improve the matching process. GIM [55] is a self-training method for matching using internet videos. Dense approaches have the capability to estimate matching pixel pairs, providing a foundation for obtaining precise affine correspondences.

2.2. Affine Correspondence Estimation

An affine correspondence consists of a pair of points along with the corresponding local affine transformation that maps the neighborhood of a point in the first image to its counterpart in the second image [5, 50]. Affine covariant detectors are commonly used to estimate affine transformations [9]. These detectors typically fall into three categories. The first category includes methods like Maximally Stable Extremal Regions (MSER) [40], which directly estimate full local affine transformations from image regions. The second category consists of detectors such as Harris-Affine [42] and Hessian-Affine [41], which refine initial estimates using iterative methods such as Baumberg iteration [9], resulting in high-quality affinities. Some methods in the third category synthesize views related by affine transformations and then apply feature detectors to these synthetic images [46, 48]. Each point correspondence gen-

erates a local affinity, which is integrated with the synthetic view transformation to form the final affine feature.

In recent years, deep learning-based feature matching has shown significant advances in performance and robustness. AffNet [47] is proposed to demonstrate that repeatability is not enough. It learns local affine-covariant regions by optimizing a descriptor-based loss. AffNet outperforms prior methods in affine shape estimation and enhances the state-of-the-art in image retrieval. LOCATE [53] incorporates local affine maps between corresponding keypoints, substantially improving the accuracy of local geometry estimation. AEU [12] is introduced to enhance feature matching accuracy by estimating relative affine transformations between features, making it more robust to disturbances.

Although these methods leverage neural networks to estimate local affine transformations and enhance image matching accuracy, their still rely on sparse detectors. As a result, they inherit limitations such as sensitivity to low-textured regions and repetitive patterns.

In contrast, our approach adopts a dense matching strategy, enabling accurate correspondences.

3. Proposed Method

To obtain accurate affine correspondences, it is essential to ensure the precision of each component involved, including high-precision point correspondences and local affine transformations. In this section, a new framework is proposed for extracting affine correspondences. We present an overview of the pipeline in Fig. 2. Taking an image pair I^A, I^B as input, the network produces reliable affine matches.

3.1. Preliminary

Let us assume that we are given a pair of images I^A and I^B and corresponding patches $patch^A$ and $patch^B$ in the images. Assuming that the objects captured in these patches are flat surfaces, there exists a linear transformation $\mathbf{A}$ satisfying $patch^A = \mathbf{A}\ patch^B$, where $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is usually called a local affine transformation, described as $(a_{11}, a_{12}, a_{21}, a_{22})$, where a_{ij} are the elements in a row-major order ($i, j \in \{1, 2\}$).

There are multiple ways to decompose an affine transformation. In this paper, we utilize the decomposition proposed in [47], which decomposes to scale, orientation, and affine shape matrix $\mathbf{A}'$. $\mathbf{A}'$ is the affine shape matrix with det($\mathbf{A}'$) = 1, which could be decomposed into identity matrix $\mathbf{I}$ and the residual shape $\mathbf{A}''$. The parameterizations of the affine transformation have a significant impact on the performance of local geometric estimators, as shown in [47]. Suppose that an affine correspondence ($\mathbf{p_1}$, $\mathbf{p_2}$, $\mathbf{A}$) and fundamental matrix $\mathbf{F}$ is known. It is trivial that every affine transformation preserves the direction of the lines going through points $\mathbf{p_1}$ and $\mathbf{p_2}$ in the first and second images [5, 7], where $\mathbf{p}_1 = [x_1, y_1, 1]^T$ and $\mathbf{p}_2 = [x_2, y_2, 1]^T$ represent a homogeneous form of point correspondence.

The geometric relationship of $\mathbf{p_1}$, $\mathbf{p_2}$, $\mathbf{F}$, and $\mathbf{A}$ is as:

$$\left(\mathbf{F}^T\mathbf{p}_1\right)_{(1:2)} + \left(\mathbf{A}^T\mathbf{F}\mathbf{p}_2\right)_{(1:2)} = 0. \tag{1}$$

The above equation only holds for the upper two rows and provides two linear equations.

3.2. Affine Correspondence Estimation

Each affine correspondence consists of a point correspondence and a local affine transformation. To improve the accuracy, we design a separate two-stage framework for optimization. Our overall process is first to extract the point matches and then estimate the affine transformation of their local patches. Unlike the traditional sparse feature matching method, we extract a large number of accurate keypoints from a dense warp and add a loss function with epipolar constraints to enable the network to learn more geometric information. The first sub-network is responsible for acquiring point correspondences and is employed to learn dense matching. Then, we obtain the corresponding patches according to the corresponding points. The second sub-network runs to estimate accurate local affine transformations, which are calculated through the scale, orientation, and residual shape. The final affine correspondences consist of these point matches and their affine transformations.

