End2End Multi-View Feature Matching with Differentiable Pose Optimization Supplementary Material

Barbara Roessle and Matthias Nießner
Technical University of Munich

1. Ablation Study

Multi-View & End-to-End. The quantitative ablation results on ScanNet [2] and MegaDepth [8] confirm that the full version of our method achieves highest performance (Tabs. 1 and 2). Fig. 5 shows qualitative results of the ablation experiments on Matterport3D [1]. Clearly, multi-view matching and end-to-end training support the correspondence reasoning and improve camera alignment, despite the extreme viewpoint changes.

	Transl. error AUC [%] ↑			Rot. error AUC [%] ↑		
	@5°	@10°	@20°	@5°	@10°	@20°
Ours w/o multi-view	24.9	42.5	59.6	60.7	75.3	85.0
Ours w/o end-to-end	23.7	40.4	56.8	57.5	73.7	84.4
Ours	**26.9**	**45.6**	**63.0**	**64.2**	**78.8**	**87.7**

Table 1. Ablation study on multi-view indoor pose estimation on ScanNet.

	Transl. error AUC [%] ↑			Rot. error AUC [%] ↑		
	@5°	@10°	@20°	@5°	@10°	@20°
Ours w/o multi-view	50.2	60.9	70.5	64.4	75.7	84.1
Ours w/o end-to-end	49.9	60.8	70.5	61.6	74.7	84.2
Ours	**52.1**	**63.0**	**72.5**	**66.7**	**77.8**	**85.9**

Table 2. Ablation study on multi-view outdoor pose estimation on MegaDepth.

Variable Image Overlap. Tab. 3 extends the multi-view pose estimation evaluation to a setting with reduced image overlap. It shows that our method achieves better pose estimation results than the baselines also in this setting.

2. Qualitative Results

Figs. 3 to 5 show additional qualitative results on ScanNet, MegaDepth and Matterport3D. Lower reprojection errors demonstrate that our matches give rise to more accurate pose estimation, even in texture-less areas (e.g., Fig. 3 sample 2) or across strong appearance changes (e.g., Fig. 4 sample 1).

		Transl. error AUC [%] ↑			Rot. error AUC [%] ↑		
		@5°	@10°	@20°	@5°	@10°	@20°
Overlap 1	Mutual nearest neighbor	8.5	17.8	31.0	33.0	48.4	62.8
	SuperGlue [11]	21.3	37.5	53.7	54.2	71.0	82.6
	LoFTR [12]	20.6	36.9	53.7	57.3	72.0	82.0
	COTR [5] cross-dataset	10.9	22.4	36.9	38.8	53.6	66.3
	3DG-STFM [10]	22.0	38.7	55.5	57.0	72.7	83.0
	Ours	**26.9**	**45.6**	**63.0**	**64.2**	**78.8**	**87.7**
Overlap 2	Mutual nearest neighbor	3.4	8.1	16.9	12.7	23.6	38.1
	SuperGlue [11]	15.8	29.1	44.3	34.6	52.1	67.3
	LoFTR [12]	15.8	28.5	43.1	35.6	51.6	65.1
	COTR [5] cross-dataset	5.4	11.9	22.2	17.4	29.0	42.6
	3DG-STFM [10]	15.4	28.1	43.0	34.3	50.3	64.5
	Ours	**20.9**	**36.6**	**53.0**	**42.8**	**60.0**	**73.6**

Table 3. Multi-view indoor pose estimation using variable image overlap (range 1: $[0.4, 0.8]$, range 2: $[0.25, 0.5]$) on ScanNet; "cross-dataset" indicates that COTR was trained on MegaDepth.

3. Cross-Dataset Results

	Pose error AUC [%] ↑		
	@5°	@10°	@20°
SuperGlue [11]	38.7	59.1	75.8
LoFTR [12]	43.5	63.5	78.6
COTR [5]	34.4	54.7	71.8
3DG-STFM [10]	43.4	63.4	78.4
Ours	**46.7**	**65.4**	**79.3**

Table 4. Cross-dataset evaluation on two-view pose-estimation on YFCC100M. Models trained on MegaDepth.

