

# Supplementary Materials: Semantic-aware Representation Learning for Homography Estimation

Anonymous Authors

## 1 IMPLEMENTATION DETAILS

### 1.1 Architecture of Semantic-guide Interactions Block (SGIB)

In order to integrate image features and fine-grained semantic features, we propose the Semantic-guide Interactions Block (SGIB). The specific network architecture is shown in Figure 1. In stage 1, we first perform convolution layers to transform the semantic features to the same dimension as image features. The semantic features  $S_i \in \mathbb{R}^{H_i^S \times W_i^S \times D_S}$ ,  $i = 0, 1$  are going through the self-attention module and then fed into the cross-attention as query  $Q$ . The image features  $\hat{C}_j \in \mathbb{R}^{H_j^C \times W_j^C \times D_C}$ ,  $j = 0, 1$  are transmitted into the cross-attention treated as key  $K$  and value  $V$  to produce fusion features  $\tilde{C}_j' \in \mathbb{R}^{H_i^S \times W_i^S \times D_C}$ , note that  $H_i^S = H_i^C$ ,  $W_i^S = W_i^C$ ,  $i = 0, 1$ . The calculation process of stage 2 is similar to stage 1.

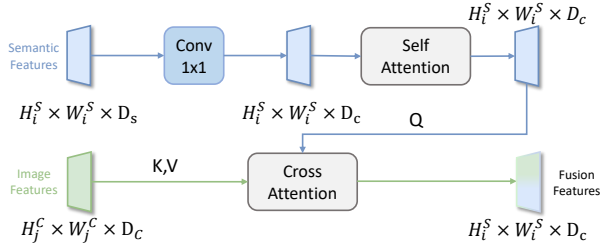


Figure 1: Architecture of Semantic-guide Interactions Block (SGIB). During stage 1,  $i = j$ ; while in stage 2,  $i \neq j$ .

## 2 MORE EXPERIMENTS

### 2.1 Performance of different methods

Table 1 shows the performance of various methods for homography estimation on HPatches [1]. We utilize the MegaDepth [12] dataset as the training dataset to retrain our SRMatcher following [21]. We compare two groups of the methods, Detector-based and Detector-free methods. Relative pose estimation results on the MegaDepth are given in Table 2.

### 2.2 Ablations about fusion strategies

Within SGIB, a variety of fusion techniques are available to incorporate priors, such as spatial attention and channel attention. We compare the cross-attention between semantic feature and image feature with these fusion strategies. As shown in Table 3, channel attention compromises the spatial integrity of semantics, resulting in the poorest performance. The performances of the spatial attention are also noticeably inferior to our proposed SGIB. This suggests that Semantic-Guide Interactions Block (SGIB) effectively implements cross-attention between semantic and image features, optimally utilizing the extensive semantic information available in Vision Foundation Models (VFM) while maintaining spatial integrity.

Table 1: Evaluation on HPatches [1] for homography estimation. For each method, a star symbol (\*) denotes the official version trained on the Oxford-Paris dataset, whereas versions without a star indicate its official release trained on the author-preferred dataset. The best and second results are highlighted.

Method	Homography est. AUC↑			
	@3px	@5px	@10px	mAUC
<i>Detector-based matching :</i>				
Superpoint [5] <small>CVPRW'18</small>	43.4	57.6	72.7	57.9
SIFT [16]	46.3	57.4	70.3	58.0
R2D2 [19] <small>NIPS'19</small>	50.6	63.9	76.8	63.8
SuperGlue [20] <small>CVPR'20</small>	53.9	68.3	81.7	68.0
<i>Detector-free matching :</i>				
LoFTR* [21] <small>CVPR'21</small>	58.5	69.8	81.1	69.8
LoFTR [21] <small>CVPR'21</small>	65.9	75.6	84.6	75.4
QuadTree [22] <small>ICLR'22</small>	66.3	76.2	84.9	75.8
ASpan [4] <small>ECCV'22</small>	67.4	76.9	85.6	76.6
TopicFM [9] <small>AAAI'23</small>	67.3	77.0	85.7	76.7
GeoFormer* [14] <small>ICCV'23</small>	68.0	76.8	85.4	76.7
<b>SRMatcher_LoFTR</b> , trained on MegaDepth	68.9	76.9	84.9	76.9
SEM [3] <small>CVPRW'23</small>	69.6	79.0	87.1	78.6
MESA [24] <small>CVPR'24</small>	71.1	78.6	86.0	78.6
<b>SRMatcher_GeoFormer*</b>	71.2	79.3	87.0	79.2
CasMTR-2c [2] <small>ICCV'23</small>	71.4	80.2	87.9	79.8
GeoFormer, trained on MegaDepth [14] <small>ICCV'23</small>	72.1	79.9	87.7	79.9
ASTR [23] <small>CVPR'23</small>	71.7	80.3	88.0	80.0
DKM [7] <small>CVPR'23</small>	71.3	80.6	88.5	80.1
PMatch [25] <small>CVPR'23</small>	71.9	80.7	<u>88.5</u>	80.4
RoMa [8] <small>CVPR'24</small>	<u>72.2</u>	<u>81.2</u>	<b>89.1</b>	<u>80.8</u>
<b>SRMatcher_GeoFormer</b> , trained on MegaDepth	73.5	<b>81.3</b>	88.0	<b>80.9</b>

