Supplementary Materials

Runyu Mao[1], Chen Bai[2], Yatong An[2], and Fengqing Zhu[1] and Cheng Lu[2]

[1] Purdue University {mao111, zhu0}@purdue.edu
[2] XPeng Motors {chenbai, yatongan, luc}@xiaopeng.com

1 Homography Estimation

We take a more detailed look at the homography evaluation from Section 4.3. We follow [3] to select 108 image sequences. There are 52 sequences under significant illumination changes (illumination set) and 56 sequences that have significant viewpoint variations (viewpoint set). All images are resized with shorter dimensions equal to 480. The model we used to evaluate is the student model we trained on Megadepth dataset [4], the training setting is reported in Section 3.5. We select top 1000 matches and vary the threshold from 1 pixel to 10 pixels for Mean Match Accuracy (MMA) [5] calculation. As shown in Table 1, we evaluate our model on illumination set and vewpoint set seperately and finally report the overall performance on the entire 540 image pairs.

Table 1: **Evaluation on HPatches [1].** The Mean Match Accuracy (MMA) at different thresholds.

Test set	MMA (%) threshold									
	≤1px	≤2px	≤3px	≤4px	≤5px	≤6px	≤7px	≤8px	≤9px	≤10px
Illumination	79.50	95.18	98.58	98.79	98.87	98.98	99.14	99.33	99.48	99.56
Viewpoint	55.78	74.14	79.01	80.98	81.85	82.29	82.54	82.74	82.90	83.02
Overall	67.53	84.57	88.71	89.81	90.28	90.56	90.77	90.96	91.12	91.22

2 Model Compression

Our student-teacher learning architecture can also be generalized to model compression tasks. A slim model with half attention layers on the coarse-level transformer L_c is introduced. In this section, we compare our full-size model and slim model on other metrics, i.e. FLOPs and total parameters, to see the compression effects. The model contains Feature Pyramid Networks (FPN), coarse-level transformer, and fine-level transformer. The coarse-level threshold is removed so that all the coarse matching pairs will be fed to the fine-level transformer. The input is a random tensor with $3 \times 640 \times 480$ dimension.

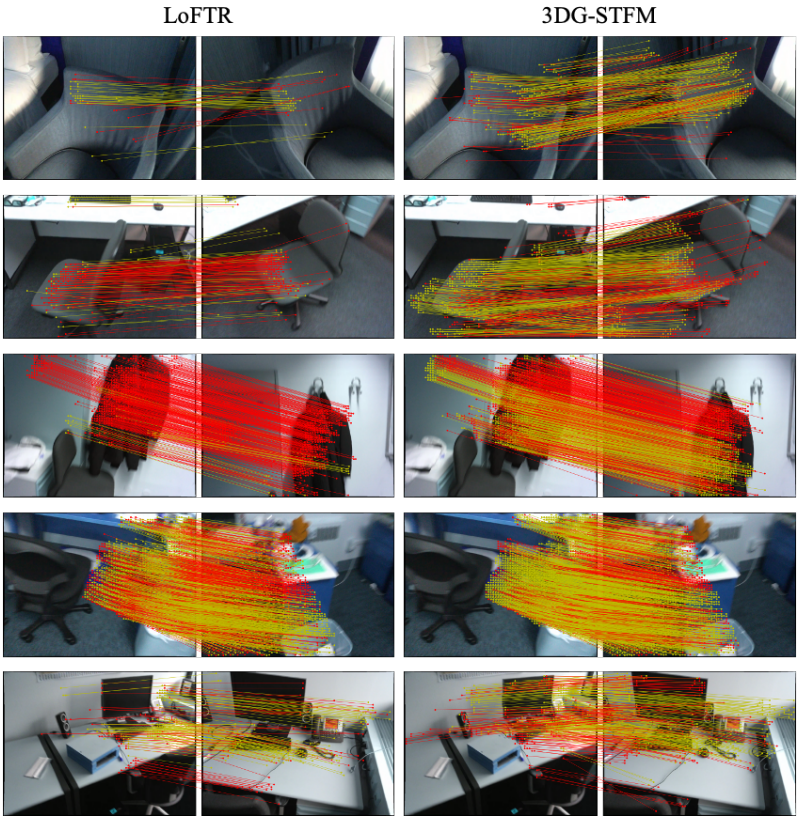
As shown in Table 2, we reduce the coarse-level transformer layer by half for our slim model. The overall FLOPs are reduced by 10.82% and the total parameters are reduced by 22.64%.