	Pose error AUC [%] ↑		
	@5°	@10°	@20°
SuperGlue [11]	16.7	33.7	51.1
LoFTR [12]	17.7	34.7	51.1
COTR [5]	11.8	26.5	42.5
3DG-STFM [10]	16.1	32.3	49.2
Ours	**18.8**	**36.4**	**52.8**

Table 5. Cross-dataset evaluation on two-view pose-estimation on ScanNet. Models trained on MegaDepth.

Tabs. 4 and 5 list cross-dataset results on two-view pose estimation, where the models are trained on MegaDepth and tested on YFCC100M [13] and ScanNet. It shows that our

method is able to transfer to different datasets.

4. Matching Metrics

Following the detector-based method SuperGlue, we compute precision (P) and matching score (MS) [11]. Our end-to-end approach learns matching and outlier filtering in one step, hence, in contrast to the baselines, it does not need outlier filtering with RANSAC to estimate poses. Tab. 6 shows that we achieve comparable or higher precision and matching score than SuperGlue with RANSAC.

		RANSAC	P [%] ↑	MS [%] ↑
SuperGlue [11]	2-view	✓	93.8 (91.3)	19.3 (38.6)
Ours	4-view	✗	**94.0**	19.6
Ours	5-view	✗	**94.0**	19.4
Ours	6-view	✗	93.9	**19.8**

Table 6. Matching metrics on ScanNet. Our end-to-end method learns feature matching and outlier filtering in one step, hence, it does not require RANSAC and yields matches of similar or higher precision and matching score compared to SuperGlue with RANSAC. Parentheses indicate SuperGlue metrics w/o RANSAC.

This evaluation (Tab. 6) is not defined for the detector-free methods (as explained in [12]), therefore, we provide an alternative evaluation, which is applicable to the detector-free methods: Fig. 1 visualizes the trade-off between the precision of matches and the pose estimation performance for increasing confidence thresholds (lower bound) starting at 0 until precision saturates. The curves are computed on the ScanNet image pairs from two-view pose estimation (main paper Section 4.1). Clearly, our method produces matching configurations with the best trade-off between precision and value for pose estimation. The baseline COTR does not provide confidences, hence its curve boils down to a point: 76.8% precision at AUC@20°of 42.5%.

5. Matching Runtime

Tab. 7 lists the matching runtime for increasing number of views, measured on a Nvidia GeForce RTX 2080. It shows that joint multi-view matching is faster than matching the corresponding pairs with SuperGlue. The savings stem from fewer intra-frame, self-attention GNN messages in multi-view matching compared to pairwise (see Sec. 8).

	2-view ≙ 1 pair	4-view ≙ 6 pairs	5-view ≙ 10 pairs	6-view ≙ 15 pairs	8-view ≙ 28 pairs
SuperGlue [11]	**45**ms	190ms	315ms	470ms	849ms
Ours	**45**ms	**181**ms	**260**ms	**352**ms	**589**ms

Table 7. Matching runtime (excluding SuperPoint) for variable number of views on ScanNet.

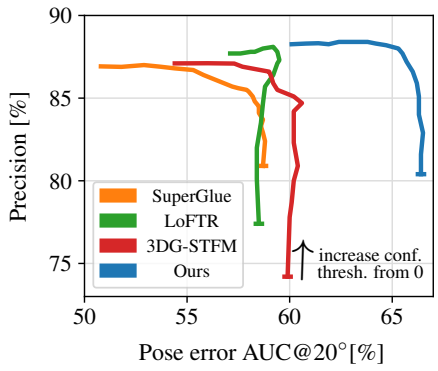

Figure 1. Trade-off between matching precision and pose estimation performance for variable confidence thresholds on ScanNet. Our matching results are both, of high precision and of high value for pose estimation.