Table 2: Relative pose estimation results (%) on MegaDepth-1500 benchmark. Training data: MegaDepth

Pose estimation AUC	MegaDepth1500 benchmark		
	AUC@5° ↑	AUC@10° ↑	AUC@20° ↑
TopicFM [9] <small>AAAI'23</small>	54.1	70.1	81.6
CasMTR-2c [2] <small>ICCV'23</small>	59.1	74.3	84.8
RoMa [8] <small>CVPR'24</small>	62.6	76.7	86.3
LoFTR [21] <small>CVPR'21</small>	52.8	69.2	81.2
SRMatcher_LoFTR	53.8	70.4	82.5
GeoFormer [14] <small>ICCV'23</small>	51.7	68.3	80.2
SRMatcher_GeoFormer	53.2	70.0	81.8

### 2.3 Ablations about Semantic Extractors

To investigate whether fine-grained semantic features enhance the efficacy of matching results, we employ three distinct semantic extractors. The first one is ResNet-50 [10] pre-trained on the ImageNet. Another one is the pre-trained CLIP [18] which has a strong semantic extraction ability due to its text-image pairs training method. As Table 4 shows, the features generated by DINOv2 [17] are more effective than the others. This superior performance

**Table 3: Ablations about different fusion strategies in SFB.**

Modification	Homography est. AUC			
	@3px	@5px	@10px	@mAUC
Channel Attention	69.7	78.0	85.7	77.8
Spatial Attention	70.1	78.5	86.0	78.2
SRMatcher_GeoFormer	<b>71.2</b>	<b>79.3</b>	<b>87.0</b>	<b>79.2</b>

stems from DINOv2's self-supervised training method, which compels the model to learn image features that are consistent across various transformations and inherently possess a high semantic value.

**Table 4: Ablations about different semantic extractors. 'R', 'C', 'D' denote the ResNet-50, CLIP and DINOv2.**

Modification	Homography est. AUC			
	@3px	@5px	@10px	@mAUC
SRMatcher-R	69.7	78.3	86.2	78.0
SRMacther-C	70.5	78.5	86.6	78.5
SRMatcher-D	<b>71.2</b>	<b>79.3</b>	<b>87.0</b>	<b>79.2</b>

## 2.4 Ablations about Different Layers

It is important to note that the semantic content extracted from different layers of DINOv2 varies. As the number of layers increases, the semantics become more representative. The key to successful semantic guidance is to extract semantic features that are both deep and capable of retaining critical spatial information, which is vital for ensuring effective semantic guidance. In our initial experimental design, following previous methods [11, 13] that utilized vision foundation models (VFMs), we opted not to use features from the last layer of VFMs. This issue was made because vision foundation models (VFMs), being pretrained for specific downstream tasks, may not be well-suited for task transfer. Therefore, in the main text, we use features from the third-to-last layer of DINOv2 as semantic priors. However, as shown in Tabel 5 we tried using the features from different layers as the semantics. We find that the the features from last layer lead to the performance increase, this finding differs from previous methods. We believe this is due to DINOv2 being pretrained through image-level and patch-level discriminative self-supervised learning, which enables it to extract all-purpose visual features, thus facilitating zero-shot patch-level feature matching capability [15].