3.2.1. Feature Matching Module

In the first sub-network, we consider the task of estimating the accurate point correspondences from two images $\left(I^A, I^B\right)$. Inspired by DKM [16], we choose the dense matching paradigm to estimate a dense warp $W^{A\to B}$ and a dense certainty $p^{A\to B}$, which represents a dense warp and the probability of correct matches, respectively. Feature maps are extracted from I^A and I^B with a ResNet50 [29] encoder, which is pre-trained on ImageNet-1K [13]. The initial encoding process could be described as

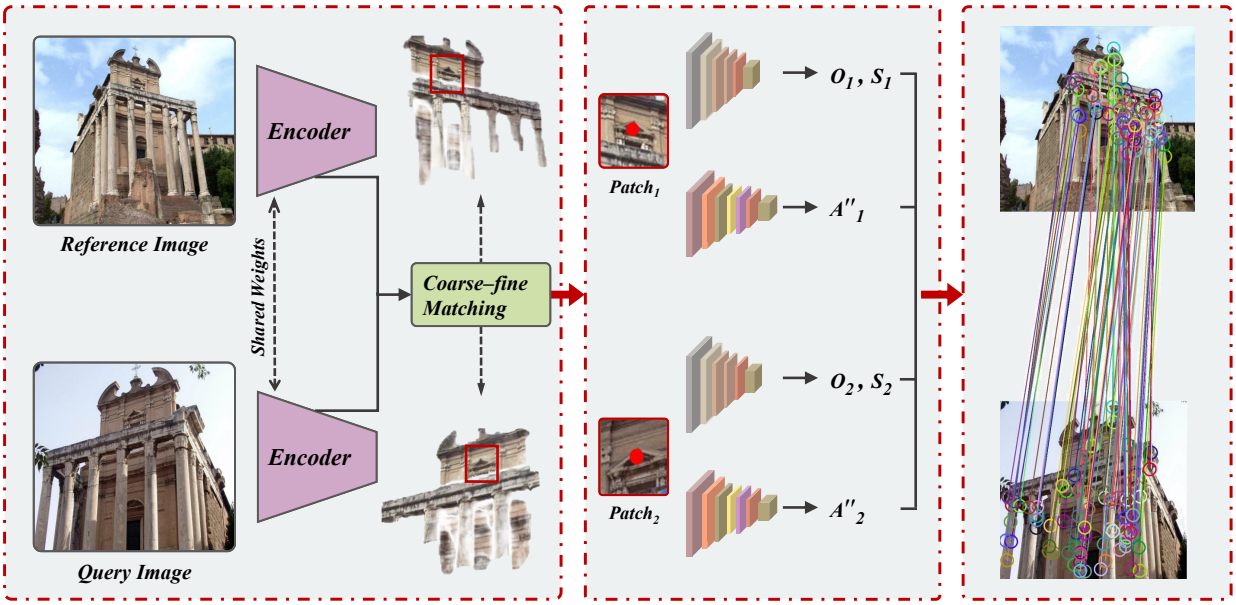
$$\left(\Phi^A_{coarse}, \Phi^A_{fine}\right) = E_\theta\left(I^A\right), \tag{2}$$

$$\left(\Phi^B_{coarse}, \Phi^B_{fine}\right) = E_\theta\left(I^B\right), \tag{3}$$

where Φ^A_{coarse}, Φ^A_{fine}, Φ^B_{coarse}, and Φ^B_{fine} are feature maps obtained at different network depths. Function E_θ represents the encoder. After obtaining feature maps, a suitable regression framework is used to infer the mapping relationship of pixels. Through a global matcher, we can obtain the coarse-level warp and certainty, which is written as follows:

$$\left(\hat{W}^{A\to B}_{\text{coarse}}, \hat{p}^{A\to B}_{\text{coarse}}\right) = G_\theta\left(\Phi^A_{\text{coarse}}, \Phi^B_{\text{coarse}}\right), \tag{4}$$

where G_θ is the kernel regression global matcher, which generates robust coarse matches, using an embedded Gaussian process regression. Parameter $\hat{W}^{A\to B}_{\text{coarse}}$ and $\hat{p}^{A\to B}_{\text{coarse}}$ are the coarse-level warp and certainty.



(a) Dense Matching Subnetwork (b) Affine Shape Subnetwork (c) Affine Correspondences

Figure 2. The overview of our method. (a) Abundant accurate point correspondences, encouraged to comply with epipolar constraints through the training loss, are obtained via a dense matching sub-network. (b) The second sub-network is used to estimate the orientation $\mathbf{O}_i$ and scale $\mathbf{S}_i$ of each patch and estimate the residual shape $\mathbf{A}_i''$. (c) Affine correspondences between the two images are estimated.

Meanwhile, to utilize the global features of the image, cosine encoding is used to enhance the ability to match weak textures and repetitive structures. The refining process predicts a residual discrepancy for the projected warp and a logit discrepancy for certainty. This process is reiterated until the highest resolution is attained as follows:

$$\left(\hat{W}_{\text{fine}}^{A\to B}, \hat{p}_{\text{fine}}^{A\to B}\right) = R_\theta\left(\Phi_{fine}^{A}, \Phi_{fine}^{B}, \hat{W}_{\text{coarse}}^{A\to B}, \hat{p}_{\text{coarse}}^{A\to B}\right), \tag{5}$$

where $\hat{W}_{\text{fine}}^{A\to B}$ and $\hat{p}_{\text{fine}}^{A\to B}$ are the predicted warp and certainty in fine-level. Function R_θ is a set of refiners that uses stacked feature maps and depthwise convolution kernels. Finally, the point correspondences with their patches are obtained as follows:

$$\left(Patch_l^A, Patch_l^B\right) = S_\theta\left(\hat{W}_{\text{fine}}^{A\to B}, \hat{p}_{\text{fine}}^{A\to B}\right), \tag{6}$$

where Patch_l^A and Patch_l^B are the corresponding patches cropped from $\left(I^A, I^B\right)$ with the size l. Function S_θ is a sampler, selecting the matches. We obtain point correspondences by using warp and probability on each pixel. The final patch pairs are generated around these corresponding points, with a fixed patch size of 32*32 pixels.

