ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095240

MSFORMER: MULTI-SCALE TRANSFORMER WITH NEIGHBORHOOD CONSENSUS FOR FEATURE MATCHING

Dongyue Li[1], Yaping Yan[1], Dong Liang[2], and Songlin Du[1,]*

1. Southeast University, Nanjing 210096, China
2. Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ABSTRACT

Existing feature matching methods tend to extract feature descriptors by feeding down-sampled feature maps into a Transformer that is unable to extend feature scales, leading to false correspondences between small-size objects. This paper proposes MSFormer, which uses Transformers situated in different branches to obtain feature descriptors. In one branch, convolutions are integrated into self-attention layers elegantly to compensate for the lack of the local structure information. In another branch, a multi-scale Transformer is proposed through injecting heterogeneous receptive field sizes into tokens. Additionally, a neighborhood consensus mechanism is proposed by re-ranking initial matches to make a constraint of geometric consensus on neighborhood feature descriptors. Extensive experiments on indoor and outdoor pose estimations show that MSFormer outperforms existing state-of-the-art methods by a large margin.

***Index Terms*—** Feature matching, Transformer, Neighborhood consensus, Convolutional neural network

1. INTRODUCTION

Building correspondences between two views of the same scene is the cornerstone of many downstream 3D computer vision tasks, including 3D reconstruction, visual localization, structure from motion (SfM), simultaneous localization and mapping (SLAM), *etc*. Given a pair of images, the traditional pipeline of feature matching is: (1) feature detection (2) feature description (3) feature matching (4) outlier rejection. Before the deep learning era, many researchers design novel handcrafted detectors and local descriptors such as ORB [1] and SIFT [2]. Based on co-occurrence probability-based pixel pairs, [3] is proposed, which has the robustness for severe imaging conditions. In the past few years, many works [4, 5, 6] tend to use CNNs to extract local features which are more robust to illumination and viewpoint changes. To enlarge the receptive field, Transformer [7] is used to capture long-term feature dependencies and form a global representation. SuperGlue [8] enhances feature descriptors by using a graph neural network (GNN), which is a general form of Transformers. Since it's hard to extract repeatable keypoints due to poor texture and repetitive patterns, some methods like [9, 10] tend to first build dense matches and then refine them to the sub-pixel level. However, there are three problems with prior semi-dense methods. First, the image local structure like edges and lines can't be modeled well due to the straightforward tokenization of input features by hard split. Second, down-sampled feature maps from CNNs are taken as inputs of a Transformer. Therefore, the transformed features are less distinctive since the fine-grained information of small objects isn't well preserved. Finally, a key geometric constraint termed neighborhood consensus [11] isn't taken into account, causing uncertainty in feature matching. Recently, the underlying relations between self-attention and convolution have been exploited by [12]. Meanwhile, [13] proposes a new scheme which can model objects of various scales simultaneously at different attention heads. In this paper, MSformer is proposed, which can obtain better feature descriptors and extract geometrically consistent matches. Our main contributions are summarized as follows:

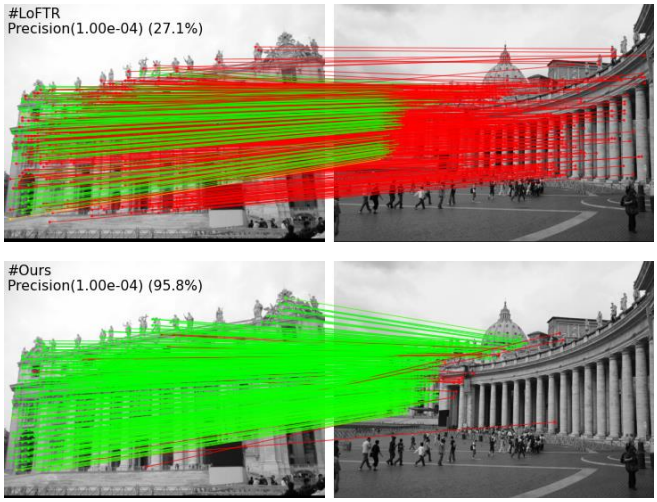

Fig. 1. Comparison between MSFormer (below) and LoFTR (above). Green color indicates epipolar error below 1×10^{-4}.

