Multi-View Optimization of Local Feature Geometry Supplementary Material

Mihai Dusmanu[1], Johannes L. Schönberger[2], and Marc Pollefeys[1,2]

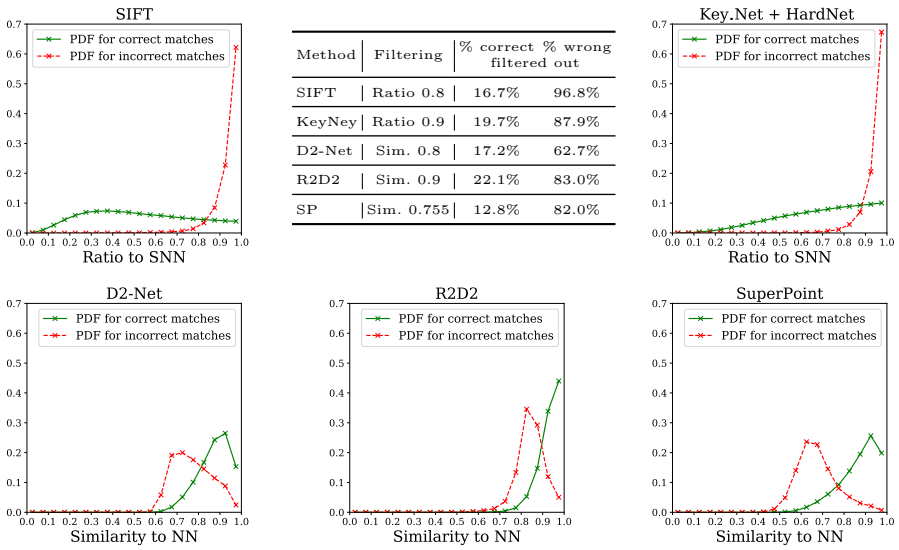
[1] Department of Computer Science, ETH Zürich [2] Microsoft

This supplementary material provides the following information: Section 1 explains how we determined the match filtering thresholds for the learned methods. Section 2 contains the additional results mentioned in the main paper (*e.g.*, ETH3D [11] triangulation results on each individual dataset and independent results of each method on the Local Feature Evaluation Benchmark [10]) as well as some qualitative examples before and after refinement. Section 3 presents an ablation study for both the two-view and the multi-view refinement procedure. Section 4 details the query keypoint refinement protocol used for camera localization on the ETH3D dataset. Section 5 describes the filtering steps used during the generation of the two-view training dataset.

1 Match filtering

Match filtering is an essential step before large-scale SfM because it significantly reduces the number of wrong registrations due to repetitive structures and semantically similar scenes. To determine a good threshold (either for similarity or ratio to the second nearest neighbor), we adopt the methodology suggested by Lowe [8] – we plot the probability distribution functions for correct and incorrect mutual nearest neighbors matches on the sequences from the HPatches dataset [1]. A match is considered correct if its projection error, estimated using the ground-truth homographies, is below 4 pixels. To have a clear separation, the threshold for incorrect matches is set to 12 pixels. All matches with errors in-between are discarded. Figure 1 shows the plots for all learned methods as well as SIFT (used as reference).

For SIFT [8], the ratio threshold traditionally used (0.8) filters out 16.7% of correct matches and 96.8% of wrong ones. For SuperPoint [4], we use the cosine similarity threshold suggested by the authors (0.755) which filters out 82.0% of wrong matches. For Key.Net [2] and R2D2 [9], we empirically determine thresholds with a similar filtering performance to the ones used for SIFT and SuperPoint. The only method that is not compatible with either the ratio test or similarity thresholding is D2-Net [5]. Thus, for it, we settle on a conservative similarity threshold of 0.8, filtering out only 62.7% of incorrect matches.



Method	Filtering	% correct	% wrong filtered out
SIFT	Ratio 0.8	16.7%	96.8%
KeyNey	Ratio 0.9	19.7%	87.9%
D2-Net	Sim. 0.8	17.2%	62.7%
R2D2	Sim. 0.9	22.1%	83.0%
SP	Sim. 0.755	12.8%	82.0%

Fig. 1: **Match filtering.** Following the protocol of Lowe [8], we plot the probability distribution function (PDF) of the correct and incorrect mutual nearest neighbors matches. The horizontal axis represents either the ratio to the second nearest neighbor or the cosien similarity to the first nearest neighbor.

2 Additional results

For the Local Feature Evaluation Benchmark [10], the results reported in the main paper show the sparse 3D reconstruction statistics on the images registered by both the refined and unrefined versions of each feature - this was done in order to allow a fair comparison in terms of number of observations, track length, and reprojection error. Nevertheless, we also provide the independent results for each local feature in Table 1.

Due to space constraints, in the main paper, we only reported the average results on indoor and outdoor scenes for the ETH3D triangulation evaluation [11]. Tables 3 and 4 show the results for each of the 13 datasets. For the learned features, the results with refinement are always better. For SIFT, the only scene where the results after refinement are worse is Meadow; this is a textureless scene where SIFT has troubles correctly matching features. Due to the low number of matches passed to COLMAP, its triangulation results are very sensitive to small changes in the input. Some qualitative examples are shown in Figures 2 and 3. A short video with additional examples is available at `https://www.youtube.com/watch?v=eH4UNwXLsyk`.