3.2.2. Local Affine Transformation Estimation Module

After obtaining point correspondences and patches centered on these points using the above module, local affine transformations for these regions are estimated. The affine transformation is decomposed according to [47]. The process of obtaining angles and scales is described as follows:

$$\left(O_i^A, S_i^A\right) = E_{o,s}\left(patch_l^A\right), \tag{7}$$

$$\left(O_i^B, S_i^B\right) = E_{o,s}\left(patch_l^B\right), \tag{8}$$

$$O_i^{A\to B} = O_i^B - O_i^A, \quad S_i^{A\to B} = S_i^B / S_i^A, \tag{9}$$

where O_i^A, S_i^A, O_i^B, and S_i^B represent the orientations and scales of patch pairs. $E_{o,s}$ consists of two independent fully connected networks, by which we compute the relative scale and orientation of each patch correspondence. By discretizing angles and scales, this network can predict the probability distribution of patches at discrete angles and scales. The scale and angle with the highest probability are considered the predicted results. The extraction of angles and scales is based on the SoTA method [64]. Through probabilistic covariant loss, the prediction accuracy of scale and direction is higher than that of traditional methods.

The final part is to calculate the residual shape, which is described as follows:

$$\mathbf{A}_i'' = E_{\text{aff}}\left(patch_l^A, patch_l^B\right), \tag{10}$$

where the residual shape $\mathbf{A}_i''$, as described in [47], is computed via an independent fully connected network E_{aff} used to regress the final residual shape. Finally, the affine correspondences are computed as represented through

$$ACs = \Psi_{syn}(P_i^A, P_i^B, O_i^{A\to B}, S_i^{A\to B}, A_i''), \tag{11}$$

where Ψ_{syn} is the process of synthesizing affine correspondences according to [47] .

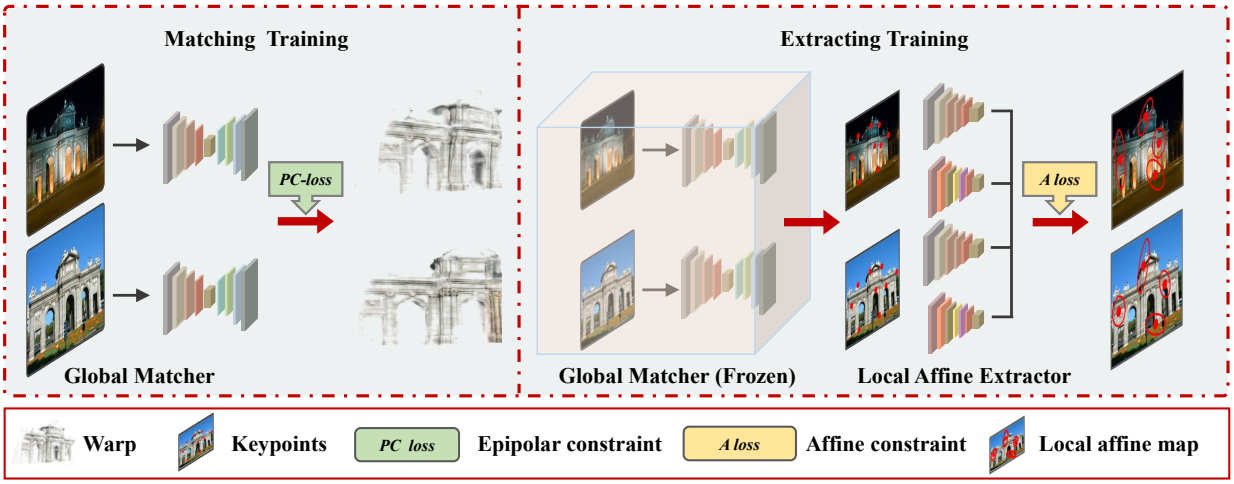


Figure 3. **The training pipeline.** The network starts with training only the dense point matcher supervised by the proposed Sampson distance-based point correspondence loss. This network extracts a dense warp between the images. Next, the point matcher sub-network is frozen, and we train the affine shape extractor to minimize the proposed affine loss, leveraging the epipolar geometry-based constraints.

3.3. Loss Function

The epipolar loss has been demonstrated to be an effective way to improve matching performance [61]. To facilitate accurate detection, we further improve it by using the Sampson distance and also a loss based on affine features.

For training the matching, we make sure that the point correspondences agree with the epipolar geometry, and use loss L_{pc} that is as follows:

$$L_{\text{pc}}\left(\hat{PC}^{A\to B}\right) = \frac{1}{N}\sum_{i=1}^{N} SD_P(E^i_{PC}), \tag{12}$$

$$E_{PC} = \mathbf{p}_2^T \mathbf{F} \mathbf{p}_1, \tag{13}$$

where $\mathbf{p}_1 = (x_1, y_1, 1)^T$, $\mathbf{p}_2 = (x_2, y_2, 1)^T$, and the $\mathbf{F}$ represent the fundamental matrix, i is the index of the point correspondence, function *SD* calculates the Sampson Distance, $\hat{PC}^{A\to B}$ is the point correspondences, *N* is the number of the correspondences, and $SD_P(E_{PC})$ is the Sampson Distance calculated from the epipolar constraint. The derivation of the Sampson Distance metric for epipolar constraints is put in the supplementary material.