∗ Corresponding authors: Songlin Du (sdu@seu.edu.cn)

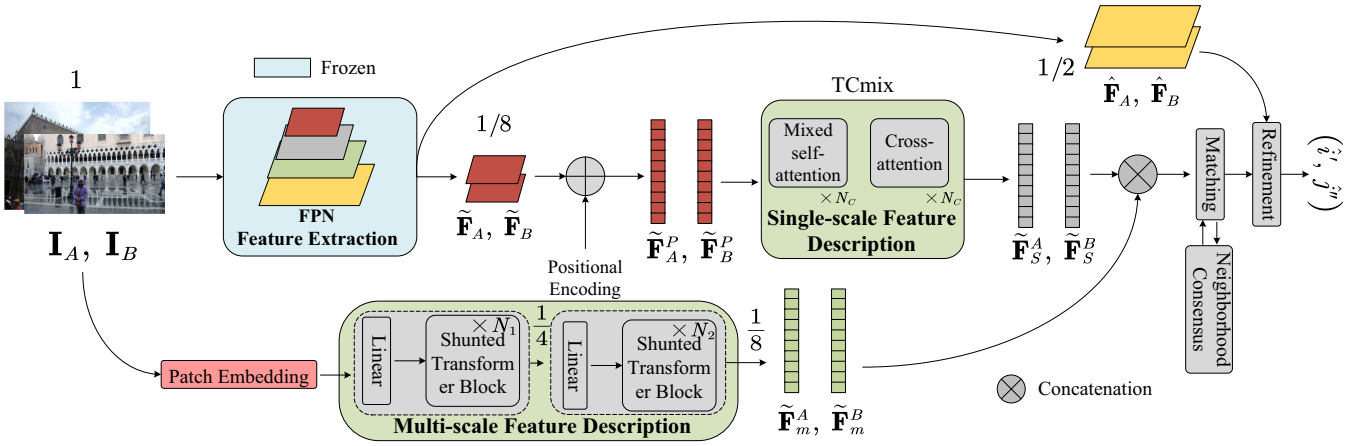

Fig. 2. The overall architecture of MSFormer. Single-scale features and multi-scale features are fused to generate feature descriptors. A neighborhood consensus module is proposed to ensure the geometrical consistency of correspondences.

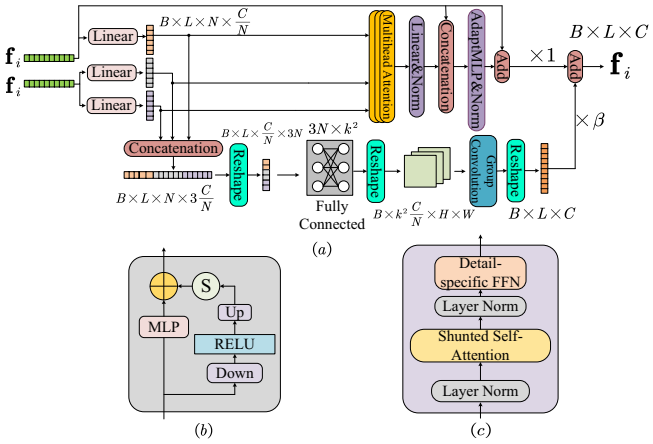

Fig. 3. (a) The architecture of a mixed self-attention layer in TCmix. (b) The architecture of AdaptMLP. (c) The architecture of a Shunted Transformer block.

1) A mixed model entitled TCmix is proposed, in which the additional convolution path is added to the self-attention layer while the cross-attention layer is kept unchanged. The 2D structure and spatial local information within each feature map can be preserved in features transformed by our TCmix.

2) A two-stage multi-scale Transformer is proposed to capture both coarse-grained and fine-grained features directly from the original images. Visualization results demonstrate that features transformed in this way can retain lots of fine-grained details.

3) The neighborhood consensus module is proposed to re-rank initial matches to reach a matching consensus in local neighborhoods between graphs. Matches violating the neighborhood consensus criterion are filtered to improve the matching precision.