**Table 5: Ablations about different layers of DINOv2.**

Modification	Homography est. AUC			
	@3px	@5px	@10px	@mAUC
Third to last ( <b>main text</b> )	71.2	79.3	87.0	79.2
Second to last	71.4	79.7	87.4	79.5
Last	<b>71.8</b>	<b>79.9</b>	<b>87.6</b>	<b>79.8</b>

## 2.5 Ablations about Semantic Extractor parameter

The DINOv2 is employed as semantic extractor to obtain various and available semantic information. Specifically, we use DINOv2 with a ViT-B/14 [6] with registers as the default semantic extractor of SRMatcher. We also conduct comparison experiments on ViT-S/14 and ViT-L/14 with different parameters shown in Table 6.

**Table 6: Ablations about DINOv2 parameters.**

Modification	Params	Homography est. AUC			
		@3px	@5px	@10px	mAUC
DINOv2_ViT-S/14	21M	70.5	78.9	86.5	78.6
DINOv2_ViT-B/14	86M	71.2	79.3	87.0	79.2
DINOv2_ViT-L/14	300M	<b>71.3</b>	<b>79.4</b>	<b>87.3</b>	<b>79.3</b>

## 3 QUALITATIVE RESULTS

We provide additional qualitative comparisons of SRMatcher and baseline methods on the Hpatches [1], ISC-HE [14] and MegaDepth [12] datasets. In Figure 2 and Figure 4, we illustrate inlier and outlier matches with various projection thresholds to evaluate the matching precision of different methods on the Hpatches dataset and ISC-HE dataset. Figure 3 and Figure 5 display further qualitative results of homography estimation, the methods being compared include LoFTR [21], GeoFormer [14], MESA [24], and our SRMatcher\_GeoFormer. Figure 6 offers more qualitative insights on the MegaDepth dataset, the methods being compared include LoFTR, SRMatcher\_LoFTR, GeoFormer, SRMacther\_GeoFormer trained on MegaDepth dataset.



