Two-view Geometry Scoring Without Correspondences

Axel Barroso-Laguna[1] Eric Brachmann[1] Victor Adrian Prisacariu[1,2]

Gabriel Brostow[1,3] Daniyar Turmukhambetov[1]

[1]Niantic [2]University of Oxford [3]University College London

www.github.com/nianticlabs/scoring-without-correspondences

Abstract

Camera pose estimation for two-view geometry traditionally relies on RANSAC. Normally, a multitude of image correspondences leads to a pool of proposed hypotheses, which are then scored to find a winning model. The inlier count is generally regarded as a reliable indicator of "consensus". We examine this scoring heuristic, and find that it favors disappointing models under certain circumstances.

As a remedy, we propose the Fundamental Scoring Network (FSNet), which infers a score for a pair of overlapping images and any proposed fundamental matrix. It does not rely on sparse correspondences, but rather embodies a two-view geometry model through an epipolar attention mechanism that predicts the pose error of the two images. FSNet can be incorporated into traditional RANSAC loops. We evaluate FSNet on fundamental and essential matrix estimation on indoor and outdoor datasets, and establish that FSNet can successfully identify good poses for pairs of images with few or unreliable correspondences. Besides, we show that naively combining FSNet with MAGSAC++ scoring approach achieves state of the art results.

1. Introduction

How to determine the relative camera pose between two images is one of the cornerstone challenges in computer vision. Accurate camera poses underpin numerous pipelines such as Structure-from-Motion, odometry, SLAM, and visual relocalization, among others [26, 37, 47, 48, 56]. Much of the time, an accurate fundamental matrix can be estimated by existing means, but the failures are prevalent enough to hurt real-world tasks, and are hard to anticipate [23]. Where are the mistakes coming from?

Traditional approaches first detect then describe a set of interest points in each image, and establish correspondences between the two sets while possibly filtering them, *e.g.*, checking for mutual nearest neighbors or applying Lowe's ratio test [36]. Then, random subsets of correspondences are sampled and a 5-point or 7-point algorithm is
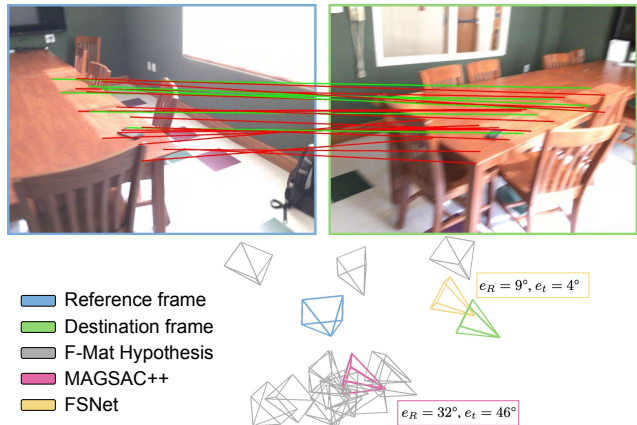
Figure 1. Example where SuperPoint-SuperGlue [19, 46] correspondences are highly populated by outliers, but there are still enough inliers to produce a valid fundamental matrix hypothesis. In such scenarios with unreliable correspondences, current top scoring methods fail (MAGSAC++ [5]), while our proposed FSNet model, a correspondence-free scoring approach, is able to pick out the best fundamental matrix.

used to estimate many essential or fundamental matrix hypotheses, respectively, (*i.e.*, two-view geometry models). A RANSAC [25] loop iterates over the generated hypotheses, and ranks them. Conventionally, the ranking is scored by counting inliers, *i.e.* the number of correspondences within a threshold of that two-view geometry hypothesis. Finally, the top-ranked hypothesis is further refined by using all inlier correspondences.

As the research in robust model estimation advances [2, 5, 13, 15, 41, 58], the different stages of the pipeline are being revisited, *e.g.*, local feature detection and description is learned with neural networks, outlier correspondences are filtered with learned models, hypotheses are sampled more efficiently, or the inlier threshold is optimized. Although the latest matching pipelines produce very accurate and robust correspondences [19, 46, 50], correspondence-based scoring methods are still sensitive to the ratio of inliers, number of correspondences, or the accuracy of the keypoints

[2, 5, 10, 23]. Incorrect two-view geometry estimation can lead to invalid merges in 3D reconstruction models [48], bad localization services [1], or more expensive steps when finding outliers in pose graphs [14].

A second family of approaches emerged in recent years, where a neural network is trained to directly regress two-view geometry from the input images [1, 40, 42, 68]. Thus, such approaches replace all the components of the RANSAC pipeline. This can be a viable approach when two views are extremely difficult, even when they do not overlap [11]. However, challenging scenarios, *e.g.*, wide-baseline, or large illumination changes, can lead to incorrect predictions [32]. Typically, poses directly regressed this way have fewer catastrophic relative pose predictions, but they have difficulty in estimating precise geometry [1]. On the other hand, correspondence-based hypotheses can be very precise, if estimated correspondences are of sufficiently high quality. Our approach uses correspondences to generate model hypotheses, but does not use correspondences to score them during the RANSAC loop.

We propose a fundamental matrix scoring network that leverages epipolar geometry to compare features of the images in a dense manner. We refer to our method as the Fundamental Scoring Network, or FSNet for short. Inspired by the success of Vision Transformers [61], and detector-free matchers [31, 50], we define an architecture that incorporates the epipolar geometry into an attention layer, and hence the quality of the fundamental matrix hypothesis conditions the coherence of the computed features. Figure 1 shows an example where correspondences are highly populated by outliers. However, there are still enough inliers to generate a good fundamental matrix, and FSNet was able to select it from the hypothesis pool.

Our contributions are 1) an analysis of the causes of scoring failures, as well as more insights into the traditional RANSAC approach of relative pose estimation; 2) FSNet, a network that predicts angular translation and rotation errors for a given image pair and a fundamental matrix hypothesis; 3) an image order-invariant design of FSNet that outputs the same values for (Image A, Image B, F) and (Image B, Image A, F^T) inputs; 4) a solution that can be combined with state-of-the-art methods to cope with current failure cases.