2. METHODOLOGY

2.1. Single-scale Feature Description

Given two images $\mathbf{I}_A$ and $\mathbf{I}_B$, a frozen feature pyramid network (FPN) [14] is used as the backbone to extract feature maps $\widetilde{\mathbf{F}}_A$, $\widetilde{\mathbf{F}}_B$ ($\frac{1}{8}$ of the original image resolution) and $\hat{\mathbf{F}}_A$, $\hat{\mathbf{F}}_B$ ($\frac{1}{2}$ of the original image resolution). Coarse feature maps $\widetilde{\mathbf{F}}_A$, $\widetilde{\mathbf{F}}_B$ are flattened into 1-D vectors and added with the positional encoding. As shown in Fig. 2, the position-dependent features are denoted as $\widetilde{\mathbf{F}}_A^P$, $\widetilde{\mathbf{F}}_B^P$. The positional encoding used here has a sinusoidal form and is often referred to as "absolute positional encoding". Assuming inputs of the l^{th} mixed self-attention layer are $\mathbf{f}_i^l, \mathbf{f}_j^l$ ($\mathbf{f}_i^l = \mathbf{f}_j^l$), the final output $\mathbf{f}_i^{l+1}$ of the l^{th} mixed self-attention layer can be attained by

$$\begin{aligned}
&\mathbf{Q}^l = \mathrm{Prj_q}(\mathbf{f}_i^l), \mathbf{K}^l = \mathrm{Prj_k}(\mathbf{f}_j^l), \mathbf{V}^l = \mathrm{Prj_v}(\mathbf{f}_j^l),\\
&\mathbf{M} = \mathrm{norm}(\mathrm{Linear}(\mathrm{MHA}(\mathbf{Q}^l, \mathbf{K}^l, \mathbf{V}^l))),\\
&\mathbf{f}_{i,att}^l = \mathrm{norm}(\mathrm{AdaptMLP}(\mathrm{concat}(\mathbf{f}_i^l, \mathbf{M}))) + \mathbf{f}_i^l,\\
&\mathbf{f}_i^{l+1} = \mathbf{f}_{i,att}^l + \beta \mathbf{f}_{i,conv}^l,
\end{aligned} \tag{1}$$

where $\mathrm{Prj_q}(\cdot)$, $\mathrm{Prj_k}(\cdot)$, $\mathrm{Prj_v}(\cdot)$ are linear projections from $\mathbf{f}_i^l, \mathbf{f}_j^l$ to $\mathbf{Q}^l, \mathbf{K}^l, \mathbf{V}^l$ and $\mathbf{M}$ is retrieved message. $\mathrm{MHA}(\cdot)$ is a multi-head linear attention block in [15]. $\mathrm{AdaptMLP}(\cdot)$ is an enhanced MLP block and the structure of AdaptMLP is present in Fig. 3(b). β is a learnable parameter, $\mathbf{f}_{i,att}^l$ is the output of the self-attention path and $\mathbf{f}_{i,conv}^l$ is the output of the convolution path. The output of the l^{th} cross-attention layer can be obtained precisely as described above, except that the convolution path is removed and the inputs are different. We interleave the mixed self-attention and cross-attention layers in our TCmix by N_C times, generating single-scale feature descriptors $\widetilde{\mathbf{F}}_s^A$, $\widetilde{\mathbf{F}}_s^B$.

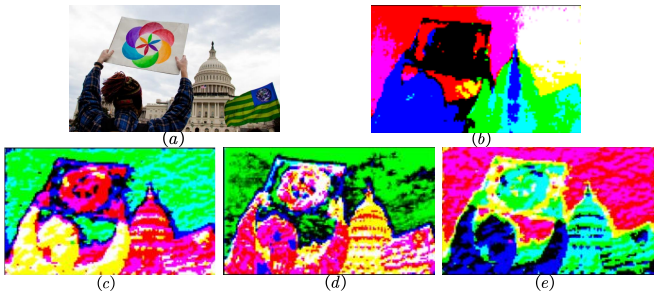

Fig. 4. Visualization of features transformed by different ways. (a) The original image. (b) The transformed feature in LoFTR. (c) The single-scale feature in our method. (d) The multi-scale feature in our method. (e) The enhanced feature used as one of matching descriptors. The dimension of transformed features is reduced by PCA and results are visualized with RGB color.

