SC-Net: Robust Correspondence Learning via Spatial and Cross-Channel Context

Shuyuan Lin[1], **Hailiang Liao**[1], **Qiang Qi**[2], **Junjie Huang**[1], **Taotao Lai**[3], **Jian Weng**[1*]

[1]College of Cyber Security, Jinan University, Guangzhou, China
[2]School of Data Science, Qingdao University of Science and Technology, Qingdao, China
[3]School of Computer and Data Science, Minjiang University, Fuzhou, China
swin.shuyuan.lin@gmail.com

Abstract

Recent research has focused on using convolutional neural networks (CNNs) as the backbones in two-view correspondence learning, demonstrating significant superiority over methods based on multilayer perceptrons. However, CNN backbones that are not tailored to specific tasks may fail to effectively aggregate global context and oversmooth dense motion fields in scenes with large disparity. To address these problems, we propose a novel network named SC-Net, which effectively integrates bilateral context from both spatial and channel perspectives. Specifically, we design an adaptive focused regularization module (AFR) to enhance the model's position-awareness and robustness against spurious motion samples, thereby facilitating the generation of a more accurate motion field. We then propose a bilateral field adjustment module (BFA) to refine the motion field by simultaneously modeling long-range relationships and facilitating interaction across spatial and channel dimensions. Finally, we recover the motion vectors from the refined field using a position-aware recovery module (PAR) that ensures consistency and precision. Extensive experiments demonstrate that SC-Net outperforms state-of-the-art methods in relative pose estimation and outlier removal tasks on YFCC100M and SUN3D datasets.

Code — http://www.linshuyuan.com

Introduction

Two-view correspondence, a core task in computer vision, aims to establish reliable matches between image pairs for recovering camera geometry. It underpins many downstream applications such as panoramic stitching (Brown and Lowe 2007), simultaneous localization and mapping (Placed et al. 2023) and structure-from-motion (Schonberger and Frahm 2016). However, due to the limited discriminative power of local descriptors (Lin et al. 2024b), putative matches often contain outliers, especially under repetitive patterns (Mousavi et al. 2022), large viewpoint changes (Jin et al. 2021), motion blur and occlusion (Lin et al. 2024c).

Recently, ConvMatch (Zhang and Ma 2023a) introduced a convolutional neural network (CNN)-based framework with stacked ResBlocks (He et al. 2016) to extract deep features for correspondence learning, replacing earlier multilayer perceptron (MLP)-based designs. While effective, it struggles to capture long-range dependencies and global context due to the limited receptive field of convolutions (Zhao et al. 2021; Li et al. 2025). ConvMatch$^+$ (Zhang and Ma 2023b) improves upon this by using bilateral convolutions and SE modules (Hu, Shen, and Sun 2018) to preserve channel-level information. However, it still lacks spatially-aware global reasoning and remains constrained by local convolutional operations. Both versions employ a graph attention network (GAT) (Veličković et al. 2017) for structuring motion vectors, but repeated message passing leads to fine-grained spatial information loss and further exacerbates over-smoothing (Chen et al. 2020).

To address the challenges, we propose a new framework, **SC-Net**, which partitions space into grids to explicitly encode spatial priors, and utilizes classification priors and shared positional bias to prevent the loss of spatial information in deeper GATs (as discussed in (Lindenberger, Sarlin, and Pollefeys 2023)) — an issue not addressed by previous methods. Specifically, SC-Net mainly benefits from two key modules. The first is the **Adaptive Focus Regularization module (AFR)**, which facilitates sharper and cleaner message passing along all paths from sample to field. The second is the **Bilateral Field Adjustment module (BFA)** , which establishes bilateral global dependencies across both spatial and channel dimensions, thereby enhancing information interaction between local motion regions and dynamically refining their central motion features. Finally, to ensure consistency and precision, the position-aware recovery module (PAR) is applied as a standard refinement step to recover the final motion vectors from the smoothed field. Collectively, these modules enable each location in the motion field to move beyond local receptive field constraints, attend to relevant positions and better preserve discontinuities across motion patterns at varying scales (Zhang et al. 2024). As shown in Fig. 1, our main contributions are summarized as follows:

- We present SC-Net, a novel network that integrates position-aware attention and cross-channel context modeling to preserve fine-grained local structures for correspondence pruning. In contrast to methods like OANet, SC-Net explicitly encodes spatial priors and achieves superior performance across diverse scenarios.
- We propose AFR to obtain a clean initial motion field

*Corresponding Author.

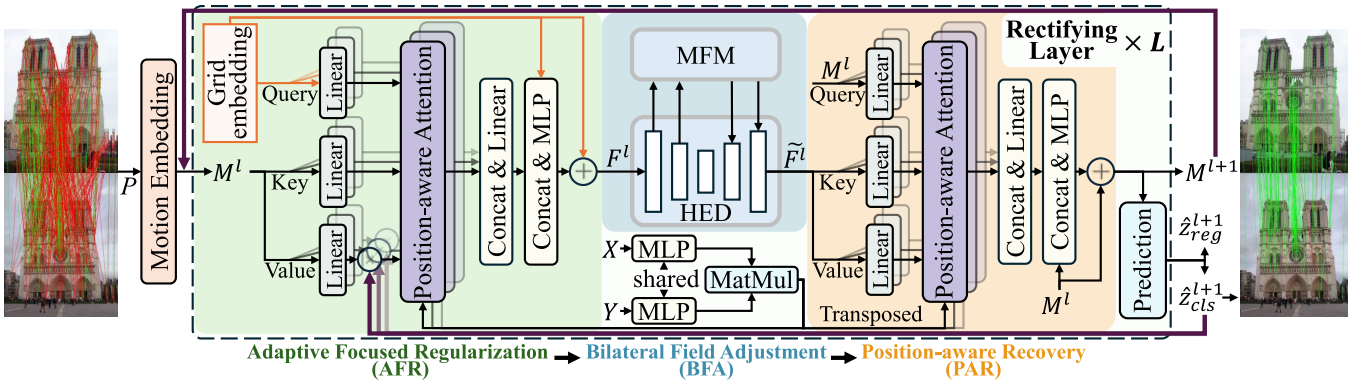


Figure 1: Architecture of the proposed SC-Net. It consists of L rectifying layers, each including AFR, BFA, and PAR. SC-Net takes putative correspondences $P \in \mathbb{R}^{N\times 4}$ as input and predicts the inlier probabilities $\hat{z}_{cls}^{l+1}$ and regression weights $\hat{z}_{reg}^{l+1}$. Purple lines denote propagation to subsequent layer.

from contaminated motion samples, simplifying the complexity of subsequent rectification while overcoming the over-smoothing issue encountered in GNNs.

- We propose BFA, which models long-range relationships across spatial and channel dimensions, enhancing the network's matching capability in complex scenarios.