2. Related Work

Establishing correspondences. Previous correspondence estimation built around SIFT [36] has largely been superseded by learned methods. Keypoint detectors [6, 62], patch-based descriptors [53, 54], joint detector-descriptors [8, 19, 22, 43], or shape estimators [7, 38, 65] are some of the steps that have benefited from data-driven techniques. To find correspondences, the mutual nearest neighbors and Lowe's ratio test [36] approach has also been revisited, and learned matchers have pushed forward the matching capa-
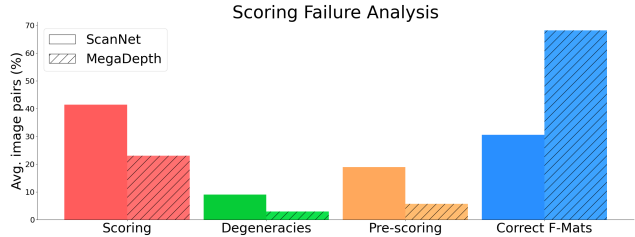

Figure 2. Failure cases analysis. We plot for ScanNet and MegaDepth datasets the percentage of image pairs for which a good fundamental matrix (fundamental matrix with pose error below 10°) was not selected due to scoring failure, degenerate case, or pre-scoring failure, *i.e.*, there was not a valid fundamental matrix in the hypothesis pool. We see that the failures center on the scoring function, it is especially difficult to score hypotheses in the indoor scenario (ScanNet), where local features suffer more [50].

bility of feature extractors [46]. Complementary to matchers, additional filtering methods learn to detect and reject outlier correspondences [12,51,64,66,67]. Once correspondences are established, the RANSAC loop finds the best hypothesis among the pool. Multiple works aim at sampling an all-inlier correspondence minimal-set sooner than RANSAC [10, 15, 55, 57, 58]. Combining that with early termination techniques and detection of degenerate configurations [13, 17, 30] can significantly improve the results and run-time. More recently, alternative methods have been proposed to improve upon the classical detect-then-describe approach, *e.g.*, detector-free matchers [31, 50, 60], or direct relative-pose regressor networks [1, 11, 42, 44, 68].

Model quality. Model quality research focuses mainly on improving the heuristics for classical inlier counting [49, 58, 59]. LO-RANSAC [16] applies a local optimization step to promising models generated from the RANSAC sampling. GC-RANSAC [3] extended previous local optimization and uses graph techniques to infer spatial structures and mask out outliers. MAGSAC++ [4, 5] proposes an iterative inlier counting score over a range of inlier thresholds, which reduces the sensitivity to the inlier-outlier threshold parameter. MQ-Net [2] combines the inlier counting score with a neural network that predicts the quality of a hypothesis from correspondence residuals.

Transformers in vision. Since its introduction, the transformer architecture [61] has become the standard in natural language processing due to its performance and simple design. Transformers are getting attention in the vision community, and have been applied successfully to image matching [31, 50], multi-view stereo [20, 29, 63], or depth estimation [27, 34], among others [21, 44]. Moreover, different works have been proposed to guide the cross-attention mechanism with epipolar supervision [20,27,29,63], where the most popular strategies use the epipolar attention to limit
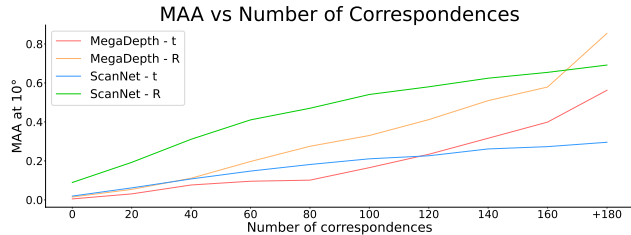
Figure 3. MAA@10° vs number of correspondences. We show that the number of correspondences has a high impact on the precision of the fundamental matrix. The latest matchers already filter out the correspondence outliers [46], and naturally, the number of correspondences correlates with the difficulty of correctly matching two images. Thus, we see that the number of correspondences is a good indicator to decide when to trust the correspondences.

| | mAA at 10° | | | |
| | ScanNet | | MegaDepth | |
	R	t	R	t
GT Pose error	**0.75**	**0.72**	**0.94**	**0.89**
SED	0.57	<u>0.48</u>	0.86	<u>0.78</u>
RE1	<u>0.74</u>	0.41	<u>0.91</u>	<u>0.78</u>
Epi. Distance	0.57	0.40	0.89	0.76
MAGSAC++	0.51	0.22	0.80	0.55

Table 1. Error criteria evaluation. Fundamental matrices are generated with SP-SG [19, 46] and MAGSAC++ [5], and evaluated under densely projected correspondences with ground-truth camera poses and depth maps on our validation set. GT Pose error uses directly the error associated to the fundamental matrix to rank them, being the upper-bound of the scoring function.

the matching search space [29], or to check the consistency of a depth map [27].

3. Analysis

In this section we explore shortcomings of the traditional RANSAC-based pose estimation approach. First, we show that correspondence-based scoring of fundamental matrix hypotheses is the major reason for RANSAC returning incorrect relative poses. Next, we show that the low number of correspondences often leads to scoring function failures motivating correspondence-free approach. To predict fundamental matrix quality without correspondence-based heuristics, we need a good training signal. We consider alternatives to correspondence-based scoring functions that can be used to quantify the quality of fundamental matrix hypotheses. For our analyses, we use SuperPoint-SuperGlue (SP-SG) [19, 46] correspondences with MAGSAC++ [5], a top-performing publicly available feature extraction, feature matching and robust estimation model. We opt for a top performing combination instead of a classical baseline, *i.e.*, SIFT with RANSAC, to analyse a more realistic use case. We then mine image pairs with low overlapping views from validation splits of ScanNet [18] and MegaDepth [35] datasets as in [46, 50] and study MAGSAC++'s behavior for 500 iterations, *i.e.*, 500 fundamental matrix hypotheses were generated and scored. See Section 4.6 for more details on the validation set generation.

Where do wrong solutions come from? In Figure 2, we show the number of times a good fundamental matrix (fundamental matrix with pose error $< 10°$) was selected. We also count failures: (i) pre-scoring, *i.e.*, no good fundamental matrix is among hypotheses, (ii) degeneracy cases, *e.g.*, inlier correspondences can be explained with a homography, (iii) scoring failures, where bad fundamental matrix was chosen by MAGSAC++ heuristic

while a good fundamental matrix was among the 500 hypotheses but had lower score. As seen in the figure, 42% (ScanNet) and 23% (MegaDepth) of the image pairs generate a valid fundamental matrix but the scoring method is not able to select it.