3 Ablation study

In this section, an ablation study for the proposed refinement procedure will be presented. We will first start by studying the effect of training data on the two-

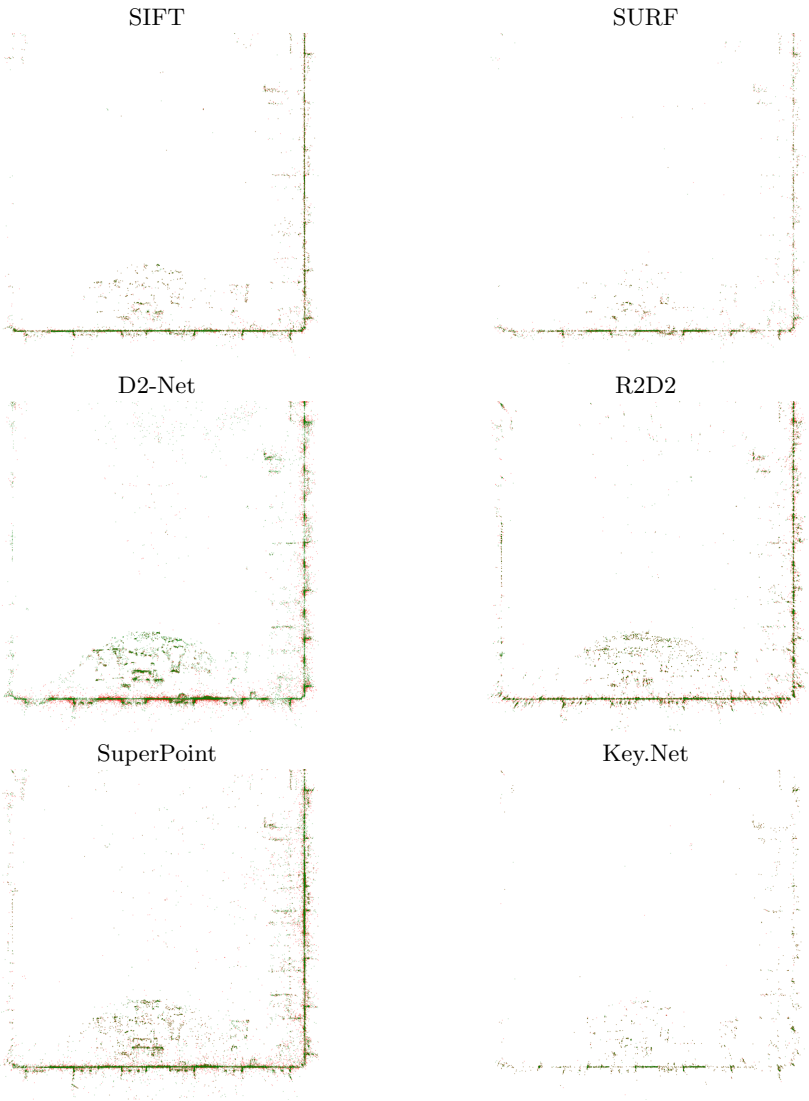


Fig. 2: **Courtyard.** We show top-down partial views of point clouds triangulated on the Courtyard scene. We overlap the point-cloud obtained from refined keypoints and the point-cloud from raw keypoints. The noise levels are drastically reduced nearby planar surfaces. Best viewed on a monitor.

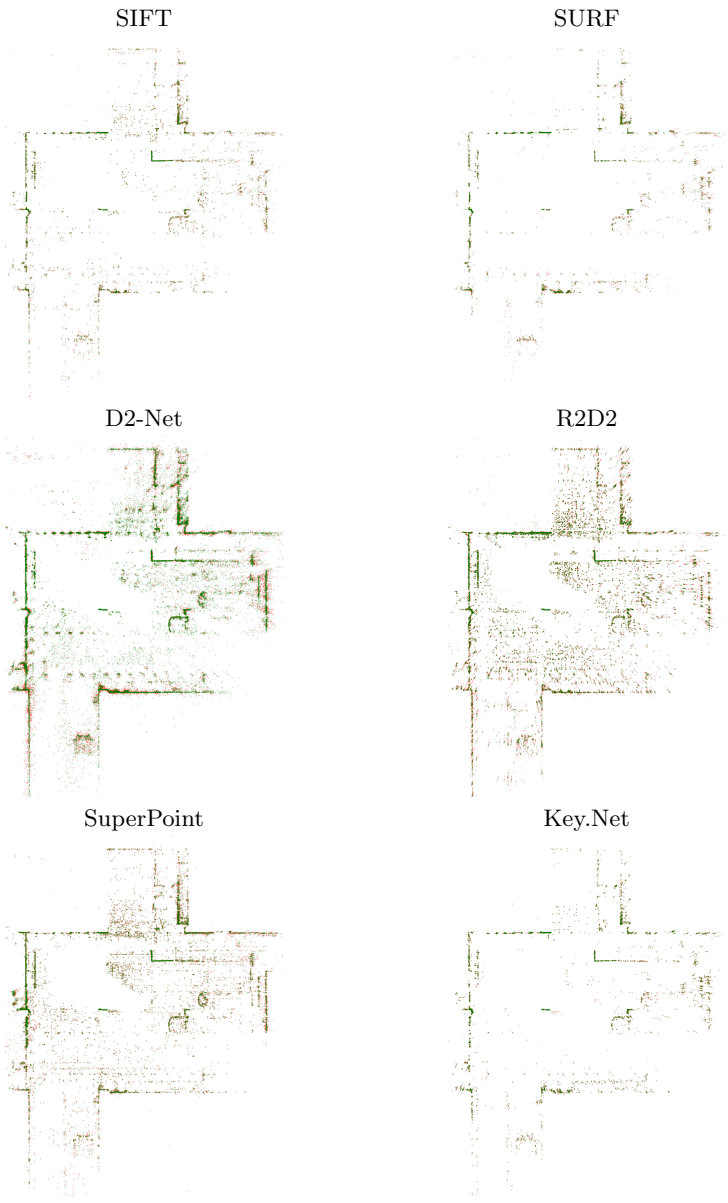


Fig. 3: **Delivery Area.** We show top-down partial views of point clouds triangulated on the Courtyard scene. We overlap the point-cloud obtained from refined keypoints and the point-cloud from raw keypoints. The noise levels are drastically reduced nearby planar surfaces. Best viewed on a monitor.

