

# Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses

Eric Brachmann and Carsten Rother  
 Visual Learning Lab  
 Heidelberg University (HCI/IWR)

<http://vislearn.de>

## Abstract

We present Neural-Guided RANSAC (NG-RANSAC), an extension to the classic RANSAC algorithm from robust optimization. NG-RANSAC uses prior information to improve model hypothesis search, increasing the chance of finding outlier-free minimal sets. Previous works use heuristic side-information like hand-crafted descriptor distance to guide hypothesis search. In contrast, we learn hypothesis search in a principled fashion that lets us optimize an arbitrary task loss during training, leading to large improvements on classic computer vision tasks. We present two further extensions to NG-RANSAC. Firstly, using the inlier count itself as training signal allows us to train neural guidance in a self-supervised fashion. Secondly, we combine neural guidance with differentiable RANSAC to build neural networks which focus on certain parts of the input data and make the output predictions as good as possible. We evaluate NG-RANSAC on a wide array of computer vision tasks, namely estimation of epipolar geometry, horizon line estimation and camera re-localization. We achieve superior or competitive results compared to state-of-the-art robust estimators, including very recent, learned ones.

## 1. Introduction

Despite its simplicity and time of invention, Random Sample Consensus (RANSAC) [12] remains an important method for robust optimization, and is a vital component of many state-of-the-art vision pipelines [39, 40, 29, 6]. RANSAC allows accurate estimation of model parameters from a set of observations of which some are outliers. To this end, RANSAC iteratively chooses random sub-sets of observations, so called minimal sets, to create model hypotheses. Hypotheses are ranked according to their consensus with all observations, and the top-ranked hypothesis is returned as the final estimate.

The main limitation of RANSAC is its poor performance in domains with many outliers. As the ratio of outliers increases, RANSAC requires exponentially many iterations to find an outlier-free minimal set. Implementations of RANSAC therefore often restrict the maximum number of iterations, and return the best model found so far [7].

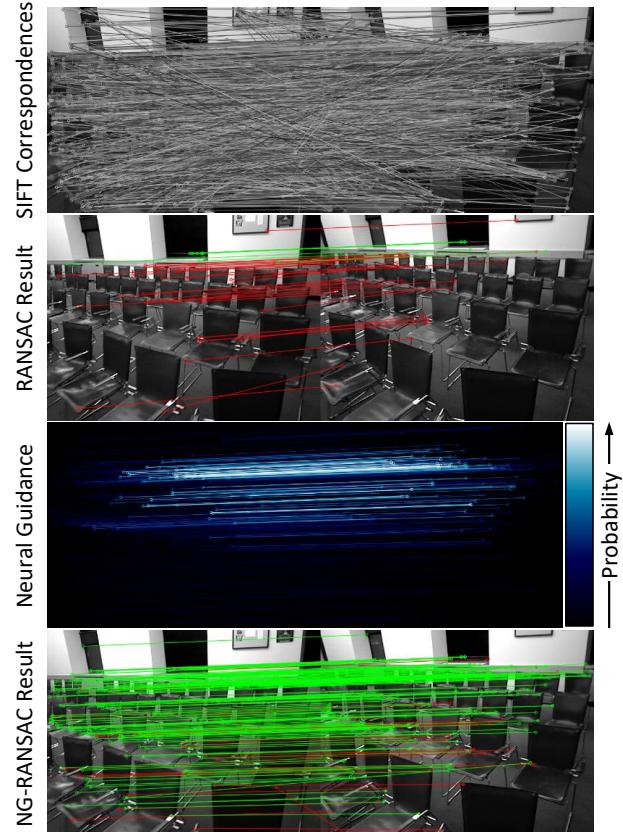


Figure 1. RANSAC vs. NG-RANSAC. We extract 2000 SIFT correspondences between two images. With an outlier rate of 88%, RANSAC fails to find the correct relative transformation (green correct and red wrong matches). We use a neural network to predict a probability distribution over correspondences. Over 90% of the probability mass falls onto 239 correspondences with an outlier rate of 33%. NG-RANSAC samples minimal sets according to this distribution, and finds the correct transformation up to an angular error of less than  $1^\circ$ .