Why are wrong solutions selected? Inlier heuristics find fewer inliers for a good fundamental matrix than a bad fundamental matrix. So, what leads to a low number of inliers for a good fundamental matrix? Intuitively, feature matching methods return fewer correspondences when the image pair is difficult to match, due to repetitive patterns, lack of textures, or small visual overlap, among the possible reasons. In such scenarios, correspondences can also be highly contaminated by outliers, making it even harder to select the correct fundamental matrix. Figure 3 shows the MAA@10° as in [32] *w.r.t* the number of correspondences generated by SP-SG and MAGSAC++. We see the strong correlation between the number of correspondences and the accuracy of the selected model. Furthermore, the SuperGlue correspondences in this experiment are already of high-quality, as bad correspondences were filtered out. In supplementary materials, we show that loosening the correspondence filtering criteria, and hence, increasing the number of correspondences, does not lead to improved results. This observation motivates our correspondence-free fundamental matrix hypothesis scoring approach, which leads to improvements in relative pose estimation task when correspondences are not reliable.

How should we select good solutions? If we want to train a model that predicts fundamental matrix quality what should this model predict? As we are interested in ranking hypotheses, or potentially discard all of them if they are all wrong, the predicted value should correlate with the *error in the relative pose* or the *mismatch of epipolar constraint*.

The quality of relative pose is measured for rotation and translation separately. The angle of rotation between the estimated and the ground truth rotation matrices provides the rotation error. As fundamental and essential matrices
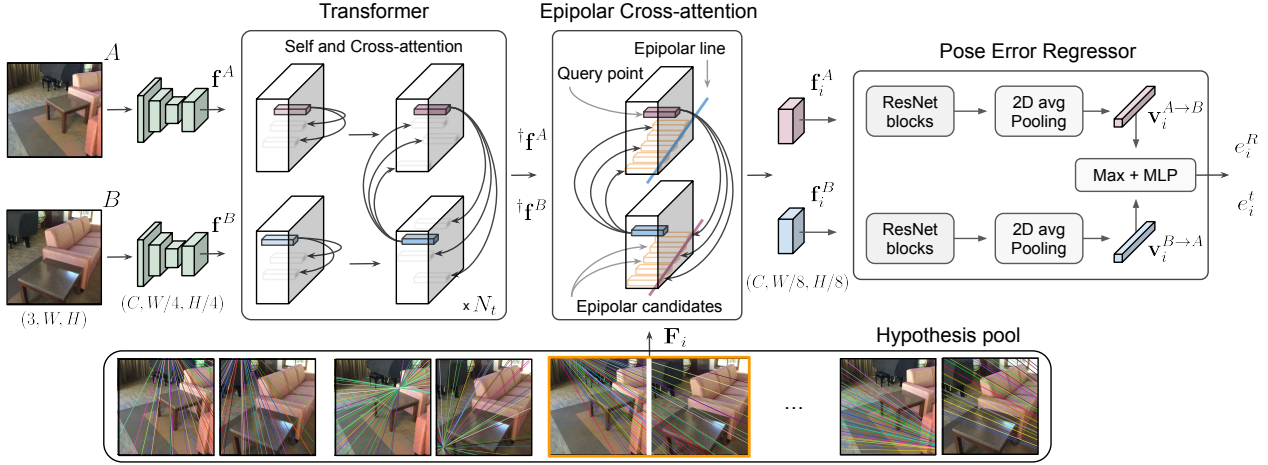
Figure 4. FSNet architecture has four components. 1. A CNN feature extractor computes the feature maps $\mathbf{f}^A$ and $\mathbf{f}^B$ from input images A and B at 1/4 of input resolution (Section 4.1). 2. The extracted features are then processed by the transformer block, which contains N_t self and cross-attention layers. The transformer outputs $^\dagger\mathbf{f}^A$ and $^\dagger\mathbf{f}^B$, which are stored and reused for every $\mathbf{F}_i$ hypothesis (Section 4.2). 3. The epipolar cross-attention layer applies cross-attention along the epipolar lines, and embeds $\mathbf{F}_i$ into the feature maps $\mathbf{f}_i^A$ and $\mathbf{f}_i^B$. Epipolar attention is done every two positions, reducing the final feature maps to an 1/8 of the input resolution (Section 4.3). 4. The pose regressor applies a ResNet and a 2D average pooling block, and outputs $\mathbf{v}_i^{A\to B}$ and $\mathbf{v}_i^{B\to A}$. The vectors are combined through a max pooling operator, and the final MLP layer predicts the e_i^R and e_i^t errors associated to $\mathbf{F}_i$ (Section 4.4).

are scale-invariant, the scale-invariant metric of translation error is the angular distance between the estimated and the ground truth translation vectors. Thus, relative pose error requires estimation of two values.

There are multiple error criteria [24] for fitting fundamental matrices. Reprojection error (RE) measures the minimum distance correction needed to align the correspondences with the epipolar geometry. Symmetric epipolar (SED) measures geometric distance of each point to its epipolar line. Finally, Sampson (RE1) distance is a first-order approximation of RE. The benefit of these metrics is that the quality of fundamental matrices is measured with one scalar value.

So, how do we choose the best error criteria as training signal for FSNet? Should it be relative pose error, or one of the fundamental matrix error criteria? Let us assume that we have access to multiple oracle models, one oracle for each of different error criteria in Table 1. So, one oracle model that predicts the relative pose error perfectly, one oracle model that predicts SED error perfectly, *etc*. Which of these oracle models provides the best fundamental matrix scoring approach? To simulate evaluation of these oracle models, we use SP-SG and MAGSAC++ to mine fundamental matrix hypotheses from validation datasets, and use the ground truth depth maps and camera poses to generate dense correspondences between image pairs to exactly compute all error criteria. In Table 1, we show the MAA@10° of different scoring criteria when evaluating mined fundamental matrices using oracle models in ScanNet and MegaDepth image pairs. As can

be seen in the table, all the fundamental matrix error criteria under-perform compared to relative pose error metrics, hence FSNet should be supervised by the relative rotation and translation errors.

