Correspondence Learning via Correspondence Embedded and Channel Recalibration Network

Yizhang Liu[1], *Shengjie* Zhao[1*], *Hao* Deng[1], *Fuqiang* Ding[2]

[1]School of Software Engineering, Tongji University, Shanghai, 201804, China
[2]Shanghai ideal Information Industry (Group) Co., LTD, Shanghai, 201315, China

Abstract. Correspondence learning is pivotal to many computer vision-based tasks. Existing methods regard each correspondence equally along the channel dimension, which weakens the feature representation capability of the network. To alleviate this problem, we propose a Correspondence Embedded and Channel Recalibration Network, named CECR-Net, to predict the inlier probability of each correspondence and recover camera poses. The proposed CECR-Net is designed to explore the potential impact of correspondences on the channel dimension, and recalibrate the weight of each channel, so that our CECR-Net can capture more exact contextual information. Experiments show that our CECR-Net is effective in outlier removal and camera pose estimation tasks on challenging public datasets.

1 Introduction

Correspondence learning aims to establish reliable correspondences between two images with overlapped regions or similar patterns, which is the basic step of many advanced computer vision tasks, such as 3D reconstruction [1], image registration [2-4], image fusion [5], etc. Existing methods solve the correspondence learning problem by constructing the initial correspondence set using an off-the-shelf feature detector and descriptor (e.g. SIFT [6], SuperPoint [7]). However, due to the complexity of the matching scenes, the correspondence set inevitably contains a high proportion of incorrect correspondences (a.k.a, outliers). Thus, a robust outlier removal method is indispensable as a post-processing to obtain accurate correspondences.

Ma et al. separate traditional outlier removal methods into three types: resampling-based, non-parametric model-based, and relaxed methods [8]. The typical representative of resampling-based methods is RANSAC [9], which introduces a hypothesize-and-verify strategy to learn inliers from the initial correspondence set. Pilet et al. [10] model the transformation by the triangulated 2-D mesh, which can reduce the negative impact of outliers. VFC [11] restricts the transformation function in the reproducing kernel Hilbert space with Tikhonov regularization and estimates it in a Bayesian model. Both of them are non-parametric outlier removal methods. GMS [12] and LPM [13] are relaxed methods, which use simple geometric constraints to adapt to complex scenes. In GMS, it firstly segments the image pair into $n \times n$ meshes, then assigns each correspondence to a specific mesh, and finally compares with the predefined threshold and chooses inliers. LPM determines the correctness of each correspondence according to the number of neighboring correspondences. The aforementioned handcrafted methods are applicable to special scenes, but they may be unsuitable for the explosive growth of general datasets with an extremely low inlier ratio.

Recently, learning-based methods are emerged to deal with the outlier removal problem. Since these methods are data-driven, they are able to obtain rich relationships among correspondences. DSAC [14] alleviates the non-differentiable defect of RANSAC and applies its idea to the deep learning network to remove outliers. Motivated by Point-Net [15], CNe [16] introduces PointNet-like architectures, named ResNet Block, to process input correspondences independently and predict the inlier probability. It brings performance improvement compared to previous traditional methods, but ignoring relationships among correspondences also results in detrimental effects.

Hence, to deal with this issue, some researchers introduce local information of correspondences in various ways. For example, OA-Net [17] proposes a clustering layer, with a combination of differentiable pooling and unpooling operation and an order-aware operation, to catch local and global contextual information of sparse correspondences. LMR [18] builds several consistency properties of correspondences and feeds them into the traditional SVM classifier for the classification of inliers and outliers. NM-Net [19] puts forward a compatibility-specific neighbor mining method, using the local affine information of feature points as prior information, to find more consistent neighbors. ACNe [20] proposes local and global attention based on the Attentive Context Normalization (ACN) mechanism to simultaneously capture local and global contextual information. The

* Corresponding author: shengjiezhao@tongji.edu.cn

above networks integrate local information and also achieve performance improvement.

