Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

E3CM: Epipolar-constrained cascade correspondence matching☆

Chenbo Zhou, Shuai Su, Qijun Chen, Rui Fan *

Tongji University, Shanghai, 201804, China

ARTICLE INFO

Communicated by Y. Liu

Keywords:
Correspondence matching
3D computer vision
Deep learning
Convolutional neural networks
Epipolar constraints

ABSTRACT

Accurate and robust correspondence matching is of utmost importance for various 3D computer vision tasks. However, traditional explicit programming-based methods often struggle to handle challenging scenarios, and deep learning-based methods require large well-labeled datasets for network training. In this article, we introduce Epipolar-Constrained Cascade Correspondence (E3CM), a novel approach that addresses these limitations. Unlike traditional methods, E3CM leverages pre-trained convolutional neural networks to match correspondence, without requiring annotated data for any network training or fine-tuning. Our method utilizes epipolar constraints to guide the matching process and incorporates a cascade structure for progressive refinement of matches. We extensively evaluate the performance of E3CM through comprehensive experiments and demonstrate its superiority over existing methods. To promote further research and facilitate reproducibility, we make our source code publicly available at https://mias.group/E3CM/.

1. Introduction

Correspondence matching forms the foundation for a variety of 3D computer vision tasks, such as simultaneous localization and mapping (SLAM) [1–3], structure from motion (SfM) [4–6], dense disparity estimation and transformation [7–9], and online stereo rig self-calibration [10,11]. Currently, popular correspondence matching algorithms are classified into two categories: (i) explicit programming-based and (ii) deep learning-based.

Explicit programming-based correspondence matching methods typically extract keypoints based on human-defined local features, such as rotated binary robust independent elementary features (ORB) [12] and scale-invariant feature transform (SIFT) [13], followed by making pairs between correspondences with nearest neighboring (NN) matching [14]. In contrast, recent deep learning-based methods [15–17] learn to detect and describe local features with neural networks, resulting in significantly enhanced robustness in correspondence matching when compared to explicit programming-based techniques. Furthermore, matchers based on graph neural networks [18] have been developed and utilized to predict correct matches while effectively filtering out incorrect ones.

In scenarios involving large perspective changes or repetitive textures, explicit programming-based methods tend to exhibit subpar performance. However, deep learning-based methods, as demonstrated in [15–17], have made remarkable strides in improving matching accuracy within such challenging scenarios. Nevertheless, it is worth noting that the existing deep learning-based methods often rely on a large amount of labeled training data, which typically includes precise information regarding camera positions and poses. To address the limitations associated with these methods, a novel approach, referred to as deep feature matching (DFM), was introduced [19]. DFM employs a hierarchical matching strategy based on the Visual Geometry Group (VGG) network, pre-trained on the ImageNet [20] database. An intriguing aspect of DFM is that it does not necessitate additional training using data annotated with correspondences, thus mitigating the need for a large amount of labeled data. Furthermore, DFM offers significantly improved accuracy and robustness compared to explicit programming-based methods, surpassing even some approaches trained with correspondences. However, it is important to note that DFM's hierarchical correspondence matching process is based on homography matrices [21], which imposes a limitation on its applicability. Specifically, DFM is most effective when dealing with cases that involve only planar surfaces. In scenarios where the scene contains non-planar or complex surfaces, the performance of DFM may be compromised. The reliance on homography matrices can lead to suboptimal performance in more intricate stereoscopic scenes, especially indoor scenes with short sight distances.

☆ This work was supported by the National Key R&D Program of China under Grant 2020AAA0108100, the National Natural Science Foundation of China under Grant 62233013, the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, and the Fundamental Research Funds for the Central Universities.

* Corresponding author.
E-mail address: rui.fan@ieee.org (R. Fan).

https://doi.org/10.1016/j.neucom.2023.126788
Received 20 March 2023; Received in revised form 1 September 2023; Accepted 11 September 2023

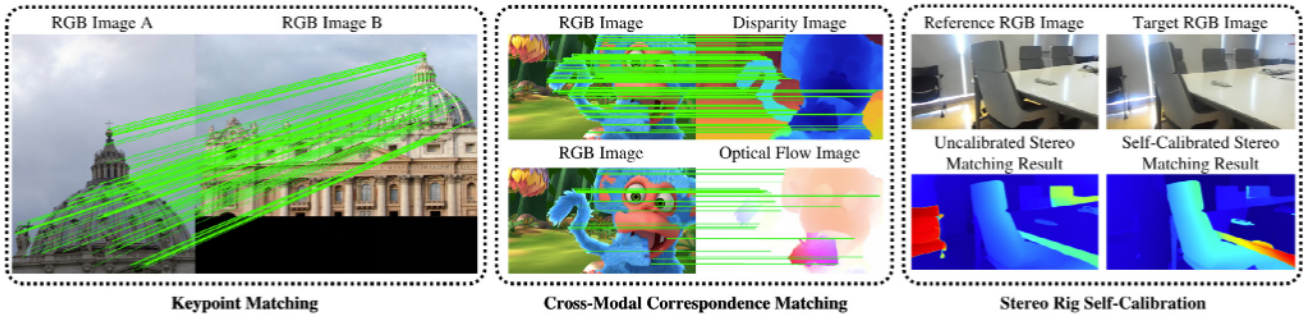


Fig. 1. Our proposed E3CM algorithm can be applied to various 3D computer vision applications, including: (a) keypoint matching, (b) cross-modal correspondence matching, and (c) online stereo rig self-calibration.

Hence in this paper, we introduce **E**pipolar-**C**onstrained **C**ascade **C**orrespondence **M**atching (**E3CM**), a plug-and-play solution designed specifically for stereo rigs. Our method leverages feature maps derived from pre-trained backbones (on the ImageNet database), incorporating a cascade outlier rejection module that relies on pose estimation and epipolar constraints. This combination of techniques enables E3CM to effectively handle the typical and prevalent scenarios encountered in real-world stereoscopic scenes. The matching process begins from the final layer of multiple selected feature maps. We utilize the obtained matches to estimate the camera's pose. Using the estimated pose, we then apply the epipolar constraint to eliminate outliers in the previous layer. Subsequently, a new pose is estimated based on the matches after outlier removal. This iterative process continues until reaching the first layer, gradually enhancing the pose estimation accuracy and increasing the number of correct matches. Additionally, we extend the utilization of pre-trained backbones, enabling a comprehensive comparison among various backbones. To further enhance the reliability of matches within the feature maps, we introduce a novel confidence score, which effectively decreases the probability of incorrectly estimating camera poses, further enhancing the overall performance of E3CM. As depicted in Fig. 1, E3CM can be effectively utilized for cross-modal correspondence matching as well as various 3D computer vision tasks, exemplified with online stereo rig self-calibration in this paper.

