NeFSAC: Neurally Filtered Minimal Samples

Luca Cavalli[1], Marc Pollefeys[1,2], and Daniel Barath[1]

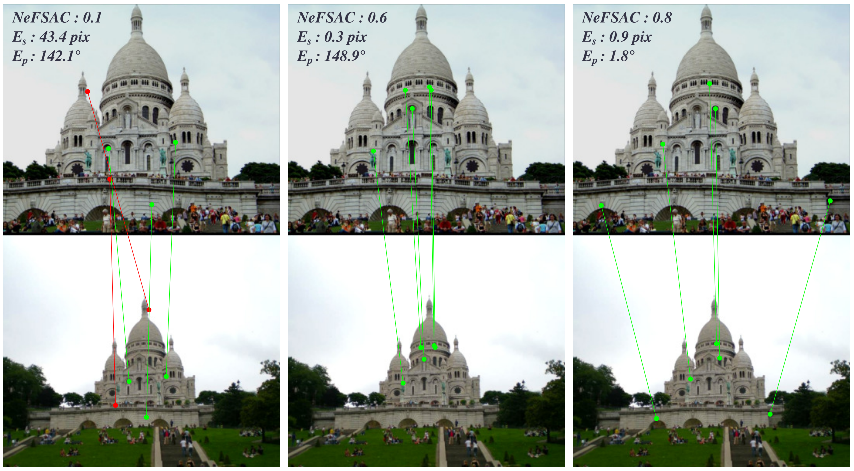
[1] Department of Computer Science, ETH Zurich, Zurich, Switzerland
[2] Microsoft Mixed Reality and AI Zurich Lab
luca.cavalli@inf.ethz.ch

Abstract. Since RANSAC, a great deal of research has been devoted to improving both its accuracy and run-time. Still, only a few methods aim at recognizing invalid minimal samples early, before the often expensive model estimation and quality calculation are done. To this end, we propose NeFSAC, an efficient algorithm for neural filtering of motion-inconsistent and poorly-conditioned minimal samples. We train NeFSAC to predict the probability of a minimal sample leading to an accurate relative pose, only based on the pixel coordinates of the image correspondences. Our neural filtering model learns typical motion patterns of samples which lead to unstable poses, and regularities in the possible motions to favour well-conditioned and likely-correct samples. The novel lightweight architecture implements the main invariants of minimal samples for pose estimation, and a novel training scheme addresses the problem of extreme class imbalance. NeFSAC can be plugged into any existing RANSAC-based pipeline. We integrate it into USAC and show that it consistently provides strong speed-ups even under extreme train-test domain gaps – for example, the model trained for the autonomous driving scenario works on PhotoTourism too. We tested NeFSAC on more than 100k image pairs from three publicly available real-world datasets and found that it leads to *one order of magnitude* speed-up, while often finding more accurate results than USAC alone. The source code is available at `https://github.com/cavalli1234/NeFSAC`.

Keywords: RANSAC, epipolar geometry estimation, minimal samples, machine learning, motion prior, autonomous driving

1 Introduction

Robust model estimation is a cardinal problem in Computer Vision. RANSAC [17] has been a very successful and widely applied approach to robust model estimation since the early days of Computer Vision, and a great research effort has been devoted to improving it. While initial efforts [11,12,14,15,17,23] and some recent works [2,5,10,28] are aimed at improving its accuracy, run-time, and robustness attacking well-understood challenges with hand-engineered techniques, more recently we see substantial advancements in augmenting RANSAC for robust model estimation with learning-based techniques [8,25,29,30,38] that



(a) With outliers (b) Poorly conditioned (c) Well conditioned

Fig. 1: **Minimal sample filtering**. We show three example minimal samples for Essential matrix estimation, with Sampson error E_s in pixels (maximum Sampson error of correspondences with respect to the ground truth Essential matrix), pose error (maximum of rotation and translation error) E_p in degrees, and our model's predicted quality score. *(a)*: a minimal sample encoding unlikely depth and motion due to outliers, easily recognized by NeFSAC only using the pixel's coordinates; *(b)*: an all-inlier sample but with two very close correspondences that lead to estimating a poorly conditioned model and, thus, high pose error; *(c)* a minimal sample with widely spaced correspondences. It leads to accurate pose estimation and is strongly preferred by NeFSAC.

derive implicit models directly from data. Since one of the biggest challenges in RANSAC is handling large outlier ratios, most of the existing learning-based works are framed as an outlier rejection problem. This task is naturally good for learning, since in real scenes correct correspondences are strongly correlated, thus the global set of correspondences can be used to predict which subset is likely to be correct. However, we argue that there is more to be learned from real scenes than recognizing individual outliers. Another important challenge in RANSAC comes with degenerate and ill-conditioned minimal samples, such as in Figure 1b. These configurations often occur in real scenes, e.g., in case of short baseline or when the epipole falls inside the image, when observing close-to-planar scenes or small textured areas leading to localized groups of inlier correspondences. This is problematic in RANSAC since it means that the often *expensive* model estimation and quality calculation is done unnecessarily on many samples which inherently lead to inaccurate models. Also, such models tend to have lots of inliers [15,21], thus misleading the quality calculation. Ideally, the perfect minimal sample filter would be able to recognize and avoid such samples to directly examine well-conditioned ones like the one in Figure 1c.