However, existing learning-based methods equally regard different channels of correspondences, which weakens the feature representation capability of the network. To alleviate this problem, motivated by SE [21], we design a Correspondence Embedded and Channel Recalibration Network, termed CECR-Net, without losing permutation-equivariant. CECR-Net first uses the potential relationship among correspondences to determine the importance of each channel, and then recalibrates the weight of each channel, so that it can capture more accurate contextual information.

The contributions of this paper are summarized as:

i) We design a Correspondence Embedded and Channel Recalibration Network (CECR-Net), which first determines the importance of each channel by the potential relationship among correspondences, and then recalibrates the weight of each channel, to obtain more accurate contextual information.

ii) Our CECR-Net obtains good performance on difficult datasets in camera pose estimation and outlier removal tasks.

2 Methods

In this section, the problem formulation is first introduced. Then, we describe the proposed Correspondence Embedded and Channel Recalibration Module and loss function. After that, we elaborate implementation details.

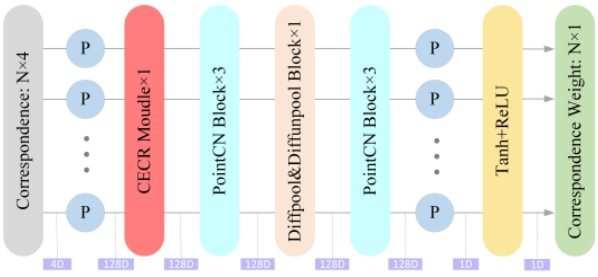

Fig. 1. The architecture of CECR-Net.

2.1 Problem Formulation

Given a pair image (I,I'), off-the-shelf feature detection and description methods (e.g., SIFT or SuperPoint) are used to establish the initial correspondence set $\mathbf{S}$ based on the nearest neighbor matching strategy:

$$\mathbf{S} = \{s_i\} \in \mathbb{R}^{N\times 4}, i = 1,2,\cdots N, \quad (1)$$

where $s_i = (x_1^i, y_1^i, x_2^i, y_2^i)$ is the i-th correspondence; N is the number of correspondences; (x_1^i, y_1^i) and (x_2^i, y_2^i) are the coordinates of i-th correspondence in two images.

As shown in Fig. 1, initial correspondences are taken as the input into a permutation-equivariant network. Then, the network outputs a logic value set $\boldsymbol{z}$ for the correspondence classification. To generate a weight set $\mathbf{w} = \{w_i\} \in \mathbb{R}^{N\times 1}$, we perform a tanh and a ReLU activation function on the logic value set $\boldsymbol{z}$, where w_i represents the inlier probability of the correspondence s_i.

After that, we use the weighted eight-point algorithm to calculate the estimated essential matrix $\hat{\mathbf{E}}$ according to the initial correspondence set $\mathbf{S}$ and the weight set $\mathbf{w}$. The above process can be formulated as:

$$\boldsymbol{z} = f_\Phi(\mathbf{S}), \quad (2)$$

$$\mathbf{w} = \tanh(\mathrm{ReLU}(\boldsymbol{z})), \quad (3)$$

$$\hat{\mathbf{E}} = g(\mathbf{w}, \mathbf{S}), \quad (4)$$

where $f_\Phi$is a permutation-equivariant network; Φ is the network parameter set; $g(\cdot,\cdot)$ is the weighted eight-point algorithm.

2.2 Correspondence Embedded and Channel Recalibration Module

In this section, as shown in Fig. 2, we detail the Correspondence Embedded Channel Recalibration Module (CECR Module), which consists of a Correspondence Embedded Block (CE Block) and a Channel Recalibration Block (CR Block). CECR Module can recalibrate channel weights, which can shift the attention of our network to important channels, and focus less on the remaining channels, to improve the feature representation capability.