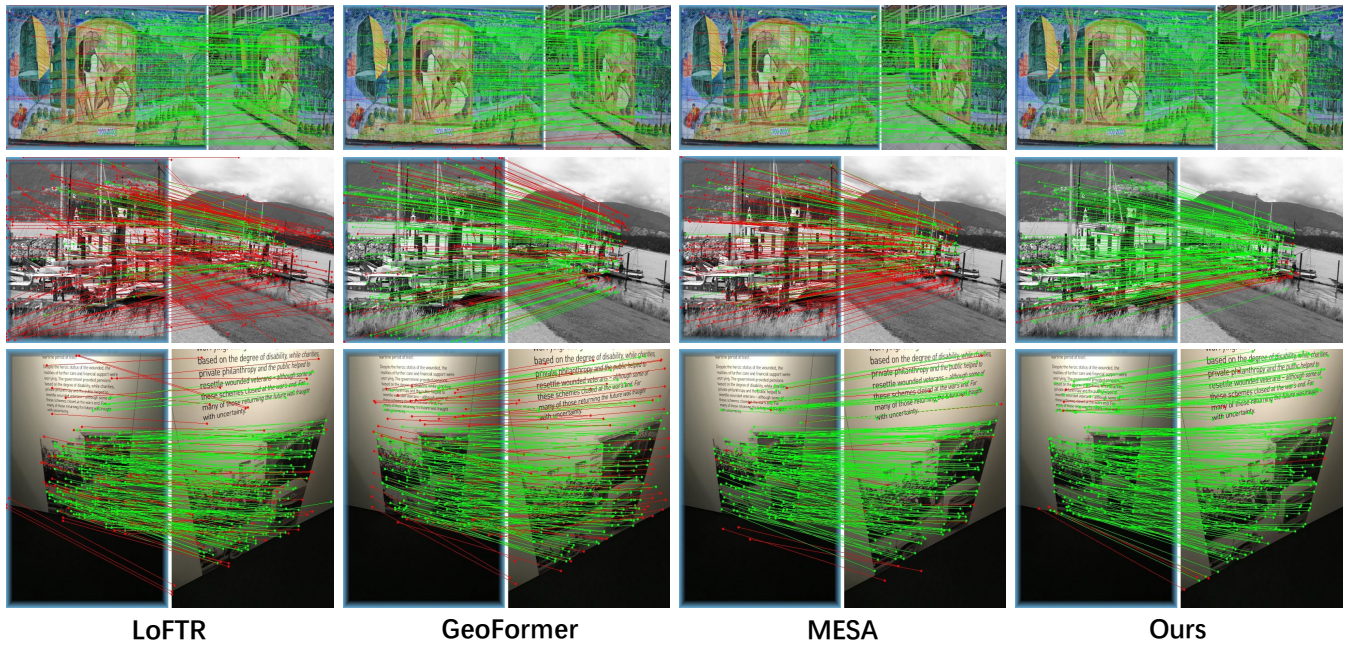


Figure 2: Qualitative of matching results with LoFTR [21], GeoFormer [14], MESA [24], and our SRMatcher on HPatches [1]. Points classified as inliers by RANSAC are displayed in green, while outliers are shown in red.

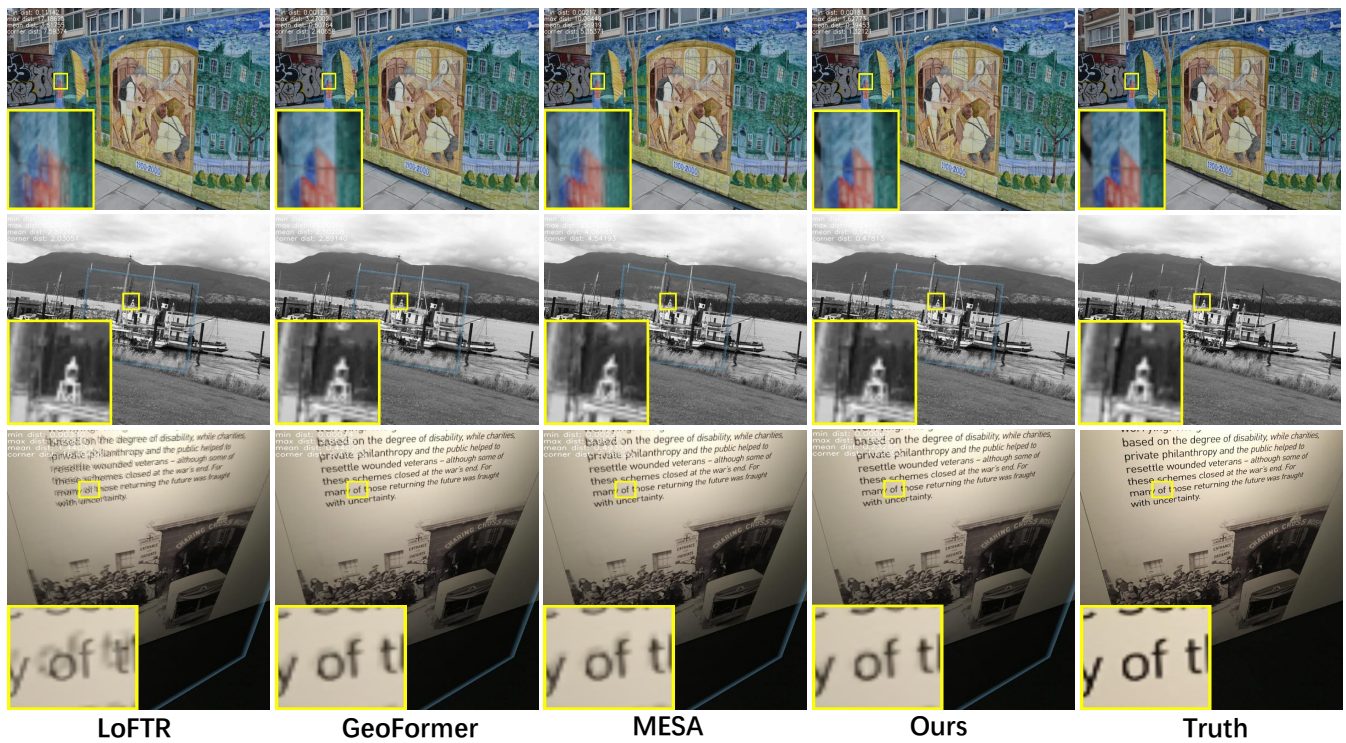


Figure 3: Qualitative of homography estimation results with LoFTR [21], GeoFormer [14], MESA [24], and our SRMatcher on HPatches [1].