Besides having inherently invalid samples, real-life images tend to follow certain motion patterns that can be also learned and used to further accelerate the robust estimation by rejecting incorrect minimal samples early. For example, in the autonomous driving scenario when the camera is mounted to a moving vehicle, it follows a distinctive motion pattern that significantly restricts the space of valid minimal sample configurations. Even without having such a strong assumption, e.g., when reconstructing internet images, people tend to take pictures approximately aligned with the gravity direction [16] that, again, gives a probabilistic constraint on the space of valid samples.

Learning a motion prior on minimal samples would allow RANSAC to find the ones that likely lead to the sought model early, spending fewer iterations on unlikely or impossible motions. For this purpose, we propose NeFSAC, which learns to filter invalid and motion-inconsistent minimal samples in RANSAC. NeFSAC can be straightforwardly integrated in any RANSAC variant, e.g. in USAC [28], and provide an important speed-up while *improving* accuracy of the estimation. We train an extremely lightweight neural network to score minimal samples prior to model estimation, thus being able to screen out thousands of minimal samples with negligible compute time. In the worst-case scenario, when the domain gap is too huge, NeFSAC degrades back to random filtering, which has no effect on the RANSAC run-time nor on the accuracy. Still, training in new domains requires only the collection of new image pairs with quasi-ground truth poses obtained from any existing pose estimation method. We integrated our approach into USAC, and measured a reduction of run-time of *one order of magnitude* with significant improvements in the estimation accuracy.

In summary, our contributions are as follows: (i) We propose NeFSAC, a novel framework to augment RANSAC by learning to efficiently distinguish good minimal samples. Our approach can be seamlessly integrated into any existing RANSAC-based pipeline. (ii) We propose a novel neural architecture for the task, and a novel training scheme for effectively learning the sample quality. (iii) We show that NeFSAC provides impressive speed-ups in RANSAC even without the need for strong motion constraints, while at the same time *improving* the accuracy. In the worst-case scenario, it degenerates to the baseline RANSAC with negligible run-time overhead and no drop in accuracy.

2 Related Works

Since RANSAC [17], great efforts from the research community have been concentrated into improving its components. Many works aim at improving the model scoring technique by modeling inlier and outlier distributions and using likelihood scores instead of the original inlier counting score [24,32,34,35]. Similarly, MAGSAC++ [3,4,5] proposes to marginalize the inlier counting score over a range of possible thresholds, reducing the sensitivity of the scores to the choice of a specific noise scale. LO-RANSAC [14] proposes to perform local optimization of promising models during the search, with later improvements on the cost function and inlier selection [22] and with graph-cut masking of outliers [2].

Many of these improvements were combined in USAC [28] and VSAC [21] to achieve state-of-the-art performance.

Closer to our work, another line of research proposes improvements on the sampling scheme to increase the likelihood of detecting an all-inlier sample early. The most widely used approach is the PROSAC [12] algorithm, where the sampling is biased by prior-established likelihoods (e.g., from ratio-test). DSAC [9] first enabled learning through a RANSAC component, followed by Neural-guided RANSAC [8] and Deep MAGSAC++ [33] which learn inlier sampling likelihoods. Other works use spatial techniques to correlate the inlier likelihoods of individual correspondences by preferring neighboring correspondences [36] or by grouping similar ones [26]. Such methods combined with early termination techniques [11,23] can lead to significant improvements in robustness and run-time. However, none of these works can identify unlikely motions or depth configurations in minimal samples, nor can detect degeneracy of minimal samples. Moreover, we consider our work to be orthogonal to these, since it can be used on top of *any* of these approaches.

The cheirality test [37] is widely used to discard some impossible depth configurations. This test discards minimal samples which imply negative depths for some triangulated points. Unfortunately, while it is possible to perform the cheirality test directly on the minimal sample for homographies, it requires expensive epipolar geometry estimation in the cases of Essential matrix or Fundamental matrix estimation.

Even an all-inlier minimal sample can be in a degenerate configuration and lead to unstable relative poses, e.g. when observing a close-to-planar scene. This problem has been recognized and addressed by DEGENSAC [15] that checks for degeneracy and planar configurations of point samples *after* estimating the related epipolar geometry by the plane-and-parallax algorithm [20]. Moreover, QDEGSAC [18] identifies quasi-degenerate solutions in RANSAC and searches for the missing constraints in the outlier set. Differently from these works, we aim to detect such cases *prior to* the expensive epipolar geometry estimation, providing a consistent speed up to the whole procedure.

Outlier filtering techniques aim to filter the set of putative correspondences to increase the inlier rate prior to robust model estimation. These techniques look for spatial patterns in correspondences and perform explicit spatial verification [6,10] or learn a spatial verification model [25,38] optionally conditioned on descriptor information to perform matching altogether [30]. Since our approach does not score individual correspondences, but rather joint minimal samples, we consider our contribution to be orthogonal to outlier filtering techniques.

Despite the enormous research devoted to improving robust model estimation with RANSAC, the early selection of minimal samples is still under-explored. Particularly, to the best of our knowledge, no existing work provides a unique solution to embed general motion and depth priors to accelerate RANSAC. Existing techniques require expensive epipolar geometry estimation to handle degeneracy. In this paper, we show that a lightweight neural network can learn such a filter sufficiently well to provide important savings in run-time *and* improvements in terms of accuracy.

