MatchFormer: Interleaving Attention in Transformers for Feature Matching
(Supplementary Materials)

Qing Wang*, Jiaming Zhang*, Kailun Yang†,
Kunyu Peng, and Rainer Stiefelhagen

Karlsruhe Institute of Technology, Germany
https://github.com/jamycheung/MatchFormer

1 Implementation Details

Transformer. We design a four-stage hierarchical Transformer, using gray-scale images as input, with an input channel of 1. Each stage contains a positional patch embedding layer and three attention layers. The channel of the feature map is gradually increased by $\{128, 192, 256, 512\}$, and the resolution is decreased by $\{1/2, 1/4, 1/8, 1/16\}$ (in the large version), or $\{1/4, 1/8, 1/16, 1/32\}$ (in the lite version). Our backbone does not contain a stem layer [3], and we use a large 7×7 convolution layer for the first patch embedding layer and a 3×3 convolution layer for the next three layers.

MLP. Inspired by the MLP design of SegFormer [8], we adopt to use the MLP layer after each attention layer in our match-aware encoder, which consists of two linear layers and a depth-wise convolution layer. To avoid excessive computation, we set the hidden features ratio [8] of all MLPs to 4. The MLP layers can enhance the features extracted by attention and introduce residual connections.

Interleaving Self-/Cross-Attention. The *extract-and-match* strategy is constructed by interleaving self- and cross-attention within our MatchFormer model. There are four stages in the match-aware encoder. As the feature map of the shallow stage (*i.e.*, stage-1 and stage-2) emphasizes textural information, more self-attention are applied to focus on exploring the feature itself. As the feature map of the deep stage (*i.e.*, stage-3 and stage-4) is biased toward semantic information, more cross-attention are applied to explore similarity cross images. The code of MatchFormer is reported in Algorithm 1.

More Structural Analysis. To explore the effect of the attention module arrangement inside the backbone of MatchFormer, we spend large effort to analyze various self-attention and cross-attention schemes at each stage, where both modules interact in a separate or interleaved manner. To be consistent with the ablation study setting, we utilize the indoor model trained on 10% of ScanNet [2] to conduct the experiment.

As shown in Table 1, the result in first row indicates that using only self-attention without cross-attention limits the matching capacity of transformer-

* Equal contribution
† Correspondence: kailun.yang@kit.edu

Algorithm 1 Code of interleaving self-/cross-attention in a PyTorch-like style.

```
# proj: channel projection
# DWConv: depth-wise convolution layer
# softmax: softmax layer
# sigmoid: sigmoid layer

import torch
import torch.nn as nn

def posPE(image):
    image = nn.Conv2D(image)
    weight = sigmoid(DWConv(image))
    image_enhance = image * weight
    return image_enhance

def interleaving_attention(image_A, image_B, cross_flags):
    seq_A, seq_B = posPE(image_A), posPE(image_B)
    Q_A, K_A, V_A = nn.Linear(seq_A).reshape()
    Q_B, K_B, V_B = nn.Linear(seq_B).reshape()

    for flag in cross_flags:
        if flag == True: # cross-attention
            attn_A = Q_A @ K_B.transpose()
            attn_A = attn_A.softmax()
            attn_B = Q_B @ K_A.transpose()
            attn_B = attn_B.softmax()
            image_A = (attn_A @ V_B).transpose().reshape()
            image_B = (attn_B @ V_A).transpose().reshape()

        else: # self-attention
            attn_A = Q_A @ K_A.transpose()
            attn_A = attn_A.softmax()
            attn_B = Q_B @ K_B.transpose()
            attn_B = attn_B.softmax()
            image_A = (attn_A @ V_A).transpose().reshape()
            image_B = (attn_B @ V_B).transpose().reshape()

    return image_A, image_B

# MatchFormer stages
# stage1: cross_flags in 3 layers = [False, False, True]
# stage2: cross_flags in 3 layers = [False, False, True]
# stage3: cross_flags in 3 layers = [False, False, True]
# stage4: cross_flags in 3 layers = [False, False, True]

def MatchFormer(image_A, image_B):
    for _ in [stage1, stage2, stage3, stage4]:
        mage_A, image_B = interleaving_attention(image_A, image_B, cross_flags)
    return image_A, image_B
```

based encoder. The results of the other separate arrangements show that arranging cross-attention modules after the self-attention stage of MatchFormer can improve the performance of pose estimation, reaching 81.8% in precision (P), when three stages are constructed with cross-attention modules. However, excessive usage of cross-attention will degrade the performance due to the lack of self-attention modules. Thus, we propose an attention-interleaving strategy for combining the self- and cross-attention within individual stage of backbone. In the experiments of the last four rows, the interleaving attention scheme of MatchFormer achieves the best performance (86.7% in P). The results indicate the effectiveness of our proposed interleaving arrangement and prove our observation that building a match-aware transformer-based encoder to perform the *extract-and-match* strategy can benefit the feature matching.