For training the matcher, the used loss is as follows:

$$L_{\text{m}} = \sum_{l=1}^{L} L_{\text{warp}}\left(\hat{W}^{A\to B}\right) + \lambda L_{\text{conf}}\left(\hat{p}^{A\to B}\right) + \gamma L_{\text{pc}}\left(\hat{PC}^{A\to B}\right) \tag{14}$$

where λ and γ are the balancing terms and set to 0.01 empirically. Specifically, for the warp loss L_{warp}, we use the L_2 distance of the predicted $W_l^{A\to B}$ and the ground truth warp $\hat{W}_l^{A\to B}$ as follows:

$$L_{\text{warp}}\left(\hat{W}^{A\to B}\right) = \sum_{\text{grid}} p_l \odot \left\| W^{A\to B} - \hat{W}^{A\to B} \right\|_2. \tag{15}$$

For confidence loss L_{conf}, we use the unweighted binary cross entropy between the prediction confidence $\hat{p}_l$ and the ground truth p_l written as follows:

$$L_{\text{conf}}\left(\hat{p}\right) = \sum_{\text{grid}} p \log \hat{p} + (1-p)\log\left(1-\hat{p}\right). \tag{16}$$

Global matchers can provide precise point correspondences, but they are insufficient to obtain accurate affine correspondences. To train the affine sub-network, we introduce the novel affine constraint loss, which is crucial for learning the correct affine shape. It is designed to quantify how well our predicted shape complies with the epipolar geometry. Minimizing the affine constraint loss enables the network to estimate local affine shapes accurately. The loss function is as follows:

$$L_{\text{aff}}\left(\hat{AC}^{A\to B}\right) = -\frac{1}{N}\sum_{i=1}^{N} SD_A(E^i_{AC}), \tag{17}$$

where $\hat{AC}^{A\to B}$ is the obtained affine correspondences in images, $SD_A(E_{AC})$ is the affine transformation constraint error represented by the affine Sampson Distance, where E_{AC} is defined by Eq. 1. More specifically, it is:

$$SD_A(E_{AC})_{(1:2)} = SD_A(\mathbf{A}^{-T}\left(\mathbf{F}^T\mathbf{p}_2\right)_{(1:2)} + (\mathbf{F}\mathbf{p}_1)_{(1:2)}). \tag{18}$$

We included the derivation of the affine Sampson Distance metric in the supplementary material.

The loss function for extracting local affine shapes is as follows:

$$L_{\text{ext}} = L_{\text{aff}}\left(\hat{AC}^{A\to B}\right) + L(P,Q)_{\text{ori}} + L(P,Q)_{\text{sca}}, \quad (19)$$

where $L(P,Q)_{\text{ori}}$ and $L(P,Q)_{\text{sca}}$ are probability covariance loss in the orientation and scale [64]. We discretize continuous angles and scales to convert regression into classification. Image patches are obtained through random scaling and rotation, using the scaling factor and rotation angle as labels. The network predicts discrete distributions of scale and orientation, and the loss function is constructed based on their discrepancy from the ideal distribution. For orientation and scale, the loss is defined as:

$$L_{\text{ori}} = \sum_{i}^{N} P_{\text{ori}} \log Q_{\text{ori}}, \quad L_{\text{sca}} = \sum_{i}^{N} P_{\text{sca}} \log Q_{\text{sca}}, \quad (20)$$

respectively, where P_i is the true discrete scale or orientation distribution, and Q_i is the predicted discrete distribution. These loss functions guide learning to ensure accurate and consistent affine correspondences.

3.4. Decoupled Training Pipeline

We employ a decoupled approach for training, as shown in Fig. 3. The two sub-networks are trained separately to reduce the loss ambiguity caused by weak supervision. Using different losses helps with convergence and performance.

During the first part of the training, only the first sub-network is optimized to learn accurate matches, and the affine transformation part is ignored. Training the network until the loss no longer decreases, we stop and freeze parameters. Then, the second sub-network is trained for local affine transformation extraction. This two-step training leads to faster convergence and increased final performance compared to training the two sub-networks together. In addition, decoupling training consumes less memory than joint training. This decoupled training approach avoids the loss of ambiguity caused by weak supervision strategies and greatly helps improve the performance of our network.

4. Experiments

In this section, we demonstrate superior performance compared to previous affine correspondence estimation and image matching methods. We validate the accuracy of the extracted matches by comparing with mainstream image-matching methods on the HPatches [2]. Additionally, we compare the proposed method on relative pose estimation on the KITTI [21] and MegaDepth [34] datasets with the state-of-the-art matchers. Finally, we conduct ablation studies to verify the effectiveness of proposed component.

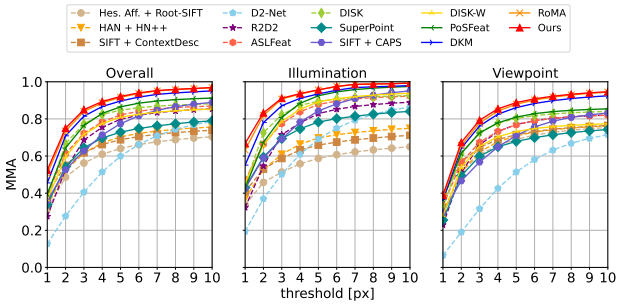


Figure 4. The mean matching accuracy (MMA; higher is better) at different thresholds (in pixels) on the HPatches Dataset [2].