The structure of the remaining paper is as follows: Section 2 provides an overview of existing works related to correspondence matching, outlining the advancements and limitations in the field. Section 3 presents the details of our proposed E3CM algorithm, explaining the key components and techniques employed. The experimental results for performance evaluation are illustrated in Section 4. Section 5 delves into a comprehensive discussion of the applicability of our work, addressing its potential applications and limitations. Finally, Section 6 serves as a summary of the paper, highlighting the key findings and contributions.

2. Related work

Correspondence matching is a crucial task in 3D computer vision applications, such as visual odometry, image stitching, and online stereo rig self-calibration. Traditional explicit programming-based correspondence matching approaches [12,22–24] are typically based on hand-crafted techniques. These approaches primarily rely on local gradients, local corners, and local blob features to detect and describe keypoints. However, they often struggle to perform well in low-texture scenes, which makes them less reliable for downstream applications.

Recent advancements in deep learning-based approaches for keypoint detection, description, and matching have demonstrated superior accuracy and robustness compared to traditional methods. SuperPoint [15] is a notable method in the field of keypoint detection and description. It is a self-supervised detector–descriptor framework that initially introduces a detector referred to as MagicPoint using synthetic images and then trains a descriptor by generating random homography matrices as ground truth. This training strategy has been widely adopted in many other works. Another deep learning-based method, D2-Net [16], is a trainable convolutional neural network (CNN) that performs joint detection and description of local features. It extracts keypoints by computing the maximum values on a feature map that is four times smaller than the source image. However, D2-Net prioritizes repeatability, which leads to reduced matching performance in regions with high texture repetition. In response to this limitation, R2D2 [17] builds upon D2-Net by emphasizing the reliability of feature points.

While NN matching has been widely utilized in the matching stage, it often overlooks the assignment structure and disregards visual information. SuperGlue [18] presents a novel approach to tackle this issue. It leverages a graph neural network (GNN) with an attention mechanism to integrate positional information and keypoint descriptions, and computes the correspondences using the Sinkhorn algorithm [25,26].

In addition to the traditional two-stage pipeline consisting of both a detector and descriptor, several existing works have adopted end-to-end frameworks for correspondence matching. Neighborhood consensus networks (NCNet) [27] were proposed to match dense correspondences without the need for a separate feature detector. However, due to intensive correlation score computation on down-scaled feature maps, NCNet's performance in camera pose estimation tasks is suboptimal. To address this limitation, efficient neighborhood consensus network via submanifold sparse convolutions (SparseNCNet) [28] employs a sparse representation of the correlation tensor by storing a portion of the scores and replacing dense 4D convolutions with sparse convolutions. Densely connected recurrent convolutional neural network (DRC-Net) [29] follows SparseNCNet and introduces a hierarchical framework to generate dense matches with higher accuracy. An epipolar-guided pixel-level correspondence matching approach, referred to as Patch2pix [30], leverages pre-trained backbones to extract potential patch-level matches and refines the matches to pixel-level accuracy using two-stage regressors. Another detector-free framework, local feature matching with Transformers (LoFTR) [31], builds upon the Transformer network and achieves superior performance compared to SuperPoint with SuperGlue.

Furthermore, various strategies have been proposed to improve the accuracy of matches by effectively rejecting outliers. Neural-guided random sample consensus (RANSAC) [32] utilizes probabilities to weigh the matches. AdaLAM [33] assumes that close matching pairs share the same local affine transformation and rejects outliers that deviate from the affine matrix within a neighboring area. Another approach, presented in [34], employs neural networks to predict binary labels for outlier identification.

The work presented in [19] can be considered as a baseline for the task of matching keypoints using a pre-trained backbone network. It employs a coarse-to-fine strategy to match features from deep layers to shallow layers. Additionally, it also estimates homography matrices

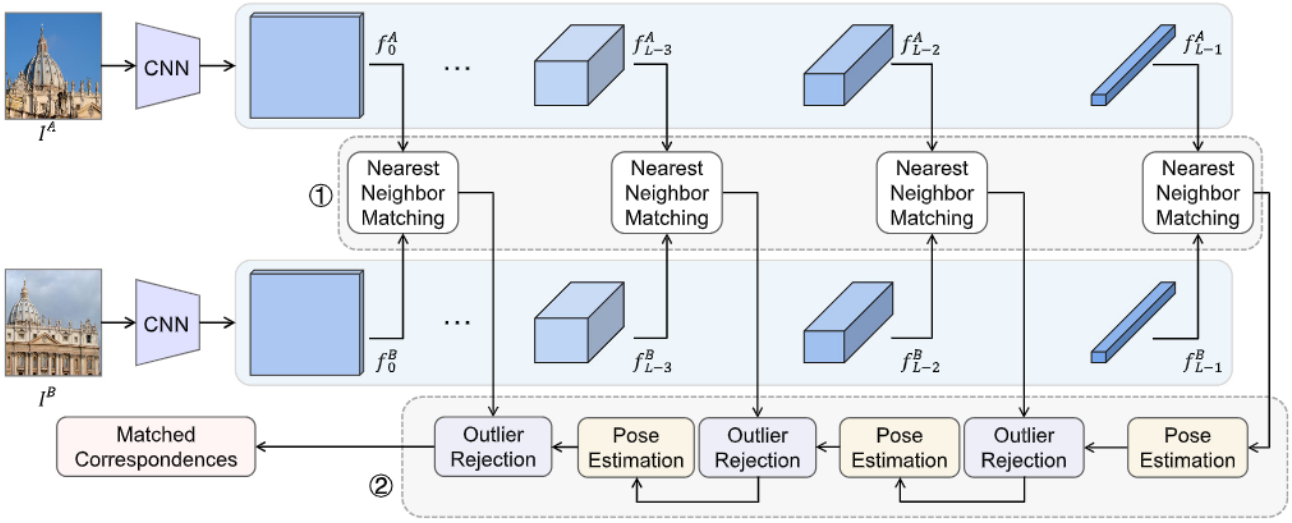


Fig. 2. The framework of our proposed epipolar-constrained cascade correspondence matching approach: ① initial deep correspondence matching; ② epipolar-constrained cascade refinement.

using feature maps to refine correspondences. However, their approach is limited to cases that involve only planar surfaces. In our work, we address this limitation by introducing the epipolar-constrained cascade refinement strategy, which replaces the two-stage method and enables matching in more general scenarios. Additionally, this paper evaluates the effectiveness of our proposed E3CM approaches using various popular backbone networks, including VGG [35], ResNet [36], DenseNet [37], MobileNet [38], and GoogleNet [39]. These backbone networks have all been pre-trained on the ImageNet database.

3. Methodology

3.1. Initial deep correspondence matching

Our proposed initial deep correspondence matching approach is developed based on DFM [19], where a VGG-19 [35] network pre-trained on the ImageNet database is employed to extract deep feature maps, $\boldsymbol{f}^A$ and $\boldsymbol{f}^B$, from a given pair of color images, I^A and I^B. Based on the hypothesis that the given image pairs can be linked by a homography matrix $\boldsymbol{H}^{BA}$, DFM warps the color images using an estimated homography matrix before the second-stage correspondence matching. However, this limits its applicability to scenarios that primarily involve a planar surface. Additionally, DFM only demonstrates its compatibility with pre-trained VGG models, as it requires the resolution of the first feature map to be identical to that of the input images. To address these limitations, we extend the applicability of the method to general cases by removing the redundant image warping step. Moreover, we enhance its compatibility with other state-of-the-art CNNs to broaden its usage and accommodate different network architectures.