2.2.1 Correspondence Embedded Block

Learning-based methods often ignore the importance of each channel, which is irrationality. Hence, to alleviate the problem, we design a CE Block, which can mine potential channel information, to make the network distinguish channel importance and gain more robust feature representation abilities. Firstly, a Feature Map Layer (FML), which consists of a Shared Perceptron, a Instance Normalization, a Batch Normalization and a ReLU activation function, further transforms the feature map $\mathbf{F}$ into another feature map $\mathbf{U} = \{u_i\} \in \mathbb{R}^{N\times C}, i = 1,2,\cdots N$, which exhibits useful information for estimating channel importance. Then, we take average pooling to capture potential information among correspondences, but it is not limited to this operation. For example, standard pooling and maximum pooling methods are also feasible. Specifically, we embed correspondence potential information of the feature map $\mathbf{U}$ into a channel weight vector $\mathbf{cs} = \{cs_j\} \in \mathbb{R}^{1\times C}, j = 1,2,\cdots C$. The above operations can be written as:

$$\mathbf{U} = FML(\mathbf{F}), \quad (5)$$

$$\mathbf{cs} = \mathrm{ave}(\mathbf{U}), \quad (6)$$

where $FML(\cdot)$ refers to a Feature Map Layer; $\mathrm{ave}(\cdot)$ refers to average pooling.

2.2.2 Channel Recalibration Block

We present a CR Block, which can learn nonlinear interactions among channels, to utilize the channel weight vector $\mathbf{cs}$ to recalibrate each channel weight. As shown in Fig. 2, CR Block is composed of a Linear Connected Layer, a ReLU activation function, another Linear Connected Layer and a sigmoid activation function. Firstly, using a Linear Connected operation compresses the channel weight vector $\mathbf{cs}$ to $\mathbf{cs}_1 \in \mathbb{R}^{1\times\frac{C}{r}}$

based on reduction ratio r. Next, to excitate $\mathbf{cs}_1$, we perform a ReLU activation function on $\mathbf{cs}_1$ and gain $\mathbf{cs}_2 \in \mathbb{R}^{1\times\frac{C}{r}}$. Then, to restore the channel weight $\mathbf{cs}_3 \in \mathbb{R}^{1\times\frac{C}{r}}$, we perform a Linear Connected operation on $\mathbf{cs}_2$, followed by a sigmoid activation function, so that a weight similarity vector $\mathbf{v} \in \mathbb{R}^{1\times C}$ can be obtained. Finally, to obtain the recalibration feature map $\tilde{\mathbf{x}} = \{\tilde{x}_c\} \in \mathbb{R}^{N\times C}, c = 1,2,\cdots C$, we make a Hadamard product between the feature map $\mathbf{U}$ and the weight similarity vector $\mathbf{v}$.

$$\mathbf{v} = \sigma(LC_2(\delta(LC_1(\mathbf{cs})))), \tag{7}$$

$$\tilde{\mathbf{X}} = \mathbf{U} \odot \mathbf{v}, \tag{8}$$

where σ is a ReLU activation function; δ is a sigmoid activation function; $LC_1(\cdot)$ and $LC_2(\cdot)$ are two linear connected layers; $\odot$ refers to a Hadamard product.

CE Block can distinguish the importance of each channel through the potential relationship among correspondences. CR Block is used to recalibrate the weight of all channels. In our CECR Module, we combine both of them, so that it can capture more accurate contextual information and make our CECR-Net perform better.

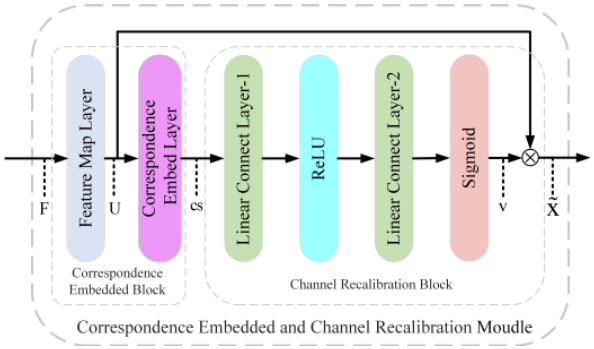

Fig. 2. The architecture of CECR module.