4.1. Model Implementation

Dataset. We utilize the MegaDepth dataset [34] and ScanNet [11] for training, using the same training and test split as in baseline approaches [16, 56].

Implementation details. We use the matching sub-network to extract point correspondences and then select the exact corresponding points for the patches, whose size is set to 32*32 pixels. We train the patches for the subsequent affine shape extraction network with different loss functions, as mentioned in Section 3.3. The model is trained on ScanNet (indoor) and MegaDepth (outdoor) datasets separately. The pre-trained model is used for weight initialization, and only the weights of the Refiner module are fine-tuned during training. We trained the model using PC-loss by randomly sampling 2k points. The AdamW optimizer with a weight-decay of 10^{-2} is used. Then, the first sub-network is frozen while training the second one from scratch. The loss function for extracting affine shapes is employed. We employ SGD optimization with an initial learning rate of 0.0001 and adopt a learning rate decay strategy.

4.2. Image Matching with Affine Correspondences

As our first experiment, we evaluate our method on the widely used HPatches dataset [2]. Consistent with the approach in D2-Net [15, 33], we exclude 8 high-resolution scenes, leaving 52 scenes with illumination variations and 56 scenes with viewpoint changes for evaluation.

Evaluation protocol. We follow the setup proposed in Posfeat [33] and report the Mean Matching Accuracy (MMA) [43] under thresholds varying from 1 to 10 pixels. We use a weighted sum of MMA at different thresholds for overall evaluation [33] as follows:

$$\text{MMAscore} = \frac{\sum_{thr\in[1,10]}(2-0.1\cdot thr)\cdot \text{ MMA } @thr}{\sum_{thr\in[1,10]}(2-0.1\cdot thr)}. \quad (21)$$

Result. As shown in Fig. 4 and Table 1, the proposed method achieves the highest MMA scores on both the illumination and viewpoint sequences. Our method out-

Table 1. The weighted sum of mean matching accuracies at multiple thresholds (MMAscore, Eq. 21; higher is better) obtained by the baseline methods and the proposed one on the illumination and viewpoint sequences of the HPatches dataset [2] separately, and on all. The best results are shown in bold.

Methods	MMAscore ↑ Overall	MMAscore ↑ Illumination	MMAscore ↑ Viewpoint
Hes. Aff. [9] + Root-SIFT [1]	0.584	0.544	0.624
HAN [9] + HN++ [44]	0.633	0.634	0.633
SIFT [35] + ContextDesc [37]	0.636	0.613	0.657
D2Net [15]	0.519	0.605	0.440
R2D2 [51]	0.695	0.727	0.665
ASLFeat [38]	0.739	0.795	0.687
DISK [59]	0.763	0.813	0.716
DELF [49]	0.571	0.903	0.262
SuperPoint [14]	0.658	0.715	0.606
SIFT [35] + CAPS [61]	0.699	0.764	0.639
DISK-W [59]	0.719	0.803	0.649
PoSFeat [33]	0.775	0.826	0.728
DKM [16]	0.819	0.869	0.772
RoMA [18]	0.843	0.901	0.789
Ours	**0.851**	**0.908**	**0.798**

performs the previous SoTA [18] and achieves substantially better overall performance of 0.851 MMAscore. This demonstrates that accurate local affine shapes can provide additional cues for matching that further improve accuracy.

4.3. Improvement in the Affine Frames

While in the previous experiment, we focused on showcasing the improved accuracy of the point correspondences, now we demonstrate the accuracy of the affine frames themselves on the HPatches dataset [2]. We extract ground truth affine shapes at each point location from the ground truth homography as proposed in [4].

We compare the proposed method with the view-synthesis-based Affine-SIFT (ASIFT) [48], the VLFeat library [60], and the learning-based AffNet [47]. We evaluate the similarity of the ground truth affine matrix and the estimated one by the Euclidean distance and cosine similarity.

Table 2 shows that the estimated affine matrices exhibit higher cosine similarity and smaller Euclidean distance than the baselines compared to the ground truth affine shapes. This demonstrates that the proposed method estimates not only accurate keypoints but also precise affine shapes.

Table 2. The accuracy of the affine shapes estimated by the VLFeat library [60], ASIFT [48], AffNet [47] and the proposed method on the HPatches dataset [2]. Reported metrics (bold=best): Euclidean distance and cosine similarity of affine matrices vs. ground truth. The best results are in bold.

	VLFeat [60]	AffNet [47]	ASIFT [48]	**Ours**
Euclidean-Distance ↓	0.202	0.264	0.329	**0.123**
Cosine-Similarity ↑	0.988	0.973	0.894	**0.994**

Table 3. Relative pose Accuracy Under the recall Curve (AUC; higher is better) thresholded at 5°, 10°, and 20° on the MegaDepth [34] dataset. All methods run RANSAC-based essential matrix estimation, except for the last row, where we run the affine-based GC-RANSAC [6, 8], benefiting from the affine correspondences that we obtain. The best results are in bold.