Correspondence matching at layer l can be solved with NN matching. Let $\boldsymbol{f}_l^A$ and $\boldsymbol{f}_l^B$ be the feature maps (size: $H/2^l \times W/2^l \times C_l$) at layer $l \in [0, L-1]$ extracted from a given pair of images I^A and I^B (resolution: $H \times W$ pixels). Given a pair of points $\boldsymbol{p}_l^A = (h_l^A; w_l^A)$ in $\boldsymbol{f}_l^A$ and $\boldsymbol{p}_l^B = (h_l^B; w_l^B)$ in $\boldsymbol{f}_l^B$, where $0 < h_l^{A,B} < H/2^l$ and $0 < w_l^{A,B} < W/2^l$, we measure the distance d between their representations $\boldsymbol{c}(\boldsymbol{p}_l^A)$ and $\boldsymbol{c}(\boldsymbol{p}_l^B)$ of size: $C_l \times 1$ as follows:

$$d(\boldsymbol{p}_l^A, \boldsymbol{p}_l^B) = 1 - \Phi(\boldsymbol{c}(\boldsymbol{p}_l^A), \boldsymbol{c}(\boldsymbol{p}_l^B)), \tag{1}$$

where

$$\Phi(\boldsymbol{c}(\boldsymbol{p}_l^A), \boldsymbol{c}(\boldsymbol{p}_l^B)) = \frac{\boldsymbol{c}(\boldsymbol{p}_l^A)^\top \boldsymbol{c}(\boldsymbol{p}_l^B)}{\left\|\boldsymbol{c}(\boldsymbol{p}_l^A)\right\|_2 \left\|\boldsymbol{c}(\boldsymbol{p}_l^B)\right\|_2}. \tag{2}$$

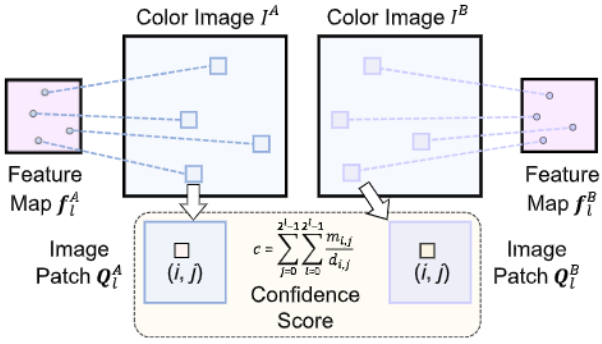


Fig. 3. Confidence score computation.

is the cosine distance between $\boldsymbol{c}(\boldsymbol{p}_l^A)$ and $\boldsymbol{c}(\boldsymbol{p}_l^B)$. A map $\boldsymbol{D}_l$ (size: $H/2^l \times W/2^l \times (H/2^l \times W/2^l)$) storing the measured cosine distances between all possible matches can thus be obtained. Given $\boldsymbol{p}_{l_k}^A$, if the ratio of its minimum distance (corresponding to $\boldsymbol{p}_{l_i}^B$) versus its second-minimum distance (corresponding to $\boldsymbol{p}_{l_j}^B$) is lower than the pre-set threshold (empirically set to 0.9), $(\boldsymbol{p}_{l_k}^A, \boldsymbol{p}_{l_i}^B)$ are considered as a pair of satisfactorily matched correspondences.

Compared to the traditional approaches that determine correspondences by detecting, describing, and matching local visual features via explicit programming, our proposed CNN-based approach, on the other hand, can directly perform correspondence matching on deep hierarchical feature maps. One of the most representative characteristics of CNNs is that the feature maps at shallow layers have higher resolutions and smaller receptive fields, while the feature maps at deeper layers have lower resolutions and larger receptive fields. Therefore, more confident but fewer correspondences can be obtained when it comes to the deeper layers of the CNNs. Based on this important characteristic, we develop an epipolar-constrained cascade refinement strategy for outlier rejection, as presented in the following subsection.

3.2. Epipolar-constrained cascade refinement

As illustrated in Fig. 3, given a pair of matched points $\boldsymbol{p}_l^A$ and $\boldsymbol{p}_l^B$, respectively in the feature maps $\boldsymbol{f}_l^A$ and $\boldsymbol{f}_l^B$, they correspond to two patches $\boldsymbol{Q}_l^A = \{\boldsymbol{q}_{i,j}^A\}_{i=0,j=0}^{2^l-1,2^l-1}$ and $\boldsymbol{Q}_l^B = \{\boldsymbol{q}_{i,j}^B\}_{i=0,j=0}^{2^l-1,2^l-1}$ of size $2^l \times 2^l$ in the

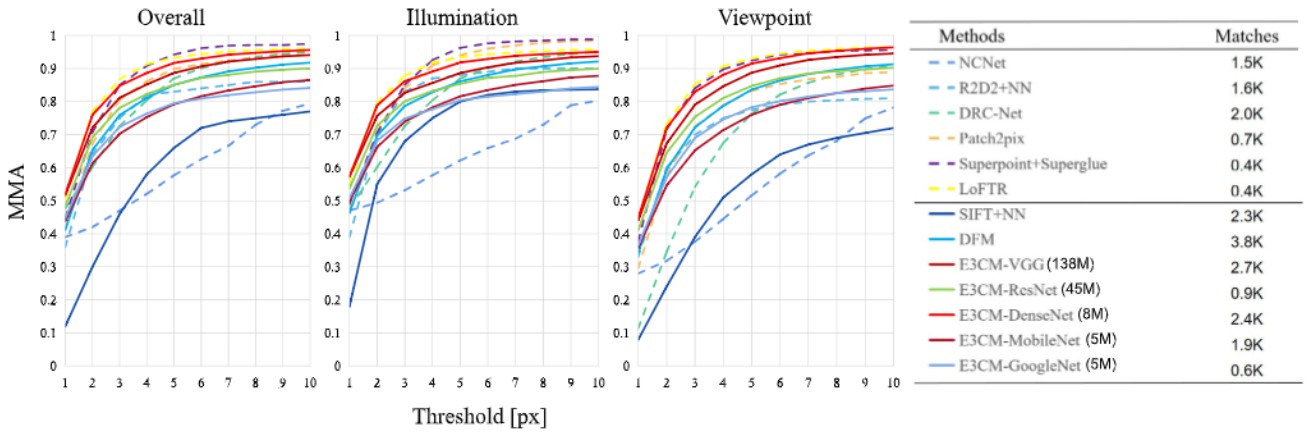


Fig. 4. MMA comparison of different methods on the HPatches dataset. The methods above the dividing line in the table on the right are fully supervised ones, while the methods below the dividing line are training-free and plug-and-play ones.