3 Neural Filtering of Minimal Samples

In robust model estimation, a set of data points $D = \{x \in \mathbb{R}^c\}$, optionally contaminated by outliers, is used to fit a model $M \in \mathbb{R}^q$ that minimizes the fitting cost $C = \sum_{x \in D} \mathcal{L}(E(M, x))$ where $E : \mathbb{R}^q \times \mathbb{R}^c \mapsto \mathbb{R}$ is a function that computes the fit error of a data point x with respect to model M, and $\mathcal{L} : \mathbb{R} \mapsto \mathbb{R}$ is a robust loss function that generally has small or zero gradients for large errors to minimize the influence of outlier data points.

Most of the real instances of the robust estimation problem are highly nonlinear, thus the approach proposed by RANSAC [17] is to discretely explore model hypotheses by successively sampling minimal sets of data points D^i_{min} such that they consist of the minimum number m of data points x that can fit exactly a finite set of models. The search is then stopped as soon as a satisfactory model has been found according to some termination criterion, and the final model is usually optimized locally to account for all of its inlier data points. The reason for using minimal sets stems from the RANSAC termination criterion where the required number of iterations depends exponentially on the sample size to provide probabilistic guarantees of finding the sought model.

In this work, we aim to drastically reduce the computational expense of such procedure by learning to pre-filter minimal samples before they are used for model estimation or compared to the rest of the data points. Notice how this is essentially different from previous works on outlier rejection, that, instead, filter out single data points with the objective of increasing the inlier ratio. While our formulation can be applied in general, in the following we will focus on the problems of Essential matrix estimation and Fundamental matrix estimation, where data points are image correspondences ($c = 4$), and minimal samples are constituted by, respectively, $m = 5$ and $m = 7$ correspondences.

In this section, we propose solutions for several challenges that come with the task of learning minimal sample filtering: how to design a lightweight neural architecture that respects all the invariants of minimal samples; how to supervise it in a context of extreme class imbalance; and how to efficiently apply it within RANSAC with the guarantee that, in the worst-case scenario of having a random filter, our method would not cause any degradation of accuracy.

3.1 Minimal Sample Filtering Network

We aim to learn a function $\mathcal{F} : \mathbb{R}^{c \times m} \mapsto [0, 1]$ to score minimal samples, where c is the dimensionality of a data point and m is the minimum number of data points required to fit a finite set of models. Particularly we are interested in Essential matrix estimation ($c = 4, m = 5$) and Fundamental matrix estimation ($c = 4, m = 7$). Note that we disregard information about the global configuration of correspondences across the two images: this simplification leads to a faster

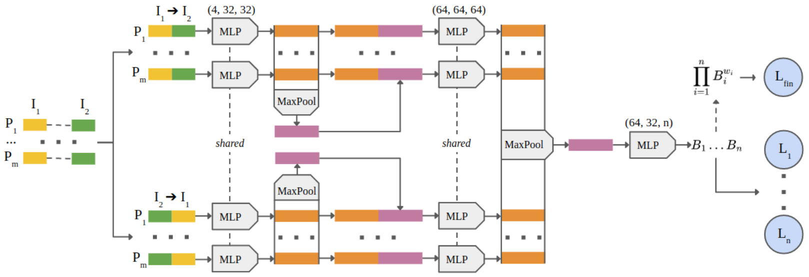


Fig. 2: **NeFSAC's network architecture for minimal samples filtering**. We predict the probability of a minimal sample leading to a good pose only taking its coordinates as input. We implement the main invariants of minimal samples with shared MLPs and channel-wise max pooling aggregation. The last MLP outputs n partial scores used during training, whose power-weighted product is the final score for use in RANSAC. The circular nodes represent the binary cross-entropy loss terms with their respective label. We do not propagate gradients across the dashed arrow.

and smaller model with very little capacity for overfitting. Particularly, since our primary objective is to reduce the computational load in RANSAC, our model of function $\mathcal{F}$ needs to be extremely lightweight. Moreover, our input is structured and the model needs to respect two main invariants: it should be invariant to the ordering of correspondences, and it should be invariant to swapping of the two images. We take inspiration from PointNet [27] and frame our main backbone encoder with shared MLPs that embed each correspondence independently in the same feature space, and then correlate them with channel-wise max pooling, thus preserving permutation invariance. Since our second invariant covers only two combinations, we run our backbone encoder with both alternatives and then max pool its features, before going through a final MLP classifier. We keep the network shallow and thin to keep its run-time negligible with respect to the subsequent RANSAC loop. An overall scheme of our network architecture is represented in Figure 2. Note that our final MLP does not output a single score, but several partial scores whose power-weighted product is the final predicted score. This is to contrast the extreme class imbalance that is encountered with random minimal samples with a novel technique that we detail in Section 3.2.