Related Work

Traditional Outlier Removal Methods

Traditional outlier removal approaches can be broadly categorized into three classes: resampling-based, non-parametric, and relaxed methods. Among them, resampling methods like RANSAC (Fischler and Bolles 1981) and its variants (Barath and Matas 2018; Barath et al. 2020) follow a hypothesize-and-verify paradigm for robust model estimation. While effective under moderate noise, they struggle with complex transformations and heavily contaminated data. To overcome the rigidity of parametric models, non-parametric methods such as VFC (Ma et al. 2014) and SparseVFC (Ma et al. 2013) learn a smooth vector field under generalized geometric constraints, with SparseVFC introducing sparse approximations for improved efficiency. However, their reliance on smoothness priors limits performance in scenes with large motion discontinuities. Relaxed methods aim to handle more diverse scenarios by loosening geometric constraints. For instance, CODE (Lin et al. 2017) integrates a coherence-aware likelihood model, while LPM (Ma et al. 2019) and GMS (Bian et al. 2017) exploit local consistency among matches. Despite their respective strengths, these methods require careful parameter tuning and often fail in the presence of strong outliers or occlusions.

Learning-based Outlier Removal Methods

Learning-based outlier rejection has advanced rapidly with the rise of deep networks (Li et al. 2024; Lin et al. 2025). As an early attempt, LFGC (Yi et al. 2018) formulates outlier rejection as a binary classification task using MLPs, but struggles to capture local context. Subsequent methods enhance contextual modeling within MLP-based frameworks. OANet (Zhang et al. 2019) incorporates spatial correlation and pooling to model local-global dependencies. T-Net (Zhong et al. 2021) and NCMNet (Liu and Yang 2023) further improve context encoding via channel-wise attention and neighbor space interaction. MSGSA (Lin et al. 2024a) extends this by leveraging inter-stage consistency. To break from the limitations of MLPs, ConvMatch (Zhang and Ma 2023a,b) introduces a CNN backbone for better geometric perception. PT-Net (Gong et al. 2024) goes further by stacking CNNs and Transformers in a pyramid to capture both local and global dependencies. DeMatch (Zhang et al. 2024) reformulates outlier rejection by decomposing disordered motions into dominant flows. While effective, existing methods still face challenges in modeling complex motion patterns and long-range dependencies. To address this, we propose a novel network that integrates spatial awareness and cross-channel context to improve both matching accuracy and motion field consistency.

Methodology

Overview

Given a pair of images (I, I'), keypoints and descriptors are first extracted by off-the-shelf methods such as SIFT (Lowe 2004). A nearest neighbor matcher is then used to generate an initial correspondence set P across images:

$$P = \{(x_i, x_i') \mid i = 1, 2, \cdots, N\} \in \mathbb{R}^{N\times 4}, \quad (1)$$

where N is the number of correspondences, and (x_i, x_i') indicates the i-th pair of normalized keypoint coordinates in I and I', respectively. In practice, P typically contains a large number of outliers. Therefore, our goal is to accurately identify inliers and estimate the relative camera motion.

We then embed the motion displacement and corresponding coordinates in the high-dimensional feature space to obtain the initial unordered motion vectors:

$$M^0 = \{\mathcal{F}_1(d_i) + \mathcal{F}_2(x_i) | i = 1, 2, \cdots, N\} \in \mathbb{R}^{N\times C}, \quad (2)$$

where $d_i = x_i' - x_i$ indicates the displacement between keypoint x_i and x_i'; $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$ denote different MLPs that

progressively elevate the dimensionality to extract deep features; C represents the dimension of motion vectors.

After that, the unordered motion vectors are fed into multiple consecutive rectifying layers to correct them and predict the inlier logits $\hat{z}^l_{cls}$ and regression weights $\hat{z}^l_{reg}$:

$$M^l, \hat{z}^l_{cls}, \hat{z}^l_{reg} = h_{\theta_l}(M^{l-1}), \quad l = 1, 2, \cdots, L, \quad (3)$$

where $h_{\theta_l}(\cdot)$ indicates the l-th rectifying layer with learnable parameters θ_l, and L denotes the total number of rectifying layers in the network. $\hat{z}^l_{cls}$ guides the next layer, whereas $\hat{z}^L_{cls}$ is used to classifies correspondences. For the regression task, we calculate the confidence scores $\hat{c}$ based on $\hat{z}^L_{cls}$ and $\hat{z}^L_{reg}$ of the last layer and apply the weighted 8-point algorithm to estimate the parametric model (Sun et al. 2020).

Adaptive Focused Regularization

To achieve learnable regularization (Zhang and Ma 2023a), we first define a bounded 2D space $\Omega = \{(u, v) \in \mathbb{R}^2 | -1 \leq u \leq 1, -1 \leq v \leq 1\}$, where the coordinates of matching pairs are normalized based on the camera intrinsic parameters or image size. This space is uniformly partitioned into a $K \times K$ grid, whose cell centers (hereafter referred to as grid coordinates) are encoded via a MLP to obtain the grid embeddings $G = \{g_k \mid k = 1, 2, \ldots, K^2\} \in \mathbb{R}^{K^2 \times C}$. We then construct a complete graph by connecting each grid embedding g_k and all motion features M^l, and apply a graph attention network $\mathcal{G}$ to allow each cell to focus on motion samples relevant to its spatial region. This enables the model to capture region-specific motion patterns and perform initial motion estimation locally, as follows:

$$F^l = \mathcal{G}(G, M^l), \quad (4)$$

where G serves as the query input; M^l acts as both the key and value inputs; F^l denotes the estimated sparse motion field in the l-th rectifying layer. It is worth noting that stacking additional rectifying layers degrades positional cues and increases sensitivity to outliers, resulting in spatial blurring and over-smoothing (Chen et al. 2020).

To address these issues, we first employ a shared-weight MLP to embed the raw positions and compute positional correlations, which are used to adjust the attention score matrix and enhance the model's spatial sensitivity. Additionally, we incorporate the classification logits $\hat{z}_{cls} \in \mathbb{R}^N$ from the preceding layer to weight the value embeddings, thereby mitigating the adverse effects of spurious motion samples. The attention paradigm in Eq. (4) can be replaced as:

$$O_i = \text{Softmax}(\frac{Q_i K_i^\mathsf{T}}{\sqrt{C_{qk}}} + \psi(\alpha_i \cdot \frac{S}{\sqrt{C}} + \beta_i))\hat{Z}_{cls} V_i, \quad (5)$$

$$S = \mathcal{F}_3(Y)\mathcal{F}_3(X)^\mathsf{T}, \quad (6)$$

$$\hat{Z}_{cls} = \text{Diag}(\sigma(\hat{z}_{cls})), \quad (7)$$

where Q_i, K_i, V_i denote the qeury, key and value matrices for the i-th head; C_{qk} is the dimension of query and key; α_i and β_i are the learnable scale factors and biases for i-th head, respectively; S is the positional correlation matrix shared with all heads; $\psi(\cdot)$ is a leaky ReLU function; X and Y represent the motion coordinates and grid coordinates, respectively; $\sigma(\cdot)$ is a sigmoid activation. Overall, the proposed AFR utilizes two improvements, namely the position-aware attention and the soft filtering, which enable each local region to focus on high-confidence motion samples from nearby locations and adaptively aggregate beneficial context to construct a cleaner motion field.