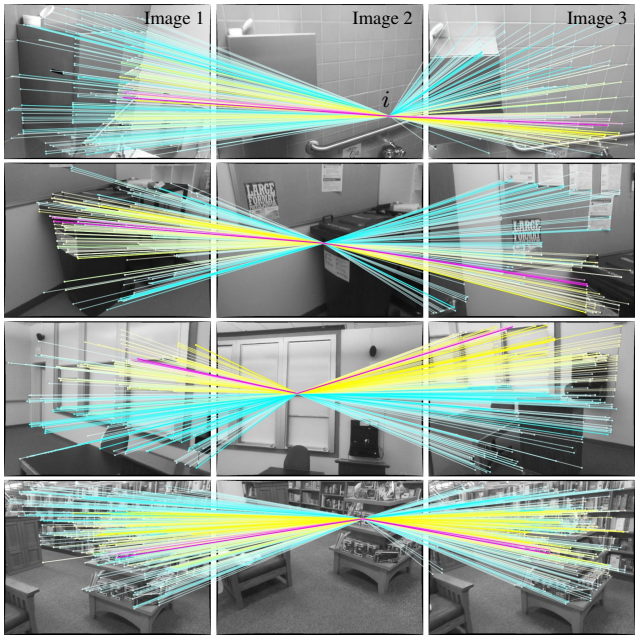

Figure 2. Early/mid/late layer cross-attention weights as opacity. Keypoint i in image 2 first interacts with spread points in images 1 and 3, then focuses around the match in middle and late cross-attention layers.

6. Cross-Attention Visualization

Fig. 2 visualizes cross-attention weights. In early layers keypoints interact with spread keypoints in the other images. In later layers, cross-attention more and more focuses on the region of the matching keypoint.

7. Training with Bundle Adjustment

We found that adding bundle adjustment in the end-to-end training, compared to training with weighted eight-

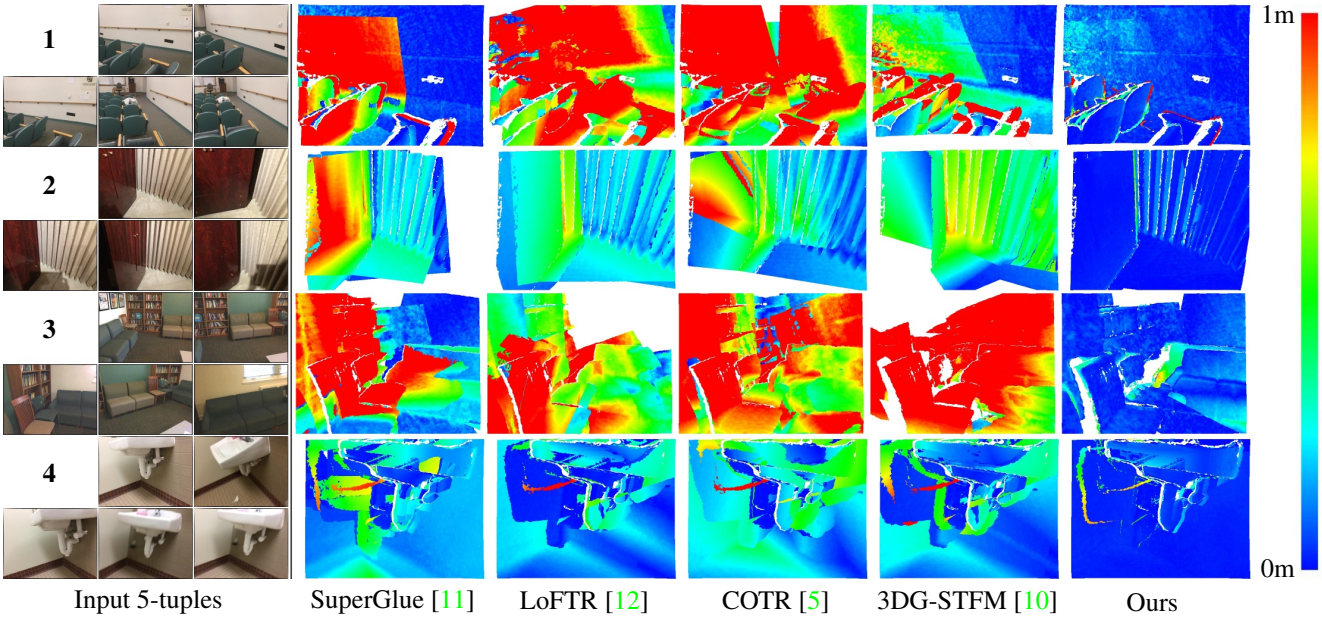

Figure 3. Reprojection error (right) for estimated camera poses on ScanNet 5-tuples (left). With multi-view matching and end-to-end training, our method successfully handles challenging pose estimation scenarios, while baselines have severe camera pose errors.

point alone, leads to a minor improvement in the pose error AUC (Tab. 8)—hence, we favored the simpler training procedure with weighted eight-point alone. At test time, however, the pose refinement with bundle adjustment is highly beneficial as shown in the experiment section of the main paper.

	weight. 8-point training	bundle adjust. training	Pose error AUC [%] ↑		
			@5°	@10°	@20°
Ours	✓	✗	25.7	47.2	66.4
Ours	✓	✓	**26.0**	**47.6**	**66.7**

Table 8. End-to-end training with weighted 8-point and bundle adjustment on ScanNet.