2.3 Loss Function

Following CNe [16], we adopt a hybrid loss function to constrain our CECR-Net, which includes a binary cross entropy loss function to classify and a loss function $L_{ess}(\cdot,\cdot)$ for regressing the essential matrix:

$$L = L_{clc}(\mathbf{w}, \mathbf{G}) + \alpha L_{ess}(\mathbf{E}, \hat{\mathbf{E}}) \tag{9}$$

where $\mathbf{w}$ and $\mathbf{G}$ represent the weight set and weakly supervised ground truth, respectively. We choose the later under the epipolar error threshold of 0.0001; $\mathbf{E}$ and $\hat{\mathbf{E}}$ are the ground truth essential matrix and the predicted essential matrix, respectively. α is a weight parameter, which is used to balance these two losses.

2.4 Implementation Details

As shown in Fig. 1, CECR-Net is orderly composed of a Perceptron Layer, a CECR Module, three PointCN Blocks, a Diifpool & Diffunpool Block, another three PointCN Blocks, another Perceptron Layer as well as a tanh and ReLU activation function. Specifically, the PointCN Block includes Perceptron, Context Normalization, Batch Normalization and ReLU activation function. And the Diifpool & Diffunpool Layer is proposed in OA-Net [17]. Following OA-Net, CECR-Net adopts two iterations. Specifically, the first Perceptron Layer maps the initial correspondence set $\mathbf{S} \in \mathbb{R}^{N\times 4}$ into a feature map set $\mathbf{F} \in \mathbb{R}^{N\times 128}$. The logit value set $\mathbf{z} \in \mathbb{R}^{N\times 1}$ and the weight set are obtained $\mathbf{w} \in \mathbb{R}^{N\times 1}$, due to the feature map set $\mathbf{F}$ handled by the remaining.

In the training phase, we use a learning rate of 0.001 with Adam, a batch size of 32. The parameter βstarts at 0 and becomes 0.1 after 20k iterations. Experiments run on Ubuntu 18.04 with NVIDIA GTX 3090 GPUs.

3 Experiments

3.1 Evaluation Protocols

Yahoo's YFCC100M dataset [22] is used for training and test. It has 72 sequences, in which 68 sequences as training sequences, and the remaining sequences as unknown scenes for test. Training sequences are separated into training (60%), validation (20%), and testing (20%), where the last one is regarded as known scenes. Following [23-26], we choose mAP5° as the metric to evaluate the estimated essential matrix, and use Precision (P), Recall (R) and F-score (F) for the outlier removal task [27].

3.2 Camera Pose Estimation

Recovering camera poses, a fundamental step for advanced computer vision tasks, requires a large number of inliers. Classical method RANSAC and five learning-based networks Point-Net++ [15], ACNe [20], CNe [16], OANet++ [17] and NM-Net [19] are chosen as comparison algorithms. We evaluate our CECR-Net on both known and unknown scenes.

In Table 1, it is evident that our CECR-Net achieves the best results among all competitors. Particularly, our CECR-Net achieves 34.90% and 43.05% mAP5° without RANSAC in the known and unknown outdoor scenes, respectively. In addition, our CECR-Net, with RANSAC for post-processing, still has obvious advantages over comparison algorithms in any case.

Table 1. Quantitative comparison for camera pose estimation with other state-of-the-art methods on known and unknown scenes. The best results are boldfaced.

Method	YFCC100M(%)	
	Known	Unknown
RANSAC	-/5.81	-/9.07
Point-Net++	10.49/33.78	16.48/46.25
ACNe	29.17/40.32	33.06/50.89
CNe	13.81/34.55	23.95/48.03
OA-Net++	32.57/41.53	38.95/52.59
NM-Net	-/-	32.93/51.90
CECR-Net	**34.90/43.33**	**43.05/53.78**