Table 2: **Model Compression Performances.**

Model	L_c	L_f	FLOPs(B)	Params (M)
Full-size model	4	1	365.23	11.57
Slim model	2	1	325.71	8.95

3 Qualitative Results

More comparison of our 3DG-STFM and baseline LoFTR [6] on ScanNet [2] and MegaDepth [4] datasets is illustrated in Fig. 1 and 2 We also include a video to demonstrate our method's performance on a challenging low texture indoor scene. The scatter indicates the feature location and the color represents the confidence score, high in red, low in blue. Benefited from the discrimination of depth modality, our method could detect more features on low texture regions since it learns the RGB-induced depth distribution from the teacher model.



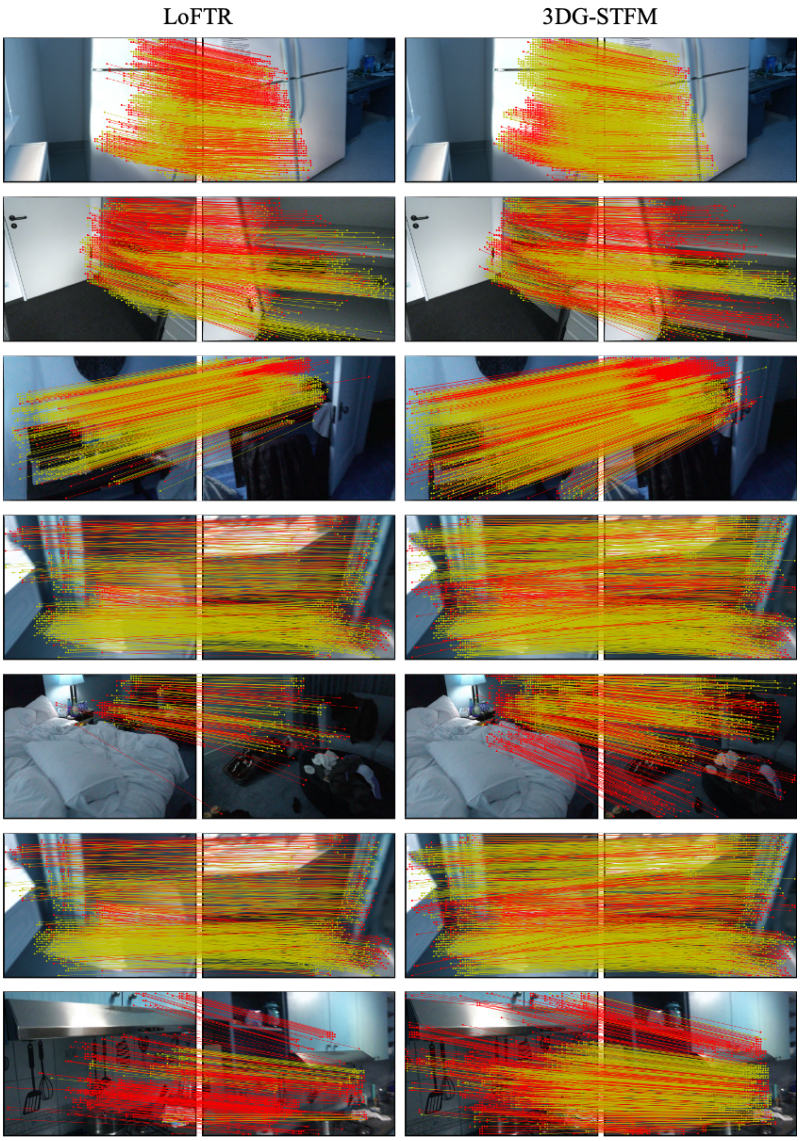


Fig. 1: **Qualitative image matches on ScanNet [2].** The red color indicates epipolar error beyond 5×10^{-4} (in the normalized image coordinates).