original images I^A and I^B. We define a confidence score

$$c = \sum_{i=0}^{2^l-1} \sum_{j=0}^{2^l-1} \frac{m_{i,j}}{d_{i,j}} \tag{3}$$

to measure the reliability of the matched image patches $\boldsymbol{Q}_l^A = \{\boldsymbol{q}_{i,j}^A\}_{i=0,j=0}^{2^l-1,2^l-1}$ and $\boldsymbol{Q}_l^B = \{\boldsymbol{q}_{i,j}^B\}_{i=0,j=0}^{2^l-1,2^l-1}$, where $d_{i,j}$ represents the distance between $\boldsymbol{q}_{i,j}^A$ and $\boldsymbol{q}_{i,j}^B$, measured using the feature maps at the shallowest layer,[1] $m_{i,j} = \{0, 1\}$ is determined by NN matching (a good match corresponds to 1, while a bad match corresponds to 0).

Since the camera pose between I^A and I^B can be estimated using at least eight pairs of matched correspondences, we choose eight pairs of image patches with the highest confidence scores to compute the fundamental matrix $\boldsymbol{F}_l$ with respect to the lth layer, where the centering point of each desired image patch is used in the eight-point algorithm [40]. The fundamental matrix $\boldsymbol{F}_l$ estimated using $\boldsymbol{f}_l^A$ and $\boldsymbol{f}_l^B$ is used to reject outliers (not satisfying the epipolar constraint) for correspondence matching using $\boldsymbol{f}_{l-1}^A$ and $\boldsymbol{f}_{l-1}^B$. In this paper, the Sampson distance [40]

$$\phi_{\boldsymbol{p}_l^A,\boldsymbol{p}_l^B} = \frac{\left((\boldsymbol{p}_l^B)^\top \boldsymbol{F}_l \boldsymbol{p}_l^A\right)^2}{\left(\boldsymbol{F}_l \boldsymbol{p}_l^A\right)_1^2 + \left(\boldsymbol{F}_l \boldsymbol{p}_l^A\right)_2^2 + \left(\boldsymbol{F}_l^\top \boldsymbol{p}_l^B\right)_1^2 + \left(\boldsymbol{F}_l^\top \boldsymbol{p}_l^B\right)_2^2} \tag{4}$$

is used to determine inliers and outliers, where $\left(\boldsymbol{F}\boldsymbol{p}^A\right)_k^2$ and $\left(\boldsymbol{F}\boldsymbol{p}^B\right)_k^2$ represent the square of the kth entry of the vector $\boldsymbol{F}\boldsymbol{p}^A$ and $\boldsymbol{F}\boldsymbol{p}^B$, respectively. As illustrated in Fig. 2, such an outlier rejection algorithm performs iteratively from the deepest layer to the shallowest layer.

4. Experiments

4.1. Image matching

We first conduct experiments on the HPatches dataset [44] to evaluate the performance of our correspondence matching method. The dataset consists of two main parts: 'Viewpoint' and 'Illumination'.

As shown in Fig. 4, our method using DenseNet161 as the backbone outperforms SuperPoint+SuperGlue (trained with correspondences) under low threshold values in all three situations: 'Overall', 'Illumination', and 'Viewpoint'. While our method may not achieve the best performance in the 'Illumination' scenario, it excels in another two situations, particularly in the 'Viewpoint' scenario. Additionally, our method produces a significantly higher number of matches compared to other methods, resulting in higher mean matching accuracy (MMA) scores.

[1] An image might have to be resampled so that its resolution is identical to that of the feature map at the shallowest layer.

We also evaluate the performance of our proposed method with respect to different backbones, including VGG19, DenseNet161, ResNet152, MobileNet-Large, and GoogleNet. Among these backbones, DenseNet161 achieved the best performance. To ensure a fair comparison with DFM, we use VGG19 as the backbone for the following experiments.

We analyze the network structures and speculated about possible reasons for the observed performance differences. ResNet, with its residual network structure, may hinder the flow of information between layers, leading to the loss of low-dimensional feature information, which is not ideal for using feature map channels directly as descriptors for feature matching. As for MobileNet-Large and GoogleNet, their lighter network structures inherently result in a loss of performance compared to more complex networks.

4.2. Homography matrix estimation

In the evaluation on the HPatches dataset, we estimate the homography matrix using our proposed E3CM method and compare it with the ground-truth homography matrix provided by the dataset. We select the four corners of the image and project them using both the ground-truth and estimated homography matrices. The average distance between the projected corner points is then calculated, and we evaluate the method using thresholds of (1, 3, 5). This allows us to determine the percentage of image pairs that fall below each threshold.

As shown in Table 1, while our E3CM method may not perform as well as the state-of-the-art methods, it outperforms many methods that require training with correspondences. Furthermore, our method achieves the best performance in homography matrix estimation among the plug-and-play and training-free methods.

4.3. Pose estimation

We also evaluate our proposed method on the MegaDepth dataset [45] in terms of pose estimation accuracy. MegaDepth consists of one million internet images from 196 different outdoor scenes and provides ground-truth poses for each image. In addition, it provides sparse reconstructions from COLMAP [4] and depth maps computed via multi-view stereo approaches.

Following the setup of DISK [46] and LoFTR, we specifically focus on the "Sacre Coeur" and "St. Peter's Square" scenes for testing. From these scenes, we select the same 1500 image pairs as LoFTR for a fair comparison. We estimate the pose using the computed matches by calculating the essential matrix. The pose error is then evaluated by calculating the area under the curve (AUC) of the pose error at thresholds of 5°, 10°, and 20°. The pose error is defined as the maximum angular error in rotation and translation. Although AUC of

Table 1
Accuracy of Homography Matrix Estimation on the HPatches dataset. The table presents the percentages of correctly estimated homographies with average corner error distances below 1/3/5 pixels.

Category	Method	Homography estimation accuracy		
		<1px	<3px	<5px
Training-free	SIFT [13] + NN [14]	0.36	0.76	0.85
Fully supervised	NCNet [27]	0.48	0.61	0.71
	LIFT [41]	0.39	0.73	0.78
	R2D2 [17] + NN [14]	0.47	0.78	0.83
	DRC-Net [29]	0.46	0.66	0.77
	SOSNet [42]	0.52	0.81	0.86
	MAGSAC [43]	0.51	0.79	0.84
	Pacth2pix [30]	0.51	0.79	0.86
	SuperPoint [15] + SuperGlue [18]	0.53	0.84	0.90
	LoFTR [31]	**0.66**	**0.86**	**0.92**
Plug-and-play	DFM [19]	0.41	0.74	0.86
	E3CM (Ours)	**0.49**	**0.78**	**0.88**

Table 2
Evaluation on MegaDepth: The table presents the percentages of correctly estimated poses with pose errors below 5/10/20 degrees. 'P' refers to the matching precision.