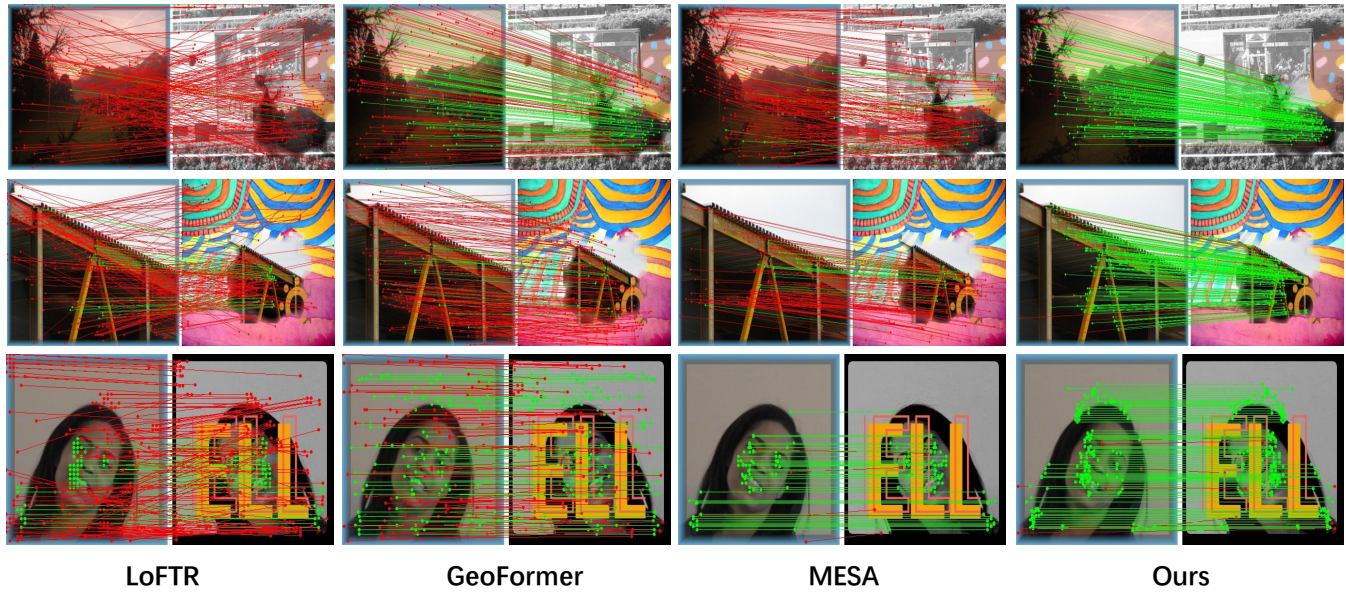


Figure 4: Qualitative of matching results with LoFTR [21], GeoFormer [14], MESA [24], and our SRMatcher on ISC-HE [14]. Points classified as inliers by RANSAC are displayed in green, while outliers are shown in red.