Table 1. More Structural Analysis of the attention arrangement in the encoder. 'S' and 'C' are short for a self-attention layer and a cross-attention layer, respectively.

	Structure				Pose estimation AUC			P
	stage1	stage2	stage3	stage4	@5°	@10°	@20°	
Separate	SS	SS	SS	SS	7.57	20.57	36.80	75.8
	SS	SS	SS	CC	10.77	24.37	42.54	78.2
	SS	SS	CC	CC	13.85	30.31	48.53	80.7
	SS	CC	CC	CC	13.58	29.57	48.12	81.8
	CC	CC	CC	CC	11.26	26.15	44.32	80.9
	SSS	SSS	CCC	CCC	12.22	27.71	45.62	81.3
Interleaving	SC	SC	SC	SC	14.04	30.57	48.31	81.1
	SSC	SSC	SSC	SSC	12.25	27.05	43.78	83.4
	SCC	SCC	SCC	SCC	14.75	31.03	48.27	85.3
	SSC	SSC	CCC	CCC	12.82	28.48	46.29	81.0
	SSC	SSC	SCC	SCC	**18.01**	**35.87**	**53.46**	**86.7**

Coarse-to-fine Matching Module. The hierarchical encoder in MatchFormer extracts multi-scale features and the decoder delivers both low- and high-resolution feature pairs ($\frac{1}{r_c}$-scaled coarse features and $\frac{1}{r_f}$-scaled fine features, *w.r.t.*, the size of input images) for coarse-to-fine matching [6].

To begin with coarse matching, the $\frac{1}{r_c}$-scaled coarse feature pair $\frac{H_1}{r_c}\times\frac{W_1}{r_c}$ and $\frac{H_2}{r_c}\times\frac{W_2}{r_c}$ is reshaped into sequences I_1^c and I_2^c to calculate the score $\boldsymbol{S}_{i,j}=\frac{1}{\tau}\cdot\langle I_1^c(i), I_2^c(j)\rangle$ of matrix $\boldsymbol{S}\in\frac{H_1W_1}{r_c}\times\frac{H_2W_2}{r_c}$, where $\langle\cdot,\cdot\rangle$ is the inner product, τ is the temperature coefficient, H and W are the image height and width. To calculate the probability of soft mutual closest neighbor matching, we use `softmax` on both dimensions of S (referred to as 2D-softmax). The coarse matching probability $\boldsymbol{P}_{i,j}^c$ is calculated via Eq. (1).

$$\boldsymbol{P}_{i,j}^c = softmax(\boldsymbol{S}_{i,j})\cdot softmax(\boldsymbol{S}_{j,i}). \tag{1}$$

To select coarse match predictions $\boldsymbol{M}^c$, $\boldsymbol{P}_{i,j}^c$ must be larger than the threshold θ and fulfill the mutual closest neighbor (MNN) criterion, as indicated in Eq. (2):

$$\boldsymbol{M}_{i,j}^c = \mathbb{1}_{(\boldsymbol{P}_{i,j}^c>\theta)\,\wedge\, MNN(\boldsymbol{P}_{i,j}^c)}. \tag{2}$$

Given a matched spot $\{(i,j)|\boldsymbol{M}_{i,j}^c=1\}$ on coarse feature maps, its paired windows are cropped as $(\boldsymbol{w}_{i'},\boldsymbol{w}_{j'})$ to conduct fine matching, where (i',j') are back-located at the $\frac{1}{r_f}$-scaled fine feature maps. The fine match probability $\boldsymbol{P}_{i,j}^f$ of the center vector $\vec{c}_i$ of $\boldsymbol{w}_i$ related to the entire $\boldsymbol{w}_j$ can be calculated by `softmax`. Solving the expectation of $\boldsymbol{P}_{i,j}^f = softmax(<\vec{c}_i,\boldsymbol{w}_j>)$ can determine the fine matching $\boldsymbol{M}_{i,j}^f$ on $\boldsymbol{w}_j$, then we map it to the original resolution to establish the final matching. Fine matching can be formulated as $\mathbb{E}_{i\to j}(\boldsymbol{P}_{i,j}^f|\vec{c}_i)$.