Category	Method	Pose estimation AUC			P
		@5°	@10°	@20°	
Training-free	SIFT [13] + NN [14]	4.78	10.71	20.44	17.19
Fully supervised	NCNet [27]	4.82	11.31	22.96	54.85
	LIFT [41]	6.03	13.71	27.96	39.97
	R2D2 [17] + NN [14]	35.07	52.83	68.03	81.33
	DRC-Net [29]	31.18	47.81	62.80	85.72
	SOSNet [42]	40.16	56.45	72.83	82.47
	MAGSAC [43]	43.98	55.74	68.18	81.03
	Patch2pix [30]	43.32	58.34	70.27	83.06
	SuperPoint [15] + SuperGlue [18]	42.28	62.36	77.86	93.34
	LoFTR [31]	**52.80**	**69.19**	**81.18**	**97.18**
Plug-and-play	DFM [19]	35.17	50.64	61.12	77.19
	E3CM(ours)	**39.85**	**54.11**	**65.86**	**91.14**

Table 3
Evaluation on YFCC100M: The table presents the percentages of correctly estimated poses with pose errors below 5/10/20 degrees. 'P' refers to the matching precision in this context.

Category	Method	Pose estimation AUC			P
		@5°	@10°	@20°	
Training-free	SIFT [13] + NN [14]	4.67	12.04	24.33	12.04
Fully supervised	NCNet [27]	2.40	7.61	17.10	35.56
	LIFT [41]	10.67	21.19	31.96	17.01
	R2D2 [17] + NN [14]	18.49	35.73	54.80	68.66
	DRC-Net [29]	20.06	38.16	57.02	62.04
	SOSNet [42]	21.12	41.80	56.96	68.70
	MAGSAC [43]	23.87	42.97	60.31	75.71
	Patch2pix [30]	26.25	43.23	59.75	72.89
	SuperPoint [15] + SuperGlue [18]	34.18	50.32	64.16	84.90
	LoFTR [31]	**37.71**	**54.69**	**67.00**	**87.08**
Plug-and-play	DFM [19]	15.30	28.57	42.91	61.68
	E3CM(ours)	**17.63**	**31.07**	**45.51**	**76.61**

pose error is influenced by RANSAC, which can discard mismatches and estimate the correct pose, it does not provide a comprehensive evaluation of the matching method. Therefore, we also calculate the matching precision (following SuperGlue [18]) as another evaluation metric for correspondence matching.

The evaluation on the YFCC100M dataset follows a similar approach to that on MegaDepth. We select the same image pairs from YFCC100M as used in SuperGlue [18] to ensure a fair comparison. We compute the AUC of pose error with thresholds of 5°, 10°, and 20° and also obtain the matching precision.

As shown in Tables 2 and 3, E3CM demonstrates advantages in pose estimation compared to traditional hand-crafted methods and some methods that require training with correspondences. Additionally, our method effectively rejects outliers and achieves high matching precision. As depicted in Fig. 5, the matches in the right column have a much higher matching accuracy rate compared to the matches in the left column, even though both columns have similar pose estimations. While matching accuracy may not be critical in pose estimation, it plays a significant role in other computer vision tasks that rely on correspondence matching, such as stereo rig self-calibration, as discussed in the following subsection.

4.4. Stereo rig self-calibration

Among the various computer stereo vision tasks that require correspondence matching, stereo rig self-calibration is of particular importance. Stereo rig self-calibration heavily relies on highly accurate correspondence matching. Even a small portion of bad matches can lead to significant deviations in the calibration results. The stereo rig self-calibration pipeline [10] typically involves correspondence matching in

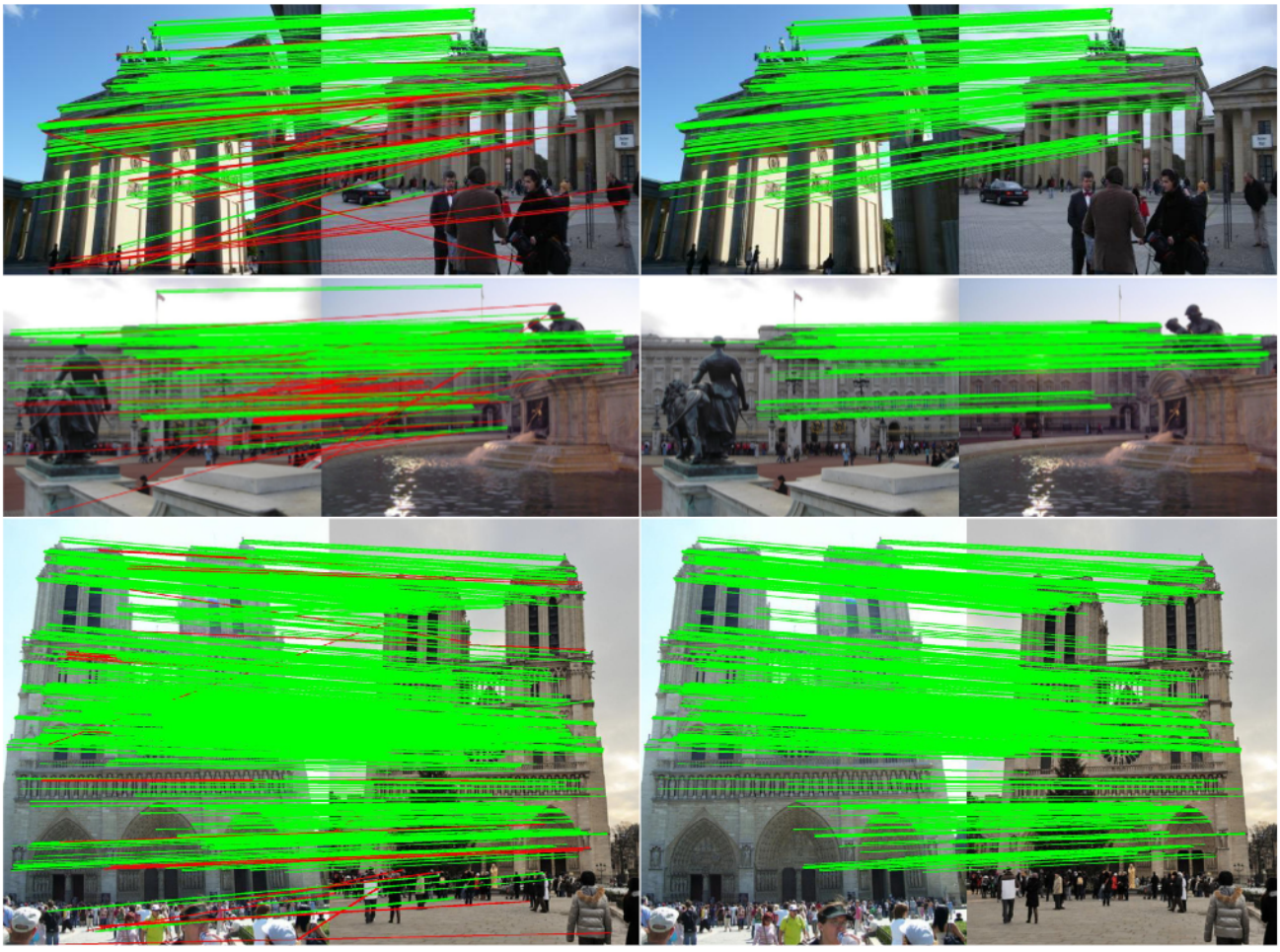


Fig. 5. Experimental results on the MegaDepth dataset. Green lines indicate good matches with a Sampson distance below 1×10^{-4}, while red lines represent bad matches. The images on the left column show the results of direct matching on feature maps without using E3CM, while the images on the right column show the matching results using E3CM.