4. Method

FSNet estimates the quality of a fundamental matrix hypothesis, $\mathbf{F}_i$, for the two input images, A and B, without relying on correspondences and processing the images directly. Besides fundamental matrix quality estimation, in a calibrated setup, FSNet could compute the score of the essential matrix, $\mathbf{E}_i$, by first obtaining $\mathbf{F}_i$ based on their relationship: $\mathbf{E}_i = \mathbf{K}_B^T \mathbf{F}_i \mathbf{K}_A$. Figure 4 shows the four major components of FSNet architecture: the feature extractor, the transformer, the epipolar cross-attention, and the pose error regressor. The supplementary materials provide more detailed description of our network.

4.1. Feature Extractor

The feature extractor is a standard convolutional architecture as in [50], it follows the Unet-style network [45] design with skip and residual connections [28] and computes feature maps at $1/4$ of the input resolution. Before feature extraction, we center-crop and resize the input images A and B to a resolution of (H, W) (in our experiments we use $(256, 256)$). Images are then processed by the convolutional feature extractor to produce the C-dimensional feature maps $\mathbf{f}^A$ and $\mathbf{f}^B$.

4.2. Transformer

We use an L multi-head attention transformer architecture following [50], and alternate between self and cross-attention blocks to exploit the similarities within and across the feature maps. We denote the transformed features as $^\dagger\mathbf{f}^A$ and $^\dagger\mathbf{f}^B$. Following the transformer nomenclature, some features are used to compute the query (Q), and potentially different features are used to compute the key (K) and the value (V). Q retrieves information from V based on the attention weight computed from the product of Q and K. In the self-attention layer, the same feature map builds Q, K, and V, meanwhile, in the cross-attention layer, Q is computed from a different feature map than K and V. We interleave the self and cross-attention block N_t times.

Self-attention and Cross-attention layers. To limit the computational complexity of our transformer block, we use a Linear Transformer [33] as in [50]. Linear Transformer reduces the computational complexity of the original Transformer from $O(N^2)$ to $O(N)$ by making use of the associativity property of matrix products and replacing the exponential similarity kernel with a linear dot-product kernel. Specifically, in a self-attention layer, the input feature map $\mathbf{f}'$ is used to compute Q, K and V. We concatenate the result of the attention layer with the input $\mathbf{f}'$ feature map and pass it through a two-layer MLP. The output of the MLP is then added to $\mathbf{f}'$ and passed to the next block. In the cross-attention layer, we repeat the previous process but compute Q from one feature map and K and V from the second.

We found that positional encoding for attention layers did not improve our results, similarly to findings in [34], and hence, we do not use positional encodings in FSNet.

4.3. Epipolar Cross-attention

Up to this point, attended feature maps, $^\dagger\mathbf{f}^A$ and $^\dagger\mathbf{f}^B$, are cached and reused. Given that FSNet computes the score for every $\mathbf{F}_i$, this design assures a more practical scenario where the overhead of computing additional fundamental matrix scores is small.

For every fundamental matrix hypothesis $\mathbf{F}_i$, our epipolar cross-attention mechanism embeds $\mathbf{F}_i$ together with feature maps $^\dagger\mathbf{f}^A$ and $^\dagger\mathbf{f}^B$. Every position $\mathbf{p}^A = [u,v]$ in feature map $^\dagger\mathbf{f}^A$ has a corresponding epipolar line in $^\dagger\mathbf{f}^B$ defined as $\mathbf{l}_{uv}^{A\to B} = \mathbf{F}_i'\bar{\mathbf{p}}^A$, where $\bar{\mathbf{p}}^A$ refers to the homogeneous coordinates of $\mathbf{p}^A$ and $\mathbf{F}_i'$ is a scaled $\mathbf{F}_i$ by a factor of $1/4$. As we consider potentially hundreds of hypotheses, the resolution of feature maps impacts run-time speed. So, we opt to define query points, $\mathbf{p}^A = [u,v]$, with a step sampling of two. This reduces even further the final feature map to a resolution of $1/8$ of the input image.

So, for every feature $^\dagger\mathbf{f}_{uv}^A \in {}^\dagger\mathbf{f}^A$ we sample $^\dagger\mathbf{f}^B$ at D equidistant locations along the epipolar line $\mathbf{l}_{uv}^{A\to B}$. We start sampling where the epipolar line meets the feature map (from left to right) and use bilinear interpolation to pro-

duce D features $^\dagger\mathbf{f}_{\mathbf{l}_{uv}}^B$. If sampling positions fall outside the image plane, or the epipolar line never crosses the image, we zero pad the features. Thus, we build feature volume $^\dagger\mathbf{f}_i^B \in [C, D, W/8, H/8]$ from feature map $^\dagger\mathbf{f}^B$ and $\mathbf{F}_i$. We use $^\dagger\mathbf{f}^A$ to compute Q, and $^\dagger\mathbf{f}_i^B$ volume to obtain the K and V, and perform attention along the epipolar candidate points. Finally, we use Q, K, and V to obtain epipolar transformed features $\mathbf{f}_i^A$. For order-invariance, we also compute $\mathbf{f}_i^B$ by repeating these operations for $(^\dagger\mathbf{f}^B, {}^\dagger\mathbf{f}^A)$ pair of feature maps and $\mathbf{F}_i^T$.

4.4. Pose Error Regressor

As seen in Figure 4, the pose error regressor uses a ResNet block to extracts features from $\mathbf{f}_i^A$ and $\mathbf{f}_i^B$. Following [1,11], we apply a 2D average pooling that results in two 1D vectors, $\mathbf{v}_i^{A\to B}$ and $\mathbf{v}_i^{B\to A}$, with size C'. Both 1D vectors are then merged by a max pooling operator, such that different order of the input images always produce the same feature vector $\mathbf{v}_i$. An MLP layer then regresses the angular translation and rotation errors, e_i^t and e_i^R, associated to $\mathbf{F}_i$.

4.5. Loss Function

Contrary to previous binary [2] or multi-class [11] formulation of the pose error, we experimentally found that FSNet is more accurate when regressing independently the translation and rotation errors (see Section 5.3). The supervision we use is angular errors, hence, the predicted error, e_i, and the ground truth error, $\bar{e}_i$, are bounded between $[0°, 180°]$. Directly using a simple L1 loss ($l = |\bar{e}_i - e_i|$) would treat all error ranges equally. However, we are interested in accurate error estimation of all good fundamental matrices, while not requiring being as precise when the fundamental matrices have high pose errors. Hence, we propose to use a soft clamping of the ground truth error as well as the network prediction, such as:

$$\mathcal{L} = |g(\bar{e}_i^t) - g(e_i^t)| + |g(\bar{e}_i^R) - g(e_i^R)|, \qquad (1)$$

where $g(x)$ refers to the $\tanh(x/t_s)$ function, and t_s is the scaling factor that adjusts the (soft) threshold after which the accuracy of the angular errors is not important.