LoFTR
3DG-STFM

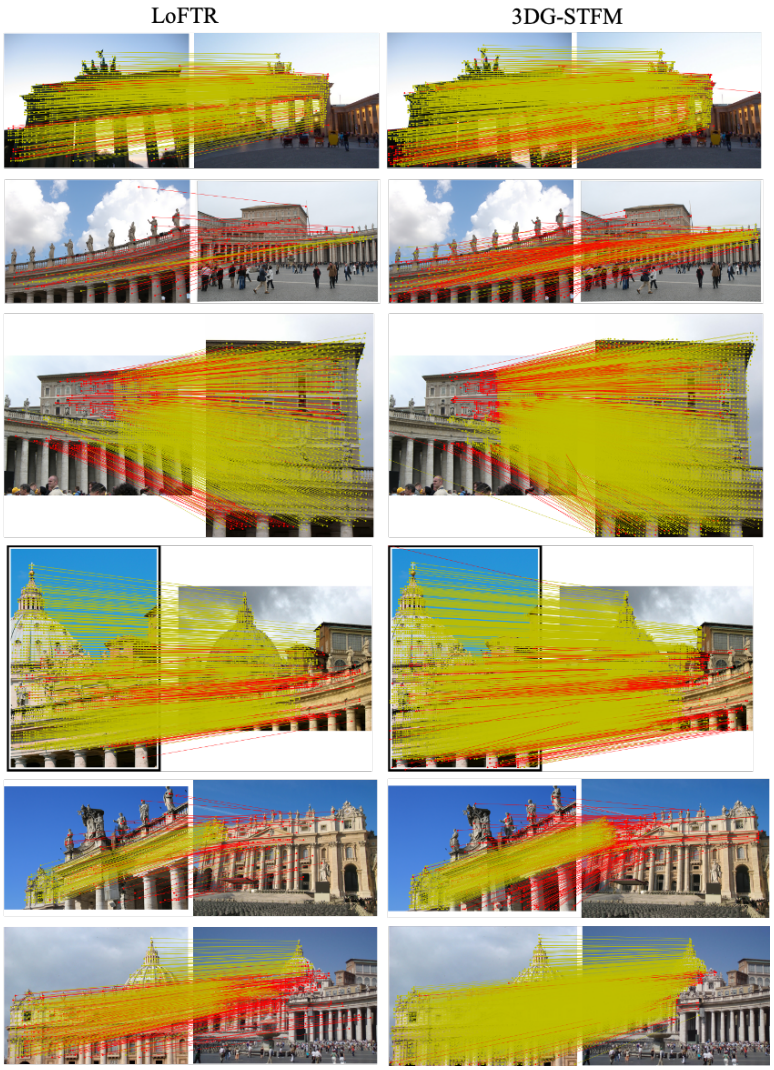


Fig. 2: **Qualitative image matches on Megadepth** [4]**.**The red color indicates epipolar error beyond 1×10^{-4} (in the normalized image coordinates).

References

1. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 5173–5182 (2017) 1
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 5828–5839 (2017) 2, 3
3. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. Proceedings of the ieee/cvf conference on computer vision and pattern recognition pp. 8092–8101 (2019) 1
4. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2041–2050 (2018) 1, 2, 5
5. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE transactions on pattern analysis and machine intelligence **27**(10), 1615–1630 (2005) 1
6. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 8922–8931 (2021) 2