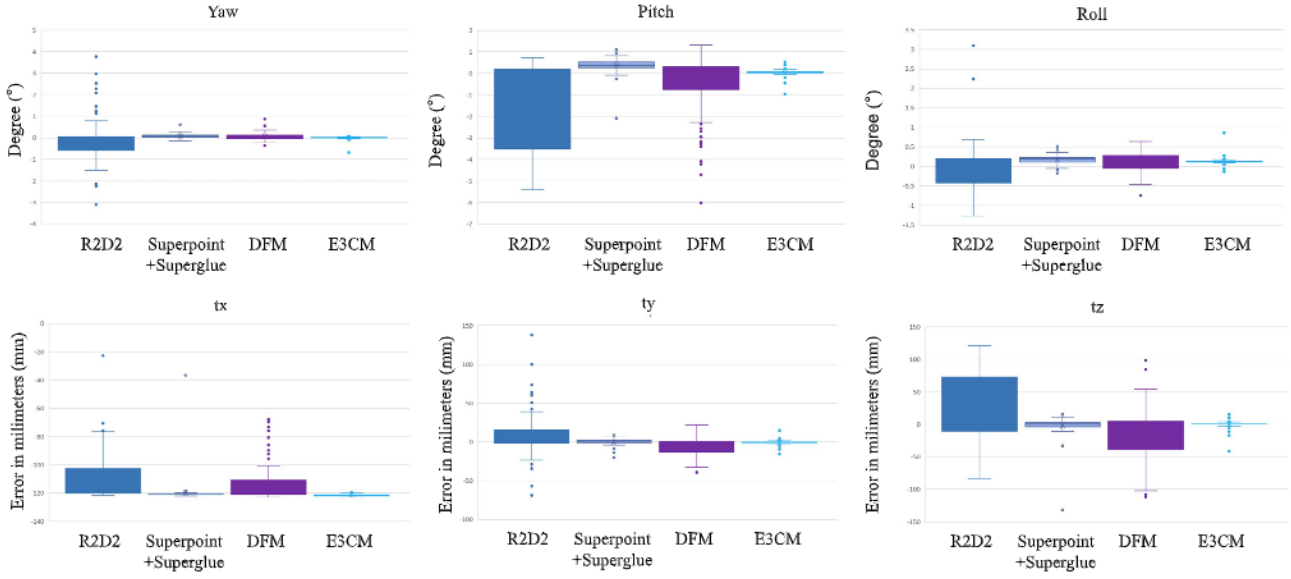


Fig. 6. Boxplots of stereo rig self-calibration using different correspondence matching methods. The results obtained using E3CM as the correspondence matching method are the most robust among the four methods. 'Yaw', 'Pitch', and 'Roll' represent the deviation of the rotation angle in three directions. 'tx', 'ty', and 'tz' represent the translation error in three directions.

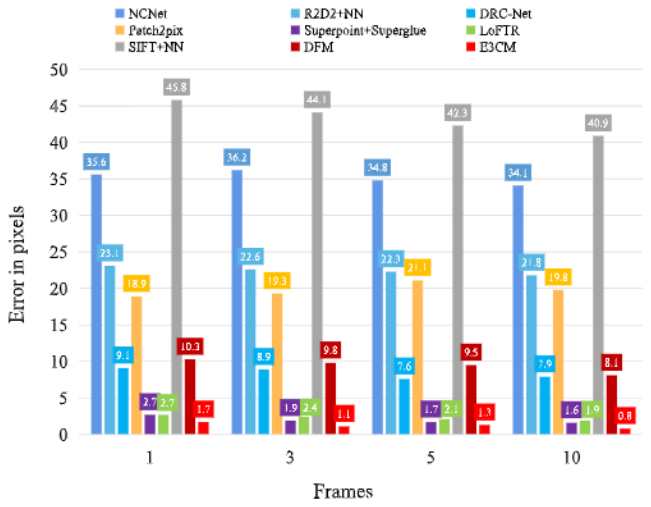


Fig. 7. Checkerboard's average reprojection errors, obtained using different correspondence matching methods.

the first stage and the optimization of extrinsics based on the matches and epipolar constraints. Hence, we can evaluate the accuracy of front-end correspondence matching methods based on the results of the back-end stereo rig self-calibration.

Following [10], we replace the corner detector [48] and BRIEF [49] with other keypoint extraction and matching methods based on deep neural networks. Additionally, we perform stereo rig self-calibration in several scenes. To enhance calibration stability, we input all matches within ten frames into the optimizer in our experiments. The calibration results are presented as boxplots, a common evaluation approach in the calibration domain, as shown in Fig. 6. Our method exhibits fewer outliers and smaller ranges in the calibration results compared to other methods, indicating that our plug-and-play method is more accurate and robust, even outperforming some fully supervised approaches.

Stereo rig calibration is a critical step in terms of achieving accurate 3D measurements [50,51]. Therefore, we further compute the reprojection error of checkerboard pattern corners with respect to different calibration results obtained using 1, 3, 5, and 10 frames, respectively. We compare the performance of several state-of-the-art methods, including our proposed E3CM and SuperPoint+SuperGlue. The results shown in Fig. 7 suggest that E3CM outperforms SuperPoint+SuperGlue in all calibration set-ups, achieving a significantly lower reprojection error.

To provide a more intuitive evaluation, we generate dense disparity maps using RAFT-Stereo [47], as shown in Fig. 8. We generate disparity maps for both rectified and unrectified stereo image pairs. It is evident from Fig. 8 that the unrectified disparity maps yield erroneous depth estimations, while the well-rectified disparity maps obtained using our proposed correspondence matching method accurately reflect the depth relationships between different objects in the scenes.