Table 1: **Evaluation on the Local Feature Evaluation Benchmark.** We report the results for each method independently, instead of considering only the commonly registered images for refined and unrefined features.

Dataset	Method	Reg. images	Num. obs.	Track length	Reproj. error	Method	Reg. images	Num. obs.	Track length	Reproj. error
Madrid Metropolis 1344 images	SIFT	**393**	188.7K	6.84	0.70	SURF	296	**121.4K**	6.22	0.76
	SIFT + ref.	390	**189.7K**	**6.90**	**0.66**	SURF + ref.	**274**	116.6K	**6.26**	**0.66**
	D2-Net	392	683.6K	6.01	1.46	R2D2	422	357.2K	**10.17**	0.90
	D2-Net + ref.	**405**	**773.4K**	**7.26**	**0.96**	R2D2 + ref.	**427**	**359.5K**	10.15	**0.76**
	SP	422	272.1K	7.64	0.98	Key.Net	317	114.4K	9.28	0.94
	SP + ref.	**425**	**279.9K**	**8.23**	**0.72**	Key.Net + ref.	**323**	**119.4K**	**9.39**	**0.75**
Gendarmen-markt 1463 images	SIFT	879	440.7K	6.34	0.82	SURF	475	164.1K	**5.45**	0.90
	SIFT + ref.	**882**	**442.2K**	**6.41**	**0.75**	SURF + ref.	**483**	**165.6K**	5.42	**0.78**
	D2-Net	865	1.482M	5.33	1.44	R2D2	**988**	**1.102M**	9.94	0.98
	D2-Net + ref.	**959**	**1.805M**	**6.38**	**1.02**	R2D2 + ref.	935	1.044M	**10.04**	**0.89**
	SP	919	627.4K	6.84	1.05	Key.Net	817	253.9K	7.08	0.99
	SP + ref.	**972**	**680.6K**	**7.07**	**0.88**	Key.Net + ref.	**828**	**260.5K**	**7.21**	**0.86**
Tower of London 1576 images	SIFT	562	448.9K	7.90	0.69	SURF	**433**	212.2K	**5.94**	0.71
	SIFT + ref.	**566**	**449.6K**	**7.96**	**0.59**	SURF + ref.	432	**212.9K**	5.92	**0.58**
	D2-Net	653	1.417M	5.93	1.48	R2D2	693	758.2K	13.44	0.92
	D2-Net + ref.	**661**	**1.568M**	**7.64**	**0.91**	R2D2 + ref.	**700**	**760.8K**	**13.73**	**0.76**
	SP	625	443.3K	8.06	0.95	Key.Net	**500**	186.9K	9.03	0.85
	SP + ref.	**633**	**458.9K**	**8.52**	**0.69**	Key.Net + ref.	495	**190.8K**	**9.18**	**0.66**

view refinement network. Secondly, we will study how each step of the multi-view refinement influences the final result.

3.1 Two-view refinement

The architecture used for the two-view refinement between tentative matches is described in Table 4. For the layers with batch normalization, we place it before the non-linearity (*i.e.*, the order is convolution followed by batch normalization and finally non-linearity) as suggested in the reference paper [6].

For this ablation study, we focus on the HPatches Sequences dataset [1], because it allows to isolate the network output. Given a tentative match u, v, we run a forward pass of the patch alignment network to predict $d_{u\to v}$ and use u and $v + d_{u\to v}$ as keypoint locations. As can be seen in Figure 4, training only with synthetic data (*i.e.*, pairs consisting of a patch and a warped version of itself) is not sufficient to achieve the final performance. By using real pairs extracted from the MegaDepth dataset [7], we allow the network to learn different illumination conditions as well as occlusions / large viewpoint changes.

3.2 Multi-view refinement

We use the largest dataset with ground-truth data available (Facade from ETH3D [11]) to study the relevance of the following steps of our pipeline: graph partitioning, inter-edges, 3×3 displacement grid. The ablation results are summarized in Table 2. For the purpose of this section, we define the set of intra-edges connecting nodes within a track as $E_{\text{intra}} = \{(u \to v) \in E | t_u = t_v\}$ and the set of inter-edges connecting nodes of different tracks as $E_{\text{inter}} = \{(u \to v) \in E | t_u \neq t_v\}$ on the entire graph G (without graph-cut).