3.2 Data Preprocessing and Network Supervision

We wish to supervise our network to predict a score for each minimal sample, such that higher scores are given to minimal samples which are more likely to lead to a good model estimation. Given a dataset with image correspondences and ground truth poses, the trivial approach would be to solve for the pose of each minimal sample, label them as positive or negative class based on the maximum between rotation and translation angular error, and train a classifier with binary cross-entropy. While extremely simple, we found that in practice this approach suffers from the extreme class imbalance. First, this is due to the fact that with an inlier ratio of r a minimal sample consisting of m data points has an exponentially lower inlier ratio of r^m. Second, not every all-inlier minimal sample leads to accurate or even meaningful models, as shown in Figure 1b. Depending on the demanded accuracy to define a positive sample, this problem can lead to imbalance rates in the order of the *hundreds* in real datasets, making traditional techniques for unbalanced classification insufficient. Our observation aligns well with the common intuition of how many RANSAC iterations are required in practice to ensure a meaningful estimation of relative pose. We tackle this challenge by proposing to split the prediction of our network into multiple branches: one branch B_1 predicts if the minimal sample is constituted of all inliers (labeled on the maximum Sampson error of its correspondences with respect to the ground truth model), and a second branch B_2 predicts if the minimal sample leads to a good estimation of the pose, *given that* it is constituted of only inliers. In this setting, the first branch learns to score down minimal samples of impossible or unlikely motions, without suffering from the extra imbalance and complexity coming from ill-conditioned samples. The second branch, trained only on full-inlier samples, learns to score down the ill-conditioned configurations leading to noisy models. We underline the importance of this second branch, since such configurations are not only common in practical scenarios, but even detrimental for RANSAC, since they can collect a large consensus over the image [15,21] and lead to erroneous early termination. For this reason, our approach not only can improve run-time, but it can also improve accuracy and robustness of the RANSAC pipeline it is used into.

In some scenarios, the possible real motions can be partly constrained with expert knowledge which could be useful as a prior to our network. For example, in an autonomous driving context we can *usually* assume that both rotation and translation only happen around the vertical axis. We propose to integrate expert knowledge into the minimal sample filter by the use of additional branches $B_3 \dots B_n$, where each branch is tasked with predicting the adherence of a minimal sample to the analytical model defined by the expert. This extra supervision biases the feature extraction network to find features that can be discriminative also for the expert guidance, thus helping every branch with generalization. We detail the expert models used for the autonomous driving application and for PhotoTourism in the supplementary materials.

Finally, since a good sample is composed of all inliers, leads to a good final pose and is conform to expert models, we predict the final score as the product of all the partial scores. Since the different predictive power of each term is not known a-priori, we weight the product of the branches $B_1 \dots B_n$ with weights $w_1 \dots w_n$ at the exponents (i.e., we make a linear combination in log space) and supervise it to predict samples which are both inliers and lead to accurate models. Moreover, we do not propagate gradients to the branch scores $B_1 \dots B_n$ to avoid unstable gradients from the power terms, therefore the branch terms are learned independently of the aggregate score. Overall, our loss function is:

$$\sum_i \mathcal{X}\left(B_i,\ l_i\right) + \mathcal{X}\left(\prod_i B_i^{w_i},\ l_1 l_2\right) \tag{1}$$

Where B_i are the output branches with respective assigned labels l_i and learned branch weights w_i, and $\mathcal{X}$ is the class-weighted cross-entropy loss. Note that, in Equation 3.2, index $i = 1$ refers to the supervision on Sampson error, and index $i = 2$ refers to the supervision on pose error which is only applied to branch B_2 when the minimal sample is outlier-free. The calculation of assigned labels l_i is detailed in the supplementary materials. We did not experiment with tuning different weights for the losses of every branch.

3.3 Filtering Minimal Samples in RANSAC

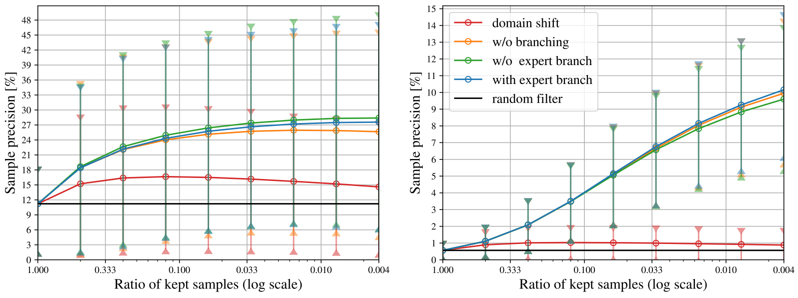
Since our model learns to score minimal samples by the probability that they will lead to a successful pose estimate, in RANSAC we are interested in exploring high-scoring minimal samples first, and have a termination criterion to stop iterating when an accurate model is found. We iteratively take N minimal samples, sort them according to the score predicted by the network, and only process in RANSAC the first $k \ll N$, after which a new batch of N minimal samples is taken only if necessary – as controlled by the RANSAC termination criterion. This procedure guarantees that even in the worst case, when the actual motion does not conform with the learned one and the model degrades back to a random filter, RANSAC still finds the sought model eventually.

We found experimentally that good values are $N = 10000$ and $k = 500$, leading to aggressive filtering, but much lower values ($N = 128$, $k = 12$) also work reasonably well for compute-constrained applications. Processing one full batch takes $1.5ms$ on a RTX2080 GPU, or $20ms$ on an i7 7700K CPU. For simplicity we did not experiment adaptive strategies.

The proposed filtering can be straightforwardly combined with the state-of-the-art pre-emptive model verification strategies and samplers. We use the ones proposed in USAC [28], i.e., Sequential Probability Ratio Test [13], PROSAC sampling [12] and, also, LO-RANSAC [14] to find accurate results.

4 Experiments

In this Section we provide experimental insights into NeFSAC and its impact when integrated into a state-of-the-art RANSAC. We first investigate the quality of its filtering on a pool of random minimal samples, and show that it can improve its average precision (as defined in Section 4.1) by over two times in photo collection scenarios (PhotoTourism [31]) and by over one order of magnitude in strongly motion-constrained scenarios (KITTI [19]). Moreover, the filtering quality generalizes well across extremely different domains. Second, we validate the performance of NeFSAC when integrated in USAC [28], and observe *one order of magnitude* speed-up in practice, together with a *significant improvement* in estimation accuracy.