5. Discussion

Our method is well-suited for stereo vision systems. However, for monocular vision systems that require correspondence matching between consecutive frames, the accuracy of our method may be significantly affected if there are dynamic objects in the scene. This is because our method relies on the assumption of a stationary background with features, such as a building, while a moving object is present. In such cases, our method may estimate incorrect poses during the epipolar-constrained cascade refinement. Specifically, in monocular cameras, if points from both stationary and dynamic objects are selected simultaneously during pose estimation, the relative pose estimation between the former and latter frames of the monocular camera will be erroneous. This discrepancy arises because monocular camera pose estimation computes the camera-to-scene relationship, and when the scene contains moving objects, it introduces changes that invalidate the pose estimation process. In contrast, stereo vision systems have relatively static cameras with respect to each other. Therefore, this problem does not arise in stereo vision systems since the pose estimation is computed from camera to camera and is independent of the scene. Although our method is somewhat limited to stereo vision systems, stereo vision remains a critical aspect of computer vision, and our method can still find broad applications in this context.

The epipolar-constrained cascade refinement method is effective for pose estimation and outlier rejection. When matching directly in the source image, the limited feature information per pixel often leads to numerous mismatches with similar features but different spatial locations. Consequently, using these matches directly for pose estimation can result in incorrect pose estimations. Moreover, removing outliers based on these estimated poses would propagate incorrect matches further. In contrast, matches in the deep feature map, which are convoluted from patches in the source image, contain more semantic information [52]. As a result, the probability of mismatches is significantly reduced compared to direct matching in the source image. Although the pose estimation based on these points, which correspond to patches in the source image, may not be highly accurate, it is generally reasonable and relatively correct. As the number of layers decreases, the coordinates of the points corresponding to the patches in the source image become more accurate. Consequently, the pose estimation based on these points becomes more precise, and the matches obtained by removing outliers based on the estimated pose also improve in accuracy. In essence, our proposed method ensures that the initial pose estimation is reasonable and progressively becomes more accurate. Therefore, outlier rejection based on the initial pose estimation is effective and accurate. Additionally, estimating a pose during the epipolar-constrained cascade refinement requires a minimum of eight pairs of matched points. However, in our experiments, we rarely encountered cases with fewer than eight pairs.

We compute the fundamental matrix by mapping the points in the feature maps to the corresponding pixels in the source image. This process does not require camera intrinsics. However, if camera intrinsics are available, we can directly calculate the essential matrix in the feature maps without the need for mapping the points to the source image pixels. In our experiments, we observe a slight improvement in accuracy using this approach, although the improvement is not significant. The reason is that mapping the points in the feature maps to the pixels in the source image and then computing the fundamental matrix is equivalent to pose estimation based solely on the pixels, which ignores the information from other pixels in the patch. On the other hand, computing the essential matrix directly in the feature maps is equivalent to pose estimation based on the patches in the source image, which includes more information and can result in improved accuracy.

6. Conclusion

In this paper, we introduced a new plug-and-play correspondence matching method that leverages an epipolar-constrained cascade refinement strategy. Our approach is compatible with various pre-trained backbone networks, allowing for flexibility in choosing the most suitable backbone for the task. We argue that pre-trained networks from other vision tasks can be effectively utilized in the correspondence matching task, often achieving comparable or even superior performance compared to fully supervised methods specifically designed for correspondence matching. By leveraging the representation learning capabilities of pre-trained networks, we can leverage the rich knowledge acquired from large-scale datasets, enabling efficient and accurate correspondence matching. Through our experiments and evaluations, we have demonstrated the effectiveness of our approach and its ability

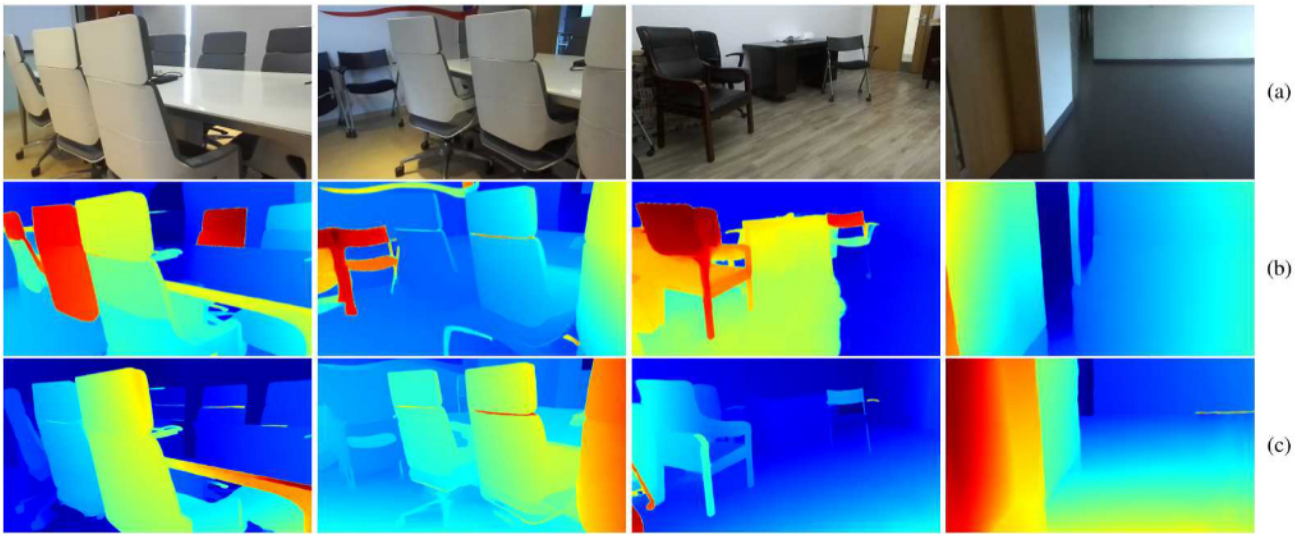


Fig. 8. Stereo rig self-calibration results: (a) reference RGB images; (b) disparity images for the unrectified stereo image pairs; (c) disparity images for the rectified stereo image pairs which are obtained via stereo rig self-calibration. The disparity maps are computed using RAFT-Stereo [47].

to leverage pre-trained networks for correspondence matching tasks. We believe that our method opens up new possibilities for utilizing pre-trained networks in a broader range of vision tasks, offering improved performance and efficiency.

CRediT authorship contribution statement

Chenbo Zhou: Writing – original draft, Validation, Formal analysis, Visualization, Software, Methodology, Investigation, Conceptualization, Data curation. **Shuai Su:** Validation, Formal analysis, Visualization, Software, Methodology, Investigation, Data curation. **Qijun Chen:** Methodology, Writing – review & editing, Funding acquisition, Resources, Supervision, Project administration. **Rui Fan:** Methodology, Writing – review & editing, Resources, Supervision, Project administration, Data curation, Formal analysis, Validation, Software, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Qijun Chen reports financial support was provided by Ministry of Science and Technology. Qijun Chen reports financial support was provided by National Natural Science Foundation of China. Rui Fan reports financial support was provided by Science and Technology Commission of Shanghai Municipality.

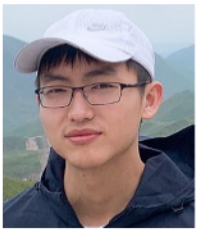
Data availability

The data that has been used is confidential.

