VSFormer: Visual-Spatial Fusion Transformer for Correspondence Pruning

Tangfei Liao[1], Xiaoqin Zhang[1]*, Li Zhao[1], Tao Wang[2], Guobao Xiao[3]*

[1]Key Laboratory of Intelligent Informatics for Safety and Emergency of Zhejiang Province, Wenzhou University, China
[2]State Key Lab for Novel Software Technology, Nanjing University, China
[3]School of Electronics and Information Engineering, Tongji University, China
{tangfeiliao, zhangxiaoqinnan, taowangzj}@gmail.com, lizhao@wzu.edu.cn, x-gb@163.com

Abstract

Correspondence pruning aims to find correct matches (inliers) from an initial set of putative correspondences, which is a fundamental task for many applications. The process of finding is challenging, given the varying inlier ratios between scenes/image pairs due to significant visual differences. However, the performance of the existing methods is usually limited by the problem of lacking visual cues (*e.g.,* texture, illumination, structure) of scenes. In this paper, we propose a Visual-Spatial Fusion Transformer (VSFormer) to identify inliers and recover camera poses accurately. Firstly, we obtain highly abstract visual cues of a scene with the cross attention between local features of two-view images. Then, we model these visual cues and correspondences by a joint visual-spatial fusion module, simultaneously embedding visual cues into correspondences for pruning. Additionally, to mine the consistency of correspondences, we also design a novel module that combines the KNN-based graph and the transformer, effectively capturing both local and global contexts. Extensive experiments have demonstrated that the proposed VSFormer outperforms state-of-the-art methods on outdoor and indoor benchmarks. Our code is provided at the following repository: https://github.com/sugar-fly/VSFormer.

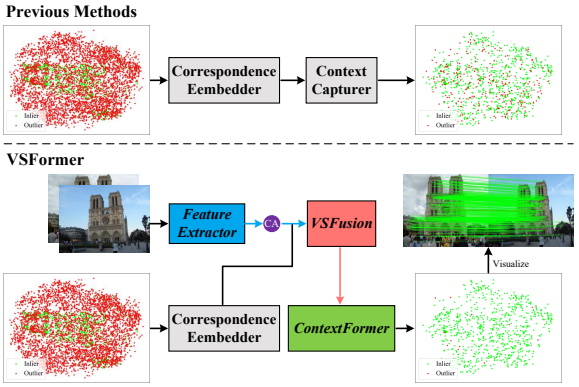


Figure 1: Comparison between previous methods and ours. Top: the architecture of previous methods, which lack the visual perception of a scene. Bottom: the architecture of our VSFormer introduces visual cues of a scene to guide correspondence pruning. For visualization purposes, the correspondences (4D) across two-view images are projected into a 2D space by t-SNE (Van der Maaten and Hinton 2008). The circle CA represents the cross-attention layer.

Introduction

Two-view correspondence learning aims to establish reliable correspondences/matches across two images and accurately recover camera poses, which is a fundamental task in computer vision and plays an important role in many applications such as simultaneous localization and mapping (Mur-Artal, Montiel, and Tardos 2015), structure from motion (Schonberger and Frahm 2016) and image registration (Xiao et al. 2020). However, the outlier (false match) ratio in initial correspondences is often over 90% due to various cross-image variations (*e.g.,* low texture, illumination changes, repetitive structures), which severely undermines the performance of downstream tasks. Therefore, much recent research has focused on pruning false matches from initial correspondences to obtain accurate two-view geometry.

Some traditional methods such as RANSAC (Fischler and Bolles 1981) and its variants (Chum and Matas 2005; Barath, Matas, and Noskova 2019) search for correct correspondences (inliers) using iterative sampling strategies, but their running time grows exponentially with outlier ratio, thus making them unsuitable for tackling high-outlier problems. Meanwhile, learning-based methods have achieved promising performance. PointCN (Yi et al. 2018) is a pioneering work, handling the disordered property of correspondences (visualization in Fig. 1) with the multilayer perceptron (MLP) architecture.

Until recently, the most popular networks commonly employed an iterative network to inherit weights from the previous iteration, which greatly improved the performance of correspondence pruning. These iterative networks typically designed some extra structures to mine the geometric consistency within correspondences, such as OANet (Zhang et al. 2019), ACNet (Sun et al. 2020), MS^{2}DG-Net (Dai et al. 2022). While such a direction deserves further exploration, these methods only considered spatial information of correspondences as input, which significantly hindered the acquisition of deep information and simultaneously dam-

*Corresponding authors

aged the network performance. Thus, there are some researches (Luo et al. 2019; Liu et al. 2021) that employ feature point descriptors of a single image to enhance the representation ability of network inputs. In this paper, we lean toward another perspective and ask the following question: *Can we provide the network with a scene-aware prior at a higher level to guide pruning?* That is, if the inlier ratio of a scene/image pair can be perceived in advance, it will facilitate the network in discriminating some ambiguous correspondences.

To this end, with the observation that the inlier ratio varies greatly between scenes/image pairs due to significant visual differences (*e.g.,* texture, illumination, occlusion), we adopt some scene visual cues as an abstract representation of the inlier ratio. As shown in Fig. 1, compared to previous methods, we add some steps for extracting and fusing scene visual cues. Specifically, the proposed Visual-Spatial Fusion Transformer (VSFormer) is mainly composed of three components: Visual Cues Extractor (VCExtractor), Visual-Spatial Fusion (VSFusion) module, and Context Transformer (ContextFormer). Firstly, the VCExtractor extracts scene visual cues with the cross attention between local features of two-view images. Then, a novel Visual-Spatial Fusion (VSFusion) module is designed to model the relationship between visual cues and spatial cues, simultaneously embedding visual cues into correspondences. The VSFusion involves two phases: i) the module adopts a transformer (Vaswani et al. 2017) to model the complex intra- and inter-modality relationships of visual and spatial cues; ii) the module encodes visual and spatial cues separately, using a simple element-wise summation operation to fuse them; Meanwhile, to facilitate fusion, VSFusion uses soft assignment manner (Zhang et al. 2019) to project spatial cues into the same space as visual cues. The proposed VSFusion effectively embeds scene visual cues into correspondences for guiding subsequent correspondence pruning.