Method AUC @ →	5° ↑	10° ↑	20° ↑
LoFTR [56] CVPR21	52.8	69.2	81.2
ASpanFormer [10] ECCV22	55.3	71.5	83.1
PDC-Net+ [58] TPAMI23	51.5	67.2	78.5
DKM [16] CVPR23	60.4	74.9	85.1
ROMA [18] CVPR24	62.6	76.7	**86.3**
Ours (RANSAC)	63.1	76.3	85.2
Ours (affine GC-RANSAC)	**65.5**	**77.9**	**86.3**

4.4. Relative Pose Estimation

In the previous sections, we demonstrate that the proposed method extracts more accurate point and local affine transformation than the state-of-the-art approaches. Here, we demonstrate that the affine correspondences obtained by our method lead to improved relative pose accuracy compared with methods obtaining point correspondences. More experiments can be found in the supplementary materials.

4.4.1. Relative Pose Estimation on MegaDepth-1500

We use the MegaDepth-1500 dataset that consists of 1500 pairs from scene 0015 (St. Peter's Basilica) and 0022 (Brandenburg Gate) [34]. We follow the evaluation protocol in [54, 56] and use a RANSAC threshold of 0.5 pixel.

Evaluation protocol. Following [54] and [56], we report the AUC of the pose error at thresholds $(5°, 10°, 20°)$. To compare with existing methods on the same baseline, we utilize RANSAC as implemented in the OpenCV library as previous methods do [56]. To demonstrate that the estimated affine frames are beneficial for pose estimation, we also run the affine correspondence-based Graph-Cut RANSAC [6, 8], designed specifically to leverage affine shapes together with the point locations.

Results. As shown in Table 3, the proposed method with RANSAC achieves similar results to the state-of-the-art RoMA [18] matcher. When leveraging the estimated affine shapes with GC-RANSAC, the proposed method achieves the best performance. It outperforms DKM [16] by a significant 5.1 AUC points at 5°, while also being better than the recent RoMA by 2.4 AUC points. These results demonstrate that the proposed geometric constraints can significantly improve matching based on feature descriptors.

4.4.2. Relative Pose Estimation on KITTI

To further verify the accuracy of the affine correspondences, we apply the method to each consecutive pair of stereo pairs in the 11 test KITTI sequences [25]. The proposed method is compared with the most widely

Table 4. The rotation and translation RMSE of the estimated relative poses on sequences 00 to 10 of the KITTI [21] dataset. Point-based methods, DKM [16] and RoMA [18], run RANSAC-based essential matrix estimation. Affine-based methods, VLFeat [60], AffNet [47], ASIFT [48], and the proposed one, run the affine-based GC-RANSAC [6, 8], directly benefiting from the affine correspondences.

	Rotation RMSE (°) ↓						Translation RMSE (°) ↓					
Solver	VLFeat [60]	AffNet [47]	ASIFT [48]	DKM [16]	RoMA [18]	**Ours**	VLFeat [60]	AffNet [47]	ASIFT [48]	DKM [16]	RoMA [18]	**Ours**
Seq. 0	0.0467	0.0625	0.0360	0.0429	0.0406	**0.0353**	0.799	0.697	0.968	0.751	0.735	**0.689**
Seq. 1	0.0433	0.0297	0.0366	0.0428	0.0399	**0.0287**	0.690	0.657	0.648	0.603	**0.590**	0.603
Seq. 2	0.0395	0.0332	0.0627	0.0389	0.0375	**0.0320**	0.732	0.678	0.979	0.739	0.726	**0.673**
Seq. 3	0.0434	0.0390	0.0766	0.0427	0.0409	**0.0378**	0.665	0.641	0.585	0.655	0.615	**0.574**
Seq. 4	0.0285	0.0213	0.0529	0.0278	0.0279	**0.0200**	0.389	0.420	0.894	0.422	0.398	**0.350**
Seq. 5	0.0567	0.0277	0.1100	0.0335	0.0312	**0.0263**	0.737	0.463	1.379	0.471	0.450	**0.407**
Seq. 6	0.0447	0.0227	0.0640	0.0290	0.0272	**0.0214**	0.486	0.362	0.715	0.376	0.371	**0.353**
Seq. 7	0.0397	0.0269	0.0650	0.0311	0.0295	**0.0262**	0.780	0.638	1.190	0.676	0.659	**0.610**
Seq. 8	0.0366	0.0281	0.0557	0.0349	0.0325	**0.0271**	0.937	0.861	1.120	0.899	0.875	**0.842**
Seq. 9	0.0369	0.0295	0.0516	0.0353	0.0329	**0.0279**	0.508	0.464	0.619	0.489	0.483	**0.439**
Seq. 10	0.0558	0.0383	0.1090	0.0394	0.0369	**0.0348**	0.733	0.599	1.180	0.558	0.564	**0.543**
Average	0.0429	0.0326	0.0655	0.0362	0.0343	**0.0289**	0.678	0.587	0.934	0.604	0.587	**0.554**

used affine correspondences extraction methods, including ASIFT [48], AffNet [47], and VLFeat [8, 60]. Also, we include the results of point-based methods, like DKM [16] and RoMA [18]. Similarly as on the MegaDepth dataset, affine-based methods use the affine GC-RANSAC method, while point-based ones use RANSAC.

Evaluation protocol. Rotation and translation errors are measured using RMSE. Rotation error is the angular difference between the ground truth and estimated rotation, while translation error is evaluated similarly by comparing angular differences, as done in baseline methods.