References

[1] J. Engel, et al., Direct sparse odometry, IEEE Trans. Pattern Anal. Mach. Intell. 40 (3) (2017) 611–625.
[2] R. Mur-Artal, et al., ORB-SLAM: A versatile and accurate monocular SLAM system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.
[3] Y. Yu, et al., Accurate and robust visual localization system in large-scale appearance-changing environments, IEEE/ASME Trans. Mechatronics 27 (6) (2022) 5222–5232.
[4] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4104–4113.
[5] C. Wu, Towards linear-time incremental structure from motion, in: 2013 International Conference on 3D Vision (3DV), IEEE, 2013, pp. 127–134.
[6] R. Fan, et al., Autonomous Driving Perception, Springer, 2023.
[7] R. Fan, et al., Pothole detection based on disparity transformation and road surface modeling, IEEE Trans. Image Process. 29 (2019) 897–908.
[8] R. Fan, et al., Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms, IEEE Trans. Image Process. 30 (2021) 8144–8154.
[9] R. Fan, M. Liu, Road damage detection based on unsupervised disparity map segmentation, IEEE Trans. Intell. Transp. Syst. 21 (11) (2020) 4906–4911.
[10] Y. Ling, S. Shen, High-precision online markerless stereo extrinsic calibration, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 1771–1778.
[11] J. Wu, et al., Simultaneous hand–eye/robot–world/camera–IMU calibration, IEEE/ASME Trans. Mechatronics 27 (4) (2021) 2278–2289.
[12] E. Rublee, et al., ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 2564–2571.
[13] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[14] M. Muja, D.G. Lowe, Scalable nearest neighbor algorithms for high dimensional data, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2227–2240.
[15] D. DeTone, et al., Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 224–236.
[16] M. Dusmanu, et al., D2-Net: A trainable CNN for joint detection and description of local features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[17] J. Revaud, et al., R2D2: repeatable and reliable detector and descriptor, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), 2019, pp. 12414–12424.
[18] P.-E. Sarlin, et al., Superglue: Learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4938–4947.
[19] U. Efe, et al., DFM: A performance baseline for deep feature matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 4284–4293.
[20] A. Krizhevsky, et al., Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 25, 2012.
[21] R. Fan, et al., Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation, IEEE/ASME Trans. Mechatronics 27 (1) (2022) 225–233.
[22] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV), Vol. 2, IEEE, 1999, pp. 1150–1157.
[23] H. Bay, et al., SURF: Speeded up robust features, in: European Conference on Computer Vision (ECCV), Springer, 2006, pp. 404–417.
[24] P.F. Alcantarilla, T. Solutions, Fast explicit diffusion for accelerated features in nonlinear scale spaces, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2011) 1281–1298.

[25] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 26, 2013.
[26] R. Sinkhorn, P. Knopp, Concerning nonnegative matrices and doubly stochastic matrices, Pacific J. Math. 21 (2) (1967) 343–348.
[27] I. Rocco, et al., Neighbourhood consensus networks, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 31, 2018.
[28] I. Rocco, et al., Efficient neighbourhood consensus networks via submanifold sparse convolutions, in: European Conference on Computer Vision (ECCV), Springer, 2020, pp. 605–621.
[29] X. Li, et al., Dual-resolution correspondence networks, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, 2020, pp. 17346–17357.
[30] Q. Zhou, et al., Patch2Pix: Epipolar-guided pixel-level correspondences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4669–4678.
[31] J. Sun, et al., LoFTR: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8922–8931.
[32] E. Brachmann, C. Rother, Neural-guided RANSAC: Learning where to sample model hypotheses, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4322–4331.
[33] L. Cavalli, et al., Adalam: Revisiting handcrafted outlier detection, 2020, arXiv preprint arXiv:2006.04250.
[34] K.M. Yi, et al., Learning to find good correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2666–2674.
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations (ICLR), Computational and Biological Learning Society, 2015.
[36] K. He, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[37] G. Huang, et al., Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
[38] A.G. Howard, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[39] C. Szegedy, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[40] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.
[41] K.M. Yi, et al., Lift: Learned invariant feature transform, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14, Springer, 2016, pp. 467–483.
[42] Y. Tian, et al., SosNet: Second order similarity regularization for local descriptor learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11016–11025.
[43] D. Barath, et al., MAGSAC: Marginalizing sample consensus, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10197–10205.
[44] V. Balntas, et al., HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5173–5182.
[45] Z. Li, N. Snavely, MegaDepth: Learning single-view depth prediction from internet photos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2041–2050.
[46] M. Tyszkiewicz, et al., DISK: Learning local features with policy gradient, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, 2020, pp. 14254–14265.
[47] L. Lipson, et al., RAFT-Stereo: Multilevel recurrent field transforms for stereo matching, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 218–227.
[48] J. Shi, et al., Good features to track, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 1994, pp. 593–600.
[49] M. Calonder, et al., BRIEF: Binary robust independent elementary features, in: European Conference on Computer Vision (ECCV), Springer, 2010, pp. 778–792.
[50] R. Fan, et al., Road surface 3D reconstruction based on dense subpixel disparity map estimation, IEEE Trans. Image Process. 27 (6) (2018) 3025–3035.
[51] R. Fan, et al., Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation, IEEE Trans. Cybern. 52 (7) (2022) 5799–5808.
[52] R. Fan, et al., SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection, in: European Conference on Computer Vision (ECCV), Springer, 2020, pp. 340–356.



Chenbo Zhou is currently an undergraduate student at Tongji University. His research interests include correspondence matching and stereo vision.

Shuai Su is currently a Ph.D. candidate with the Robotics and Artificial Intelligence Laboratory at Tongji University. His research interests include correspondence matching and simultaneous localization and mapping.

Qijun Chen received the B.S. degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999. He is currently a Full Professor in the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include robotics control, environmental perception, and understanding of mobile robots and bioinspired control.

Rui Fan received the B.Eng. degree in Automation from the Harbin Institute of Technology in 2015 and the Ph.D. degree (supervisors: Prof. John G. Rarity and Prof. Naim Dahnoun) in Electrical and Electronic Engineering from the University of Bristol in 2018. He worked as a Research Associate (supervisor: Prof. Ming Liu) at the Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoctoral Scholar-Employee (supervisors: Prof. Linda M. Zangwill and Prof. David J. Kriegman) at the University of California San Diego between 2020 and 2021. Rui began his faculty career as a Full Research Professor with the College of Electronics & Information Engineering at Tongji University in 2021, and was then promoted to a Full Professor in the same college, as well as at the Shanghai Research Institute for Intelligent Autonomous Systems in 2022. Rui served as an associate editor of ICRA'23 and IROS'23, and as a senior program committee member of AAAI'23/24. Rui is the general chair of the AVVision community and organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, and ECCV'22. Rui was named in Stanford University List of Top 2% Scientists Worldwide in 2022. His research interests include computer vision, deep learning, and robotics.