To fully mine contextual information of correspondences, we also propose a novel structure called ContextFormer for pruning, simply stacking a transformer sub-network on top of a graph neural network (GNN). In GNN, a novel graph attention block is designed to improve the representation ability of a KNN-based graph. The block adopts the squeeze-and-excitation mechanism to efficiently capture the potential spatial-, channel-, and neighborhood-wise relations inside a KNN-based graph, facilitating neighborhood aggregation. The proposed structure exploits the neighborhood information of a KNN-based graph and the global modeling ability of the transformer, explicitly capturing both local and global contexts of correspondences, thus further improving the performance of our method.

The contributions of this paper are summarized as follows:

- A visual-spatial fusion transformer is proposed to extract and embed scene visual cues into correspondences for guiding pruning. Meanwhile, we design a joint visual-spatial fusion module to fuse visual and spatial cues in the same space. To the best of our knowledge, this is the first time that scene cues have been introduced for correspondence pruning.
- A simple yet effective ContextFormer is proposed to explicitly capture both local and global contexts of correspondences. In this structure, we also design a graph attention block based on the squeeze-and-excitation mechanism to enhance the representation ability of a KNN-based graph.
- The proposed VSFormer achieves a precision increase of 15.79% and 4.45% compared with the state-of-the-art result on outdoor and indoor benchmarks, respectively.

Related Work

Traditional Methods. RANSAC (Fischler and Bolles 1981) and its variants based on iterative sampling strategies to estimate a geometry model. To be specific, these methods resample the smallest subset of input correspondences to estimate a parametric model as a hypothesis, and then verify its confidence by counting the number of consistent inliers. PROSAC (Chum and Matas 2005) can significantly expedite this process. USAC (Raguram et al. 2012) proposes a unified framework that incorporates multiple advancements for RANSAC variants. MAGSAC (Barath, Matas, and Noskova 2019) uses σ-consensus to eliminate the requirement for a predefined inlier-outlier threshold. RANSAC and its variants have been widely recognized as the standard solution for robust model estimation. However, their performance degrades severely as the outlier ratio increases (Ma et al. 2021).

Learning-Based Methods. PointCN (Yi et al. 2018) is a pioneering work that formulates correspondence pruning as both an essential matrix regression problem and a binary classification problem. It employs MLPs to effectively process the disordered property of correspondences and proposes a context normalization technique to embed global information into each correspondence. DFE (Ranftl and Koltun 2018) proposes a distinct loss function and an iterative network based on PointCN. ACNe (Sun et al. 2020) employs the attention mechanism to enhance the performance of the network. OANet (Zhang et al. 2019) performs full-size prediction for all initial correspondences and introduces a clustering layer to capture local context. MSA-Net (Zheng et al. 2022) and PGFNet (Liu et al. 2023) also propose some jointly spatial-channel attention blocks to capture the global context of correspondences. After that, there are some researches based on the graph neural network (Zhao et al. 2021; Dai et al. 2022; Liao et al. 2023). CLNet (Zhao et al. 2021) introduces a neighborhood aggregation manner and the pruning strategy to refine coarse correspondences. MS^{2}DG-Net (Zhao et al. 2021) builds KNN-based graphs at different stages and employs the multi-head self-attention mechanism to enhance the representation ability of graphs. Although these methods have achieved promising performance, they only consider spatial information of correspondences as input, which severely limits the acquisition of deep information and simultaneously impairs network performance. Therefore, LMCNet (Liu et al. 2021) exploits feature point descriptors of a single image to enhance the representation ability of correspondences. In this paper, with another perspective, jointly visual-spatial cues as a scene-aware prior to guide correspondence pruning.

Method

Problem Formulation

Given two-view images $(\mathbf{I_A}, \mathbf{I_B})$, our task is to precisely identify correct correspondences from initial correspondences and recover camera poses. To be specific, feature points and descriptors are first extracted from two-view images using a feature detector (*e.g.,* SIFT (Lowe 2004) and SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018)). Then, the initial correspondence set $\mathbf{I_C}$ is built by a nearest neighbor matching strategy:

$$\mathbf{I_C} = \{\mathbf{c_1}, \mathbf{c_2}, ..., \mathbf{c_N}\} \in \mathbb{R}^{N\times 4},\ \mathbf{c_i} = (x_i, y_i, x_i', y_i'), \quad (1)$$

where $\mathbf{c_i}$ is the i-th correspondence; (x_i, y_i) and (x_i', y_i') are the feature point coordinates of the given two-view images that have been normalized by camera intrinsics (Zhang et al. 2019).

