ETO:Efficient Transformer-based Local Feature Matching by Organizing Multiple Homography hypotheses
– Supplementary Material –

Junjie Ni[1]   **Guofeng Zhang**[1*]   **Guanglin Li**[1]   **Yijin Li**[1]
Xinyang Liu[1]   **Zhaoyang Huang**[2]   **Hujun Bao**[1⋆]
[1]State Key Lab of CAD&CG, Zhejiang University   [2]CUHK MMLab

In this supplementary document, we describe the parametric scheme of homography matrix in Sec. 1, provide an additional explanation with graph for our uni-directional cross attention in Sec. 2, discuss the details on segmentation in Sec. 3, describe more implementation details in Sec. 4, provide a proof for the use of homography hypotheses in Sec. 5 and show some qualitive results in Sec. 6.

1   Parametric Scheme

H_i can be decomposed into 2d-translation $p_i^t - p_i^s \in \mathcal{R}^2$, scale $s_i \in \mathcal{R}^1$, rotations around the z-axis $r_i \in \mathcal{R}^1$, and perspective components $q_i \in \mathcal{R}^4$. We use these attributes to calculate four imaginary points in target images to construct the system of linear equations and solve them for homography matrix:

$$
\begin{aligned}
B &= \begin{bmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \\
B_i^s &= B + p_i^s \\
q_i &= \delta_{xx}, \delta_{xy}, \delta_{yx}, \delta_{yy}, \\
Q_i &= \begin{bmatrix} -\delta_{xx} - \delta_{xy} & -\delta_{yx} - \delta_{yy} \\ \delta_{xx} - \delta_{xy} & +\delta_{yx} - \delta_{yy} \\ -\delta_{xx} + \delta_{xy} & -\delta_{yx} + \delta_{yy} \\ \delta_{xx} + \delta_{xy} & +\delta_{yx} + \delta_{yy} \end{bmatrix} \\
B_i^t &= p_i^t + \mathscr{R}(B + Q_i, r_i) * s_i.
\end{aligned}
\tag{1}
$$

Here Q_i represents the influence of perspective vectors q_i for B_i^t in the 1st-order of Taylor series, which behaves as the offsets on B. $\mathscr{R}$ is the operation of rotate points around their center for r_i degree. B_i^s and B_i^t are four virtual points that assist in calculating the homography matrix H_i. These operations allow each variable within the homography matrix H_i to be deduced from four projection equations $B_i^t = H_i B_i^s$.

The reason why we use this parametric scheme to solve the homography matrix instead of directly estimating the coordinates of the four imaginary points on target images is that the direct parametric scheme can easily construct singular matrices. For example, if connecting three of the four points in a line, the optimization process will fail. In the experiment of outdoor pose estimation for Megadepth dataset [1], the direct parametric scheme will induce 0.53 for the indicator of AUC@5, while our parametric scheme induce 28.5.

⋆Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2
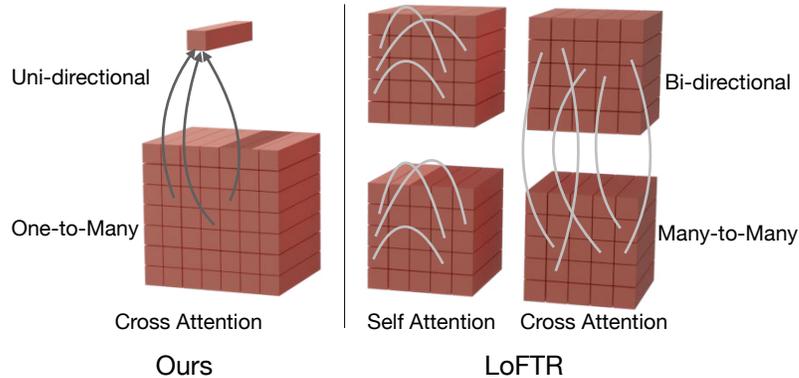
Fig. 1. Uni-directional cross attention.

2 Uni-directional Cross Attention.

As shown by Fig. 1, previous methods [2] apply self-attention and cross-attention to each feature within a 5×5 feature map, resulting in 2,500 (25×25×4) inner product calculations to gather feature information within a 4-pixel radius. In contrast, our approach conducts a uni-directional cross-attention solely at the query position on a 7×7 feature map, requiring just 49 inner product calculations to capture feature information up to a 6-pixel distance. This makes our method approximately 50 times faster than the previous approach.