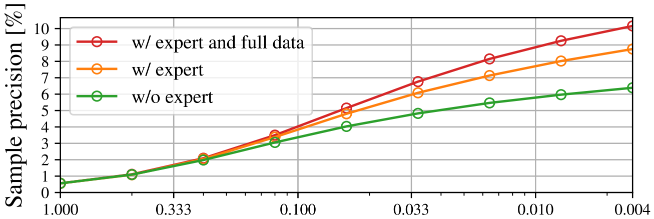
(a) Precision on PhotoTourism

(b) Precision on KITTI

Fig. 3: **Precision of neural filtering.** From a pool of minimal samples for E estimation, we keep the highest-scoring minimal samples according to our model and measure the precision of the kept set, i.e. the rate of samples with less than 10° of rotation and angular translation error, and less than 2 pixels of Sampson error. The distribution of results across images is represented with solid lines for the average and vertical lines for the two middle quartiles. Our method improves the precision of the minimal sample pool by **over one order of magnitude** in the autonomous driving scenario, and over two-fold on PhotoTourism.

4.1 Filtering Accuracy and Ablation Study

In this section, we compare several variations of our filtering network on the quality of the ordering they induce on minimal samples. In practice, for each tested baseline and for each test image, we take a pool of $N = 2^{16}$ random minimal samples, and sort them according to the predicted model score to have the best samples first. We then select the first k minimal samples for filtering rates $r \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ where $k = N/r$, and measure its precision, defined as the rate of samples leading to models with less than 10° of rotation and angular translation error, and less than 2 pixels of Sampson error. We test on a motion-constrained autonomous driving dataset and on a weakly-constrained image collection dataset. We use KITTI [19] for the motion-constrained scenario, and use sequences 0 to 4 for training, sequence 5 for validation and early stopping, and sequences 6 to 10 for testing. We train and validate using random frame differences between 1 and 7, and test with random frame differences between 1 and 5. In KITTI, we mostly see forward and turning motions with limited speed, thus a strong motion prior can be learned. For the weakly-constrained scenario, we use the PhotoTourism [31] data from the 2020 CVPR RANSAC Tutorial [1], with the suggested train-validation-test split. This dataset does not consist of image sequences but rather of crowd-sourced image collections of famous landmarks, therefore the motion prior that this data can exhibit is very limited. However, there are still distinctive motion patterns originating, e.g., from the fact that people usually align their photos with the gravity direction and, also, the range of translations along the vertical axis is limited.



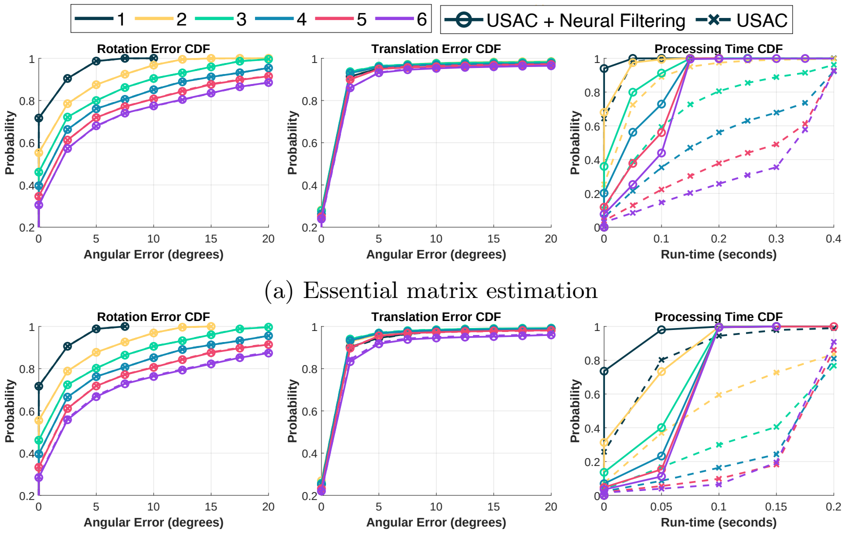
(a) Precision on PhotoTourism

Fig. 4: **Impact of expert branches.** The same evaluation as in Figure 3b, performed on KITTI and training with only 250 image pairs from sequence 0 and frame difference 4. The expert branch significantly helps preserving the filtering accuracy in data-scarce conditions. We show only averages for clarity.

We compare the following alternatives: (i) **w/o expert branch**: NeFSAC is trained only with branches B_1 and B_2 with Sampson error and pose error, as described in Section 3.2. (ii) **with expert branch**: NeFSAC, in addition to B_1 and B_2, is trained with a further branch B_3 with the expert supervision defined in the supplementary material. (iii) **w/o branching**: NeFSAC is directly trained to infer the complete label $l_1 l_2$ indicating low Sampson error and low pose error. Note that this is the same metric we use for testing, therefore this baseline has an intrinsic advantage in the testing process. Nonetheless, we show that branching leads to superior performance. (iv) **domain shift**: NeFSAC has been trained on a significantly different dataset from the one at test. We use the model trained on KITTI for the test on PhotoTourism, and the model trained on PhotoTourism for the test on KITTI. We use no expert branch for this baseline.