In our task, the correspondence pruning is typically formulated as an essential matrix regression problem and an inlier/outlier classification problem (Yi et al. 2018). Following CLNet (Zhao et al. 2021), this paper iteratively uses ContextFormer for correspondence pruning and produces the final probability set $\mathbf{P_f} = \{\mathbf{p_1}, ..., \mathbf{p_i}, ..., \mathbf{p_{N'}}\}$ of candidates, which indicates the probability of each candidate as an inlier. The above process can be formulated as follows:

$$(\mathbf{C_f}, \mathbf{W_f}) = f_\phi(\mathbf{I_A}, \mathbf{I_B}, \mathbf{I_C}), \quad (2)$$

$$\mathbf{P_f} = \mathrm{Softmax}(\mathbf{W_f}), \quad (3)$$

where $\mathbf{W_f} = \{\mathbf{w_1}, ..., \mathbf{w_i}, ..., \mathbf{w_{N'}}\}$ represents the weights of final candidates; $\mathbf{C_f} = \{\mathbf{c_1}, ..., \mathbf{c_i}, ..., \mathbf{c_{N'}}\}$ represents the final candidate set; $f_\phi(\cdot)$ indicates our proposed VSFormer; ϕ indicates the network parameters.

Then, the final candidate set $\mathbf{C_f}$ and the probability set $\mathbf{P_f}$ are taken as input, and a weighted eight-point algorithm (Yi et al. 2018) is applied to regress the essential matrix. The process is presented as:

$$\widehat{\mathbf{E}} = g(\mathbf{C_f}, \mathbf{P_f}), \quad (4)$$

where $g(\cdot,\cdot)$ represents a function of the weighted eight-point algorithm, and the matrix $\widehat{\mathbf{E}}$ indicates the predicted essential matrix.

In addition, following (Zhao et al. 2021), this paper also adopts the full-size verification approach to deal with the inlier/outlier classification problem. Specifically, the matrix $\widehat{\mathbf{E}}$ and the initial correspondence set $\mathbf{I_C}$ are taken as inputs to produce the predicted symmetric epipolar distance set $\widehat{\mathbf{D}}$. Note that an empirical threshold (10^{-4}) of epipolar distance is used criterion to discriminate outliers from inliers (Hartley and Zisserman 2003). The process can be formulated as follows:

$$\widehat{\mathbf{D}} = h(\widehat{\mathbf{E}}, \mathbf{I_C}), \quad (5)$$

where $h(\cdot,\cdot)$ represents a function of the full-size verification.

Visual-Spatial Fusion Module

On the one hand, the vast majority of methods only use the spatial information of correspondences as input, which is still challenging for datasets with a large number of outliers. On the other hand, some researches (Luo et al. 2019; Liu et al. 2021) employ feature point descriptors of a single image to further improve the network performance. However, existing methods lack a scene-aware prior to guide correspondence pruning. In this paper, we first base on the fact that the inlier ratio varies greatly between scenes/image pairs due to significant visual differences (*e.g.,* texture, illumination, occlusion). Then, we extract some visual cues of a scene/image pair to abstractly represent the inlier ratio, which is beneficial for the network to distinguish some ambiguous correspondences. To this end, as illustrated in Fig. 2, we propose Visual Cues Extractor (VCExtractor) and Visual-Spatial Fusion (VSFusion) module for extracting and fusing visual cues into correspondences.

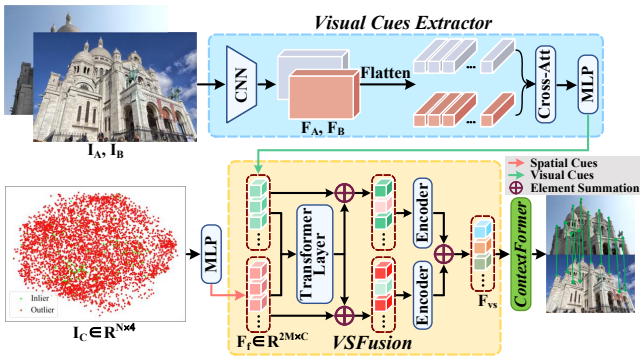


Figure 2: The architecture of our VSFormer mainly contains Visual Cues Extractor (VCExtractor), Visual-Spatial Fusion (VSFusion) Module, and Context Transformer (ContextFormer). Note that we omit the inlier predictor after ContextFormer for simplicity.

Visual Cues Extractor. VCExtractor is used to extract scene visual cues, and when significant visual differences such as occlusions, large viewpoint changes, and illumination changes between two-view images, the attention map scores in the cross-attention layer tend to be generally low. This message is passed through visual cues to the VSFusion and embedded into each correspondence. To be specific, in our VCExtractor, a standard convolution architecture with ResNet34 (He et al. 2016) is first used to extract high-dimensional local features $\{\mathbf{F_A}, \mathbf{F_B}\} \in \mathbb{R}^{C_F\times\frac{H}{4}\times\frac{W}{4}}$ from two-view images $\{\mathbf{I_A}, \mathbf{I_B}\} \in \mathbb{R}^{3\times H\times W}$. Then, local features are flattened into 1-D vectors and delivered into the cross-attention layer (Sun et al. 2021) to produce initial visual cues of a scene. Subsequently, these initial visual cues are embedded with an MLP to obtain visual cues $\mathbf{F_v} \in \mathbb{R}^{M\times C}$ for fusing. In addition, the initial correspondences $\mathbf{I_C} \in \mathbb{R}^{N\times 4}$ are also embedded with an MLP to extract the deep feature $\mathbf{F_s} \in \mathbb{R}^{N\times C}$ as spatial cues.

Visual-Spatial Fusion. The VSFusion is responsible for fusing visual and spatial cues in the same space, and projects jointly visual-spatial cues into the original space. Firstly, since the fusion between two modalities is beneficial in the

same space, VSFusion projects spatial cues into a unified space with visual cues through a learnable soft assignment manner (Zhang et al. 2019). The above process can be formulated as follows:

$$\mathbf{F}_{\mathbf{s}}^{'} = (\mathbf{W})^{T}\mathbf{F}_{\mathbf{s}}, \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{N\times M}$ is a learnable matrix.