4.6. Implementation Details

We train indoor and outdoor FSNet models on ScanNet [18] and MegaDepth [35] datasets. To generate training and validation sets, we first extract SP-SG [19,46] correspondences using appropriate indoor and outdoor pre-trained models. We then draw minimal subsets of correspondences randomly and extract 500 two-view hypotheses for every image pair. For each hypothesis, we compute the angular translation (e_t) and rotation (e_R) errors using the ground truth extrinsic and intrinsic parameters. During training, we ensure that the ground truth hypothesis is among the 500

	0-100		100-Inf		All	
	MAA@10° ↑	**Median (°) ↓**	**MAA@10° ↑**	**Median (°) ↓**	**MAA@10° ↑**	**Median (°) ↓**
	R / t / max(R, t)	e_R / e_t	R / t / max(R, t)	e_R / e_t	R / t / max(R, t)	e_R / e_t
Fundamental						
RANSAC [25]	0.15 / 0.07 / 0.04	17.97 / 30.28	0.24 / 0.06 / 0.05	11.91 / 37.03	0.17 / 0.07 / 0.04	15.54 / 32.18
MAGSAC++ [5]	0.28 / 0.12 / 0.08	9.19 / 24.38	0.63 / 0.29 / 0.27	2.80 / 8.66	0.38 / 0.17 / 0.14	6.35 / 17.89
FSNet (F)	0.29 / **0.19** / 0.12	**7.98** / **13.30**	0.52 / 0.25 / 0.21	4.16 / 9.62	0.36 / 0.21 / 0.15	6.52 / 12.05
FSNet (F + E)	0.29 / 0.18 / 0.11	8.01 / 14.05	0.50 / 0.24 / 0.20	4.35 / 10.26	0.35 / 0.20 / 0.14	6.64 / 12.78
w/ Corresp. filter	0.29 / **0.19** / 0.12	**7.98** / **13.30**	0.63 / 0.29 / 0.27	2.80 / 8.66	0.39 / 0.22 / 0.16	6.09 / 11.59
w/ Candidate filter	**0.33** / 0.18 / **0.13**	**7.48** / 14.91	**0.66** / **0.36** / **0.32**	**2.69** / **6.50**	**0.43** / **0.23** / **0.19**	**5.38** / **11.39**
Essential						
EssNet [68]	-	-	-	-	0.01 / 0.02 / 0.01	48.64 / 52.95
Map-free [1]	-	-	-	-	0.39 / 0.13 / 0.09	5.75 / 14.20
RANSAC [25]	0.27 / 0.16 / 0.13	9.78 / 16.64	0.63 / 0.38 / 0.33	3.03 / 6.15	0.37 / 0.24 / 0.19	6.55 / 11.68
MAGSAC++ [5]	0.29 / 0.19 / 0.14	8.90 / 16.00	0.65 / 0.41 / 0.38	2.57 / 5.65	0.40 / 0.26 / 0.21	5.95 / 10.96
FSNet (E)	**0.36** / **0.25** / **0.18**	**6.34** / **10.51**	0.61 / 0.35 / 0.31	3.22 / 6.78	**0.44** / 0.28 / 0.22	5.13 / 8.95
FSNet (F + E)	0.35 / 0.24 / 0.17	6.70 / 10.65	0.60 / 0.33 / 0.30	3.41 / 7.15	0.43 / 0.27 / 0.21	5.26 / 9.21
w/ Corresp. filter	**0.36** / **0.25** / **0.18**	**6.34** / **10.51**	0.65 / 0.41 / 0.38	2.57 / 5.65	**0.44** / **0.29** / **0.23**	**5.07** / **8.75**
w/ Candidate filter	0.33 / 0.22 / 0.16	7.71 / 13.41	**0.69** / **0.44** / **0.40**	**2.40** / **5.26**	**0.44** / 0.28 / **0.23**	5.14 / 9.48

Table 2. **Fundamental and essential matrix estimation on ScanNet**. We compute the MAA@10° and Median error (°) metrics in the test split of ScanNet, and divide the pairs of images based on the number of SP-SG correspondences to highlight the benefits of FSNet, which results in 3,522 (0-100) and 1,478 (100-Inf) image pairs. In fundamental and essential estimation, we see that when number of correspondences is small (0-100), FSNet provides a more robust solution than competitors. In the overall split (All), FSNet obtains more precise rotation errors for the fundamental estimation tasks, while outperforming in all metrics in essential estimation. Moreover, we also show how FSNet and MAGSAC++ can be easily combined to obtain more reliable approaches.

hypotheses. In batch generation, we cluster the hypotheses into bins based on the pose error and randomly select a bin, from which one hypothesis is uniformly sampled.

For the indoor model, we use the training splits proposed in [50]. For our outdoor model, we use the MegaDepth training and validation scenes proposed in [50], except for scenes in the test set of the CVPR IMW 2019 PhotoTourism dataset [32, 52] as in [46]. For both datasets, we generate a total of 90,000 and 30,000 image pairs with 500 fundamental and essential matrix hypotheses for each image pair. To highlight the benefit of FSNet over hard image pairs, we sample image pairs in both datasets that have a visual overlapping score between 10% and 40% of the image.

We indicate the training source of FSNet by (F) and (E) for fundamental and essential matrices, respectively. We also found that the distribution of essential matrices is very different from distribution of fundamental matrices (see supplementary), so, we combine fundamental and essential matrix hypotheses into (F + E) training dataset.
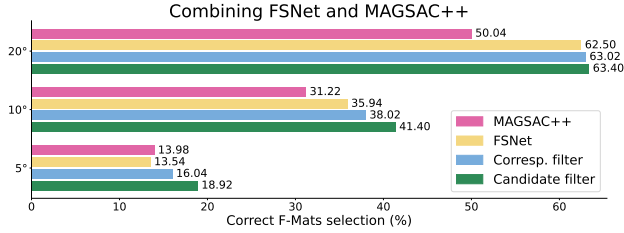
FSNet is trained end-to-end with randomly initialized weights, a learning rate of 1×10^{-4} and a batch size of 56 image pairs. We train on four V100 GPUs and the network converges after 72 hours. For the feature extractor, we use a ResNet-18 [28] as in [50]. The transformer block uses three attention layers ($N_t = 3$), and we sample $D = 45$ matching candidates along the epipolar line. The loss formulation uses $t_s = 25$, such that the pose errors above 40°are (soft) clamped. At test time, FSNet selects the model hypothesis

that returns the minimum pose error, where the pose error for $\mathbf{F}_i$ is computed as $e_i = \max(e_i^R, e_i^t)$.