Layer	Batch Norm.	ReLU	Output shape
input, RGB			33 × 33 × 3
conv1_1, 3 × 3		✓	33 × 33 × 64
conv1_2, 3 × 3		✓	33 × 33 × 64
max_pool1, 3 × 3, stride 2			17 × 17 × 64
conv2_1, 3 × 3		✓	17 × 17 × 128
conv2_2, 3 × 3		✓	17 × 17 × 128
correlation			17 × 17 × 289
reg_conv1, 5 × 5	✓	✓	13 × 13 × 128
reg_conv2, 5 × 5	✓	✓	9 × 9 × 128
reg_conv3, 5 × 5	✓	✓	5 × 5 × 64
reg_conv4, 5 × 5	✓	✓	1 × 1 × 64
reg_fc			2

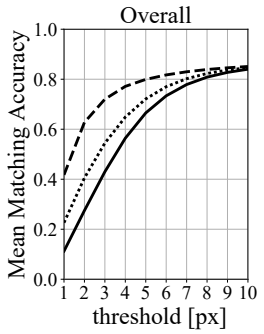


Fig. 4: **Two-view refinement.** *Left – architecture:* We use a slightly modified version of VGG16 up to `conv2_2` for feature extraction. The results of the dense matching are processed by a sequence of convolutional and fully connected layers. *Right – ablation:* The results for D2-Net without refinement are reported by a solid line. We compare a network trained on synthetic pairs only (dotted) and one trained with both synthetic and real data (dashed).

Without any graph partitioning, the optimization can be formulated as:

$$\begin{aligned} &\min_{x_p} \sum_{(u \to v) \in E} s_{u \to v} \rho(\|\bar{x}_v - \bar{x}_u - T_{u \to v}(\bar{x}_u)\|^2) \\ &\text{s.t.} \|\bar{x}_p\|_1 = \|x_p - x_p^0\|_1 \leq K, \forall p \ . \end{aligned} \tag{1}$$

Despite the long track length, the reprojection error is generally larger and the point clouds are less accurate - this is mainly due to wrong tentative matches. Moreover, this formulation has one of the highest optimization runtimes.

After partitioning the graph into tracks, one could ignore the inter-edges:

$$\begin{aligned} &\min_{x_p} \sum_{(u \to v) \in E_{\text{intra}}} s_{u \to v} \rho(\|\bar{x}_v - \bar{x}_u - T_{u \to v}(\bar{x}_u)\|^2) \\ &\text{s.t.} \|\bar{x}_p\|_1 = \|x_p - x_p^0\|_1 \leq K, \forall p \ . \end{aligned} \tag{2}$$

This formulation can be solved independently for each track and is thus the fastest. However, detectors often fire multiple times for the same visual feature. Since we restrict the tracks to only contain one feature from each image, these multiple detections will never be merged into a single point.

To address this, the inter-edges must be considered:

$$\begin{aligned} \min_{x_p} &\sum_{(u \to v) \in E_{\text{intra}}} s_{u \to v} \rho(\|\bar{x}_v - \bar{x}_u - T_{u \to v}(\bar{x}_u)\|^2) + \\ &\sum_{(u \to v) \in E_{\text{inter}}} s_{u \to v} \psi(\|\bar{x}_v - \bar{x}_u - T_{u \to v}(\bar{x}_u)\|^2) \\ \text{s.t.} &\|\bar{x}_p\|_1 = \|x_p - x_p^0\|_1 \leq K, \forall p \ . \end{aligned} \tag{3}$$

The main issue with this formulation is its runtime due to having the same number of residuals as Equation 1. However, it generally achieves similar accuracy and reprojection error to Equation 2 while having a better track length.

Table 2: **Multi-view refinement ablation study.** Reconstruction statistics are reported for the Facade scene of ETH3D [11] consisting of 76 images for different formulations of the multi-view optimization problem.

Method		Comp. (%)			Accuracy (%)			Track length	Reproj. error	Optim. runtime
		1cm	2cm	5cm	1cm	2cm	5cm			
SIFT	no refinement	0.06	0.36	3.08	36.04	52.10	73.28	5.42	1.07	
	no graph partitioning	0.09	0.50	3.85	44.52	62.26	82.74	5.86	0.81	49.7s
	intra-edges	0.09	0.51	3.84	45.16	62.56	82.44	5.80	0.80	10.2s
	+ inter-edges	0.09	0.50	3.83	45.46	62.58	82.65	5.82	0.80	54.1s
	+ graph-cut *(full)*	0.09	0.50	3.82	45.19	62.32	82.38	5.81	0.80	13.3s
	full (constant flow)	0.09	0.49	3.72	44.12	61.33	80.65	5.75	0.84	11.9s
D2-Net	no refinement	0.02	0.18	2.26	7.56	14.21	29.90	3.20	1.60	
	no graph partitioning	0.11	0.71	5.50	28.20	43.64	67.50	5.64	1.09	317.7s
	intra-edges	0.16	1.01	8.17	34.85	53.05	75.80	5.05	0.85	20.0s
	+ inter-edges	0.16	1.01	8.16	34.88	53.18	75.90	5.06	0.85	223.6s
	+ graph-cut *(full)*	0.16	1.01	8.18	34.86	53.32	76.02	5.06	0.85	31.1s
	full (constant flow)	0.13	0.87	7.53	29.37	46.05	69.51	4.78	0.99	28.1s
SuperPoint	no refinement	0.07	0.49	4.95	18.82	32.21	54.72	4.21	1.54	
	no graph partitioning	0.09	0.62	5.54	25.77	41.67	66.16	4.95	1.28	246.7s
	intra-edges	0.14	0.90	7.21	35.16	53.12	74.72	5.23	0.94	32.4s
	+ inter-edges	0.14	0.89	7.21	35.73	53.36	75.32	5.31	0.93	255.8s
	+ graph-cut *(full)*	0.14	0.90	7.23	35.73	53.49	75.48	5.25	0.94	47.6s
	full (constant flow)	0.12	0.78	6.66	30.41	47.57	69.90	5.12	1.06	43.2s

By using recursive graph cut to split the connected components into smaller sets and solving on each remaining component independently, we strike a balance between the performance of Equation 3 and the efficiency of Equation 2.