Then, a transformer is adopted to robustly model the relationship between visual cues and spatial cues. The transformer layer takes the concatenated feature $\mathbf{F_f} \in \mathbb{R}^{2M\times C}$ of visual and spatial cues as input. For each head in the multi-headed self-attention layer, three learnable matrices $\mathbf{W_Q}$, $\mathbf{W_K}$ and $\mathbf{W_V}$ project the concatenated feature to query $\mathbf{Q} \in \mathbb{R}^{2M\times d}$, key $\mathbf{K} \in \mathbb{R}^{2M\times d}$ and value $\mathbf{V} \in \mathbb{R}^{2M\times d}$, where $d = C/h$ and h is the number of heads. Subsequently, the attention matrix $\mathbf{A} \in \mathbb{R}^{2M\times 2M}$ is calculated by applying a row-wise softmax function on $\mathbf{QK}^{T}$. The messages $\mathbf{F}_{\mathbf{f}}^{'} \in \mathbb{R}^{2M\times C}$ are formulated as $\mathbf{AV}$, which fuse the complex relations between visual and spatial cues. After that, these messages are processed through a feed-forward network and split into $\mathbf{F}_{\mathbf{v}}^{'} \in \mathbb{R}^{M\times C}$ and $\mathbf{F}_{\mathbf{s}}^{''} \in \mathbb{R}^{M\times C}$. The above process can be simply described as:

$$\mathbf{F}_{\mathbf{f}}^{'} = \mathrm{MHSA}(\mathbf{F_f}), \tag{7}$$

$$(\mathbf{F}_{\mathbf{v}}^{'}, \mathbf{F}_{\mathbf{s}}^{''}) = \mathrm{Split}(\mathrm{FFN}(\mathbf{F}_{\mathbf{f}}^{'})), \tag{8}$$

where $\mathrm{MHSA}(\cdot)$ denotes the multi-headed self-attention layer described above; $\mathrm{FFN}(\cdot)$ represents the feed-forward network.

Next, an element-wise summation is employed to obtain jointly visual-spatial cues $\mathbf{F_{vs}} \in \mathbb{R}^{M\times C}$. Meanwhile, skip connections and resnet-like encoders (Yi et al. 2018) are used to rebuild the intra-modality context. The above process can be expressed as:

$$\mathbf{F_{vs}} = \mathrm{R}_1(\mathbf{F_v} + \mathbf{F}_{\mathbf{v}}^{'}) + \mathrm{R}_2(\mathbf{F}_{\mathbf{s}}^{'} + \mathbf{F}_{\mathbf{s}}^{''}), \tag{9}$$

where $\mathrm{R}_1(\cdot)$ and $\mathrm{R}_2(\cdot)$ represent the encoders with different parameters. Finally, similar to Eq. 6, the jointly visual-spatial cues $\mathbf{F_{vs}} \in \mathbb{R}^{M\times C}$ is projected into the original space $\mathbf{F}_{\mathbf{vs}}^{'} \in \mathbb{R}^{N\times C}$.

Context Transformer

In our task, mining the consistency within correspondences is important to search for correct matches. In this paper, to fully capture contextual information of joint visual-spatial correspondences, a Context Transformer (ContextFormer) structure is designed. Intuitively, correct correspondences should be consistent in both their local and global contexts, thus ContextFormer explicitly captures local and global contexts by stacking the graph neural network and transformer, as shown in Fig. 3.

Local Context Capturer. Following (Wang et al. 2019; Zhao et al. 2021), our ContextFormer first builds a KNN-based graph according to the Euclidean distances between each jointly visual-spatial correspondences:

$$\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i), i \in [1, N], \tag{10}$$

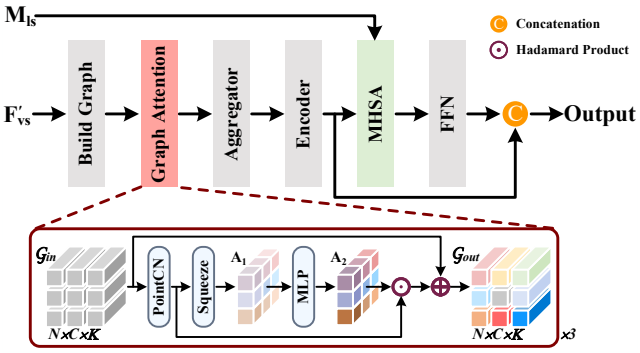


Figure 3: Illustration of our proposed ContextFormer. Meanwhile, we design a novel graph attention block to mine potential relationships along three different dimensions.

where $\mathcal{V}_i = \{f_{i1}, ..., f_{ik}\}$ represents the k-nearest neighbors of f_i in the feature space; $\mathcal{E}_i = \{\boldsymbol{e_{i1}}, ..., \boldsymbol{e_{ik}}\}$ represents the set of directed edges connecting f_i and its neighbors. The construction of edges can be formulated as follows:

$$\boldsymbol{e_{ij}} = [f_i, f_i - f_{ij}], \tag{11}$$

where f_i and f_{ij} represent the i-th jointly visual-spatial correspondence and its j-th neighbor, respectively; $[\cdot, \cdot]$ denotes the concatenate operation along the channel dimension.