5. Experiments

This section presents results for different fundamental and essential hypothesis scoring methods. We first compute correspondences with SP-SG [19, 46], and then use the universal framework USAC [41] to generate hypotheses with a uniformly random minimal sampling. Besides the inlier counting of RANSAC [25], we also use the state-of-the-art MAGSAC++ scoring function [5], and compare them both to FSNet scoring method. To control for randomness, we use the same set of hypotheses generated for each image pair for the evaluation of all methods. We refine all the selected hypotheses by applying LSQ fitting followed by the Levenberg-Marquardt [39] optimization to the correspondences that agree with the best hypothesis model.

To evaluate the scoring methods, we decompose the hypothesis (fundamental or essential matrix) into rotation matrix and translation vectors, and compute the angular errors (e_R and e_t) w.r.t ground truth rotation and translation. We report the mean Average Accuracy (mAA), which thresholds, accumulates, and integrates the errors up to a maximum threshold (in our experiments 10°) [32]. The integration over the accumulated values gives more weight to accurate poses and discards any camera pose with a relative angular error above the maximum threshold. Besides the

Combining FSNet and MAGSAC++

| | | | | | | | |
50.04 — 20°
62.50
63.02
63.40

31.22 — 10°
35.94
38.02
41.40

13.98 — 5°
13.54
16.04
18.92

Correct F-Mats selection (%) 0 10 20 30 40 50 60

Legend: MAGSAC++ · FSNet · Corresp. filter · Candidate filter

Figure 5. Combining FSNet and MAGSAC++. Complementing both scoring methods retrieves the highest number of correct fundamental matrices under all error thresholds (5°, 10° and 20°).

mAA, we report the median errors of the selected models.

5.1. Indoor Pose Estimation

Evaluations on the Scannet test set are shown in Table 2. Motivated by the analysis in Section 3, and to emphasize the benefits of FSNet, we also report results on two types of image pairs: image pairs with up to 100 input correspondences and image pairs with more than 100 input correspondences. The threshold is chosen based on our validation, where 64% of the image pairs do not have more than 100 correspondences, and hence, top-performing methods can struggle.

We see that in the few-correspondences split (0-100), all FSNet models outperform RANSAC and MAGSAC++, being especially effective on the angular translation error metrics. Overall, FSNet returns the best mAA score and median error for translation, while having comparable results to MAGSAC++ on rotation, both for fundamental and essential matrix evaluations. We also observe that FSNet (F+E) does not score as well as models trained on F or E hypotheses only, further indicating differences in the distribution of fundamental and essential matrices.

For essential matrix evaluations, we also report scores of relative pose regression methods (RPR), *i.e.*, EssNet [68] and Map-free [1], which also do not rely on correspondences. FSNet (E) obtains the best mAA scores and median errors in the full test set (All). RPR methods do not compute as accurate poses as RANSAC-based methods or FSNet.

As MAGSAC++ is very effective when SP-SG is able to extract sufficient number of correspondences (100-Inf), we propose two approaches to combine both methods. In *Corresp. filter*, we use the number of correspondences such that if the number is below 100, we use FSNet to score, otherwise, we use MAGSAC++. In *Candidate filter*, we first select top hypotheses using MAGSAC++, and then we use FSNet to choose the best hypothesis out of the selected hypotheses. We use FSNet to score the top 10 hypotheses for fundamental, and top 20 hypotheses for essential based on a hyperparameter search on the validation set. Moreover, in Figure 5, we show the percentage of correct fundamental matrices, *i.e.*, hypotheses with a pose error un-

	MAA@10° ↑		Median (°) ↓
	0-100	All	All
	R / t	R / t	e_R / e_t
Fundamental			
RANSAC [25]	0.12 / 0.11	0.55 / 0.21	3.72 / 17.75
MAGSAC++ [5]	0.24 / 0.17	0.80 / 0.44	1.39 / 5.42
FSNet (F)	0.21 / 0.19	0.71 / 0.35	2.08 / 7.22
w/ Corresp. filter	0.21 / 0.19	0.80 / 0.44	1.39 / 5.42
w/ Candidate filter	**0.27 / 0.21**	**0.82 / 0.47**	**1.32 / 4.78**
Essential			
RANSAC [25]	0.25 / 0.31	0.81 / 0.59	1.13 / 3.06
MAGSAC++ [5]	0.24 / 0.30	0.85 / 0.65	0.92 / 2.18
FSNet (E)	**0.30** / 0.31	0.81 / 0.55	1.39 / 3.42
w/ Corresp. filter	**0.30** / 0.31	0.85 / 0.65	0.92 / 2.18
w/ Candidate filter	0.29 / **0.35**	**0.87 / 0.67**	**0.88 / 2.08**

Table 3. Fundamental and essential matrix estimation on PhotoTourism dataset. We show the mAA (10°) and median error when using SP-SG correspondences. Contrary to indoor datasets [18], in the outdoor scenario, current feature extractors provide very accurate and robust features, and in the test split only 1.5% of the image pairs returned fewer than 100 correspondences. Nevertheless, combinations of FSNet and MAGSAC++ provide the most reliable methods for fundamental and essential matrix estimation.

der given thresholds (5°, 10°, and 20°), and see how combining both methods always maximizes the number of correct fundamental matrices. Moreover, we observe that in the more flexible regime (20°) the FSNet performance is almost identical to the combinations of both methods, indicating that although FSNet estimations are not as accurate as MAGSAC++, there are fewer catastrophic predictions. We link this behaviour with the nature of our method, which uses low resolution feature maps to predict the pose errors, and hence, lacks precision. On the other hand, MAGSAC++ relies on potentially ill-conditioned sets of correspondences, resulting in occasional catastrophic poses.