2 Indoor Pose Estimation.

Robustness evaluation. To evaluate the robustness with less training data, we further compare MatchFormer-large-LA and LoFTR in different percentages of datasets in Table 2. The different sizes of training data are selected from the

Table 2. Indoor pose estimation on ScanNet with less training data. The AUC of three different thresholds and the average matching precision (P) are evaluated.

Method	Data percent	Pose estimation AUC @5°	@10°	@20°	P
LoFTR [6]	10%	15.47	31.72	48.63	82.6
MatchFormer-large-SEA	10%	18.01 (+2.54)	35.87 (+4.15)	53.46 (+4.83)	86.7 (+4.1)
LoFTR [6]	30%	18.20	35.54	52.58	84.1
MatchFormer-large-SEA	30%	21.20 (+3.00)	39.65 (+4.11)	57.16 (+4.58)	88.5 (+4.4)
LoFTR [6]	50%	19.65	37.48	53.89	86.3
MatchFormer-large-SEA	50%	21.10 (+1.45)	39.91 (+2.43)	57.36 (+3.47)	89.0 (+2.7)
LoFTR [6]	70%	19.55	37.82	54.77	85.7
MatchFormer-large-SEA	70%	21.34 (+1.79)	41.08 (+3.26)	58.97 (+4.20)	88.8 (+3.1)
LoFTR [6]	100%	22.06	40.80	57.62	87.9
MatchFormer-large-SEA	100%	24.31 (+2.25)	43.90 (+3.10)	61.41 (+3.79)	89.5 (+1.6)

first $x \in \{10,30,50,70,100\}$ percentages of the original dataset. With different sizes of training data, MatchFormer has maintained consistent performance. Hence it has tremendous potential for data-constrained real-world scenarios.

Qualitative Comparisons. The visualizations of indoor matching qualitative comparisons can be found in Fig. 1. From top to bottom are the matching results from SuperGlue [5], LoFTR [6] with 10% training data, MatchFormer-large-SEA with 10% training data, LoFTR and MatchFormer-large-SEA with all training data. Due to the captured long-range dependency, MatchFormer achieves dense feature matching in such challenging indoor scenes with large viewing angle changes, such as the first and the second column in Fig. 1. In the low-texture scene of the third column, our method can provide more matches compared to SuperGlue and LoFTR. Additionally, the performance of MatchFormer-large-SEA is significantly better than LoFTR, when they are trained on the same 10% data of ScanNet, which indicates that our model is more flexible when transferred to a moderate dataset.

3 Outdoor Pose Estimation

Qualitative Comparisons. As shown in Fig. 2, we visualize the qualitative comparisons of the outdoor model at MegaDepth [4]. In outdoor scene matching, MatchFormer-large-LA outperforms LoFTR and SuperGlue in matching performancec. The matching performance of MatchFormer-lite-SEA and MatchFormer-lite-LA are on par with that of LoFTR and SuperGlue.

4 Homography Estimation

Qualitative Comparisons. To evaluate the feature matching in the benchmark for geometric relations estimation, we perform Homography Estimation on HPatches [1] with the MatchFormer-large-LA. In Fig. 3, we visualize more qualitative comparison based on the matching results of MacthFormer-large-LA, LoFTR [6], and SuperGlue [5]. MatchFormer can perform more dense and confident matching than SuperGlue. Besides, MatchFormer has further improvements