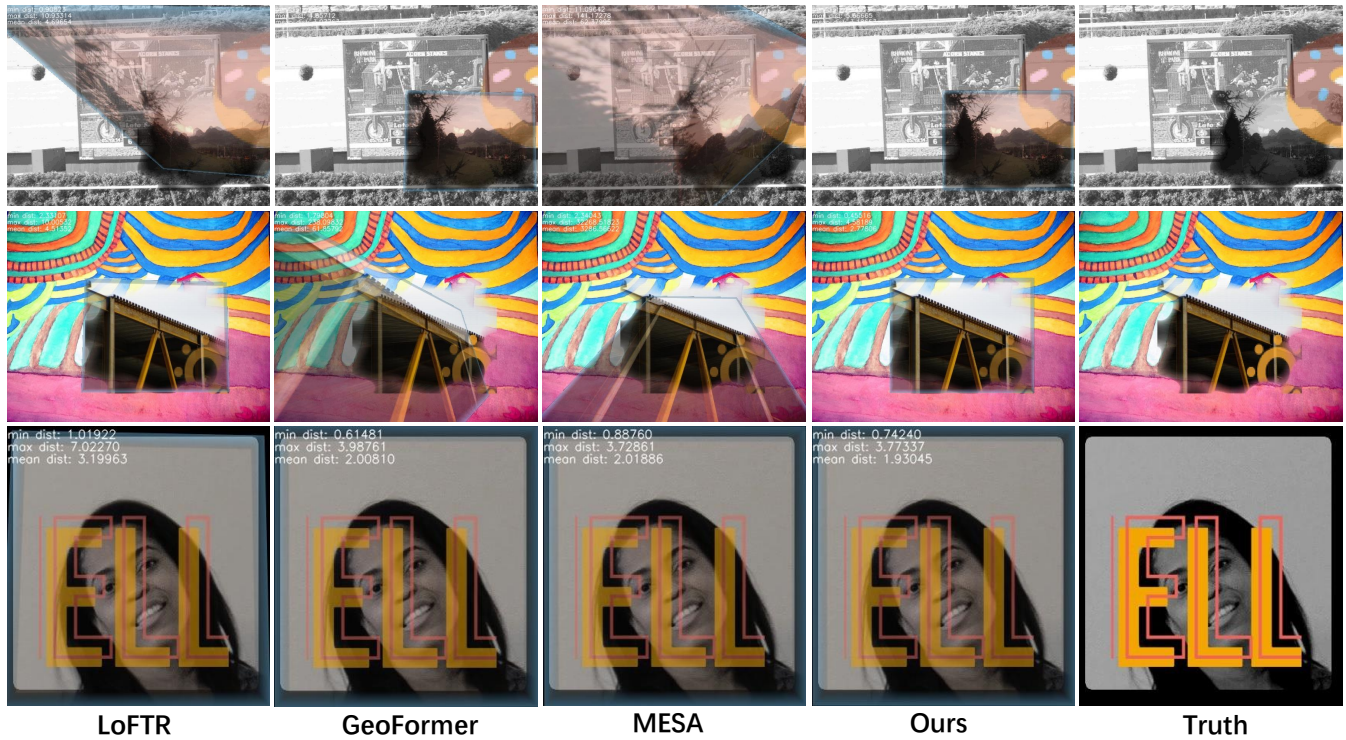
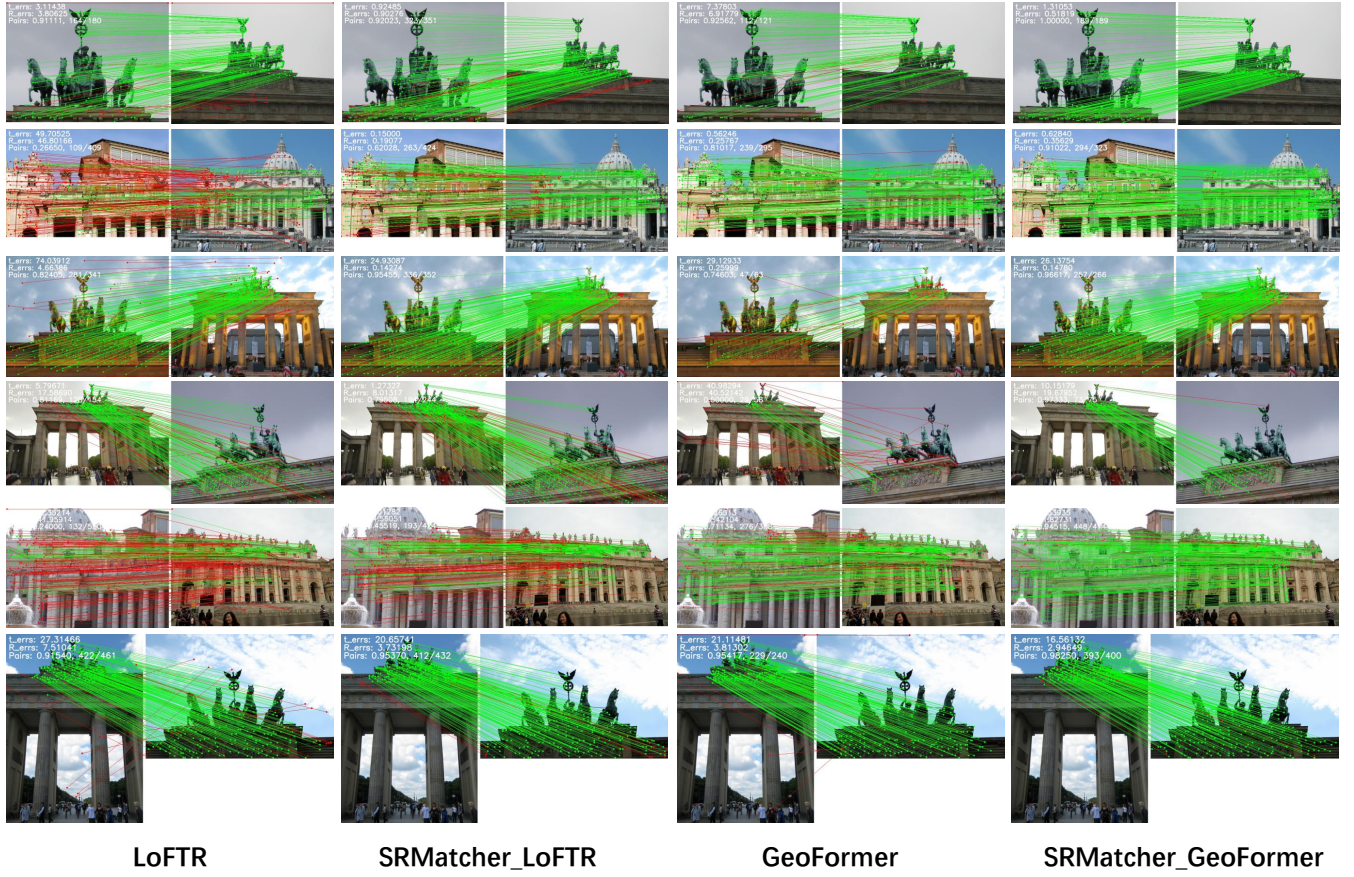