5.2. Outdoor Pose Estimation

In Table 3, we show the mAA and median errors of the fundamental and essential matrix estimation tasks on the PhotoTourism test set. Analogous to the indoor evaluation, we split the image pairs based on the number of SP-SG correspondences. Contrary to the indoor scenario, there are only 1.5% of the image pairs in the test set that have fewer than 100 SP-SG correspondences. This is in line with the analysis of Figure 2, where scoring failures are not as common in outdoor dataset as in indoor dataset. There are multiple explanations of these phenomena: (i) walls and floors are mostly untextured surfaces which lead to fewer reliable correspondences for indoor images, (ii) indoor surfaces are closer to the camera, so small camera motion can result in large parts of the scene not being visible from both

	MAA@10° ↑		Median (°) ↓
	0-100	**All**	**All**
	R / t	R / t	e_R / e_t
Supervision			
RE1	0.23 / 0.12	0.28 / 0.14	8.53 / 19.42
CE as in [2]	0.27 / <u>0.18</u>	0.34 / <u>0.20</u>	6.94 / 13.21
$L1$	0.25 / 0.16	0.32 / 0.18	7.52 / 14.07
FSNet (Soft-$L1$)	**0.29** / **0.19**	<u>0.36</u> / **0.21**	<u>6.52</u> / **12.05**
MAGSAC++ [5]	<u>0.28</u> / 0.12	**0.38** / 0.17	**6.35** / 17.89

Table 4. Loss function and generalization ablation results on the fundamental matrix estimation task in the test set of ScanNet. Supervising the training of FSNet directly with the camera pose error provides the best results, while the CE approach outperforms L1 norm and Sampson distance (RE1) prediction models.

cameras, which could result in fewer correspondences, (iii) outdoor datasets are generated with SfM methods that rely on 2D-2D correspondences estimated with RANSAC, while indoor datasets rely on RGBD SLAM with dense geometry alignment, so inherent algorithmic bias could result in fewer difficult image pairs in outdoor datasets [9].

In the fundamental estimation task, FSNet scoring approach obtains the most accurate poses in the low number of correspondences split (0-100), while still outperforming RANSAC inlier counting approach in the full test set. In the essential estimation task, even though we observe that correspondence-based scoring approaches, *i.e.*, RANSAC or MAGSAC++, outperform FSNet, they still benefit from combining their predictions with FSNet.

5.3. Understanding FSNet

Loss supervision ablation. In Table 4, we experimentally verify our analysis of supervision signals in Section 3. We show that FSNet trained with Eq. 1 outperforms FSNet trained using RE1 supervision (best epipolar constraint error criteria among oracle models in Table 1) and FSNet trained using $L1$ norm without soft clamping. We also train FSNet using binary cross-entropy (CE), where fundamental matrix hypothesis $\mathbf{F}_i$ is deemed correct if $e_i < 10°$ and incorrect otherwise. We follow [2] and incorporate the network uncertainty in the cross-entropy loss; please see the supplementary materials for details.

Failures. Figure 6 shows failure cases for FSNet hypothesis selection. As FSNet needs to run on hundreds of hypotheses in a reasonable time, the network is designed to work on low resolution images. Unfortunately, at low resolution, matching features can seem degenerate. Looking into the failures, we observe that the fundamental matrices selected by FSNet have a valid epipolar geometry, *i.e.*, epipolar lines in image B cross the surrounding regions of corresponding points from image A, and hence, the low resolution features $((W/8, H/8) = (32, 32))$ given to the pose error regressor lack the precision to produce a good estimate.
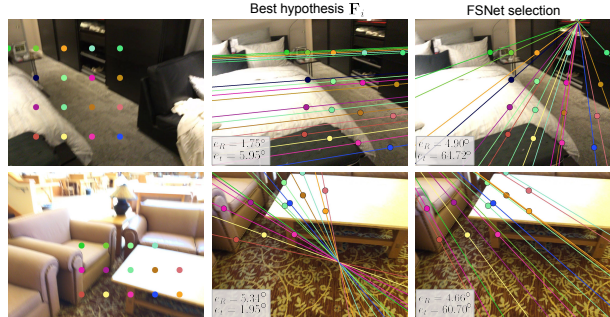
Figure 6. Failure examples. We plot the co-visible points and corresponding epipolar lines of the fundamental matrix selected by FSNet and the best hypothesis in the pool. Besides, we also mark the intersection of the best and FSNet epipolar lines. We see that the lines intersect at positions close to neighbor regions of the corresponding points, and hence, the low-resolution design of FSNet limits its capability to select the optimal hypothesis.

Time analysis. Given a pair of images and 500 hypotheses, FSNet scores and ranks them in 0.97s on a machine with one V100 GPU. Running the feature extraction and transformer subnetworks takes 138.28ms. This is precomputed once and reused for every hypothesis. The hypotheses are queried in batches, so the epipolar cross-attention and pose error regressor subnetworks run in 167.13ms for a batch of 100 fundamental matrices. Besides, strategies such as *Candidate filter* can reduce FSNet time by only scoring promising hypotheses, *e.g.*, for fundamental estimation it only adds 150ms on top of MAGSAC++ runtime (71ms).

6. Conclusions

We introduce FSNet, a correspondence-free two-view geometry scoring method. Our experiments show that our learned model achieves SOTA results, performing exceptionally well in challenging scenarios with few or unreliable correspondences. These results suggest exciting avenues for future research. We mention some of them here.

Naive combinations of FSNet with traditional inlier counting lead to even higher reliability. This suggests that more sophisticated combinations of scoring approaches could lead to further improvements. Alternatively, a better network design may improve prediction accuracy, which could supersede inlier counting in all scenarios.

The first stages of our pipeline (feature extraction, self and cross-attention blocks) form a subnetwork that precomputes feature maps for each image pair. This subnetwork architecture is similar to the latest feature extraction methods, *e.g.* [50]. This means that feature map precomputation, local feature extraction, feature matching, and hypothesis generation can all be performed by this subnetwork. Training the whole network on two tasks: estimating good correspondences and two-view geometry scoring could lead to synergistic improvements on both tasks.