8. Number of GNN Messages

Tab. 9 shows that jointly matching N images in a single graph reduces the number of GNN messages along self-edges compared to separately matching the corresponding $P = \sum_{n=1}^{N-1} n$ pairs. E.g., consider matching 5 images with K keypoints each, either (A) jointly in a single match graph or (B) matching the 10 possible pairs. In each layer, (A) computes self-attention for 5 images, hence $5K^2$ GNN messages (B) computes self-attention for 10 pairs, i.e., 20 images, hence $20K^2$ GNN messages. The number of messages along cross-edges is the same in pairwise and joint matching.

	Number of GNN messages	
	along self-edges	along cross-edges
Pairwise matching	$2PK^2$	$N(N-1)K^2$
Joint matching	NK^2	$N(N-1)K^2$

Table 9. Number of GNN messages per layer for matching N images, each with K keypoints, as P individual image pairs versus joint matching in a single graph.

9. Architecture Details

Our multi-view matching network is inspired by the SuperGlue [11] architecture.

Keypoint Encoder. The input visual descriptors from SuperPoint [3] have size $D = 256$. The graph nodes equally have an embedding size of D. Hence, the keypoint encoder F_{encode} maps a keypoint's image coordinates and confidence score to D dimensions. It is a MLP, composed of five layers with 32, 64, 128, 256 and D channels. Each layer, except the last, uses batch normalization and ReLU activation.

Graph Attention Network. We found that multi-view matching benefits from more information flow along cross-edges compared to self-edges. Hence, the GNN has 7 self-attention layers, each followed by three cross-attention layers. In the two-view setting and on MegaDepth—due to limited amount of data—we use a smaller network size with 9 self- and 9 cross-attention layers in alternating fash-