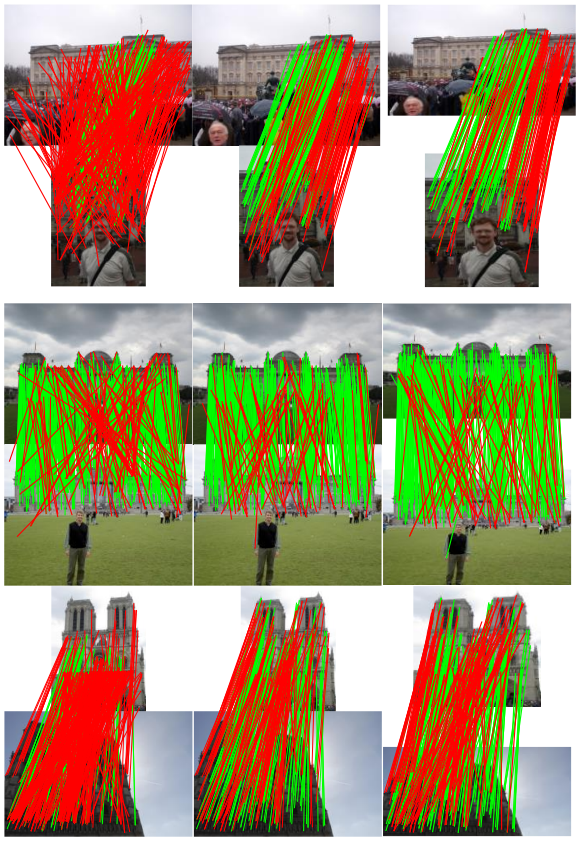

Fig. 3. Partial typical visualization results of RANSAC, OA-Net++ and our CECR-Net (from left to right). The green lines and red lines are inliers and outliers, respectively.

3.3 Outlier Removal

Outlier removal is an important step for many high-level computer vision tasks. Hence, we further evaluate our CECR-Net and comparative algorithms on the outlier removal task, where comparative algorithms are the same as that in the camera pose estimation task. As shown in Table 2, our CECR-Net outperforms other state-of-the-arts on P, R and F in most cases. Besides, compared with traditional RANSAC, learning-based methods have obvious advantages on public datasets with a high outlier rate in the outlier removal.

Beyond that, partial visualization results (RANSAC, OANet++ and CECR-Net) are displayed in Fig. 3. We can find that the visualization results of our proposed CECR-Net are better than others in all scenarios.

Table 2. Quantitative comparison for outlier removal with other state-of-the-art methods on known and unknown scenes. The best results are boldfaced.

Dataset	YFCC100M(%)					
Method	Known			Unknown		
	P	R	F	P	R	F
RANSAC	47.35	52.39	49.74	43.55	50.65	46.83
Point-Net++	49.62	86.19	62.98	46.39	84.17	59.81
ACNe	60.02	88.99	71.69	55.62	85.47	67.39
CNe	54.43	86.88	66.93	52.84	85.68	65.37
OA-Net++	60.03	**89.31**	71.80	55.78	**85.93**	67.65
NM-Net	-	-	-	55.30	85.80	64.71
CECR-Net	**61.97**	88.74	**72.98**	**58.54**	85.36	**69.45**

3.4 Ablation Studies

In this section, we conduct ablation studies about the reduction ratio presented, which is used in the CR Block. We set the reduction rate to {1, 2, 4, 8, 16, 32} for training and test, respectively.

As shown in Table 3, we can find that although the reduction ratio r has little effect on the network performance, when $r = 16$, the model achieves the best performance. Meanwhile, we can also find that with the increasing of the reduction ratio r, the amount of parameters gradually decreases, and when $r = 8$, the parameter quantity hardly decreases. Therefore, considering the effect and efficiency, we set the reduction rate to 16 as default.

Table 3. The performance statistics of different reduction ratio r in CR block. The mAP5° without/with RANSAC is reported. The best results are boldfaced.

Reduction Ratio (r)	Known		Unknown		Param (M)
	-	RANSAC	-	RANSAC	
1	34.49	42.93	41.55	53.08	2.57
2	34.00	42.21	40.63	53.25	2.54
4	35.40	43.18	40.03	52.98	2.52
8	34.49	43.06	41.38	53.73	**2.51**
16	**34.90**	**43.33**	**43.05**	**53.78**	**2.51**
32	34.46	42.91	41.95	52.90	**2.51**

4 Conclusion

In this paper, we propose Correspondence Embedded and Channel Recalibration (CECR-Net), to find reliable correspondences. Our CECR-Net firstly learns the importance of each channel through the potential relationship among correspondences, and afterwards reassigns weights to all channels, to acquire stronger expression contextual information. Comparative experimental results between our CECR-Net and state-of-the-art methods prove that our CECR-Net is effective in camera pose estimation and outlier rejection tasks, even under challenging datasets. However, our CECR-Net also results in some inliers being misjudged, leading to a small decrease in Recall metric. In the future, we will try to use a full-size verification that recover the inliers in the initial correspondence set that are misjudged.