While the constant flow assumption also improves the performance of local features, it is not sufficient to explain all structures. The 3×3 deformation grid is better suited and achieves a superior performance across the board.

The runtime of the proposed graph optimization procedure is shown in Figure 5 for all datasets of our evaluation. The top-right points correspond to the internet reconstruction from the Local Feature Evaluation Benchmark [10] (Madrid Metropolis, Gendarmenmarkt, Tower of London). For these datasets, the runtime remains low ($1 - 5$ minutes depending on the method) compared to the runtime of the sparse 3D reconstruction ($15 - 30$ minutes).

4 Query refinement

For the localization experiments, we used the tentative matches $\{u_1, u_2, \dots\}$ of each query feature q to refine its location. First, all matches corresponding to non-triangulated features are discarded since they cannot be used for PnP. For each remaining match $u_i \leftrightarrow q$, let π_i be the 3D point associated to u_i and $\hat{u}_i$ be the reprojection of π_i to the image of u_i.

Since these matches are purely based on appearance, the points u_i might correspond to different 3D points of the partial model. Each matching 3D location Π is considered as an independent hypothesis. Given that the reprojected locations are fixed, the optimization problem can be simplified by considering only one-

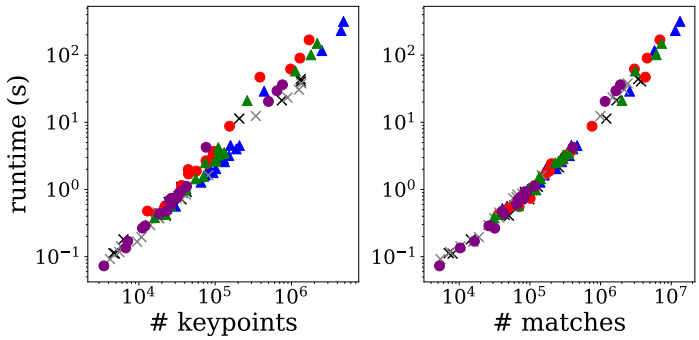


Fig. 5: **Graph optimization runtime.** The runtime is plotted as a function of the number of keypoints and matches. We respect the color coding from the main paper: SIFT [8], SURF [3], D2-Net [5], R2D2 [9], SuperPoint [4], and Key.Net [2].

directional $u_i \rightarrow q$ edges:

$$\begin{aligned} &\min_{x_q} \sum_{u_i \text{ s.t. } \pi_i = \Pi} s_{u_i \rightarrow q} \rho(\|\bar{x}_q - d_{\hat{u}_i \rightarrow q}\|^2) \\ &\text{s.t.} \|\bar{x}_q\|_1 = \|x_q - x_q^0\|_1 \leq K \quad . \end{aligned} \tag{4}$$

The central point flow in the above formulation is considered from the reprojected feature $\hat{u}_i$ in a view of the partial model to the query feature q.

After removing the robustifier and supposing that the two-view displacements are always smaller than K, the problem can be rewritten as follows:

$$\min_{x_q} \sum_{u_i \text{ s.t. } \pi_i = \Pi} s_{u_i \rightarrow q} \|\bar{x}_q - d_{\hat{u}_i \rightarrow q}\|^2 \quad . \tag{5}$$

This formulation has a closed-form solution:

$$x_q^{\Pi} = x_q^0 + \frac{\sum\limits_{u_i \text{ s.t. } \pi_i = \Pi} s_{u_i \rightarrow q} d_{\hat{u}_i \rightarrow q}}{\sum\limits_{u_i \text{ s.t. } \pi_i = \Pi} s_{u_i \rightarrow q}} \quad . \tag{6}$$

Thus, for each query feature q with triangulated tentative matches, we obtain one or more refined 2D-3D correspondences (x_q^{Π}, Π) which can be used for pose estimation.

5 Training dataset

As mentioned in the main paper, several steps were taken to improve the quality of the training data extracted from MegaDepth [7].

Scene filtering. We discarded 16 scenes due to inconsistencies between sparse and dense reconstructions. This was done automatically using the following heuristic: from each scene, 100000 random pairs of matching 2D observations

part of the 3D model were selected; for each such pair (k_1, k_2), the Multi-View Stereo (MVS) depth was used to warp k_1 to the other image obtaining $\hat{k}_2$; a keypoint is inconsistent if its reprojection $\hat{k}_2$ is more than 12 pixels away from its feature position k_2, *i.e.*, $|\hat{k}_2 - k_2| > 12$. The following scenes were removed for having a low number of consistent points: `0000`, `0002`, `0011`, `0020`, `0033`, `0050`, `0103`, `0105`, `0143`, `0176`, `0177`, `0265`, `0366`, `0474`, `0860`, `4541`.

Depth consistency. We enforce depth consistency to make sure that the central pixel is not occluded. The MVS depth D_1 of a source image is used to back-project a keypoint k_1 to 3D and obtain p. We then reproject this 3D point to the target image to obtain $\hat{k}_2$ and depth d. The depth consistency verifies that the MVS depth from the second image D_2 is consistent with the 3D point p, *i.e.*, $|D_2(\hat{k}_2) - d| < 10^{-2}$.