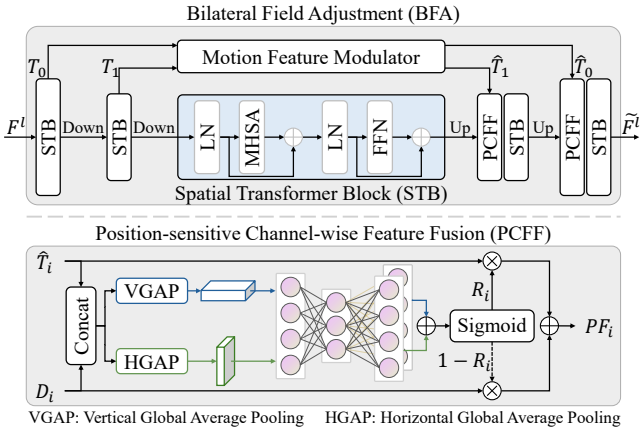


Figure 2: Structure of Bilateral Field Adjustment (BFA).

Bilateral Field Adjustment

To avoid the over-smoothing effect of convolution-based rectification (Zhang and Ma 2023a,b; Gong et al. 2024), we propose BFA, which combines a hierarchical encoder-decoder and a motion feature modulator to better preserve motion discontinuities, as illustrated in Fig. 2.

Hierarchical Encoder-Decoder. As shown in Fig. 2 (a-b), the encoder-decoder primarily consists of spatial transformer blocks, with patch merging and patch expanding layers used for down-sampling and up-sampling (Cao et al. 2022), respectively. Based on these, we construct a hierarchical spatial encoder-decoder to simultaneously capture both low-level motion details and high-level motion pattern features, thereby enabling refined and piecewise smoothing across the motion field. To address the inconsistency between upsampled features and those from skip connections, we propose a novel Position-sensitive Channel-wise Feature Fusion (PCFF) module, which facilitates more accurate feature integration and improves transformer-based decoding. Specifically, PCFF first upsamples the extracted deep feature, denoted as D_i, and concatenates it with the corresponding skip-connected feature $\hat{T}_i$ to form $U_i = [D_i \| \hat{T}_i]$. Then, two 1D global average pooling operations (Hou, Zhou, and Feng 2021) are applied to extract direction-specific statistical cues, which are passed through two MLPs (denoted as $\mathcal{F}_h$ and $\mathcal{F}_w$) with shared nonlinear transformations. The resulting pixel-wise attention maps are used to recalibrate $\hat{T}_i$ and D_i, and to guide their weighted fusion, as follows:

$$\hat{U}^h_i = \mathcal{F}_h(\text{Pool}_h(U_i)), \hat{U}^w_i = \mathcal{F}_w(\text{Pool}_w(U_i)), \quad (8)$$

$$R_i = \sigma(\mathcal{S}(\hat{U}^h_i, \hat{U}^w_i)), \quad (9)$$

$$PF_i = R_i \odot \hat{T}_i + (\mathbf{1} - R_i) \odot D_i, \quad (10)$$

where $\mathrm{Pool}_h(\cdot)$ and $\mathrm{Pool}_w(\cdot)$ represent horizontal and vertical pooling operations, respectively; $\mathcal{S}(\cdot)$ denotes the broadcasting sum operation; $\odot$ indicates the Hadamard product.

Motion Feature Modulator. As shown in Fig. 2, the multi-scale encoded features T_i $(i = 0, 1)$ are fed into the proposed Motion Feature Modulator (MFM), which modulates the channel-wise feature distribution to reduce the motion ambiguities across different scales. To explain further (see Fig. 3), MFM consists of two components: a cross-scale channel attention (CSCA) and a multi-scale feed-forward network (MSFFN). Given the multi-scale encoded features T_i, we first conduct spatial alignment by performing patch embedding on them, thereby obtaining flattened 2D patches $E_i \in \mathbb{R}^{P \times C_i}$. We then take these tokens as queries, treating their concatenation $E_{\sum} \in \mathbb{R}^{P \times \sum_i C_i}$ as keys and values. Similar to self-attention (Vaswani 2017), we generate $Q_i^{CA} \in \mathbb{R}^{P \times C_i}$ and $\{K^{CA}, V^{CA}\} \in \mathbb{R}^{P \times C_\Sigma}$ via layer normalization followed by linear projection. To facilitate the implicit modeling of the contextualized global relationships during the computation of covariance-based attention maps, we embed spatial context information into them using depth-wise convolution. This operation yields $\hat{Q}_i^{CA} \in \mathbb{R}^{P \times C_i}$ and $\{\hat{K}^{CA}, \hat{V}^{CA}\} \in \mathbb{R}^{P \times C_\Sigma}$. Using these, we calculate the attention matrix to weight the values:

$$E_i^{CA} = \mathrm{Linear}(\mathrm{Softmax}(\mathrm{IN}(\frac{(\hat{Q}_i^{CA})^\intercal \hat{K}^{CA}}{\sqrt{C_\Sigma}}))(\hat{V}^{CA})^\intercal), \tag{11}$$

and apply a skip connection to obtain $\hat{E}_i = E_i + E_i^{CA}$, where $\mathrm{IN}(\cdot)$ denotes instance normalization used for steadily propagating the gradient during training (Wang et al. 2022). Traditional feed-forward networks (FFNs) enhance representational capabilities by independently applying non-linear transformations to each pixel position, but they neglects the spatial dependencies between pixels. Therefore, the proposed MSFFN extends the original single path into a dual-path structure and integrates spatial embedding layers with diverse kernel sizes to capture multi-scale local context and model global spatial dependencies (as shown in Fig. 3). It also employs recursive skip connections and layer normalization to enhance feature discrimination and enable adaptive input-residual fusion for better optimization (Liu et al. 2020). Given the partitioned feature CF_i in MSFFN, the recursive fusion process is formulated as follows:

$$CF_i{}^0 = \mathrm{Dconv}(CF_i), \tag{12}$$

$$CF_i{}^{r+1} = \mathrm{LN}(CF_i + CF_i{}^r), r = 0, 1, 2, \tag{13}$$

where $\mathrm{Dconv}(\cdot)$ represents the depth-wise convolution, and $\mathrm{LN}(\cdot)$ denotes the layer normalization. To maintain spatial consistency, MSFFN output is first fused with $\hat{E}_i$, then upsampled to original resolution via bilinear interpolation and refined by a convolution, together forming the patch recovery layer before entering PCFF. Following this, we obtain a smoothed motion field $\tilde{F}^l$, which is further refined by PAR to recover implicitly rectified motion features M^{l+1}. Based on the residual information between M^l and M^{l+1}, we predict $\hat{z}_{cls}^L$ and $\hat{z}_{reg}^L$ using two ResBlocks and a linear layer.