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 6, 7

[2] Daniel Barath, Luca Cavalli, and Marc Pollefeys. Learning to find good models in RANSAC. In *CVPR*, pages 15744–15753, 2022. 1, 2, 5, 8

[3] Daniel Barath and Jiří Matas. Graph-cut RANSAC. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6733–6741, 2018. 2

[4] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10197–10205, 2019. 2

[5] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. MAGSAC++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1304–1312, 2020. 1, 2, 3, 6, 7, 8

[6] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint detection by hand-crafted and learned CNN filters. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5836–5844, 2019. 2

[7] Axel Barroso-Laguna, Yurun Tian, and Krystian Mikolajczyk. ScaleNet: A shallow architecture for scale estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12808–12818, 2022. 2

[8] Axel Barroso-Laguna, Yannick Verdie, Benjamin Busam, and Krystian Mikolajczyk. HDD-Net: Hybrid detector descriptor with mutual interactive learning. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[9] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. On the limits of pseudo ground truth in visual camera re-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6218–6228, 2021. 8

[10] Eric Brachmann and Carsten Rother. Neural-guided RANSAC: Learning where to sample model hypotheses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4322–4331, 2019. 2

[11] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14566–14575, 2021. 2, 5

[12] Luca Cavalli, Viktor Larsson, Martin Ralf Oswald, Torsten Sattler, and Marc Pollefeys. AdaLAM: Revisiting hand-crafted outlier detection. *arXiv preprint arXiv:2006.04250*, 2020. 2

[13] Luca Cavalli, Marc Pollefeys, and Daniel Barath. NeFSAC: Neurally filtered minimal samples. In *ECCV*, 2022. 1, 2

[14] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2021. 2

[15] Ondrej Chum and Jiri Matas. Matching with PROSAC-progressive sample consensus. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 220–226. IEEE, 2005. 1, 2

[16] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *Joint Pattern Recognition Symposium*, pages 236–243. Springer, 2003. 2

[17] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 772–779. IEEE, 2005. 2

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3, 5, 7

[19] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 2, 3, 5, 6

[20] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. TransMVS-Net: Global context-aware multi-view stereo network with transformers. In *CVPR*, pages 8585–8594, 2022. 2

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[22] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019. 2

[23] Hongyi Fan, Joe Kileel, and Benjamin Kimia. On the instability of relative pose estimation and RANSAC's role. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8935–8943, 2022. 1, 2

[24] Mohammed E Fathy, Ashraf S Hussein, and Mohammed F Tolba. Fundamental matrix estimation: A study of error criteria. *Pattern Recognition Letters*, 32(2):383–391, 2011. 4

[25] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 6, 7

[26] Debabrata Ghosh and Naima Kaabouch. A survey on image mosaicing techniques. *Journal of Visual Communication and Image Representation*, 34:1–11, 2016. 1

[27] Vitor Guizilini, Rareș Ambruș, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with

transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 160–170, 2022. 2, 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[29] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *CVPR*, pages 7779–7788, 2020. 2, 3

[30] Maksym Ivashechkin, Daniel Barath, and Jiří Matas. VSAC: Efficient and accurate estimator for h and f. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15243–15252, 2021. 2

[31] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 2

[32] 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. 2, 3, 6

[33] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 5

[34] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *ICCV*, pages 6197–6206, 2021. 2, 5

[35] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 3, 5

[36] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 1, 2

[37] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004. 1

[38] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–300, 2018. 2

[39] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978. 6

[40] Omid Poursaeed, Guandao Yang, Aditya Prakash, Qiuren Fang, Hanqing Jiang, Bharath Hariharan, and Serge Belongie. Deep fundamental matrix estimation without correspondences. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[41] Rahul Raguram, Ondrej Chum, Marc Pollefeys, Jiri Matas, and Jan-Michael Frahm. USAC: A universal framework for random sample consensus. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):2022–2038, 2012. 1, 6

[42] René Ranftl and Vladlen Koltun. Deep fundamental matrix estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 284–299, 2018. 2

[43] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019. 2

[44] Chris Rockwell, Justin Johnson, and David F. Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *T3DV*, 2022. 2

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[46] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2, 3, 5, 6

[47] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 1

[48] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2

[49] Charles V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):925–938, 1995. 2

[50] 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. 1, 2, 3, 4, 5, 6, 8

[51] Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Moo Yi. ACNe: Attentive context normalization for robust permutation-equivariant learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11286–11295, 2020. 2

[52] 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. *Communications of the ACM*, 59(2):64–73, 2016. 6

[53] Yurun Tian, Axel Barroso Laguna, Tony Ng, Vassileios Balntas, and Krystian Mikolajczyk. HyNet: Learning local descriptor with hybrid similarity measure and triplet loss. *Advances in Neural Information Processing Systems*, 33:7401–7412, 2020. 2

[54] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second order similarity

regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019. 2

[55] Wei Tong, Jiri Matas, and Daniel Barath. Deep MAGSAC++. *arXiv preprint arXiv:2111.14093*, 2021. 2

[56] Philip HS Torr and David W Murray. Outlier detection and motion segmentation. In *Sensor Fusion VI*, volume 2059, pages 432–443. SPIE, 1993. 1

[57] Philip Hilaire Torr, Slawomir J Nasuto, and John Mark Bishop. NAPSAC: High noise, high dimensional robust estimation-it's in the bag. In *British Machine Vision Conference (BMVC)*, volume 2, page 3, 2002. 2

[58] Philip HS Torr and Andrew Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer vision and image understanding*, 78(1):138–156, 2000. 1, 2

[59] Philip H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, 50(1):35–61, 2002. 2

[60] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2021. 2

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[62] Yannick Verdie, Kwang Yi, Pascal Fua, and Vincent Lepetit. TILDE: A temporally invariant learned detector. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5279–5288, 2015. 2

[63] Xiaofeng Wang, Zheng Zhu, Fangbo Qin, Yun Ye, Guan Huang, Xu Chi, Yijia He, and Xingang Wang. MVSTER: Epipolar transformer for efficient multi-view stereo. *arXiv preprint arXiv:2204.07346*, 2022. 2

[64] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2666–2674, 2018. 2

[65] Kwang Moo Yi, Yannick Verdie, Pascal Fua, and Vincent Lepetit. Learning to assign orientations to feature points. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 107–116, 2016. 2

[66] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5845–5854, 2019. 2

[67] Chen Zhao, Zhiguo Cao, Chi Li, Xin Li, and Jiaqi Yang. NM-Net: Mining reliable neighbors for robust feature correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 215–224, 2019. 2

[68] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3319–3326. IEEE, 2020. 2, 6, 7