Acknowledgement

This work is supported in part by the National Key Research and Development Project under Grant 2019YFB2102300, in part by the National Natural Science Foundation of China under Grant 61936014, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, in part by the Shanghai Science and Technology Innovation Action Plan Project No. 22511105300, in part by the Fundamental Research Funds for the Central Universities.

References

1. Z. Kang, J. Yang, Z. Yang, and S. Cheng, "A review of techniques for 3d reconstruction of indoor environments," ISPRS Int. Geo-Inf., vol. 9, no. 5, p. 330, 2020.
2. Y. Liu, B. N. Zhao, S. Zhao, and L. Zhang, "Progressive Motion Coherence for Remote Sensing Image Matching," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–13, 2022.
3. Y. Liu et al., "Motion Consistency-Based Correspondence Growing for Remote Sensing Image Matching," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2021.
4. Y. Liu, B. N. Zhao, and S. Zhao, "Rectified Neighborhood Construction for Robust Feature Matching With Heavy Outliers," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
5. J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," Inf. Fusion, vol. 45, pp. 153–178, 2018.
6. D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int. J. Comput. Vis., vol. 60, pp. 91–110, 2004.
7. D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit,, pp. 337–33712, 2017.
8. J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image Matching from Handcrafted to Deep Features: A Survey," Int. J. Comput. Vis., vol. 129, pp. 23–79, 2020.
9. M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981.
10. J. Pilet, V. Lepetit, and P. Fua, "Fast non-rigid surface detection, registration and realistic augmentation," Int. J. Comput. Vis., vol. 76, no. 2, pp. 109–122, 2008.
11. J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust Point Matching via Vector Field Consensus," IEEE Trans. Image Process., vol. 23, pp. 1706–1721, 2014.
12. J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T. D. Nguyen, and M.-M. Cheng, "GMS: Grid-Based Motion Statistics for Fast, Ultra-robust Feature Correspondence," Int. J. Comput. Vis., vol. 128, pp. 1580–1593, 2017.
13. J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," Int. J. Comput. Vis., vol. 127, no. 5, pp. 512–531, 2019.
14. E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit, 2017, pp. 6684–66921
15. C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 652–660.
16. K. Moo Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2666–2674.
17. J. Zhang et al., "OANet: Learning Two-View Correspondences and Geometry Using Order-Aware Network," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 3110–3122, 2022.
18. J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "Lmr: Learning a two- class classifier for mismatch removal," IEEE Trans. Image Process., vol. 28, no. 8, pp. 4045–4059, 2019.
19. C. Zhao, Z. Cao, C. Li, X. Li, and J. Yang, "Nm-net: Mining reliable neighbors for robust feature correspondences," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 215–224.
20. W.Sun,W.Jiang,E.Trulls,A.Tagliasacchi,andK.M. Yi,"Acne:Attentive context normalization for robust permutation-equivariant learning," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 11 286–11 295.
21. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.
22. B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," Communications of the ACM, vol. 59, no. 2, pp. 64–73, 2016.
23. X. Liu and J. Yang, "Progressive Neighbor Consistency Mining for Correspondence Pruning," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 9527–9537.
24. L. Dai et al., "MS2DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph," in Proceedings of IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 8973–8982.
25. X. Liu, G. Xiao, R. Chen, and J. Ma, "Pgfnet: Preference-guided filtering network for two-view correspondence learning," IEEE Trans. Image Process., vol. 32, pp. 1367–1378, 2023.
26. L. Dai et al., "Enhancing two-view correspondence learning by local-global self-attention," Neurocomputing, vol. 459, pp. 176–187, 2021.
27. Y. Liu et al., "Robust feature matching via advanced neighborhood topology consensus," Neurocomputing, vol. 421, pp. 273–284, 2021.