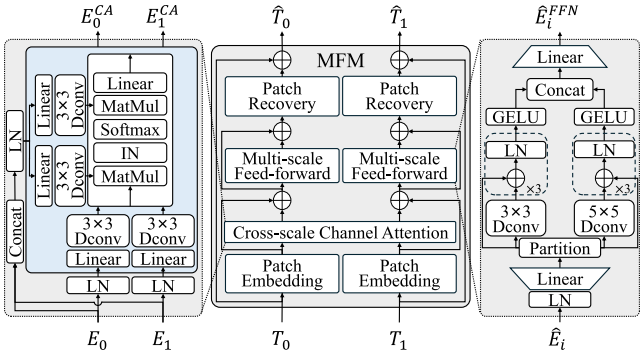


Figure 3: Structure of Motion Feature Modulator (MFM).

Loss Function

We optimize SC-Net using a hybrid loss, which combines the correspondence classification $\mathcal{L}_{cls}$ and the parametric model regression $\mathcal{L}_{reg}$ (Yi et al. 2018; Zhang et al. 2019):

$$\mathcal{L} = \sum\nolimits_{l=0}^{L-1} (\mathcal{L}_{cls}(\hat{z}_{cls}^l, z_{cls}) + \lambda \mathcal{L}_{reg}(\hat{E}^l, E)), \tag{14}$$

where L is the number of rectifying layers; λ is a balance weight; z_{cls} denotes weakly correspondence labels derived by the epipolar distance threshold 10^{-4}; $\hat{E}^l$ and E represent the predicted and the ground-truth essential matrix, respectively. To alleviate label ambiguity, we adopt the adaptive temperature τ from (Zhao et al. 2021) in the classification loss:

$$\mathcal{L}_{cls}(\hat{z}_{cls}^l, z_{cls}) = \mathcal{H}(\tau \odot \hat{z}_{cls}^l, z_{cls}), \tag{15}$$

where $\mathcal{H}(\cdot, \cdot)$ is the binary cross-entropy. The regression loss follows (Ranftl and Koltun 2018; Zhang et al. 2019).

Experiments

Implement Details

We establish up to $N = 2000$ initial correspondences using SIFT and nearest neighbor matching and additionally use RootSIFT (Arandjelović and Zisserman 2012) and SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018) to further validate the generalization ability of the comparison methods. The model takes normalized input correspondences, scaled to $[-1, 1]$ according to the image size or intrinsics. We employ six rectifying layers ($L = 6$) with $K = 16$ for efficiency and four heads in all GATs. In MFM, input features are split into $P = 16$ patches. We use the Adam optimizer (Kingma and Ba 2015) with an initial learning rate of 10^{-4} , decaying by a factor of 9.6×10^{-5} over the first $80k$ steps. SC-Net is trained for $500k$ steps, with λ increasing from 0 to 0.5 after $20k$ steps. All experiments are conducted on Ubuntu 18.04 with an RTX3090.

Datasets and Evaluation Protocols

Following (Zhang et al. 2019), we evaluate SC-Net on both outdoor and indoor scenes. For the outdoor scenes, we use a subset of the YFCC100M dataset (Thomee et al. 2016), which includes 72 image sequences of tourist landmarks, with 68 used as the known scenes and 4 as the unknown scenes. For the indoor scenes, we use the SUN3D

Method	YFCC100M (%)				SUN3D (%)			
	Known		UnKnown		Known		Unknown	
	5°	20°	5°	20°	5°	20°	5°	20°
RANSAC	5.74	16.67	9.05	22.71	4.43	15.38	2.85	11.23
LFGC	14.51	35.82	23.71	50.57	11.93	36.03	9.73	33.09
DFE	19.27	42.14	30.55	59.15	14.18	39.14	12.13	26.26
ACNe	29.63	52.71	34.00	62.98	19.08	46.32	14.27	39.29
OANet	33.50	57.53	41.33	68.79	22.41	49.23	17.57	42.61
T-Net	40.86	63.81	46.74	73.11	23.55	50.99	17.69	44.03
MSA-Net	37.40	60.16	48.45	73.23	18.51	45.74	15.26	41.00
MS^{2}DG-Net	37.78	62.78	46.98	75.13	22.93	50.67	17.34	43.41
NCMNet	50.12	71.15	63.85	82.44	25.68	52.20	20.64	46.24
U-Match	46.03	67.60	60.25	79.70	26.40	53.55	22.38	48.62
ConvMatch	43.12	65.57	55.45	77.49	27.45	54.65	22.52	48.65
ConvMatch$^+$	46.26	68.45	57.08	79.17	28.21	55.74	23.08	49.12
PT-Net	47.41	68.90	57.58	79.39	27.23	54.38	22.88	48.52
DeMatch	47.53	69.08	59.98	79.97	28.50	55.61	23.51	49.84
MSGSA	45.94	67.67	57.93	78.78	26.47	54.26	21.01	47.43
BCLNet	53.21	73.48	67.85	84.57	24.32	51.24	20.06	45.83
CGR-Net	52.40	72.97	65.97	83.62	26.48	53.41	21.69	47.94
DeMo	49.78	71.33	63.45	82.91	30.16	57.33	24.00	50.44
SC-Net (ours)	**64.32**	**80.40**	**71.75**	**86.49**	**33.56**	**59.74**	**25.94**	**52.05**

Table 1: Quantitative results of the relative pose estimation task on YFCC100M and SUN3D, presented as mAP@5° (%) and mAP@20° (%). The optimal and suboptimal indicators are shown in **bold** and underlined, respectively.

dataset (Xiao, Owens, and Torralba 2013), which is split into 239 sequences for the known scenes and 15 for the unknown scenes. The known scenes are divided into training (60%), validation (20%), and testing (20%), while the unknown scenes are used for generalization evaluation. SC-Net is evaluated on the following two tasks. For the relative pose estimation task, we report the mean Average Precision (mAP) (Zhang et al. 2019) of the maximum angular error in rotation and translation under 5° and 20° thresholds, along with the area under the cumulative error (AUC) (Zhang and Ma 2023a) at the same thresholds for comprehensive comparisons. For the outlier removal task, we use *Precision* (P), *Recall* (R) and *F-score* (F) as the metrics.