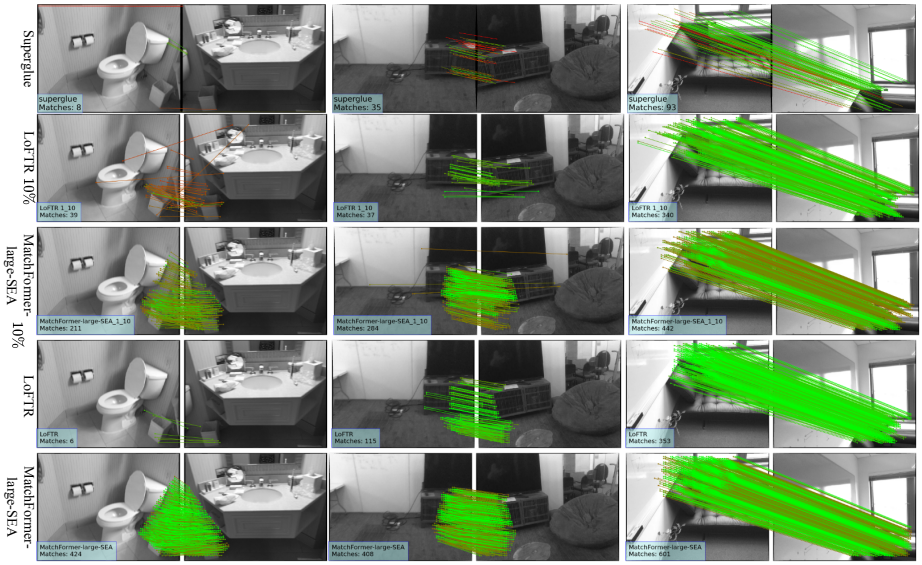

Fig. 1. More Qualitative Comparisons in Indoor Scene Matching of MatchFormer, LoFTR, and SuperGlue. The color represents matching confidence, where green represents more correct matches, and red represents uncertain matches. Models (10%) represent indoor models trained on 10% of the ScanNet dataset [2].

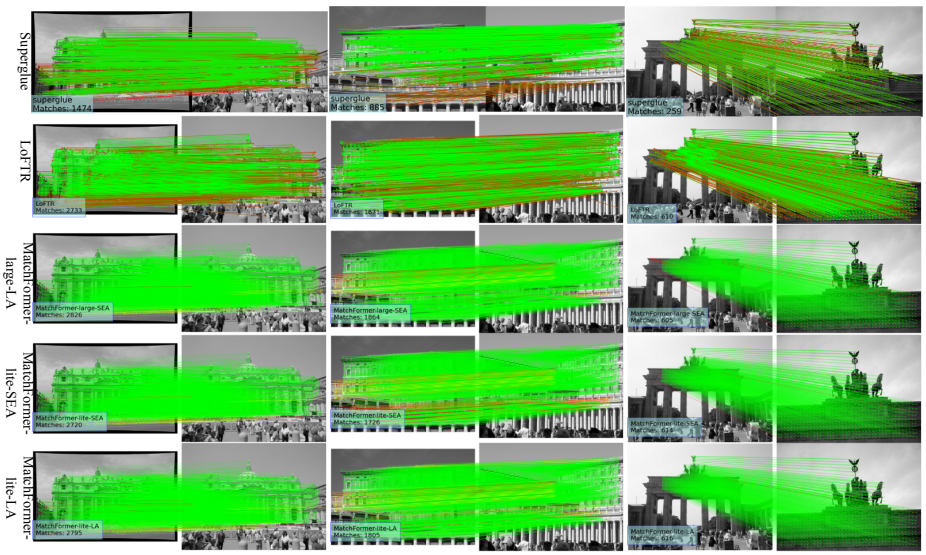

Fig. 2. More Qualitative Comparisons in Outdoor Scene Matching of MatchFormer, LoFTR, and SuperGlue. The color represents matching confidence, where green represents more correct matches, and red represents uncertain matches. lite represents the model outdoor model for outputting low-resolution matching feature maps. LA represents linear attention. SEA represents spatial efficient attention.

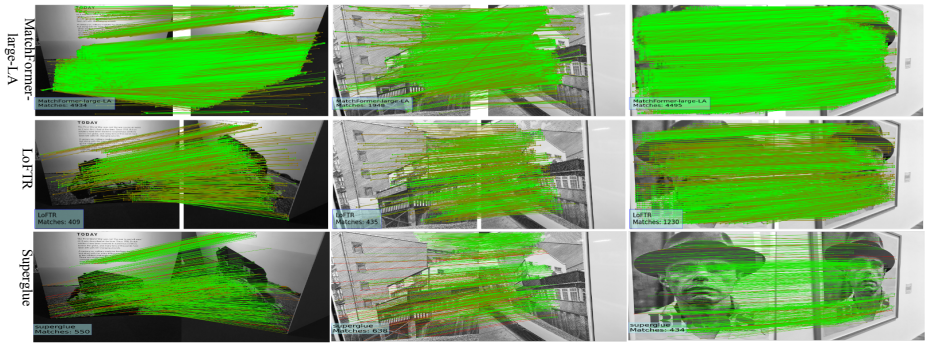

Fig. 3. Qualitative Comparisons on HPatches.