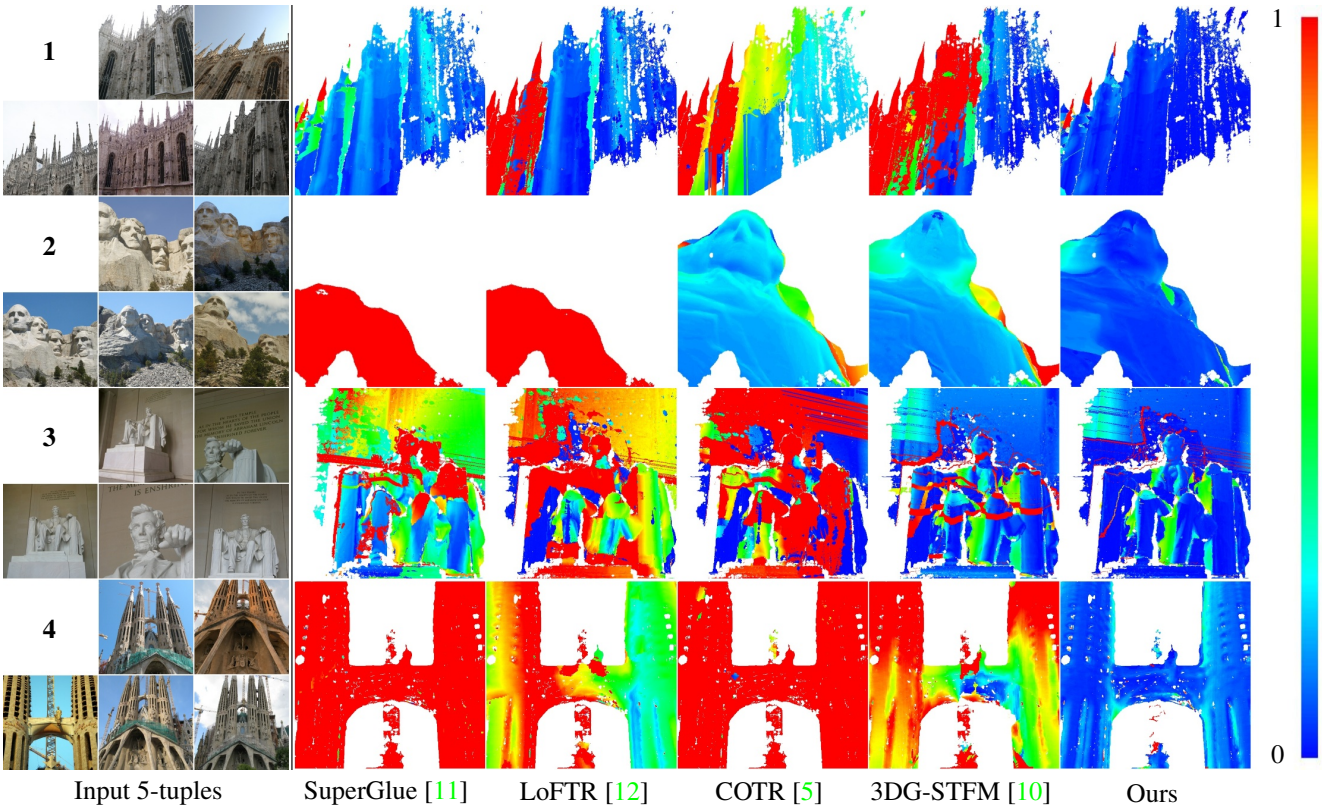

Figure 4. Reprojection error (right) for estimated camera poses on MegaDepth 5-tuples (left). Through multi-view matching and end-to-end training, our method successfully estimates camera poses in challenging outdoor scenarios, while baselines show misalignment. Reprojection errors are visualized in the MegaDepth scaling.

ion. The attentional aggregation of incoming messages from other nodes uses multi-head attention with four heads. The resulting messages have size D, like the node embeddings. The MLP F_{update}, which computes the update to the receiving node, operates on the concatenation of the current node embedding with the incoming message. It has two layers with $2D$ and D channels. Batch normalization and ReLU activation are employed between the two layers.

Partial Assignment. We use 100 iterations of the Sinkhorn algorithm to determine the partial assignment matrices.

Confidence MLP. $F_{\text{conf\_3}}$ merges the final node descriptors of matching keypoints—i.e., it operates on the concatenated match descriptors and applies two linear layers with $2D$ and D channels. $F_{\text{conf\_2}}$ lifts the corresponding partial assignment score to descriptor space through two linear layers with D channels each. The D-dimensional output embeddings of $F_{\text{conf\_2}}$ and $F_{\text{conf\_3}}$ are summed and fed into $F_{\text{conf\_1}}$, which is a final linear layer with sigmoid activation that reduces to a single channel, the matching confidence. All layers in $F_{\text{conf\_2}}$ and $F_{\text{conf\_3}}$ use batch normalization and ReLU activation.

Pose Optimization. The camera poses are optimized by conducting $T = 5$ Gauss-Newton updates at training time and $T = 10$ at test time. The damping factor β is initially set to 0.1. It is divided by a factor of 3.5 if the magnitude of the residual vector decreases, conversely, it is multiplied by a factor of 1.5 if the magnitude of the residual vector increases.

10. Training Details

Two-Stage Training. Our end-to-end pipeline is trained in two stages. The first stage uses the loss term on the matching result $\mathcal{L}_{\text{match}}$. The second stage additionally applies the pose loss $\mathcal{L}_{\text{pose}}$. Stage 1 is trained until the validation match loss converges, stage 2 until the validation pose loss converges. On ScanNet/ Matterport3D/ MegaDepth the training takes 32/ 343/ 143 epochs for stage 1 and 40/ 365/ 126 epochs for stage 2. We found that the training on Matterport3D and MegaDepth benefits from initializing the network weights to the weights after the first training stage on ScanNet, where most data is available. During stage 2 we linearly increase the weight of $\mathcal{L}_{\text{pose}}$ from 0 to 242/ 585/ 345 on ScanNet/ Matterport3D/