Then, the global graph $\mathcal{G}_{in} = \{\mathcal{G}_1, ..., \mathcal{G}_i, ..., \mathcal{G}_N\}$ has rich contextual information, but capturing them only by neighborhood aggregation is not robust. To this end, a novel graph attention block adopts the squeeze-and-excitation mechanism to efficiently capture the potential spatial-, channel-, and neighborhood-wise relations inside the global graph. To be specific, as shown in Fig. 3, the global graph $\mathcal{G}_{in}$ is embedded by a PointCN block (Yi et al. 2018); then, the max-pooling and average-pooling operations along channel dimension to squeeze the global graph, and the element-wise summation operation is applied to produce an initial attention map $\mathbf{A_1} \in \mathbb{R}^{N\times K}$; subsequently, the initial attention map is excited with an MLP to capture neighborhood-wise relations $\mathbf{A_2} \in \mathbb{R}^{N\times K}$ of the global graph. finally, these relationships are embedded into the global graph via Hadamard product, adding a residual to obtain $\mathcal{G}_{out}$. The above operations can be described as:

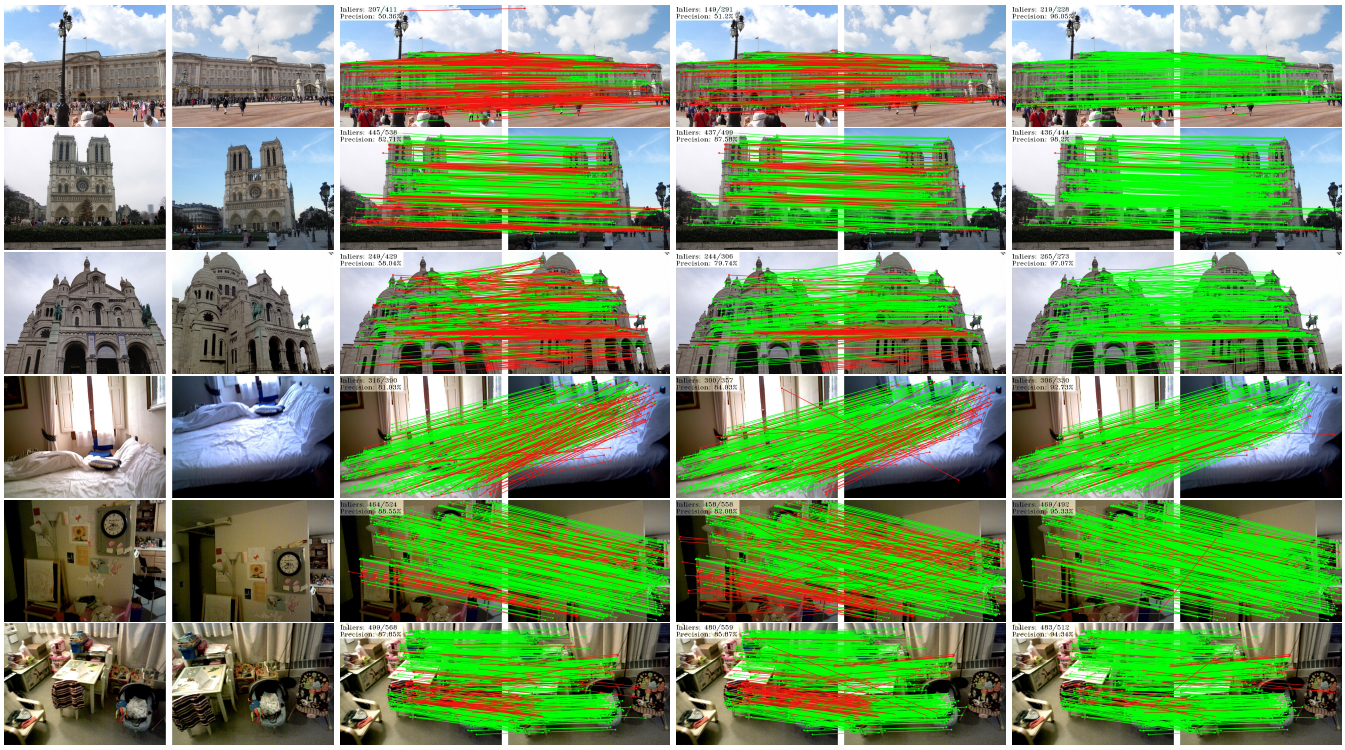
$$\mathcal{G}_{in}^{'} = \mathrm{PointCN}(\mathcal{G}_{in}), \tag{12}$$

$$\mathbf{A_2} = \mathrm{MLP}(\mathrm{AvgPool}(\mathcal{G}_{in}^{'}) + \mathrm{MaxPool}(\mathcal{G}_{in}^{'})), \tag{13}$$

$$\mathcal{G}_{out} = (\mathcal{G}_{in}^{'} \odot \mathbf{A_2}) + \mathcal{G}_{in}. \tag{14}$$

Similar to the neighborhood-wise attention (as described above), the operations of channel- and spatial-wise attention are omitted for simplicity. Despite its simplicity, the graph attention block effectively improves the representation ability of the global graph.

Next, following (Zhao et al. 2021), we perform neighborhood aggregation on the enhanced global graph $\mathcal{G}_{out}$ to obtain the correspondence feature $\mathbf{F_{local}} \in \mathbb{R}^{N\times C}$ embedded with both global and local contexts.



Plain image pair (a) OANet++ (b) CLNet (c) Ours

Figure 4: Partial typical visualization results on two challenging datasets, i.e., YFCC100M, SUN3D. From left to right: the results of OANet++, CLNet, and our VSFormer. From top to bottom: the top three results come from unknown outdoor scenes and the rest come from unknown indoor scenes. The correspondence is drawn in green if it represents the true-positive and red for the false-positive. Best viewed in color.