Relative Pose Estimation

The relative pose estimation task aims to estimate the positional changes (rotation and translation) between two cameras capturing an image pair. We compare SC-Net with previous state-of-the-art MLP-based methods including LFGC (Yi et al. 2018), DFE (Ranftl and Koltun 2018), ACNe (Sun et al. 2020), OANet (Zhang et al. 2019), T-Net (Zhong et al. 2021), MSA-Net (Zheng et al. 2022), MS^{2}DG-Net (Dai et al. 2022), NCMNet (Liu and Yang 2023), U-Match (Li, Zhang, and Ma 2023), MSGSA (Lin et al. 2024a), BCLNet (Miao et al. 2024), and CGR-Net (Yang et al. 2024), and motion-based methods including ConvMatch (Zhang and Ma 2023a), ConvMatch$^+$ (Zhang and Ma 2023b), PT-Net (Gong et al. 2024), DeMatch (Zhang et al. 2024) and DeMo (Lu et al. 2025) on both known and unknown scenes from YFCC100M and SUN3D. The comparative results are shown in Table 1. We observe that the proposed SC-Net achieves exceptional performance across all scenarios. In terms of mAP at 5°, our method improves performance by 11.11%, 3.90%, 3.40%, and 1.94%, respectively, across four different scenes compared to the suboptimal results. It is evident that, although motion-based methods do not perform as well as several recent MLP-based methods (e.g., BCLNe and CGR-Net) on YFCC100M, they generally exhibit superior performance on SUN3D, demonstrating the robustness of the motion-based framework. By the improvements of spatially adaptive focusing and spatial-channel interaction, SC-Net further releases the potential of this framework, achieving more accurate pose estimation. AUC results for several representative learning-based methods are reported in Table 2. SC-Net consistently outperforms all the competing methods across all evaluation metrics.

Method	YFCC100M (%)				SUN3D (%)			
	Known		UnKnown		Known		Unknown	
	5°	20°	5°	20°	5°	20°	5°	20°
OANet	14.49	48.04	17.00	58.61	8.04	40.14	6.09	34.08
T-Net	18.20	54.48	20.70	62.94	9.03	42.08	6.62	35.66
MSA-Net	15.69	49.68	20.95	62.31	7.24	38.54	6.07	34.32
MS^{2}DG-Net	15.36	52.18	18.84	63.27	8.51	41.81	6.24	35.22
NCMNet	25.02	61.31	32.40	71.82	9.96	42.88	7.92	38.16
U-Match	21.62	57.77	29.57	68.95	10.20	44.43	8.40	40.08
ConvMatch	19.95	55.71	26.68	66.82	10.84	45.43	8.68	40.13
ConvMatch$^+$	21.40	58.26	27.31	68.18	11.24	46.44	8.84	40.53
PT-Net	22.30	59.07	27.62	68.52	10.67	45.11	8.59	39.85
DeMatch	22.40	59.02	30.00	69.35	11.36	46.34	9.20	41.24
MSGSA	21.99	58.26	28.23	68.31	10.43	44.99	7.81	39.00
BCLNet	27.48	63.63	37.22	74.44	9.64	42.37	7.72	37.62
CGR-Net	26.45	62.97	35.01	73.17	10.60	44.29	8.50	39.50
DeMo	24.10	61.15	31.65	71.97	12.17	47.89	9.35	41.74
SC-Net (ours)	**35.27**	**70.47**	**40.30**	**76.44**	**14.47**	**50.34**	**10.36**	**43.28**

Table 2: Quantitative results of the relative pose estimation task on YFCC100M and SUN3D, presented as AUC@5° (%) and AUC@20° (%). The optimal and suboptimal indicators are shown in **bold** and underlined, respectively.

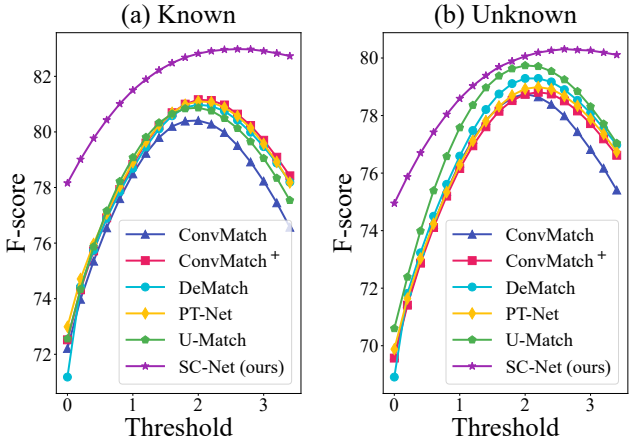


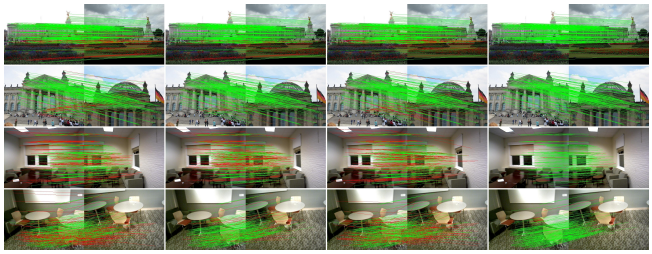
Figure 4: Comparison of F-score results for different models under the different logit thresholds on the known (a) and unknown (b) scenes in YFCC100M.

Outlier Removal

The outlier removal task aims to identify and eliminate incorrect correspondences between image pairs, improv-