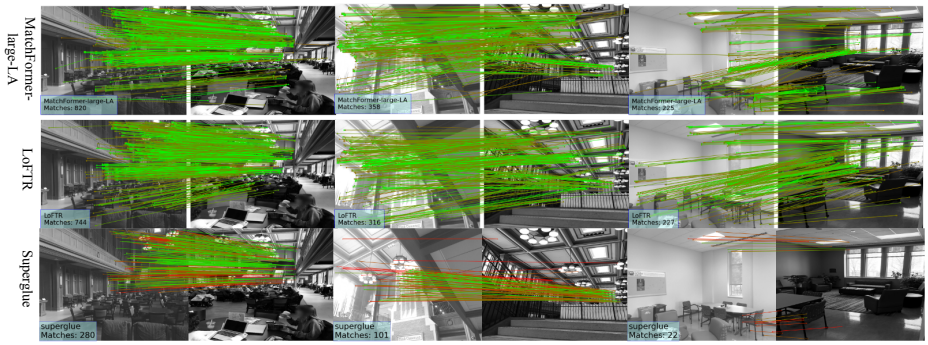

Fig. 4. Qualitative Comparisons on InLoc.

by yielding more matches compared to LoFTR, such as an improvement with more than $4.5K$ matches in the first column of Fig. 3.

5 Image Matching

Following the experimental setup of Patch2Pix [9], we choose the same 108 HPatches sequences, including 52 sequences with illumces with viewpoint change. Each sequence contains six images. To match the first with all others, we report the mean matching accuracy (MMA) at thresholds from [1,10] pixels, and the number of matches and features. The input size of the image is set to 1024, the matching threshold is set to 0.2, and RANSAC threshold as 2 pixels.

6 InLoc Visual Localization

Detailed Settings. On the InLoc [7] benchmark, we follow Patch2pix [9] to evaluate the same first 40 retrieval pairs. The same temporal consistency check is performed to limit the retrievals, and the RANSAC threshold is set to 48 pixels for pose estimation. We adjust the images to 1024 on the long side.

Qualitative Comparisons. To evaluate the effectiveness of MatchFormer in the visual localization task, we evaluate MatchFormer-large-LA on the InLoc [7] benchmark. The visualizations of InLoc visual localization results can be found in Fig. 4. In comparison to the detector-based MatchFormer method, MatchFormer

has a greater and more accurate number of matches. MatchFormer performs at a level comparable to the detector-free method LoFTR.

7 Limitations and Future Work

For indoor scenes and outdoor scenes, MatchFormer employs two kinds of attention, *i.e.*, spatial efficient attention (SEA) and linear attention (LA), which have varying degrees of computational reductions and different abilities for feature extraction. They are appropriate for either indoors or outdoors. In our experiments, LA proved to be more suitable for outdoor scenes with dense high-resolution input. In contrast, SEA was more appropriate for indoor scenes with sparse low-resolution input. Exploring a uniform efficient attention module to handle both indoor and outdoor inputs with different resolutions, we leave it as the future work. Besides, in MatchFormer, we introduce an efficient FPN-like decoder that can combine match-aware feature maps generated by interleaving attention. It is potential to adapt an alternative decoder to the feature fusion task, such as MLP-decoder.

8 Acknowledgments

This work was supported in part by the Federal Ministry of Labor and Social Affairs (BMAS) through the AccessibleMaps project under Grant 01KM151112, in part by the University of Excellence through the "KIT Future Fields" project, in part by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition, and in part by Hangzhou SurImage Technology Company Ltd.

References

1. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: CVPR (2017) 4
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017) 1, 5
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 1
4. Li, Z., Snavely, N.: MegaDepth: Learning single-view depth prediction from internet photos. In: CVPR (2018) 4
5. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020) 4
6. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR (2021) 3, 4
7. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: InLoc: Indoor visual localization with dense matching and view synthesis. In: CVPR (2018) 6
8. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021) 1
9. Zhou, Q., Sattler, T., Leal-Taixe, L.: Patch2Pix: Epipolar-guided pixel-level correspondences. In: CVPR (2021) 6