In Figure 3a, we show results on 4000 image pairs from the PhotoTourism validation set. Our neural filtering model without any expert branch improves precision (as defined above) by 2.5 times on average at peak filtering rates compared to the original minimal sample pool. Expert knowledge, being inaccurate and hard to formulate in this context, has a slightly detrimental influence. The baseline without any branching, even though it is the only one trained end-to-end to optimize the test metric, does not keep up with the branched alternatives on higher filtering rates. Even with an extreme domain gap, the model trained on KITTI still manages to improve the quality of the original sample pool, showing that NeFSAC is very robust to distribution shifts. We attribute this robustness to the limited information that the neural filtering has at inference time, since our model never observes the global configuration of correspondences.

In Figure 3b, we show results on 4000 image pairs from the KITTI sequences 6 to 10. The strong motion statistics on this dataset allow NeFSAC to improve the precision of the minimal sample pool **by over one order of magnitude** (18x) at peak filtering rates. Filtering is close-to-perfect up until 25% keep rate. On this dataset, we do not observe significant differences between the three main variants, likely due to the presence of a very simple and discriminative motion



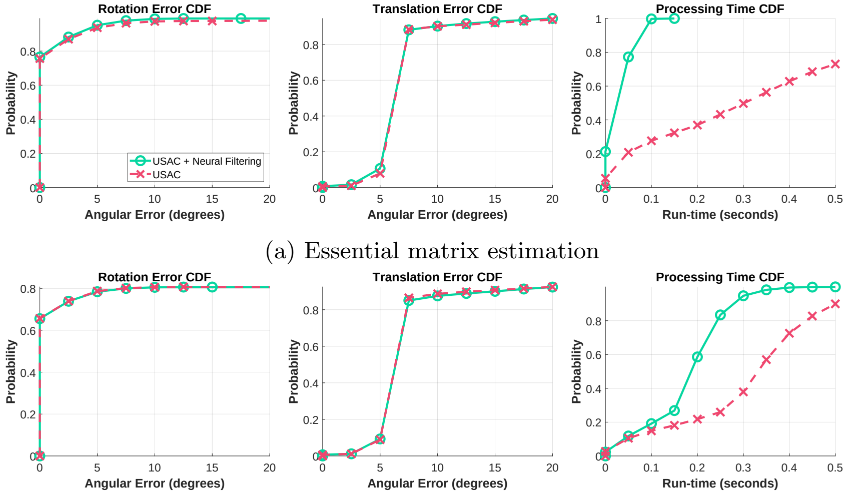
(a) Essential matrix estimation

(b) Fundamental matrix estimation

Fig. 5: **CDF on KITTI**. The cumulative distribution functions (CDF) of the rotation and translation errors (degrees) and run-time (seconds) of epipolar geometry estimation by USAC with and without the proposed neural sample filtering on 47700 image pairs from sequences 6–10 from the KITTI dataset [19]. Colors indicate the frame difference, e.g. the red curve uses a frame difference of 5. The corresponding mean values are in Table 1.

model that is well learned by all the baselines trained on KITTI. However, we can observe that the model with expert branch performs the best on this domain, where the expert supervision is more adherent to the real dataset statistics. Interestingly, we found that the learned weight w_3 on the expert branch of this model is close to zero: this branch is not playing a significant role in the final prediction of the network, even though it still had a positive impact as a prior in learning good features for the other branches during training. The model trained on PhotoTourism, not taking advantage of the restricted motion statistics of this test set, still improves the precision of the original minimal sample pool by a factor up to 2, very close to its original performance on the PhotoTourism test.

Finally, in Figure 4a, we test the influence of our expert branch in conditions of data scarcity. We keep the same evaluation protocol and test set as in Figure 3b, but we train NeFSAC only on 250 image pairs from KITTI, all taken from sequence 0 and fixed frame difference of 4. We also report the baseline NeFSAC trained on the full training set for comparability. We observe that the impact of the expert branch is very significant when little training data is available, halving the gap to the baseline filtering accuracy. This scenario can be very important when training NeFSAC in a new domain with limited data.



(a) Essential matrix estimation

(b) Fundamental matrix estimation

Fig. 6: **CDF on Malaga**. The cumulative distribution functions (CDF) of the rotation and translation errors (degrees) and run-time (seconds) of epipolar geometry estimation by USAC [28] with and without the proposed neural minimal sample filtering on 27147 image pairs from the Malaga dataset [7]. The corresponding mean values are in Table 1.

4.2 Comparative Experiments within RANSAC

The interaction between several components in RANSAC is non-trivial: having observed an improvement of ten times in the average precision of a pool of random minimal samples does not necessarily translate into the equivalent speed-up of ten times in any RANSAC. In this section, we examine the effect of NeFSAC on a representative RANSAC variant with state-of-the-art components. We choose USAC [28] with cheirality tests, PROSAC sampling [12], LO-RANSAC [14], and SPRT [13] as preemptive verification. While there might be variants leading to better accuracy, e.g. MAGSAC++ [3], their low run-times still come from SPRT and PROSAC, therefore similar speed-ups are expected. We show additional experiments in the supplementary material.

In the following, we compare the rotation error ϵ_R, translation error ϵ_t and run-time t of USAC with and without NeFSAC filtering for the case of autonomous driving and for the case of unstructured image collections.