Table 3: **ETH3D triangulation evaluation - Indoors.** We report triangulation statistics on each indoor dataset for methods with and without refinement.

Dataset	Method	Comp. (%) 1cm	2cm	5cm	Accuracy (%) 1cm	2cm	5cm	Method	Comp. (%) 1cm	2cm	5cm	Accuracy (%) 1cm	2cm	5cm
Deliv. Area 44 images	SIFT	0.06	0.34	2.29	61.59	74.40	86.98	SURF	0.03	0.20	1.35	53.91	70.15	83.18
	SIFT + ref.	**0.09**	**0.44**	**2.66**	**71.65**	**82.47**	**91.64**	SURF + ref.	**0.06**	**0.30**	**1.76**	**68.67**	**80.26**	**89.72**
	D2-Net	0.08	0.53	3.53	30.99	47.16	67.35	R2D2	0.17	0.86	5.26	52.09	66.80	82.29
	D2-Net + ref.	**0.40**	**1.93**	**9.87**	**65.00**	**77.26**	**88.51**	R2D2 + ref.	**0.27**	**1.11**	**5.81**	**70.57**	**81.63**	**91.29**
	S	0.15	0.80	5.36	56.80	71.42	85.10	Key.Net	0.05	0.28	1.78	52.08	69.11	85.33
	SP + ref.	**0.22**	**1.07**	**6.36**	**74.39**	**85.36**	**93.89**	Key.Net + ref.	**0.09**	**0.38**	**2.15**	**74.01**	**84.16**	**92.56**
Kicker 31 images	SIFT	0.27	1.29	5.64	71.78	82.69	91.63	SURF	0.22	1.08	4.78	65.20	77.94	90.31
	SIFT + ref.	**0.33**	**1.44**	**5.92**	**77.32**	**86.61**	**93.90**	SURF + ref.	**0.31**	**1.34**	**5.31**	**77.43**	**86.12**	**93.82**
	D2-Net	0.20	1.16	6.18	38.41	56.54	75.83	R2D2	0.46	1.87	8.41	68.08	80.30	89.91
	D2-Net + ref.	**0.87**	**3.51**	**11.20**	**69.53**	**79.72**	**88.16**	R2D2 + ref.	**0.56**	**2.12**	**8.89**	**75.12**	**84.28**	**91.48**
	SP	0.44	2.08	9.24	67.88	79.43	89.01	Key.Net	0.18	0.84	4.28	62.94	79.51	90.44
	SP + ref.	**0.57**	**2.46**	**10.05**	**79.23**	**87.04**	**92.01**	Key.Net + ref.	**0.25**	**1.07**	**4.85**	**72.73**	**84.31**	**92.92**
Office 26 images	SIFT	0.11	0.53	**2.72**	75.48	84.81	93.30	SURF	0.06	0.26	1.36	70.50	**86.47**	95.17
	SIFT + ref.	**0.12**	**0.55**	2.64	**77.27**	**86.98**	**94.69**	SURF + ref.	**0.07**	**0.34**	**1.58**	**70.72**	85.57	**96.01**
	D2-Net	0.12	0.76	3.76	38.78	57.62	82.58	R2D2	0.33	1.45	6.02	54.97	70.53	87.52
	D2-Net + ref.	**0.54**	**2.08**	**6.21**	**65.46**	**79.30**	**91.47**	R2D2 + ref.	**0.42**	**1.66**	**6.53**	**61.07**	**75.67**	**89.38**
	SP	0.27	1.19	5.27	75.36	85.47	95.46	Key.Net	0.11	0.53	2.73	63.14	77.22	90.43
	SP + ref.	**0.34**	**1.37**	**5.46**	**84.05**	**91.69**	**96.96**	Key.Net + ref.	**0.17**	**0.71**	**3.19**	**80.63**	**90.29**	**95.87**
Pipes 14 images	SIFT	0.06	0.27	1.11	73.23	80.52	87.53	SURF	0.02	0.10	0.52	66.90	74.19	90.30
	SIFT + ref.	**0.08**	**0.34**	**1.50**	**80.66**	**86.61**	**93.52**	SURF + ref.	**0.03**	**0.14**	**0.64**	**77.65**	**84.04**	**91.84**