Method	YFCC100M (%)						SUN3D (%)					
	Known			Unknown			Known			Unknown		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
RANSAC (Fischler and Bolles 1981)	47.35	52.39	49.47	43.55	50.65	46.83	51.87	56.27	53.98	44.87	48.82	46.76
LFGC (Yi et al. 2018)	54.43	86.88	66.93	52.84	85.68	65.37	53.70	87.03	66.42	46.11	83.92	59.52
DFE (Ranftl and Koltun 2018)	56.72	87.16	68.72	54.00	85.56	66.21	53.96	87.23	66.68	46.18	84.01	59.60
ACNe (Sun et al. 2020)	60.02	88.99	71.69	55.62	85.47	67.39	54.11	88.46	67.15	46.16	84.01	59.58
OANet (Zhang et al. 2019)	61.14	88.16	69.73	57.90	85.07	66.53	54.43	88.08	63.72	46.50	83.83	56.32
T-Net (Zhong et al. 2021)	61.18	89.94	70.47	57.18	87.01	66.73	55.01	88.36	64.18	46.50	83.98	56.33
MSA-Net (Zheng et al. 2022)	59.27	90.28	68.92	56.49	88.60	66.46	56.09	87.57	64.71	48.64	83.81	57.89
MS^{2}DG-Net (Dai et al. 2022)	64.24	89.31	72.49	60.38	86.71	68.96	55.58	89.01	64.63	47.42	84.50	57.12
U-Match (Li, Zhang, and Ma 2023)	63.29	92.09	72.56	61.02	90.67	70.61	55.29	89.35	64.53	47.69	85.60	57.53
ConvMatch (Zhang and Ma 2023a)	63.15	91.19	72.21	60.22	89.48	69.66	55.79	89.23	64.89	48.13	85.54	57.87
ConvMatch$^+$ (Zhang and Ma 2023b)	63.24	91.90	72.52	60.10	89.35	69.58	55.29	89.33	64.56	47.22	85.42	57.15
MSGSA (Lin et al. 2024a)	63.48	91.04	74.80	60.43	89.01	71.98	55.92	88.56	68.55	47.99	84.32	61.22
PT-Net (Gong et al. 2024)	63.89	91.57	72.99	60.45	89.29	69.88	55.41	89.17	64.66	47.45	85.52	57.39
DeMatch (Zhang et al. 2024)	61.32	**92.76**	71.18	58.74	91.04	68.91	56.00	88.90	65.10	48.27	85.24	58.08
NCMNet (Liu and Yang 2023)	77.69	81.27	79.05	76.58	78.58	77.33	66.23	74.69	69.46	61.19	69.07	64.25
BCLNet (Miao et al. 2024)	78.36	82.23	79.87	77.90	80.07	78.73	66.20	74.12	69.19	61.14	68.33	63.92
CGR-Net (Yang et al. 2024)	77.88	81.41	79.22	77.06	79.11	77.84	66.46	74.46	69.51	61.24	68.85	64.18
DeMo (Lu et al. 2025)	64.42	92.68	73.71	61.41	**91.28**	71.22	56.08	**89.66**	65.28	47.87	**85.74**	57.75
SC-Net (ours)	**83.91**	82.96	**82.82**	**82.14**	79.05	**80.06**	**73.13**	74.01	**72.49**	**68.49**	69.88	**68.25**

Table 3: Quantitative results of the outlier removal task on YFCC100M and SUN3D, presented as *Precision* (%), *Recall* (%) and *F-score* (%). The optimal and suboptimal indicators are shown in **bold** and underlined, respectively.



ConvMatch MSGSA PT-Net SC-Net (Ours)

Figure 5: Qualitative results of outlier removal (1st and 2nd rows show the outdoor scenes from YFCC100M, while 3rd and 4th rows show the indoor scenes from SUN3D). False matches are marked in red and correct matches in green.

ing the accuracy of subsequent processing steps. As shown in Table 3, these methods (i.e., NCMNet (Liu and Yang 2023), BCLNet (Miao et al. 2024), and CGR-Net (Yang et al. 2024)) perform correspondence classification based on the epipolar distance of each correspondence with respect to the estimated parametric model, whereas other methods (i.e., LFGC (Yi et al. 2018), DFE (Ranftl and Koltun 2018), ACNe (Sun et al. 2020), OANet (Zhang et al. 2019), T-Net (Zhong et al. 2021), MSA-Net (Zheng et al. 2022), MS^{2}DG-Net (Dai et al. 2022), U-Match (Li, Zhang, and Ma 2023), MSGSA (Lin et al. 2024a), DeMatch (Zhang et al. 2024), ConvMatch (Zhang and Ma 2023a), ConvMatch$^+$ (Zhang and Ma 2023b), and PT-Net (Gong et al. 2024)) identify inliers based on the predicted logits of correspondences. It is obvious that the logit-based methods generally achieve higher recall but relatively lower precision compared to distance-based ones. As a result, they often fail to achieve F-scores comparable to those of distance-based methods. However, we empirically observe that the inlier threshold of logits significantly affects the classification performance. Therefore, we test different thresholds for validating the effectiveness of our method. The visualization results of the F-score under different thresholds are presented in Fig. 4. Our method outperforms other recent logit-based methods across all threshold settings and achieves the highest peak F-score, even significantly surpassing the distance-based ones at certain thresholds. Compared to the suboptimal results of other alternatives, SC-Net obtains an improvement of 1.33% to 4.00% in the F-score metric. We further present a visualization of several matching results in Fig. 5. Our method effectively combines strong mismatch filtering capabilities with the retention of a substantial number of correct matches, leading to more precise pose estimation as in the last section.

Discussions

Generalization Ability

We combine other keypoint detectors (Arandjelović and Zisserman 2012; DeTone, Malisiewicz, and Rabinovich 2018) to evaluate the generalization ability of our method. The results are shown in Table 4. All models are trained with SIFT correspondences but tested with correspondences extracted by other detectors. We can observe that our method outperforms other alternatives in all cases, which effectively demonstrates its robustness and generalization ability. We further evaluate SC-Net on the image matching benchmark PhotoTourism (Jin et al. 2021), with results summarized in Table 5. Table 5 reports cross-dataset generalization results on PhotoTourism and SUN3D using both SIFT and SuperPoint matches. SC-Net achieves the best performance under both descriptors, surpassing previous methods such as BCLNet, ConvMatch$^+$ and DeMo. In particular, its strong performance on SuperPoint—characterized by dense and irregular keypoints—demonstrates SC-Net's superior spatial awareness and generalization. In terms of computational

Method	YFCC100M (%)		SUN3D (%)	
	RootSIFT	SP	RootSIFT	SP
OANet (Zhang et al. 2019)	41.73	19.12	17.41	6.60
T-Net (Zhong et al. 2021)	49.35	23.15	17.93	7.47
MSA-Net (Zheng et al. 2022)	48.45	19.78	16.85	6.40
MS^{2}DG-Net (Dai et al. 2022)	45.95	19.12	17.50	7.08
NCMNet (Liu and Yang 2023)	63.38	23.55	21.44	9.57
U-Match (Li, Zhang, and Ma 2023)	60.85	26.58	22.47	8.46
ConvMatch (Zhang and Ma 2023a)	56.20	28.15	22.77	12.95
ConvMatch$^+$ (Zhang and Ma 2023b)	58.60	30.85	23.54	13.77
PT-Net (Gong et al. 2024)	58.25	30.60	22.96	13.28
DeMatch (Zhang et al. 2024)	61.20	27.15	24.09	14.22
MSGSA (Lin et al. 2024a)	54.42	26.35	21.42	9.14
BCLNet (Miao et al. 2024)	68.37	24.37	20.66	8.17
CGR-Net (Yang et al. 2024)	67.53	21.30	22.42	8.99
DeMo (Lu et al. 2025)	63.98	20.85	24.51	9.92
SC-Net (ours)	**71.97**	**44.08**	**26.34**	**23.52**

Table 4: Intra-dataset generalization evaluation. All models are trained with SIFT matches and then tested with RootSIFT and SuperPoint (SP). mAP@5° is reported.