Result. The performance is evaluated based on the median error for each image sequence. Table 4 presents the per-frame error in rotation and translation direction for all tested KITTI sequences, calculated according to the provided ground truth. The proposed method improves the rotation accuracy in all test sequences. We also improve the estimated translations in all but one sequence, where the proposed method secures the second lowest errors.

4.5. Ablation Study

To validate the effectiveness of designed components, we conduct ablation experiments on the HPatches [2] and MegaDepth [34] datasets.

First, We first compared the performance of different affine 558 transformation synthesis methods. we compare the results from direct affine shape regression, approximating the affine shape from orientation and scale, and, finally, using direction, scale and residual shape to estimate the affine transformation. As shown in Table 5, among the several ways to calculate the affine transformation matrix, the way we adopt leads to the highest degree of similarity to the ground truth. This is consistent with the conclusion in AffNet [47]. Second, we demonstrate the improvement caused by our proposed PC-loss, affine extractor, and affine loss when added to DKM [16] on relative pose estimation on the MegaDepth dataset [34]. Table 6 demonstrates that each proposed component leads to improvements in all accuracy metrics compared to the original DKM method.

Table 5. Ablation study on affine shape parameterizations on the HPatches [2] dataset. The reported metrics are the Euclidean distance and cosine similarity of the estimated affine matrices w.r.t. the ground truth ones. n/c – did not converge.

Estimated parameters	Euclidean-Distance	Cosine-Similarity
$(a_{11}, a_{12}, a_{21}, a_{22})$	n/c	n/c
(O, S)	0.346	0.987
(O, S, A'')	**0.123**	**0.994**

Table 6. Ablation study on MegaDepth-1500 [34]. We report the relative pose Accuracy Under the recall Curve (AUC; higher is better) thresholded at $5°$, $10°$, and $20°$. The best results are in bold.

Method	AUC @ →	5° ↑	10° ↑	20° ↑
DKM		60.4	74.9	85.1
DKM + PC-loss		60.7	75.1	**85.4**
Ours (DKM + Affine Extractor + PC-Loss + AC-Loss)		**63.1**	**76.3**	85.2

5. Conclusion

We propose a new framework designed for affine correspondence extraction. The geometric constraints that the points and affine shapes induce are formalized as losses used to supervise the network, learning geometry to improve matching accuracy. The experiments demonstrate that the proposed method surpasses existing affine shape estimators in terms of accuracy, while also improving upon state-of-the-art point-based approaches. We believe this research advances accurate affine correspondence extraction.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant No. 12372189) and the Hunan Provincial Natural Science Foundation for Excellent Young Scholars (Grant No. 2023JJ20045).

References

[1] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012. 7

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2017. 6, 7, 8

[3] Daniel Barath. On making sift features affine covariant. *International Journal of Computer Vision*, 131:2316–2332, 2023. 1, 2

[4] Daniel Barath and Levente Hajder. A theory of point-wise homography estimation. *Pattern Recognition Letters*, 94: 7–14, 2017. 7

[5] Daniel Barath and Levente Hajder. Efficient recovery of essential matrix from two affine correspondences. *IEEE Transactions on Image Processing*, 27(11):5328–5337, 2018. 1, 2, 3

[6] Daniel Barath and Jiri Matas. Graph-Cut RANSAC: Local optimization on spatially coherent structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4961–4974, 2022. 7, 8

[7] Daniel Barath, Tekla Tóth, and Levente Hajder. A minimal solution for two-view focal-length estimation using two affine correspondences. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2557–2565, 2017. 1, 3

[8] Daniel Barath, Michal Polic, Wolfgang Förstner, Torsten Sattler, Tomas Pajdla, and Zuzana Kukelova. Making affine correspondences work in camera geometry computation. In *European Conference on Computer Vision*, pages 723–740, 2020. 1, 7, 8

[9] Adam Baumberg. Reliable feature matching across widely separated views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 774–781 vol.1, 2000. 2, 7

[10] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David N. R. McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36, 2022. 7

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017. 6

[12] Ji Dai, Shiwei Jin, Junkang Zhang, and Truong Q. Nguyen. Boosting feature matching accuracy with pairwise affine estimation. *IEEE Transactions on Image Processing*, page 8278–8291, 2020. 2, 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 337–33712, 2018. 2, 7

[15] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable cnn for joint description and detection of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8084–8093, 2019. 2, 6, 7

[16] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 1, 2, 3, 6, 7, 8

[17] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, don't describe—describe, don't detect for local feature matching. In *2024 International Conference on 3D Vision (3DV)*, pages 148–157. IEEE, 2024. 2

[18] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 7, 8

[19] Ivan Eichhardt and Daniel Barath. Optimal multi-view correction of local affine frames. *British Machine Vision Conference*, 2019. 1

[20] Yuan Gao, Jianfeng He, Tianzhu Zhang, Zhe Zhang, and Yongdong Zhang. Dynamic keypoint detection network for image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14404–14419, 2023. 2

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6, 8

[22] Banglei Guan and Ji Zhao. Affine correspondences between multi-camera systems for 6DOF relative pose estimation. In *European Conference on Computer Vision*, pages 634–650, 2022. 1

[23] Banglei Guan, Ji Zhao, Daniel Barath, and Friedrich Fraundorfer. Minimal cases for computing the generalized relative pose using affine correspondences. In *International Conference on Computer Vision*, pages 6068–6077, 2021. 2