	D2-Net	0.14	0.76	3.53	54.80	76.15	91.93	R2D2	0.22	0.97	4.83	68.50	79.42	87.92
	D2-Net + ref.	**0.59**	**2.08**	**5.69**	**87.10**	**93.23**	**97.50**	R2D2 + ref.	**0.31**	**1.19**	**5.21**	**75.22**	**82.71**	**88.68**
	SP	0.41	1.77	7.30	85.31	90.70	**96.23**	Key.Net	0.05	0.24	1.27	76.85	87.68	93.31
	SP + ref.	**0.55**	**2.17**	**8.25**	**91.15**	**94.15**	96.07	Key.Net + ref.	**0.07**	**0.30**	**1.55**	**82.15**	**92.89**	**95.36**
Relief 31 images	SIFT	0.30	1.35	5.19	81.88	91.02	96.29	SURF	0.09	0.46	2.20	73.72	86.39	94.79
	SIFT + ref.	**0.35**	**1.46**	**5.43**	**86.59**	**92.80**	**96.61**	SURF + ref.	**0.13**	**0.55**	**2.40**	**83.11**	**89.60**	**95.07**
	D2-Net	0.45	2.51	9.29	46.72	67.65	88.16	R2D2	0.52	2.16	9.86	71.12	85.64	95.50
	D2-Net + ref.	**1.82**	**6.45**	**16.58**	**87.71**	**92.03**	**95.33**	R2D2 + ref.	**0.70**	**2.48**	**10.45**	**87.07**	**93.13**	**96.89**
	SP	0.49	2.25	9.17	77.73	88.05	95.52	Key.Net	0.14	0.66	3.23	65.82	80.47	92.78
	SP + ref.	**0.60**	**2.49**	**9.75**	**91.01**	**94.82**	**97.07**	Key.Net + ref.	**0.18**	**0.74**	**3.41**	**83.26**	**89.70**	**94.99**
Relief 2 31 images	SIFT	0.16	0.80	3.74	76.67	86.48	93.35	SURF	0.05	0.29	1.41	64.15	82.25	93.02
	SIFT + ref.	**0.20**	**0.89**	**4.00**	**83.77**	**91.19**	**95.64**	SURF + ref.	**0.08**	**0.37**	**1.62**	**80.24**	**89.38**	**94.64**
	D2-Net	0.25	1.48	7.63	46.03	64.57	84.66	R2D2	0.49	2.10	10.16	74.70	86.28	94.43
	D2-Net + ref.	**1.36**	**5.24**	**16.12**	**86.58**	**91.56**	**95.01**	R2D2 + ref.	**0.67**	**2.47**	**10.84**	**88.42**	**93.04**	**96.73**
	SP	0.32	1.58	7.80	77.21	88.20	94.85	Key.Net	0.11	0.58	3.00	59.26	76.71	93.30
	SP + ref.	**0.41**	**1.83**	**8.42**	**89.62**	**94.49**	**97.05**	Key.Net + ref.	**0.16**	**0.70**	**3.32**	**79.91**	**90.00**	**95.35**
Terrains 42 images	SIFT	0.44	1.46	4.60	89.51	93.47	**96.76**	SURF	0.11	0.46	2.14	70.22	75.98	80.49
	SIFT + ref.	**0.50**	**1.60**	**5.01**	**90.14**	**93.81**	96.29	SURF + ref.	**0.15**	**0.58**	**2.55**	**76.13**	**82.15**	**85.45**
	D2-Net	1.99	5.59	15.11	72.96	84.66	92.26	R2D2	1.49	4.87	15.15	77.45	85.81	92.74
	D2-Net + ref.	**4.51**	**10.40**	**25.10**	**88.34**	**92.13**	**95.37**	R2D2 + ref.	**1.71**	**5.21**	**15.81**	**85.44**	**89.72**	**93.33**
	SP	2.07	5.82	17.89	86.51	93.60	96.93	Key.Net	0.51	1.64	4.77	85.47	92.35	95.69
	SP + ref.	**2.30**	**6.20**	**18.58**	**92.77**	**95.80**	**97.74**	Key.Net + ref.	**0.58**	**1.78**	**5.05**	**90.91**	**93.30**	**96.09**