Method	PhotoTourism (%)		SUN3D (%)		FLOPs (G)	Time (ms)	Memory (GB)
	SIFT	SP	SIFT	SP			
OANet	33.43	19.73	3.89	3.62	117.80	**30.29**	**0.33**
T-Net	42.27	28.24	4.12	3.75	173.10	56.36	0.43
MSA-Net	35.98	19.83	2.44	3.65	**100.22**	52.16	0.43
MS^{2}DG-Net	35.72	24.07	3.71	3.30	322.81	163.90	2.52
NCMNet	54.11	34.97	6.05	4.09	557.92	434.71	2.46
U-Match	45.48	30.62	6.77	3.61	239.40	82.49	0.61
ConvMatch	45.66	33.86	6.33	5.38	242.14	64.20	0.65
ConvMatch$^+$	48.23	39.22	6.95	6.24	344.27	73.11	0.68
PT-Net	49.81	39.19	5.49	6.00	193.25	67.04	0.66
DeMatch	49.69	35.23	7.60	4.88	150.18	37.24	0.46
MSGSA	48.37	33.74	5.98	4.20	299.43	107.65	0.92
BCLNet	56.58	32.11	5.05	2.95	644.05	290.87	1.96
CGR-Net	55.91	30.10	5.85	3.68	424.40	217.50	1.98
DeMo	51.15	33.81	7.65	3.83	231.17	113.80	0.52
SC-Net (ours)	**67.58**	**58.83**	**7.77**	**7.15**	293.84	128.55	1.13

Table 5: Cross-dataset generalization evaluation. All models are trained on YFCC100M with SIFT matches and then applied in inference to other datasets. mAP@5° is reported.

cost, SC-Net maintains moderate FLOPs and peak memory usage, while being significantly more efficient than heavier models like BCLNet and CGR-Net. Overall, SC-Net offers an effective trade-off between accuracy and efficiency.

Parameter Setting

We conduct comparative experiments on SC-Net with varying parameters, including different numbers of rectifying layers and grids. As shown in Table 6, the optimal combination of $L = 6$ and $K = 16$ yields the best performance. Increasing the number of rectifying layers initially improves pose estimation accuracy, but beyond a certain point, performance deteriorates because excessive layers tend to filter out motions with slight deviations from the ideal, reducing reliable correspondences. A larger K value helps refine the motion field but may also introduce additional noise during regularization, thereby complicating subsequent rectification.

Cases	mAP@5°	mAP@20°
$L = 4$	60.79	78.33
$\boldsymbol{L = 6}$	**64.35**	**80.38**
$L = 8$	63.96	**80.38**
$K = 12$	63.24	79.84
$\boldsymbol{K = 16}$	**64.35**	**80.38**
$K = 20$	63.40	80.04

Table 6: Comparison results of the relative pose estimation task for the different number of layers L and the number of grids K in YFCC100M.

Ablation Studies

In this section, we conduct ablation studies on the outdoor scenes from YFCC100M to verify the effectiveness of each component. Notably, by modifying ConvMatch to use soft-mask Softmax for confidence calculation in the weighted 8-point algorithm, the baseline achieves performance gains consistent with the findings of MSGSA. As a foundational component of BFA, the hierarchical encoder-decoder (HED) performs multi-level encoding and decoding of the input motion field, producing more hierarchical and enriched feature representations. This enables it to capture subtle variations and correct the motion field based on high-level motion patterns. Building upon the HED foundation, the motion feature modulator (MFM) dynamically adjusts motion features based on channel-wise feature distributions, effectively reducing ambiguity. In AFR, the soft filtering mechanism (SF) incorporates the classification prediction from the previous layer to suppress noisy motion samples, enhancing the robustness of our network. Meanwhile, position-aware attention (PA) improves spatial awareness by incorporating spatial correlations into the attention mechanism, enabling more precise motion field initialization. As shown in Table 7, the performance of our network progressively improves with the incremental addition of each component. Ultimately, the integration of all components leads to the best overall results.

Baseline	HED	MFM	SF	PA	mAP@5°	mAP@20°
✓					46.09	67.59
✓	✓				57.96	76.72
✓	✓	✓			59.60	77.78
✓	✓	✓	✓		61.96	78.89
✓	✓	✓	✓	✓	**64.35**	**80.38**

Table 7: Ablation results in YFCC100M. BFA consists of **HED** and **MFM**. AFR includes **SF** and **PA**.

Conclusion

In this paper, we propose SC-Net, an effective network designed for challenging correspondence learning tasks. To address key limitations in the existing framework, we design AFR to generate cleaner initial motion fields and construct BFA to capture long-range dependencies across spatial and channel dimensions. These components collectively enhance SC-Net's ability to refine motion fields. We validate SC-Net through extensive experiments on public benchmarks, which demonstrate superior accuracy, robustness, and generalization over existing state-of-the-art methods. Future work includes improving SC-Net's efficiency and extending it to low-overlap or dynamic scenarios.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Nos. U22A2095, 62476112, 62332007, U22B2028, 62501343, 62172197); in part by Guangdong Basic and Applied Basic Research Foundation (Nos. 2024A1515011740, 2025A1515010181); in part by Natural Science Foundation of Shandong Province (No. ZR2024QF294); in part by Fundamental Research Funds for the Central Universities (Nos. 21624404, 23JNSYS01); in part by Science and Technology Major Project of Tibetan Autonomous Region of China (No. XZ202201ZD0006G), National Joint Engineering Research Center of Network Security Detection and Protection Technology, Guangdong Key Laboratory of Data Security and Privacy Preserving, Guangdong Hong Kong Joint Laboratory for Data Security and Privacy Protection, and Engineering Research Center of Trustworthy AI, Ministry of Education.

References

Arandjelović, R.; and Zisserman, A. 2012. Three things everyone should know to improve object retrieval. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2911–2918.

Barath, D.; and Matas, J. 2018. Graph-cut RANSAC. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6733–6741.

Barath, D.; Noskova, J.; Ivashechkin, M.; and Matas, J. 2020. MAGSAC++, a fast, reliable and accurate robust estimator. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1304–1312.

Bian, J.; Lin, W.-Y.; Matsushita, Y.; Yeung, S.-K.; Nguyen, T.-D.; and Cheng, M.-M. 2017. GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4181–4190.

Brown, M.; and Lowe, D. G. 2007. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74: 59–73.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of European Conference on Computer Vision*, 205–218.

Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; and Sun, X. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 34, 3438–3445.

Dai, L.; Liu, Y.; Ma, J.; Wei, L.; Lai, T.; Yang, C.; and Chen, R. 2022. MS2DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8973–8982.

DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 224–236.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of ACM*, 24(6): 381–395.