Figure 5: Qualitative of homography estimation results with LoFTR [21], GeoFormer [14], MESA [24], and our SRMatcher on ISC-HE [14].





**Figure 6: Qualitative image matches on MegaDepth dataset. Green signifies that the epipolar error in normalized image coordinates is below  $1 \times 10^{-4}$ , whereas red denotes that this threshold has been surpassed. Training data: MegaDepth.**

## REFERENCES

- [1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. 2017. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5173–5182.
- [2] Chenjie Cao and Yanwei Fu. 2023. Improving transformer-based image matching by cascaded capturing spatially informative keypoints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12129–12139.
- [3] Jiahao Chang, Jiahuan Yu, and Tianzhu Zhang. 2023. Structured Epipolar Matcher for Local Feature Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6176–6185.
- [4] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. 2022. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*. Springer, 20–36.
- [5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 224–236.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [7] Johan Edstedt, Ioannis Athanasiadis, Márten Wadenbäck, and Michael Felsberg. 2023. DKM: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17765–17775.
- [8] Johan Edstedt, Qiyu Sun, Georg Bökman, Márten Wadenbäck, and Michael Felsberg. 2023. RoMa: Revisiting Robust Losses for Dense Feature Matching. *arXiv preprint arXiv:2305.15404* (2023).
- [9] Khang Truong Giang, Soohwan Song, and Sungho Jo. 2023. TopicFM: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2447–2455.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [11] Bingchen Li, Xin Li, Hanxin Zhu, Yeying Jin, Ruoyu Feng, Zhizheng Zhang, and Zhibo Chen. 2024. SeD: Semantic-Aware Discriminator for Image Super-Resolution. *arXiv preprint arXiv:2402.19387* (2024).
- [12] Zhengqi Li and Noah Snavely. 2018. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2041–2050.
- [13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [14] Jiazhen Liu and Xirong Li. 2023. Geometrized Transformer for Self-Supervised Homography Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9556–9565.
- [15] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. 2023. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310* (2023).
- [16] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60 (2004), 91–110.
- [17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [19] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. 2019. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems* 32 (2019).
- [20] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4938–4947.
- [21] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8922–8931.
- [22] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. 2022. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767* (2022).
- [23] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Wu Feng. 2023. ASTR: Adaptive spot-guided transformer for consistent local feature matching. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, Vol. 7.
- [24] Yeseng Zhang and Xu Zhao. 2024. MESA: Matching Everything by Segmenting Anything. *arXiv preprint arXiv:2401.16741* (2024).
- [25] Shengjie Zhu and Xiaoming Liu. 2023. Pmatch: Paired masked image modeling for dense geometric matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21909–21918.