Table 4: **ETH3D triangulation evaluation - Outdoors.** We report triangulation statistics on each outdoor dataset for methods with and without refinement.

Dataset	Method	Comp. (%) 1cm	2cm	5cm	Accuracy (%) 1cm	2cm	5cm	Method	Comp. (%) 1cm	2cm	5cm	Accuracy (%) 1cm	2cm	5cm
Courtyard 38 images	SIFT	0.08	0.47	3.72	67.94	81.80	92.04	SURF	0.06	0.31	1.88	66.40	80.04	89.58
	SIFT + ref.	**0.10**	**0.56**	**4.03**	**75.17**	**86.01**	**94.00**	SURF + ref.	**0.08**	**0.41**	**2.25**	**79.96**	**87.53**	**94.03**
	D2-Net	0.03	0.24	2.07	22.63	38.53	61.33	R2D2	0.07	0.37	2.73	45.72	62.08	79.61
	D2-Net + ref.	**0.21**	**1.14**	**5.98**	**66.78**	**79.04**	**89.40**	R2D2 + ref.	**0.10**	**0.52**	**3.33**	**63.91**	**78.18**	**90.29**
	SP	0.13	0.79	5.04	45.36	60.61	77.84	Key.Net	0.02	0.12	0.83	41.60	62.78	79.38
	SP + ref.	**0.21**	**1.12**	**6.68**	**63.98**	**77.69**	**88.95**	Key.Net + ref.	**0.03**	**0.16**	**0.99**	**63.54**	**77.83**	**89.96**
Electro 45 images	SIFT	0.03	0.15	0.94	63.76	78.46	88.84	SURF	0.01	0.07	0.48	47.54	65.22	81.48
	SIFT + ref.	**0.03**	**0.18**	**1.05**	**65.82**	**79.19**	**90.11**	SURF + ref.	**0.02**	**0.11**	**0.68**	**62.75**	**75.20**	**87.06**
	D2-Net	0.03	0.19	1.50	30.30	45.29	66.46	R2D2	0.12	0.57	3.66	57.32	73.33	87.98
	D2-Net + ref.	**0.19**	**0.95**	**4.99**	**68.36**	**79.57**	**89.56**	R2D2 + ref.	**0.17**	**0.72**	**4.00**	**70.96**	**82.32**	**91.46**
	SP	0.06	0.34	2.45	60.66	75.89	89.26	Key.Net	0.02	0.11	0.83	45.09	65.80	82.31
	SP + ref.	**0.09**	**0.44**	**2.77**	**76.96**	**87.29**	**93.75**	Key.Net + ref.	**0.03**	**0.17**	**1.01**	**65.93**	**81.83**	**91.56**
Facade 76 images	SIFT	0.06	0.36	3.08	36.04	52.10	73.28	SURF	0.05	0.36	3.18	25.17	41.25	63.75
	SIFT + ref.	**0.09**	**0.50**	**3.82**	**45.19**	**62.32**	**82.38**	SURF + ref.	**0.11**	**0.66**	**4.71**	**43.41**	**63.28**	**83.43**
	D2-Net	0.02	0.18	2.26	7.56	14.21	29.90	R2D2	0.05	0.28	2.17	25.07	40.83	64.42
	D2-Net + ref.	**0.16**	**1.01**	**8.18**	**34.86**	**53.32**	**76.02**	R2D2 + ref.	**0.08**	**0.42**	**2.91**	**37.34**	**56.66**	**78.81**
	SP	0.07	0.49	4.95	18.82	32.21	54.72	Key.Net	0.01	0.06	0.58	15.21	25.12	49.91
	SP + ref.	**0.14**	**0.90**	**7.23**	**35.73**	**53.49**	**75.48**	Key.Net + ref.	**0.01**	**0.08**	**0.74**	**29.77**	**43.53**	**71.33**
Meadow 15 images	SIFT	0.01	0.04	0.35	**60.25**	**78.01**	**89.47**	SURF	0.00	0.01	0.10	30.77	63.64	**84.62**
	SIFT + ref.	**0.01**	**0.05**	**0.40**	49.26	73.95	87.12	SURF + ref.	**0.00**	**0.02**	**0.13**	**55.56**	**65.31**	80.70
	D2-Net	0.00	0.03	0.35	21.89	34.05	57.35	R2D2	0.02	0.14	0.95	50.23	70.77	87.10
	D2-Net + ref.	**0.03**	**0.17**	**1.19**	**49.89**	**62.62**	**77.82**	R2D2 + ref.	**0.03**	**0.17**	**1.05**	**63.15**	**81.45**	**91.74**
	SP	0.02	0.12	1.06	51.05	68.91	**88.18**	Key.Net	0.00	0.01	0.06	46.67	56.25	64.71
	SP + ref.	**0.03**	**0.16**	**1.21**	**66.67**	**78.85**	88.02	Key.Net + ref.	**0.00**	**0.01**	**0.07**	**51.72**	**64.52**	**85.71**
Playground 38 images	SIFT	0.15	0.80	4.86	66.57	78.10	90.58	SURF	0.03	0.18	1.14	57.25	73.61	86.05
	SIFT + ref.	**0.18**	**0.91**	**5.27**	**70.70**	**81.76**	**91.73**	SURF + ref.	**0.06**	**0.27**	**1.57**	**74.60**	**83.76**	**92.70**
	D2-Net	0.05	0.31	2.42	28.01	46.88	69.61	R2D2	0.26	1.28	7.71	63.69	78.08	91.31
	D2-Net + ref.	**0.46**	**2.01**	**8.19**	**71.63**	**83.73**	**93.60**	R2D2 + ref.	**0.37**	**1.58**	**8.29**	**78.03**	**88.76**	**96.53**
	SP	0.19	0.97	5.63	59.09	72.42	86.01	Key.Net	0.03	0.15	1.26	45.61	59.18	80.10
	SP + ref.	**0.28**	**1.29**	**6.83**	**70.30**	**79.84**	**90.09**	Key.Net + ref.	**0.04**	**0.22**	**1.54**	**64.06**	**78.65**	**91.32**
Terrace 23 images	SIFT	0.04	0.20	1.66	55.32	70.28	83.23	SURF	0.01	0.06	0.56	38.13	54.91	72.80
	SIFT + ref.	**0.05**	**0.26**	**1.93**	**63.53**	**78.10**	**88.41**	SURF + ref.	**0.02**	**0.10**	**0.75**	**61.00**	**72.97**	**84.68**
	D2-Net	0.02	0.19	2.21	17.73	31.53	55.85	R2D2	0.12	0.64	4.46	50.46	69.33	86.43
	D2-Net + ref.	**0.22**	**1.24**	**8.26**	**62.92**	**75.78**	**87.34**	R2D2 + ref.	**0.19**	**0.84**	**4.90**	**69.73**	**81.24**	**91.69**
	SP	0.10	0.56	4.04	63.03	77.40	88.71	Key.Net	0.02	0.10	0.96	41.31	58.29	77.42
	SP + ref.	**0.14**	**0.72**	**4.75**	**77.76**	**87.87**	**93.94**	Key.Net + ref.	**0.03**	**0.14**	**1.13**	**58.70**	**70.11**	**83.46**

References

1. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: Proc. CVPR (2017)
2. Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K.: Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In: Proc. ICCV (2019)
3. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Proc. ECCV (2006)
4. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description. In: CVPR Workshops (2018)
5. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In: Proc. CVPR (2019)
6. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv (2015)
7. Li, Z., Snavely, N.: MegaDepth: Learning single-view depth prediction from internet photos. In: Proc. CVPR (2018)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
9. Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M.: R2D2: Repeatable and Reliable Detector and Descriptor. In: Advances in NeurIPS (2019)
10. Schönberger, J.L., Hardmeier, H., Sattler, T., Pollefeys, M.: Comparative evaluation of hand-crafted and learned local features. In: Proc. CVPR (2017)
11. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proc. CVPR (2017)