Gong, Z.; Xiao, G.; Shi, Z.; Wang, S.; and Chen, R. 2024. PT-Net: Pyramid transformer network for feature natching learning. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–11.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hou, Q.; Zhou, D.; and Feng, J. 2021. Coordinate attention for efficient mobile network design. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13713–13722.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Jin, Y.; Mishkin, D.; Mishchuk, A.; Matas, J.; Fua, P.; Yi, K. M.; and Trulls, E. 2021. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2): 517–547.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representation*, 115–123.

Li, C.; Wang, P.; Wang, C.; Zhang, L.; Liu, Z.; Ye, Q.; Xu, Y.; Huang, F.; Zhang, X.; and Yu, P. S. 2025. Loki's dance of illusions: A comprehensive survey of hallucination in large language models. *arXiv preprint arXiv:2507.02870*, 1–31.

Li, Q.; Cheng, J.; Gao, Y.; and Li, J. 2024. Learning Geometric Information via Transformer Network for Key-Points Based Motion Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9): 7856–7869.

Li, Z.; Zhang, S.; and Ma, J. 2023. U-Match: Two-vew correspondence learning with hierarchy-aware local context aggregation. In *Proceedings of International Joint Conference on Artificial Intelligence*, 1169–1176.

Lin, S.; Chen, X.; Xiao, G.; Wang, H.; Huang, F.; and Weng, J. 2024a. Multi-stage network with geometric semantic attention for two-view correspondence Learning. *IEEE Transactions on Image Processing*, 3031–3046.

Lin, S.; Huang, F.; Lai, T.; Lai, J.; Wang, H.; and Weng, J. 2024b. Robust heterogeneous model fitting for multi-source image correspondences. *International Journal of Computer Vision*, 132: 2907–2928.

Lin, S.; Lo, M.; Chen, H.; Liang, Y.; and Wu, Q. 2025. MGCA-Net: Multi-graph contextual attention network for two-view correspondence learning. In *Proceedings of International Joint Conference on Artificial Intelligence*, 1539–1547.

Lin, S.; Yang, A.; Lai, T.; Weng, J.; and Wang, H. 2024c. Multi-motion segmentation via co-attention-induced heterogeneous model fitting. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3): 1786–1798.

Lin, W.-Y.; Wang, F.; Cheng, M.-M.; Yeung, S.-K.; Torr, P. H.; Do, M. N.; and Lu, J. 2017. CODE: Coherence based decision boundaries for feature correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1): 34–47.

Lindenberger, P.; Sarlin, P.-E.; and Pollefeys, M. 2023. Lightglue: Local feature matching at light speed. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 17627–17638.

Liu, F.; Ren, X.; Zhang, Z.; Sun, X.; and Zou, Y. 2020. Rethinking Skip Connection with Layer Normalization. In *Proceedings of International Conference on Computational Linguistics*, 3586–3598.

Liu, X.; and Yang, J. 2023. Progressive neighbor consistency mining for correspondence pruning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9527–9537.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60: 91–110.

Lu, Y.; Le, J.; Li, Z.; Yuan, Y.; and Ma, J. 2025. Deep motion field consensus with learnable kernels for wwo-view correspondence learning. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 39, 5829–5837.

Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; and Guo, X. 2019. Locality preserving matching. *International Journal of Computer Vision*, 127: 512–531.

Ma, J.; Zhao, J.; Tian, J.; Bai, X.; and Tu, Z. 2013. Regularized vector field learning with sparse approximation for mismatch removal. *Pattern Recognition*, 46(12): 3519–3532.

Ma, J.; Zhao, J.; Tian, J.; Yuille, A. L.; and Tu, Z. 2014. Robust point matching via vector field consensus. *IEEE Transactions on Image Processing*, 23(4): 1706–1721.

Miao, X.; Xiao, G.; Wang, S.; and Yu, J. 2024. Bclnet: Bilateral consensus learning for two-view correspondence pruning. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 38, 4225–4232.

Mousavi, V.; Varshosaz, M.; Remondino, F.; Pirasteh, S.; and Li, J. 2022. A two-step descriptor-based keypoint filtering algorithm for robust image matching. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–21.

Placed, J. A.; Strader, J.; Carrillo, H.; Atanasov, N.; Indelman, V.; Carlone, L.; and Castellanos, J. A. 2023. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *IEEE Transactions on Robotics*, 39(3): 1686–1705.

Ranftl, R.; and Koltun, V. 2018. Deep fundamental matrix estimation. In *Proceedings of European Conference on Computer Vision*, 284–299.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Sun, W.; Jiang, W.; Trulls, E.; Tagliasacchi, A.; and Yi, K. M. 2020. Acne: Attentive context normalization for robust permutation-equivariant learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11286–11295.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. Yfcc100m: The new data in multimedia research. *Communications of ACM*, 59(2): 64–73.

Vaswani, A. 2017. Attention is all you need. *Proceedings of Advances in Neural Information Processing Systems*, 5998—-6008.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. 1–12.

Wang, H.; Cao, P.; Wang, J.; and Zaiane, O. R. 2022. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 36, 2441–2449.

Xiao, J.; Owens, A.; and Torralba, A. 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 1625–1632.

Yang, C.; Li, X.; Ma, J.; Zhuang, F.; Wei, L.; Chen, R.; and Chen, G. 2024. CGR-Net: Consistency guided resformer for two-view correspondence learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–16.

Yi, K. M.; Trulls, E.; Ono, Y.; Lepetit, V.; Salzmann, M.; and Fua, P. 2018. Learning to find good correspondences. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2666–2674.

Zhang, J.; Sun, D.; Luo, Z.; Yao, A.; Zhou, L.; Shen, T.; Chen, Y.; Quan, L.; and Liao, H. 2019. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 5845–5854.

Zhang, S.; Li, Z.; Gao, Y.; and Ma, J. 2024. DeMatch: Deep decomposition of motion field for two-view correspondence learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20278–20287.

Zhang, S.; and Ma, J. 2023a. ConvMatch: Rethinking network design for two-view correspondence learning. In *Proceedings of AAAI Conference on Artificial Intelligence*, 3472–3479.

Zhang, S.; and Ma, J. 2023b. ConvMatch: Rethinking network design for two-view correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 2920–2935.

Zhao, C.; Ge, Y.; Zhu, F.; Zhao, R.; Li, H.; and Salzmann, M. 2021. Progressive correspondence pruning by consensus learning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 6464–6473.

Zheng, L.; Xiao, G.; Shi, Z.; Wang, S.; and Ma, J. 2022. MSA-Net: Establishing reliable correspondences by multi-scale attention network. *IEEE Transactions on Image Processing*, 31: 4598–4608.

Zhong, Z.; Xiao, G.; Zheng, L.; Lu, Y.; and Ma, J. 2021. T-Net: Effective permutation-equivariant network for two-view correspondence learning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 1950–1959.