3 Details on Segmentation.

Segmentation refers to the classification of each unit, where we determine which homography hypothesis should be adopted for unit j on M_2 through classification. The way to obtain the classification result is by comparing the classification score matrix C_j of unit j for different hypotheses H_i , where the largest one is the result of our classification operation. This classification uses the concept of multi-label classification, a method widely applied in detection problems. Therefore, we refer to DETR and use focal loss to optimize segmentation here. We can describe the process of obtaining the classification score matrix C_j in the form of a formula: $C_{ji} = (T(f_j) + P(i), f_j)$, where C_{ji} refers to the matching score of unit j for hypothesis i. T refers to the function that converts the feature dimension of i (256 dimensions) to the feature dimension of j (128 dimensions); here, we use a 2D CNN to perform T . P refers to positional embedding, which directly represents the relative position of the unit corresponding to the hypotheses i in the local 3*3 units. And (* , *) indicates the inner product.

4 Implementation Details

Outdoor model. When training the outdoor model, in order to make our model more generalized, we introduced a data enhancement after the initial training. Specifically, we customized the collect_fn function to make the matching images in different batches have different resolutions, while the matching images in the same batch have the same resolution. In addition, we also rotate 10% of the matched images by 90 degrees to make the model more robust to extreme rotation.

Indoor model. Consider that our model is trained on 3 RTX3090, we differentiated between the training data on different GPUs when training the indoor model, specifically by using the ScanNet dataset for training on two RTX3090 and Megadepth on the third RTX3090.

3
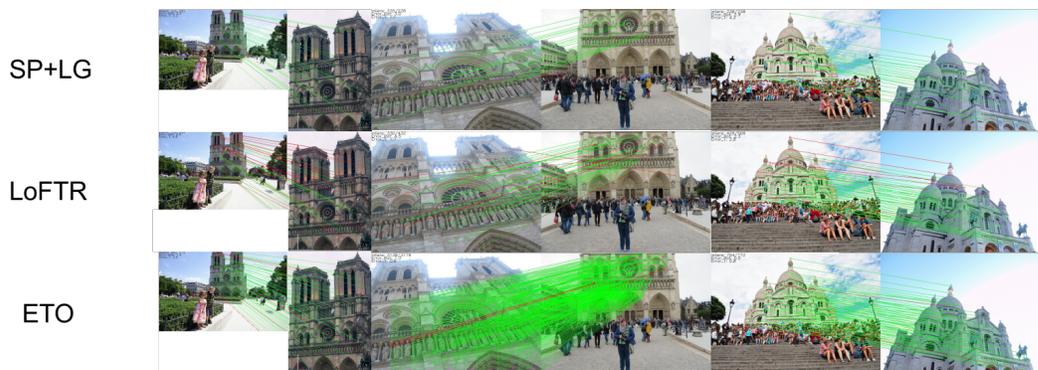
Fig. 2. Qualitative Results of Feature Matching. Inlier matches are highlighted in green and outliers in red. For visual clarity, the displayed matches are reduced to one-tenth of the actual number. As can be seen from the figure, our method is robust to various extreme scenarios and thus can achieve very superior performance.

5 Proof

According to the theory of multiple view geometry [3], the correspondence for the same plane in $\mathcal{R}^3$ from two viewpoints can be defined by a homography matrix. Here we provide the process of proof.

the correspondence function from two view points is:

$$x_2 = K_2(RK_1^{-1}x_1 + t) \tag{2}$$

where R is the rotation matrix, t is the translation vector, K is the intrinsic matrix of camera, x is the coordinates of points on images, and the plane can be defined as:

$$\frac{1}{d}n^T K_1^{-1} x_1 = 1 \tag{3}$$

where d is the distance between points and the plane and n is the normal vector of the plane. Then we can substitute Eq. 3 into Eq. 2:

$$x_2 = K_2(R + \frac{1}{d}tn^T)K_1^{-1}x_1$$
$$H = K_2(R + \frac{1}{d}tn^T)K_1^{-1} \tag{4}$$

Here H is the homography matrix. We use the homography hypothesis to represent the correspondence is the same as simplifying real world in $\mathcal{R}^3$ to many planes.

6 Qualitative Results.

We show some Qualitative Results in Fig. 2.

Bibliography

[1] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[2] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021.

[3] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.