In this work, we combine RANSAC with a neural network that predicts a weight for each observation. The weights ultimately guide the sampling of minimal sets. We call the resulting algorithm Neural-Guided RANSAC (NG-RANSAC). A comparison of our method with vanilla RANSAC can be seen in Fig. 1.

When developing NG-RANSAC, we took inspiration from recent work on learned robust estimators [56, 36]. In particular, Yi *et al.* [56] train a neural network to classify observations as outliers or inliers, fitting final model parameters only to the latter. Although designed to replace RANSAC, their method achieves best results when combined with RANSAC during test time, where it would remove any outliers that the neural network might have missed. This motivates us to train the neural network in conjunction with RANSAC in a principled fashion, rather than imposing it afterwards.

Instead of interpreting the neural network output as soft inlier labels for a robust model fit, we let the output weights guide RANSAC hypothesis sampling. Intuitively, the neural network should learn to decrease weights for outliers, and increase them for inliers. This paradigm yields substantial flexibility for the neural network in allowing a certain misclassification rate without negative effects on the final fitting accuracy due to the robustness of RANSAC. The distinction between inliers and outliers, as well as which misclassifications are tolerable, is solely guided by the minimization of the task loss function during training. Furthermore, our formulation of NG-RANSAC facilitates training with any (non-differentiable) task loss function, and any (non-differentiable) model parameter solver, making it broadly applicable. For example, when fitting essential matrices, we may use the 5-point algorithm rather than the (differentiable) 8-point algorithm which other learned robust estimators rely on [56, 36]. The flexibility in choosing the task loss also allows us to train NG-RANSAC self-supervised by using maximization of the inlier count as training objective.

The idea of using guided sampling in RANSAC is not new. Tordoff and Murray first proposed to guide the hypothesis search of MLESAC [48], using side-information [47]. They formulated a prior probability of sparse feature matches being valid based on matching scores. While this has a positive affect on RANSAC performance in some applications, feature matching scores, or other hand-crafted heuristics, were clearly *not designed* to guide hypothesis search. In particular, calibration of such ad-hoc measures can be difficult as the reliance on over-confident but wrong prior probabilities can yield situations where the same few observations are sampled repeatedly. This fact was recognized by Chum and Matas who proposed PROSAC [9], a variant of RANSAC that uses side-information only to change *the order* in which RANSAC draws minimal sets. In the worst case, if the side-information was not useful at all, their method would degenerate to vanilla RANSAC. NG-RANSAC takes a different approach in (i) learning the weights to guide hypothesis search rather than using hand-crafted heuristics, and (ii) integrating RANSAC itself in the training process which leads to self-calibration of the predicted weights.

Recently, Brachmann *et al.* proposed differentiable RANSAC (DSAC) to learn a camera re-localization pipeline [4]. Unfortunately, we can not directly use DSAC to learn hypothesis sampling since DSAC is only differentiable w.r.t. to observations, not sampling weights. However, NG-RANSAC applies a similar trick also used to make DSAC differentiable, namely the optimization of the expected task loss during training. While we do not rely on DSAC, neural guidance can be used in conjunction with DSAC (NG-DSAC) to train neural networks that predict observations and observation confidences at the same time.

We summarize our main contributions:

- We present NG-RANSAC, a formulation of RANSAC with learned guidance of hypothesis sampling. We can use any (non-differentiable) task loss, and any (non-differentiable) minimal solver for training.
- Choosing the inlier count itself as training objective facilitates self-supervised learning of NG-RANSAC.
- We use NG-RANSAC to estimate epipolar geometry of image pairs from sparse correspondences, where it surpasses competing robust estimators.
- We combine neural guidance with differentiable RANSAC (NG-DSAC) to train neural networks that make accurate predictions for parts of the input, while neglecting other parts. These models achieve competitive results for horizontal line estimation, and state-of-the-art for camera re-localization.