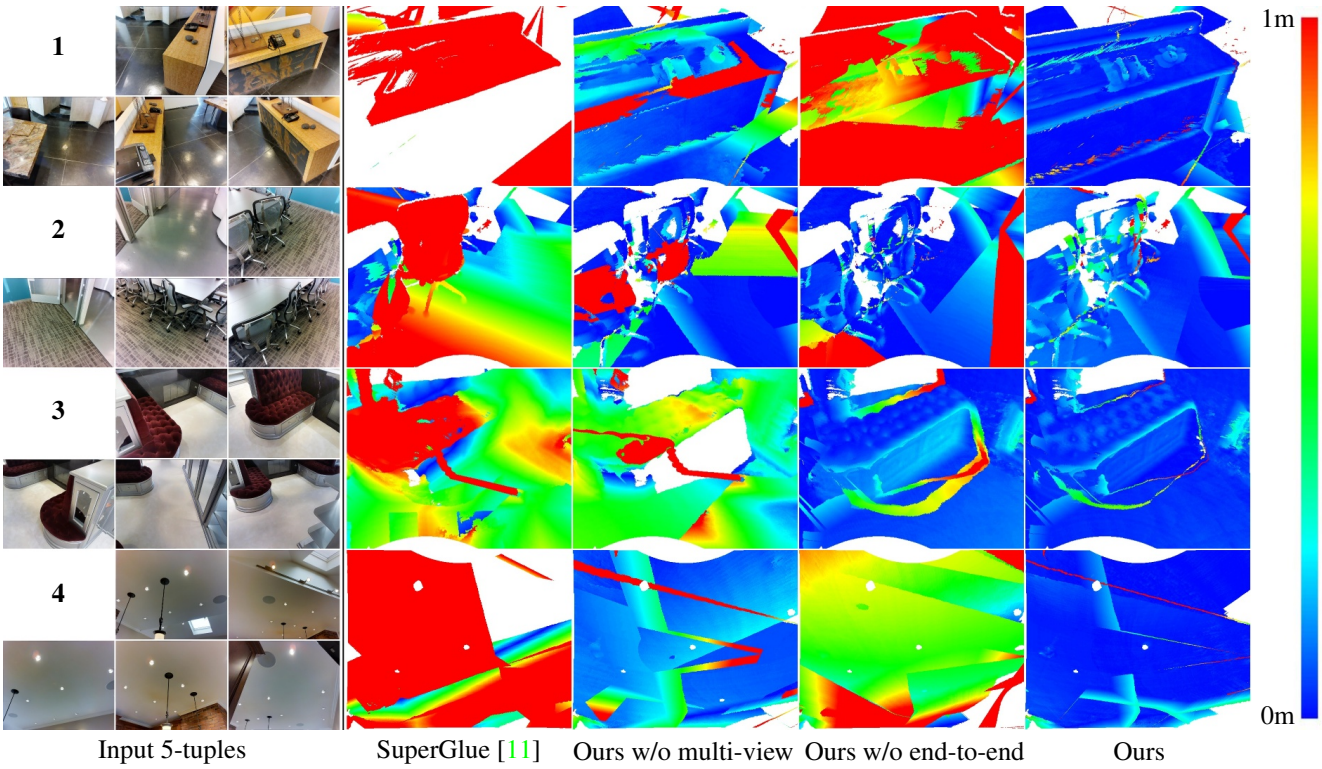

Figure 5. Reprojection error (right) for estimated camera poses on Matterport3D 5-tuples (left). Our complete method improves camera alignment over the ablated versions and SuperGlue, showing the importance of multi-view matching and end-to-end training.

MegaDepth, while linearly decreasing the weight of $\mathcal{L}_{\text{match}}$ from 1 to 0.01, over a course of 40000 iterations. The balancing factor of the rotation term λ_{rot} is set to 3.0/ 1.2/ 2.0 on ScanNet/ Matterport3D/ MegaDepth. We use the Adam optimizer [7] with learning rate 0.0001. The learning rate is exponentially decayed with a factor of 0.999992 starting after 100k iterations.

Ground Truth Generation. The ground truth matches $\mathcal{T}_{ab}$ and sets of unmatched keypoints $\mathcal{U}_{ab}$, $\mathcal{V}_{ab}$ of an image pair are computed by projecting the detected keypoints from each image to the other, resulting in a reprojection error matrix. Keypoint pairs where the reprojection error is both minimal and smaller than 5 pixels in both directions are considered matches. Unmatched keypoints must have a minimum reprojection error greater than 15 pixels on the indoor datasets and greater than 10 pixels on MegaDepth.