Autonomous driving. We train NeFSAC on KITTI [19] on sequences 0 to 4 with random frame differences between 1 and 7, and use sequence 5 for validation. We train separate models for Essential matrix estimation and Fundamental matrix estimation. We then test such models on the KITTI sequences 6 to 10 as well as on the Malaga [7] dataset to test for generalization. We report results in Figures 5 and 6 and Table 1. On KITTI, NeFSAC+USAC is over three times

	KITTI (47700 pairs)				Malaga (27147 pairs)			
USAC	$\epsilon_{\mathbf{R}}$ (°)	$\epsilon_{\mathbf{t}}$ (°)	t (ms)	# models	$\epsilon_{\mathbf{R}}$ (°)	$\epsilon_{\mathbf{t}}$ (°)	t (ms)	# models
w/o NF (**E**)	**4.3**	2.5	234.8	941	3.3	8.9	350.0	3225
w/ NF (**E**)	**4.3**	**2.3**	**69.7**	**260**	**1.9**	**8.7**	**34.0**	**753**
w/o NF (**F**)	**4.2**	2.7	213.4	1974	**1.4**	9.0	380.2	3837
w/ NF (**F**)	**4.2**	**2.3**	**85.9**	**357**	**1.4**	**8.6**	**77.1**	**467**

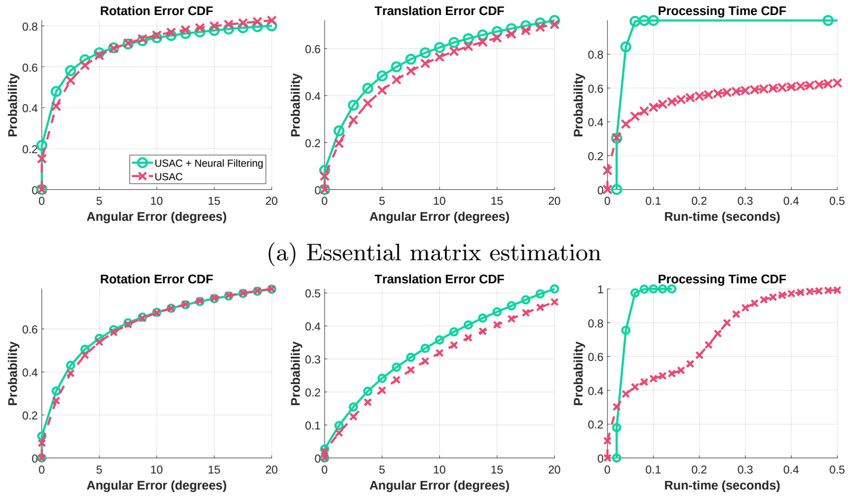
Table 1: **Results on KITTI and Malaga**. The average rotation and translation errors (degrees), the run-time (milliseconds), and the number of models tested inside USAC [28] on the KITTI [19] and Malaga [7] datasets for essential (**E**) and fundamental matrix (**F**) estimation. NeFSAC provides great speed-ups while improving accuracy as well. The corresponding CDFs are in Figs. 5,6.

	Essential matrix				Fundamental matrix			
USAC	$\epsilon_{\mathbf{R}}$ (°)	$\epsilon_{\mathbf{t}}$ (°)	t (ms)	# models	$\epsilon_{\mathbf{R}}$ (°)	$\epsilon_{\mathbf{t}}$ (°)	t (ms)	# models
w/o NF	2.7	7.9	805.1	4550	4.8	22.5	154.6	5559
w/ NF	**2.1**	**6.1**	**76.5**	**364**	**3.9**	**17.9**	61.7	764
w/ NF*	2.6	7.8	103.0	660	4.8	22.3	**61.2**	**740**

Table 2: **Results on PhotoTourism**. The median rotation and translation errors (degrees), the average run-time (milliseconds), and the number of models tested inside USAC [28] on the PhotoTourism [31] (from [1]; 52200 image pairs) dataset. NF* is trained on KITTI [19]. NeFSAC provides great speed-ups while improving accuracy as well. The corresponding CDFs are in Fig. 7.

faster on E estimation and more than twice as fast on F estimation compared to USAC, with slightly better accuracy. On Malaga, NeFSAC achieves a **ten-fold** speed-up and reduces the average rotation error by 1.4 degrees on E estimation, and a five-fold speed-up on F estimation, **despite being trained on KITTI**.

PhotoTourism. We train NeFSAC on the PhotoTourism [31] data provided by the 2020 CVPR RANSAC Tutorial [1]. We use the standard split for training, validation and test. Results are reported in Figure 7 and Table 2. NeFSAC improves run-time **again by one order of magnitude** on E estimation and two-fold on F estimation, while providing an **important improvement of accuracy** on both. Our aggressive filtering setup, tuned for challenging image pairs, causes a small overhead on the easy tail of the distribution, suggesting the use of adaptive strategies. We further perform an extreme generalization test training NeFSAC on KITTI (reported as NF* in Table 2). While the model trained on PhotoTourism is superior, NeFSAC still manages to bring a very significant speed-up and some improvement in accuracy even under such extreme domain shift. This motivates us to claim that our model is partly learning very general knowledge on the task, and can robustly stay well above the worst-case scenario where it degenerates to the baseline RANSAC.



(a) Essential matrix estimation

(b) Fundamental matrix estimation

Fig. 7: **CDF on PhotoTourism**. The cumulative distribution functions (CDF) of the rotation and translation errors (in degrees) and run-time (in seconds) of epipolar geometry estimation by USAC [28] with and without the proposed neural minimal sample filtering on a total of 52200 image pairs from the PhotoTourism dataset as used in [1]. The corresponding mean values are in Table 2.