Global Context Capturer. It mainly involves a multi-headed self-attention (MHSA) layer employed to capture and fuse global context into each correspondence. In particular, this paper introduces length similarity (Bai et al. 2021) into the MHSA layer, which produces a spatially aware attention matrix by combining length consistency and feature consistency. To be specific, given two correspondences $\mathbf{c_i} = (\mathbf{p_i^A}, \mathbf{p_i^B})$ and $\mathbf{c_j} = (\mathbf{p_j^A}, \mathbf{p_j^B})$, the length similarity between them is computed as:

$$m_{i,j} = \left| \|\mathbf{p_i^A} - \mathbf{p_j^A}\| - \|\mathbf{p_i^B} - \mathbf{p_j^B}\| \right|. \tag{15}$$

Then, the length consistency is fused into the attention matrix $\mathbf{A_3} \in \mathbb{R}^{N \times N}$ by the MHSA layer, while generating a spatial aware attention matrix $\mathbf{A_3} \in \mathbb{R}^{N \times N}$ to guide message passing. This operation can be formulated as follows:

$$\mathbf{A_4} = \mathbf{A_3} \odot \mathbf{M_{ls}}, \tag{16}$$

where $\mathbf{M_{ls}} \in \mathbb{R}^{N \times N}$ represents the length similarity matrix obtained by Eq. 15. Besides, the other operation details are similar to those of the transformer in VSFusion, thus this paper omits the rest for simplicity. Finally, we adopt the same inlier predictor as (Zhao et al. 2021) to process the concatenated feature.

Loss Function

Following (Hartley and Zisserman 2003; Ranftl and Koltun 2018), a hybrid loss function is employed to supervise the training process of our proposed method:

$$\mathcal{L} = \mathcal{L}_c(o_j, y_j) + \alpha \mathcal{L}_e(\widehat{\mathbf{E}}, \mathbf{E}), \tag{17}$$

where $\mathcal{L}_c$ denotes the classification loss; $\mathcal{L}_e$ represents the essential matrix loss; α is a hyper-parameter to balance these two losses.

Following (Zhao et al. 2021), the classification loss $\mathcal{L}_c$ can be formulated as:

$$\mathcal{L}_c(o_j, y_j) = \sum_{j=1}^{\lambda} H(\omega_j \odot o_j, y_j), \tag{18}$$

where $H(\cdot)$ denotes a binary cross entropy loss function; o_j represents the relevant weights of the j-th iteration; y_j represents the weakly supervised labels, which are chosen under the epipolar distance threshold 10^{-4} as positive (Hartley and Zisserman 2003); ω_j is an adaptive temperature vector, and $\odot$ represents the Hadamard product.

Following (Zhang et al. 2019), the essential matrix loss $\mathcal{L}_e$ can be formulated as:

$$\mathcal{L}_e = \frac{(\mathbf{p'^T}\widehat{\mathbf{E}}\mathbf{p})^2}{\|\mathbf{E}\mathbf{p}\|_{[1]}^2 + \|\mathbf{E}\mathbf{p}\|_{[2]}^2 + \|\mathbf{E}\mathbf{p'}\|_{[1]}^2 + \|\mathbf{E}\mathbf{p'}\|_{[2]}^2}, \tag{19}$$

Method	Params.	YFCC100M (%)		SUN3D (%)	
		-	RANSAC	-	RANSAC
PointNet++ (2017)	12.00M	16.48	46.25	8.10	15.29
PointCN (2018)	0.39M	23.95	48.03	9.30	16.40
DFE (2018)	0.40M	30.27	51.16	12.06	16.26
OANet++ (2019)	2.47M	38.95	52.59	16.18	17.18
ACNe (2020)	0.41M	33.06	50.89	14.12	16.99
T-Net (2021)	3.78M	48.20	55.85	17.24	17.57
LMCNet (2021)	0.93M	47.50	55.03	16.82	17.38
CLNet (2021)	1.27M	51.90	59.15	15.83	18.99
MSA-Net (2022)	1.45M	50.65	56.28	16.86	17.79
MS^{2}DG-Net (2022)	2.61M	49.13	57.68	17.84	17.79
PGFNet (2023)	2.99M	53.70	57.83	19.32	18.00
Ours	2.57M	**62.18**	**63.35**	**20.18**	**20.27**

Table 1: Quantitative comparison results of the camera pose estimation on unknown scenes. The mAP5° without/with RANSAC as a post-processing step is reported.

Method	YFCC100M (%)		SUN3D (%)	
	-	RANSAC	-	RANSAC
PointNet++ (2017)	10.49	33.78	10.58	19.17
PointCN (2018)	13.81	34.55	11.55	20.60
OANet++ (2019)	32.57	41.53	20.86	22.31
ACNe (2020)	29.17	40.32	18.86	22.12
T-Net (2021)	42.99	45.25	22.38	22.96
LMCNet (2021)	33.73	40.39	19.92	21.79
CLNet (2021)	38.75	44.88	19.20	23.83
MSA-Net (2022)	39.53	44.57	18.64	22.03
MS^{2}DG-Net (2022)	38.36	45.34	22.20	23.00
PGFNet (2023)	44.20	46.28	23.66	23.87
Ours	**48.83**	**49.03**	**24.81**	**24.76**

Table 2: Quantitative comparison results of the camera pose estimation on known scenes. The mAP5° without/with RANSAC as a post-processing step is reported.

where $\mathbf{E}$ denotes the ground truth of the essential matrix; $\mathbf{p}$ and $\mathbf{p}'$ represent virtual correspondence coordinates obtained by the essential matrix $\mathbf{E}$; $\|\cdot\|_{[i]}$ denotes the i-th element of vector.