[24] Banglei Guan, Ji Zhao, Daniel Barath, and Friedrich Fraundorfer. Minimal cases for computing the generalized relative pose using affine correspondences. In *IEEE International Conference on Computer Vision*, pages 6048–6057, 2021. 1

[25] Banglei Guan, Ji Zhao, Daniel Barath, and Friedrich Fraundorfer. Efficient recovery of multi-camera motion from two affine correspondences. In *IEEE International Conference on Robotics and Automation*, pages 1305–1311, 2021. 7

[26] Banglei Guan, Ji Zhao, Zhang Li, Fang Sun, and Friedrich Fraundorfer. Relative pose estimation with a single affine correspondence. *IEEE Transactions on Cybernetics*, 52(10): 10111–10122, 2022. 1, 2

[27] Banglei Guan, Ji Zhao, Daniel Barath, and Friedrich Fraundorfer. Minimal solvers for relative pose estimation of multi-camera systems using affine correspondences. *International Journal of Computer Vision*, 2023. 2

[28] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 1–6, 1988. 2

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[30] Petr Hruby, Marc Pollefeys, and Daniel Barath. Semi-calibrated relative pose from an affine correspondence and monodepth. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[31] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, page 517–547, 2021. 2

[32] Axel Barroso Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint detection by handcrafted and learned cnn filters. In *IEEE International Conference on Computer Vision*, pages 698–711, 2019. 2

[33] Kunhong Li, Longguang Wang, Li Liu, Qing Ran, Kai Xu, and Yulan Guo. Decoupling makes weakly supervised local feature better. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15838–15848, 2022. 2, 6, 7

[34] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 6, 7, 8

[35] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, page 91–110, 2004. 2, 7

[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 *European Conference on Computer Vision*, pages 170–185, 2018. 1

[37] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2522–2531, 2019. 7

[38] 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. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6588–6597, 2020. 7

[39] Rodríguez Mariano, Facciolo Gabriele, Rafael. Grompone. von. Gioi, Musé Pablo, and Delon Julie. Robust estimation of local affine maps and its applications to image matching. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1331–1340, 2020. 2

[40] Jiri Matas, Ondrej Chum, Martin Urban, and Tomas Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. 2

[41] Krystian Mikolajczyk and Krystian Mikolajczyk. Scale amp affine invariant interest point detectors. *International Journal of Computer Vision*, page 63–86, 2004. 1, 2

[42] Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142, 2002. 2

[43] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005. 6

[44] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, page 4829–4840, 2017. 7

[45] Dmytro Mishkin, Jiri Matas, and Michal Perdoch. Mods: Fast and robust method for two-view matching. *Computer Vision and Image Understanding*, 141:81–93, 2015. 1

[46] Dmytro Mishkin, Jiri Matas, and Michal Perdoch. MODS: Fast and robust method for two-view matching. *Computer vision and image understanding*, 141:81–93, 2015. 2

[47] Dmytro Mishkin, Filip Radenović, and Jiři Matas. Repeatability is not enough: Learning affine regions via discriminability. In *European Conference on Computer Vision*, pages 287–304, 2018. 1, 2, 3, 4, 7, 8

[48] Jean-Michel Morel and Guoshen Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, page 438–469, 2009. 2, 7, 8

[49] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *IEEE International Conference on Computer Vision*, pages 3476–3485, 2017. 7

[50] Carolina Raposo and Joao P. Barreto. Theory and practice of structure-from-motion using affine correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5470–5478, 2016. 1, 2

[51] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2D2: Reliable and repeatable detector and descriptor. In *Advances in Neural Information Processing Systems*, page 12414–12424, 2019. 2, 7

[52] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NCNet: Neighbourhood consensus networks for estimating image correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1020–1034, 2022. 2

[53] Mariano Rodríguez, Julie Delon, and Jean-Michel Morel. Covering the space of tilts. application to affine invariant image comparison. *SIAM Journal on Imaging Sciences*, page 1230–1267, 2018. 3

[54] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2020. 7

[55] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. GIM: Learning generalizable image matcher from internet videos. *ArXiv*, abs/2402.11095, 2024. 2

[56] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching

with transformers. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 1, 2, 6, 7

[57] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second order similarity regularization for local descriptor learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11008–11017, 2019. 2

[58] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. PDC-Net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10247–10266, 2023. 2, 7

[59] MichałTyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. In *Advances in Neural Information Processing Systems*, pages 14254–14265, 2020. 2, 7

[60] Andrea Vedaldi and Brian Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. *ACM international conference on Multimedia*, page 1469–1472, 2010. 2, 7, 8

[61] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *European Conference on Computer Vision*, pages 757–774, 2020. 2, 5, 7

[62] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[63] Shibiao Xu, Shunpeng Chen, Rongtao Xu, Changwei Wang, Peng Lu, and Li Guo. Local feature matching using deep learning: A survey. *Information Fusion*, 107:102344, 2024. 2

[64] Pei Yan, Yihua Tan, Shengzhou Xiong, Yuan Tai, and Yansheng Li. Learning soft estimator of keypoint scale and orientation with probabilistic covariant loss. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19384–19393, 2022. 4, 6

[65] Zhenbao Yu, Banglei Guan, Shunkun Liang, Zibin Liu, Yang Shang, and Qifeng Yu. Globally optimal solution to the generalized relative pose estimation problem using affine correspondences. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):12568–12580, 2024. 1

[66] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2Pix: Epipolar-guided pixel-level correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4667–4676, 2021. 2