5 Conclusions

In this paper we proposed NeFSAC, a novel framework for Neural Filtering of minimal samples in RANSAC that can be seamlessly integrated into any existing RANSAC pipeline. NeFSAC learns to predict the quality of minimal samples by their crude pixel coordinates to filter out the ones consistent with unlikely or impossible motions and common poorly-conditioned configurations. We showed that NeFSAC can learn stronger filters when a constrained motion is present in the training data, but can be very discriminative even in datasets without a strong motion prior, like in general image collections, while being very robust to domain shifts. We showed that, in practice, NeFSAC can reduce the run-time by **one order of magnitude** in modern state-of-the-art RANSAC variants on Essential and Fundamental matrix estimation while often significantly *improving* estimation accuracy.

Acknowledgments: This work was supported by the ETH Zurich Postdoctoral Fellowship and the Google Focused Research Award.

References

1. Barath, D., Chin, T.J., Chum, O., Mishkin, D., Ranftl, R., Matas, J.: RANSAC in 2020 tutorial. In: CVPR (2020), `http://cmp.felk.cvut.cz/cvpr2020-ransac-tutorial/`
2. Barath, D., Matas, J.: Graph-cut RANSAC. In: CVPR. pp. 6733–6741 (2018)
3. Barath, D., Noskova, J., Ivashechkin, M., Matas, J.: MAGSAC++, a fast, reliable and accurate robust estimator. In: CVPR (2020)
4. Barath, D., Noskova, J., Matas, J.: MAGSAC: marginalizing sample consensus. In: CVPR (2019), https://github.com/danini/magsac
5. Barath, D., Noskova, J., Matas, J.: Marginalizing sample consensus. IEEE TPAMI (2021)
6. Bian, J., Lin, W.Y., Matsushita, Y., Yeung, S.K., Nguyen, T.D., Cheng, M.M.: Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: CVPR. pp. 4181–4190 (2017)
7. Blanco-Claraco, J.L., Moreno-Duenas, F.A., González-Jiménez, J.: The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario. IJRR **33**(2), 207–214 (2014)
8. Brachmann, E., Rother, C.: Neural-guided RANSAC: Learning where to sample model hypotheses. In: CVPR. pp. 4322–4331 (2019)
9. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: Dsac-differentiable ransac for camera localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6684–6692 (2017)
10. Cavalli, L., Larsson, V., Oswald, M.R., Sattler, T., Pollefeys, M.: Handcrafted outlier detection revisited. In: ECCV. pp. 770–787. Springer (2020)
11. Chum, O., Matas, J.: Randomized RANSAC with tdd test. In: BMVC. vol. 2, pp. 448–457 (2002)
12. Chum, O., Matas, J.: Matching with PROSAC-progressive sample consensus. In: CVPR. IEEE (2005)
13. Chum, O., Matas, J.: Optimal randomized RANSAC. TPAMI **30**(8), 1472–1482 (2008)
14. Chum, O., Matas, J., Kittler, J.: Locally optimized ransac. In: Joint Pattern Recognition Symposium. Springer (2003)
15. Chum, O., Werner, T., Matas, J.: Two-view geometry estimation unaffected by a dominant plane. In: CVPR. IEEE (2005)
16. Ding, Y., Barath, D., Kukelova, Z.: Minimal solutions for panoramic stitching given gravity prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5579–5588 (2021)
17. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM (1981)
18. Frahm, J.M., Pollefeys, M.: Ransac for (quasi-) degenerate data (qdegsac). In: CVPR. vol. 1, pp. 453–460. IEEE (2006)
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: CVPR. IEEE (2012)
20. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
21. Ivashechkin, M., Barath, D., Matas, J.: VSAC: Efficient and accurate estimator for h and f. ICCV (2021)

22. Lebeda, K., Matas, J., Chum, O.: Fixing the locally optimized RANSAC. In: BMVC. Citeseer (2012)
23. Matas, J., Chum, O.: Randomized RANSAC with sequential probability ratio test. In: ICCV. vol. 2, pp. 1727–1732. IEEE (2005)
24. Moisan, L., Moulon, P., Monasse, P.: Automatic homographic registration of a pair of images, with a contrario elimination of outliers. Image Processing On Line **2**, 56–73 (2012)
25. Moo Yi, K., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: CVPR. pp. 2666–2674 (2018)
26. Ni, K., Jin, H., Dellaert, F.: GroupSAC: Efficient consensus in the presence of groupings. In: ICCV. pp. 2193–2200. IEEE (2009)
27. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017)
28. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: USAC: a universal framework for random sample consensus. TPAMI (2013), `https://www.cs.unc.edu/~rraguram/usac`
29. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: ECCV. pp. 284–299 (2018)
30. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR. pp. 4938–4947 (2020)
31. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM siggraph 2006 papers, pp. 835–846 (2006)
32. Stewart, C.V.: MINPRAN: A new robust estimator for computer vision. TPAMI **17**(10), 925–938 (1995)
33. Tong, W., Matas, J., Barath, D.: Deep magsac++. arXiv preprint arXiv:2111.14093 (2021)
34. Torr, P.H.S.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. IJCV (2002)
35. Torr, P.H.S., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. CVIU (2000)
36. Torr, P.H., Nasuto, S.J., Bishop, J.M.: Napsac: High noise, high dimensional robust estimation-it's in the bag. In: BMVC. vol. 2, p. 3 (2002)
37. Werner, T., Pajdla, T.: Cheirality in epipolar geometry. In: ICCV. vol. 1, pp. 548–553. IEEE (2001)
38. Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H.: Learning two-view correspondences and geometry using order-aware network. In: CVPR. pp. 5845–5854 (2019)