2.2. Multi-scale Feature Description

As shown in Fig. 2, the two-stage multi-scale Transformer used here is composed of the patch embedding module, linear projections and interleaved Shunted Transformer blocks [13]. The structure of one Shunted Transformer block is shown in Fig. 3(c). In contrast to the standard attention layer, the length of $\mathbf{K}, \mathbf{V}$ is not uniform across different heads of the same shunted self-attention layer, enabling the two-stage multi-scale Transformer to capture multi-granularity features. Multi-scale features directly extracted from original images $\mathbf{I}_A, \mathbf{I}_B$ are denoted as $\widetilde{\mathbf{F}}_m^A$, $\widetilde{\mathbf{F}}_m^B$. $\widetilde{\mathbf{F}}_m^A$, $\widetilde{\mathbf{F}}_m^B$ and $\widetilde{\mathbf{F}}_s^A$, $\widetilde{\mathbf{F}}_s^B$ are concatenated along the channel dimension, respectively, forming enhanced descriptors $\widetilde{\mathbf{F}}_{tr}^A$ and $\widetilde{\mathbf{F}}_{tr}^B$ for feature matching. As illustrated in Fig. 4, it is obvious that the single-scale feature retains the information of local structure while the multi-scale feature embeds rich low-level details.

2.3. Neighborhood Consensus Module

In the matching module, the similarity matrix $\mathbf{P}(i,j)$ is calculated by the inner product of $\widetilde{\mathbf{F}}_{tr}^A$, $\widetilde{\mathbf{F}}_{tr}^B$

$$\mathbf{P}(i,j) = \tau\langle\widetilde{\mathbf{F}}_{tr}^A(i), \widetilde{\mathbf{F}}_{tr}^B(j)\rangle, \tag{2}$$

where τ is a temperature parameter. By using $-\mathbf{P}(i,j)$ as the cost matrix of the partial assignment problem as in [8], the confidence matrix $\mathbf{C}$ is given, from which we obtain coarse matches $(\widetilde{i}, \widetilde{j})$. The corresponding descriptors $\widetilde{\mathbf{F}}_{tr}^A(\widetilde{i})$, $\widetilde{\mathbf{F}}_{tr}^B(\widetilde{j})$ are extracted and the partial similarity matrix $\mathbf{S}$ is computed by $\mathbf{S} = \text{softmax}(\langle\widetilde{\mathbf{F}}_{tr}^A(\widetilde{i}), \widetilde{\mathbf{F}}_{tr}^B(\widetilde{j})\rangle)$.

A pair of coarse matches can be viewed as a pair of node correspondences between graphs $\mathbf{G}_a$ and $\mathbf{G}_b$, which correspond to images $\mathbf{I}_A$ and $\mathbf{I}_B$, respectively. In each graph, each node establishes edges with 2 other nodes based on attention

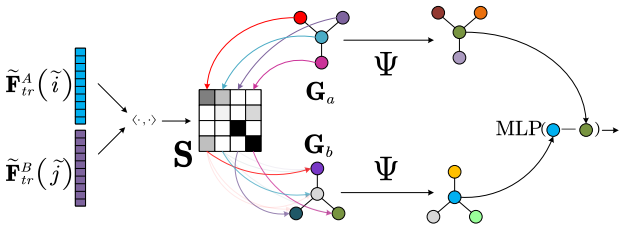

Fig. 5. The neighborhood consensus module.

weights, which means that each node is connected to 2 other nodes it is most similar with. The edge matrix of $\mathbf{G}_a$ and $\mathbf{G}_b$ is defined as $\mathbf{E}_a$ and $\mathbf{E}_b$, respectively. A randomly colored node matrix $\mathbf{R}_a \in \mathbb{R}^{M\times r}$ is proposed as the color of $\mathbf{G}_a$, where M is the number of nodes in graph $\mathbf{G}_a$. r is the number of colors and $r \ll M$. Another colored node matrix $\mathbf{R}_b$ can be obtained by $\mathbf{R}_b = \mathbf{S}^T\mathbf{R}_a$, mapping the color of $\mathbf{G}_a$ to $\mathbf{G}_b$ with $\mathbf{S}$. We distribute this coloring in corresponding neighborhoods by performing synchronous message passing on both graphs via a shared graph neural network Ψ

$$\mathbf{d}_0 = \Psi(\mathbf{R}_a, \mathbf{E}_a), \quad \mathbf{d}_1 = \Psi(\mathbf{R}_b, \mathbf{E}_b). \tag{3}$$