Implementation Details

Holistically, SIFT (Lowe 2004) is adopted to establish $N = 2000$ initial correspondences, channel dimension C is 128, network iteration λ is 2, and pruning ratio r is 0.5; besides, considering reducing the training cost, this paper only uses VSFusion in the second iteration. In VCExtractor, the original images are resized to $H = 120, W = 160$, and the channel dimension C_F is 64. In ContextFormer, the number of k neighbors is set to 9 and 6 for two iterations. We adopt Adaptive Moment Estimation (Adam) with a weight decay of 0 as the optimizer to train our network, and the canonical learning rate (for batch size is 32) is set to 10^{-3}. Following (Zhao et al. 2021), the weight α in Equation 17 is set as 0 during the first $20k$ iterations and 0.5 in the remaining $480k$ iterations.

ResNet	VSFusion	ContextFormer	2x	3x	mAP5°	mAP20°
✓			✓		43.25	67.12
✓	✓		✓		49.63	72.87
		✓	✓		57.55	78.12
	✓	✓	✓		**62.18**	**80.95**
	✓	✓		✓	55.70	76.08

Table 3: Ablation study of network architecture. 2x and 3x represent the number of network iterations.

Experiments and Analysis

Evaluation Protocols

We test our method on both outdoor and indoor benchmarks to evaluate the performance on relative pose estimation (Zhang et al. 2019). Yahoo's YFCC100M (Thomee et al. 2016) dataset is used as our outdoor scenes, which is made up of 100 million outdoor photos from the Internet. The SUN3D (Xiao, Owens, and Torralba 2013) dataset is used as our indoor scenes, which consists of large-scale indoor RGB-D videos with information about camera poses. Following the data division of (Zhang et al. 2019), all methods are evaluated on both unknown scenes and known scenes. In this paper, the mAP of pose error at the thresholds (5° and 20°) are reported, where the pose error is the maximum of angular errors from rotation and translation.

Comparison Results

As shown in Table 1 and Table 2, our VSFormer outperforms other state-of-the-art (SOTA) methods on outdoor and indoor scenes. To be specific, on outdoor scenes, our method achieves a performance improvement of 15.8% over the recent MLP-based SOTA method (PGFNet) on unknown scenes at mAP5° without RANSAC. Similarly, compared to the recent Graph-based SOTA method (CLNet), our method attains a performance improvement of 19.8% at mAP5° without RANSAC. On indoor scenes, our method achieves a performance improvement of 24.7% and 27.5% over two baselines (OANet++ and CLNet) on unknown scenes at mAP5° without RANSAC, respectively. Meanwhile, our method also achieves the best performance among all methods with RANSAC. The results indicate that our proposed visual-spatial fusion and transformer-based structure can further improve the network performance. Additionally, as shown in Fig. 4, partial typical visualization results of OANet++ (Zhang et al. 2019), CLNet (Zhao et al. 2021), and our network are shown from left to right. It can be seen that our method achieves the best performance under various challenging scenes.

Ablation Studies

To deeply analyze the proposed method, we perform detailed ablation studies on YFCC100M to demonstrate the effectiveness of each component in VSFormer.

Network Architecture. As shown in Table 3, we intend to gradually add these components to the baseline. The baseline (Row-1) we used is PointCN (Yi et al. 2018) with the

ContextFormer	Cross	TR	Proj	Sum	mAP5°	mAP20°
✓					57.55/61.82	78.12/81.20
✓	✓				60.33/61.88	80.04/81.19
✓		✓			60.18/61.68	79.89/81.24
✓		✓	✓		61.23/62.43	80.60/81.20
✓		✓	✓	✓	**62.18/63.35**	**80.95/81.84**

Table 4: Ablation study for the VSFusion. The results of mAP (%) with/without RANSAC on unknown scenes are reported. Cross, TR, Proj, and Sum respectively represent the cross-attention layer, transformer layer, projection layer, and final element-wise summation.

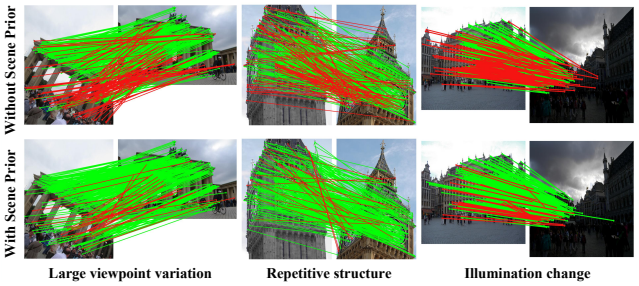


Figure 5: Qualitative comparison with/without scene prior.

pruning strategy. It can be seen that all our component combinations outperform the baseline on outdoor scenes. To be specific, in the second row, the VSFusion is first introduced, which achieves a performance improvement of 14.8% over the baseline at mAP5° without RANSAC. It indicates the importance of exploiting visual cues of a scene/image pair to guide correspondence pruning. Meanwhile, as illustrated in the third row, replacing ResNet encoders (Yi et al. 2018) with our ContextFormer, which obtained a performance improvement of 33.1% over the baseline at mAP5° without RANSAC. Moreover, when combining the proposed modules and using two iterations, the mAP5° is significantly better than those of the baselines. As shown in the last row, this paper also explores the effect of increasing the number of network iterations. The experiment demonstrates that this leads to a performance penalty of 10.4% over the proposed method. This is mainly because the redundant iterations discard some inliers, which are important for the geometric model estimation. That is, the task of camera pose estimation requires sufficient and accurate inliers.