Input Data. We train the multi-view model on 5-tuples, which are sampled based on overlap ranges. On ScanNet and Matterport3D, overlap is computed using the ground truth poses, depth maps and intrinsic parameters. Following prior work [11, 12, 10], an overlap range of $[0.4, 0.8]$ is used on ScanNet. On Matterport3D, where view capture is much more sparse, we relax the overlap criterion to $[0.25, 0.8]$. On MegaDepth, the overlap between images is the portion of co-visible 3D points of the sparse reconstruction [11, 4], thus the overlap definition is different from the indoor datasets and not comparable. Overlap ranges $[0.1, 0.7]$ and $[0.1, 0.4]$ are used at train and test time, respectively [11]. The network is trained with a batch size of 24 on indoor data and with a batch size of 4 on outdoor data. The image size is 480×640 on ScanNet, 512×640 on Matterport3D and 640×640 on MegaDepth. The SuperPoint network is configured to detect keypoints with a non-maximum suppression radius of 4/ 3 on indoor/ outdoor data. On the indoor datasets we use 400 keypoints per image during training time: first, keypoints above a confidence threshold of 0.001 are sampled, second, if there are fewer than 400, the remainder is filled with random image points and confidence 0 as a data augmentation. On MegaDepth the same procedure is applied to sample 1024 keypoints using a confidence threshold of 0.005. At test time on indoor/ outdoor data, we use up to 1024/ 2048 keypoints above the mentioned confidence thresholds.

Dataset Split. On ScanNet and Matterport3D, we use the official dataset split. On MegaDepth, we follow the data split of prior work [12, 14, 10] using scenes 0015 and 0022 for validation, scenes 0008, 0019, 0021, 0024, 0025, 0032, 0063 and 1589 for testing and the remaining scenes

for training. Scenes with low quality depth maps are filtered out [14, 12, 5, 10]. This way, on ScanNet/ Matterport3D/ MegaDepth we have 240k/ 20k/ 15k 5-tuples for training, 62k/ 2200/ 200 for validation and 1500/ 1500/ 1500 for testing.

11. Baseline Comparison Details

In the baseline comparison, we use the network weights provided by the authors of SuperGlue [11], LoFTR [12], COTR [5] and 3DG-STFM [10]. There are SuperGlue, LoFTR and 3DG-STFM models trained on ScanNet and on MegaDepth, as well as a COTR model trained on MegaDepth. We additionally train a SuperGlue model on Matterport3D and a SuperGlue model on MegaDepth using the above described dataset split, which is necessary as the provided model was trained on a train set that contains our test set, as well as the Image Matching Challenge scenes. For the baselines, SuperGlue, LoFTR, and 3DG-STFM, we use their default confidence thresholds—0.2 for all three—and verify that they benefit from this threshold. We found that our method predicts accurate confidences that do not require thresholding for weighted pose estimation. When using RANSAC for two-view pose estimation, we filter matches from our model w/o multi-view using a threshold of 0.02.

In the multi-view evaluation we found that all methods benefit from a confidence-weighted bundle adjustment formulation on the inlier matches using Ceres solver (step (iv) in Section 4.2). Following [9], we conduct the Image Matching Challenge (IMC) [6] multi-view evaluation on the scenes Reichstag, Sacre Coeur and St. Peter's Square. The above described MegaDepth dataset split ensures that these scenes do not overlap with the training set. Since the IMC protocol does not consider matches in a confidence-weighted manner, we apply a threshold of 0.06 on matches from our multi-view model.

Following [11], matches are considered correct if the symmetric epipolar distance is smaller than $5 \cdot 10^{-4}$ or $1 \cdot 10^{-4}$ in the indoor and outdoor setting, respectively.

References

[1] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *3DV*, 2017. 1

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *CVPR*, 2017. 1

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–33712, 2018. 3

[4] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8084–8093, 2019. 5

[5] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence Transformer for Matching Across Images. In *ICCV*, 2021. 1, 3, 4, 6

[6] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 6

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2015. 5

[8] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 1

[9] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In *ICCV*, 2021. 6

[10] Runyu Mao, Chen Bai, Yatong An, Fengqing Zhu, and Cheng Lu. 3dg-stfm: 3d geometric guided student-teacher feature matching. *ECCV*, 2022. 1, 3, 4, 5, 6

[11] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4937–4946, 2020. 1, 2, 3, 4, 5, 6

[12] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8918–8927, 2021. 1, 2, 3, 4, 5, 6

[13] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, 1 2016. 1

[14] Michal J. Tyszkiewicz, P. Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 2020. 5, 6