The difference between $\mathbf{d}_0$ and $\mathbf{d}_1$ can be viewed as a measurement of the neighborhood consensus and used to rectify the similarity matrix $\mathbf{P}(i,j)$

$$\mathbf{P}(\widetilde{i},\widetilde{j}) = \mathbf{P}(\widetilde{i},\widetilde{j}) + \text{MLP}(\mathbf{d}_0 - \mathbf{d}_1). \tag{4}$$

Repeating the above-mentioned procedure, then we get the refined confidence matrix $\mathbf{C}$ and obtain re-ranked matches $(\widetilde{i}', \widetilde{j}')$. Similar to LoFTR, in the refinement module, the enhanced descriptors $\widetilde{\mathbf{F}}_{tr}^A$ and $\widetilde{\mathbf{F}}_{tr}^B$ are used as context to augment fine features $\hat{\mathbf{F}}_A$ and $\hat{\mathbf{F}}_B$. The augmented fine features are then transformed by a TCmix and the transformed features are used to refine matches $(\widetilde{i}', \widetilde{j}')$ to a sub-pixel level with a correlation-based approach.

3. EXPERIMENTS

In this paper, we evaluate MSFormer on several downstream tasks such as indoor and outdoor pose estimations. We inherit the loss function in LoFTR. Our implemented LoFTR results under the same experimental conditions and with the same hardware equipment are present for fair comparison. All the experiments are conducted on a single RTX 3090. Both LoFTR and MSFormer use the optimal transport matching layer to extract coarse matches.

3.1. Outdoor Pose Estimation

Datasets. In this experiment, MSFormer is trained and tested on the MegaDepth [16] dataset. To save memory, images are resized so that their longer dimension (dubbed as SIZE)

Table 1. Evaluation on MagaDepth for outdoor pose estimation. The AUC of the pose error and precision@1e-4 are reported.

SIZE	Method	@5°	@10°	@20°	precision @1e-4
-	DRC-Net [17]	27.01	42.96	58.31	-
-	SuperPoint [6] +SuperGlue [8]	42.18	61.16	75.96	-
840	LoFTR [9]	47.54	64.54	77.33	83.88
840	Ours	**48.50**	**65.30**	**77.62**	**94.23**
640	LoFTR [9]	43.48	59.41	72.28	82.35
640	Ours	**45.35**	**62.24**	**74.96**	**92.85**

Table 2. Evaluation on ScanNet for indoor pose estimation. The AUC $\{5°, 10°, 20°\}$ and precision@5e-4 are reported. Symbol '†' means that the model is trained on MegaDepth.

Method	@5°	@10°	@20°	precision @5e-4
ORB [1]+GMS [18]	5.21	13.65	25.36	-
D2Net [4]+NN	5.25	14.53	27.96	-
ContextDesc [19]+Ratio Test [2]	6.64	15.01	25.75	-
SuperPoint [6]+NN	9.43	21.53	36.40	-
SuperPoint [6]+PointCN [20]	11.40	25.47	41.41	-
SuperPoint [6]+OANet [21]	11.76	26.90	43.85	-
DRC-Net † [17]	7.69	17.93	30.49	-
LoFTR † [9]	14.18	30.26	46.94	57.60
Ours†	**16.20**	**32.24**	**48.07**	**73.67**

equals to 640 pixels during training. However, all other baseline methods are evaluated on the test set with larger SIZE. For fair comparison, MSFormer is tested on both 840 SIZE and 640 SIZE test sets.

Results. The matching precision and the AUC of the pose error under thresholds $\{5°, 10°, 20°\}$ are reported. The pose error is defined as the maximum of angular error in translation and rotation. Similar to LoFTR, MSFormer is sensitive to the sizes of images used for training. As shown in Table 1, MSFormer outperforms LoFTR on the AUC $\{5°, 10°, 20°\}$ and achieves a +10.35% precision gain when SIZE equals to 840 pixels. MSFormer maintains its lead (+1.87%, +2.83%, +2.68%, +10.5%) when evaluated on the 640 SIZE test set, implying the validity of our method.

3.2. Indoor Pose Estimation

Datasets. In this section, MSFormer is evaluated on the ScanNet [22] dataset as in [9]. Because DRC-Net is trained on MegaDepth and the computational cost of training MSFormer on ScanNet from scratch is unbearable, we present results of MSFormer trained on MegaDepth.