VSFusion Module. As shown in Table 4, we explore the detailed design of VSFusion. Comparing the results of Row-2 and Row-5, we can see that our VSFusion can achieve better performance than using a simple cross-attention layer to fuse visual and spatial cues. Experiments in the fourth and fifth rows verify the necessity of spatial projection and refusion. As illustrated in Fig. 5, visualization results in some challenging scenes also highlight the importance of scene visual cues. In addition, as shown in Table 5, our proposed VSFusion can be used as a plug-and-play module to improve the performance of some baselines.

Method	Known Scene (%)		Unknown Scene (%)	
	mAP5°	mAP20°	mAP5°	mAP20°
PointCN	13.81	35.20	23.95	52.44
PointCN*	**24.87**$_{+11.06}$	**47.96**$_{+12.76}$	**28.18**$_{+4.23}$	**56.57**$_{+4.13}$
OANet++	32.57	56.89	38.95	66.85
OANet++*	**37.90**$_{+5.33}$	**59.97**$_{+3.08}$	**46.10**$_{+7.15}$	**70.68**$_{+3.83}$
CLNet	38.27	62.48	51.80	75.76
CLNet*	**40.58**$_{+2.31}$	**63.06**$_{+0.58}$	**55.20**$_{+3.40}$	**76.83**$_{+1.07}$

Table 5: Quantitative comparison on outdoor scenes without RANSAC. The performance of the baseline can be comprehensively improved after using VSFusion.

Encoder	Graph	NA	GAB	TR	LS	mAP5°	mAP20°
✓						49.63/57.90	72.87/77.93
✓	✓					51.53/57.78	74.60/78.43
✓	✓	✓				53.73/59.88	75.48/79.67
✓	✓	✓	✓			58.93/61.95	79.69/81.50
✓	✓	✓	✓	✓		59.58/62.65	79.68/81.48
✓	✓	✓	✓	✓	✓	**62.18/63.35**	**80.95/81.84**

Table 6: Ablation study for the ContextFormer. Encoder, Graph, NA, GAB, TR, and LS represent the ReNet encoder, KNN-based graph, neighborhood aggregation, graph attention block, transformer, and introduced length similarity matrix, respectively.

ContextFormer. In this paper, we also conduct some ablation studies to verify the effectiveness of each component in the proposed structure. As shown in Table 6, each component of the proposed ContextFormer can further improve the network performance. Among them, the proposed graph attention block is the core component of the structure. Experiments also show that our graph attention block achieves a performance improvement of 9.68%. That is, the KNN-based graph has rich context information, and our method can effectively capture these potential relationships.

Conclusion

In this paper, with another perspective, we exploit visual cues of a scene/image pair to guide correspondence pruning. To this end, we design a joint visual-spatial fusion module to fuse visual and spatial cues. Additionally, to mine consistency within correspondences, we propose a context transformer to explicitly capture both local and global contexts. Meanwhile, a graph attention block is designed to mine contextual information inside the KNN-based graph. Both comparative and ablation experiments demonstrate the effectiveness of our proposed method, which can achieve better performance with fewer parameters.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant U2033210 and Grant 62072223 and in part by the Zhejiang Provincial Natural Science Foundation under Grant LDT23F02024F02.

References

Bai, X.; Luo, Z.; Zhou, L.; Chen, H.; Li, L.; Hu, Z.; Fu, H.; and Tai, C.-L. 2021. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 15859–15869.

Barath, D.; Matas, J.; and Noskova, J. 2019. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 10197–10205.

Chum, O.; and Matas, J. 2005. Matching with PROSAC-progressive sample consensus. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, volume 1, 220–226. IEEE.

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 the 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 the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 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 the ACM*, 24(6): 381–395.

Hartley, R.; and Zisserman, A. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

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

Liao, T.; Zhang, X.; Xu, Y.; Shi, Z.; and Xiao, G. 2023. SGA-Net: A Sparse Graph Attention Network for Two-View Correspondence Learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, X.; Xiao, G.; Chen, R.; and Ma, J. 2023. PGFNet: Preference-Guided Filtering Network for Two-View Correspondence Learning. *IEEE Transactions on Image Processing*, 32: 1367–1378.

Liu, Y.; Liu, L.; Lin, C.; Dong, Z.; and Wang, W. 2021. Learnable motion coherence for correspondence pruning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 3237–3246.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2): 91–110.

Luo, Z.; Shen, T.; Zhou, L.; Zhang, J.; Yao, Y.; Li, S.; Fang, T.; and Quan, L. 2019. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2527–2536.

Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; and Yan, J. 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1): 23–79.

Mur-Artal, R.; Montiel, J. M. M.; and Tardos, J. D. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5): 1147–1163.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 652–660.

Raguram, R.; Chum, O.; Pollefeys, M.; Matas, J.; and Frahm, J.-M. 2012. USAC: A universal framework for random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 2022–2038.

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

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

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 8922–8931.

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 the 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 the ACM*, 59(2): 64–73.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics*, 38(5): 1–12.

Xiao, G.; Ma, J.; Wang, S.; and Chen, C. 2020. Deterministic model fitting by local-neighbor preservation and global-residual optimization. *IEEE Transactions on Image Processing*, 29: 8988–9001.

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

Yi, K. M.; Trulls, E.; Ono, Y.; Lepetit, V.; Salzmann, M.; and Fua, P. 2018. Learning to find good correspondences. In *Proceedings of the 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 the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 5845–5854.

Zhao, C.; Ge, Y.; Zhu, F.; Zhao, R.; Li, H.; and Salzmann, M. 2021. Progressive correspondence pruning by consensus learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 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 the IEEE/CVF International Conference on Computer Vision*, 1950–1959.