Results. In indoor pose estimation, the AUC $\{5°, 10°, 20°\}$ and precision@5e-4 are reported. Symbol '†' means that the model is trained on MegaDepth. As shown in Table 2, MSFormer outperforms LoFTR by 27.85% at precision@5e-4. We attribute the significant precision gain to our more distinctive feature descriptors and the neighborhood consensus module. For other methods, MSFormer outperforms them although it is trained on MegaDepth, which demonstrates its generalizability.

Table 3. Ablation study. Variants of MSFormer are evaluated on the MegaDepth dataset (840 SIZE).

Single-scale description	multi-scale description	Neighbors	@5°	@10°	@20°	precision @1e-4
✓			48.22	64.37	77.11	88.71
	✓		46.22	62.24	75.30	89.03
✓	✓		48.12	65.17	77.19	94.20
✓	✓	2	48.50	65.30	77.62	94.23
✓	✓	4	48.57	65.18	77.51	94.42
✓	✓	8	48.64	65.26	77.88	94.35

3.3. Ablation Study

To better understand MSFormer, we evaluate the effect of different modules respectively. Experiments are conducted under the same training and evaluation protocol as outdoor pose estimation on MeagDepth. MSFormer with the only single-scale or multi-scale descriptor can't achieve the highest performance although both of them outperform LoFTR. Single-scale descriptors are more robust to geometric transformations thanks to CNNs and cross-attention layers. Multi-scale descriptors preserve rich low-level details due to shunted self-attention layers, which can implicitly enhance the distinctiveness of descriptors. In order to simultaneously maintain the robustness and the distinctiveness of the descriptors, the single-scale and multi-scale descriptor are fused together and the increased feature dimension implies more information is embedded into the enhanced descriptors. By using the neighborhood consensus module, the overall performance can be boosted again. Increasing the number of neighbors slows down the inference speed of the model although it improves the performance. Therefore, the number of neighbors is set to 2 as a compromise. The above experiments demonstrate that each of our designs makes a contribution.

4. CONCLUSION

Feature descriptors in prior semi-dense methods lose too many low-level details, incurring inaccurate matches. Therefore, TCmix and an additional multi-scale Transformer are proposed to learn distinctive feature descriptors. We also propose a neighborhood consensus module for filtering false correspondences. Experiments show that our MSFormer achieves state-of-the-art performances on several tasks.

5. REFERENCES

[1] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] D. Liang, S. Kaneko, and Y. Satoh, "A Robust Appearance Model and Similarity Measure for Image Matching," *J. Robot. Mechatronics*, vol. 27, no. 2, pp. 126–135, 2015.

[4] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-Net: A Trainable CNN for Joint Description and Detection of Local Features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8084–8093.

[5] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2D2: Reliable and Repeatable Detector and Descriptor," in *Proc. Advances Neural Inf. Process. Syst.*, 2019.

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 224–236.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. Advances Neural Inf. Process. Syst.*, 2017.

[8] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching With Graph Neural Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4937–4946.

[9] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-Free Local Feature Matching with Transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8918–8927.

[10] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "MatchFormer: Interleaving Attention in Transformers for Feature Matching," in *Proc. Asian Conf. Comput. Vis.*, 2022.

[11] M. Fey, J. E. Lenssen, C. Morris, J. Masci, and N. M. Kriege, "Deep Graph Matching Consensus," in *Proc. Int. Conf. Learn. Representations*, 2020.

[12] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the Integration of Self-attention and Convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 815–825.

[13] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted Self-Attention via Multi-Scale Token Aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10853–10862.

[14] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[15] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5156–5165.

[16] Z. Li and N. Snavely, "MegaDepth: Learning Single-View Depth Prediction from Internet Photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2041–2050.

[17] X. Li, K. Han, S. Li, and V. Prisacariu, "Dual-Resolution Correspondence Networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2020, pp. 17346–17357.

[18] J. Bian, W. Lin, Y. Matsushita, S. Yeung, T. Nguyen, and M. Cheng, "GMS: Grid-Based Motion Statistics for Fast, Ultra-Robust Feature Correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2828–2837.

[19] Z. Luo, T. Shen, L. Zhou, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "ContextDesc: Local Descriptor Augmentation With Cross-Modality Context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2522–2531.

[20] K. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to Find Good Correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2666–2674.

[21] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, H. Liao, and L. Quan, "Learning Two-View Correspondences and Geometry Using Order-Aware Network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5844–5853.

[22] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner, "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2432–2443.