## 2. Related Work

RANSAC was introduced in 1981 by Fischler and Bolles [12]. Since then it was extended in various ways, see *e.g.* the survey by Raguram *et al.* [35]. Combining some of the most promising improvements, Raguram *et al.* created the Universal RANSAC (USAC) framework [34] which represents the state-of-the-art of classic RANSAC variants. USAC includes guided hypothesis sampling according to PROSAC [9], more accurate model fitting according to Locally Optimized RANSAC [11], and more efficient hypothesis verification according to Optimal Randomized RANSAC [10]. Many of the improvements proposed for RANSAC could also be applied to NG-RANSAC since we do not require any differentiability of such add-ons. We only imposes restrictions on how to generate hypotheses, namely according to a learned probability distribution.

RANSAC is not often used in recent machine learning-heavy vision pipelines. Notable exceptions include geometric problems like object instance pose estimation [3, 5, 21], and camera re-localization [41, 51, 28, 8, 46] where RANSAC is coupled with decision forests or neural networks that predict image-to-object correspondences. However, in most of these works, RANSAC is not part of the training process because of its non-differentiability. DSAC [4, 6] overcomes this limitation by making the hypothesis

selection a probabilistic action which facilitates optimization of the expected task loss during training. However, DSAC is limited in *which* derivatives can be calculated. DSAC allows differentiation w.r.t. to observations. For example, we can use it to calculate the gradient of image coordinates for a sparse correspondence. However, DSAC does not model observation selection, and hence we cannot use it to optimize a matching probability. By showing how to learn neural guidance, we close this gap. The combination with DSAC enables the full flexibility of learning both, observations and their selection probability.

Besides DSAC, a *differentiable* robust estimator, there has recently been some work on *learning* robust estimators. We discussed the work of Yi *et al.* [56] in the introduction. Ranftl and Koltun [36] take a similar but iterative approach reminiscent of Iteratively Reweighted Least Squares (IRLS) for fundamental matrix estimation. In each iteration, a neural network predicts observation weights for a weighted model fit, taking into account the residuals of the last iteration. Both, [56] and [36], have shown considerable improvements w.r.t. to vanilla RANSAC but require differentiable minimal solvers, and task loss functions. NG-RANSAC outperforms both approaches, and is more flexible when it comes to defining the training objective. This flexibility also enables us to train NG-RANSAC in a self-supervised fashion, possible with neither [56] nor [36].

### 3. Method

**Preliminaries.** We address the problem of fitting model parameters  $\mathbf{h}$  to a set of observations  $\mathbf{y} \in \mathcal{Y}$  that are contaminated by noise and outliers. For example,  $\mathbf{h}$  could be a fundamental matrix that describes the epipolar geometry of an image pair [16], and  $\mathcal{Y}$  could be the set of SIFT correspondences [27] we extract for the image pair. To calculate model parameters from the observations, we utilize a solver  $f$ , for example the 8-point algorithm [15]. However, calculating  $\mathbf{h}$  from all observations will result in a poor estimate due to outliers. Instead, we can calculate  $\mathbf{h}$  from a small subset (minimal set) of observations with cardinality  $N$ :  $\mathbf{h} = f(\mathbf{y}_1, \dots, \mathbf{y}_N)$ . For example, for a fundamental matrix  $N = 8$  when using the 8-point algorithm. RANSAC [12] is an algorithm to chose an outlier-free minimal set from  $\mathcal{Y}$  such that the resulting estimate  $\mathbf{h}$  is accurate. To this end, RANSAC randomly chooses  $M$  minimal sets to create a pool of model hypotheses  $\mathcal{H} = (\mathbf{h}_1, \dots, \mathbf{h}_M)$ .

RANSAC includes a strategy to adaptively choose  $M$ , based on an online estimate of the outlier ratio [12]. The strategy guarantees that an outlier-free set will be sampled with a user-defined probability. For tasks with large outlier ratios,  $M$  calculated like this can be exponentially large, and is usually clamped to a maximum value [7]. For notational simplicity, we take the perspective of a fixed  $M$  but do not restrict the use of an early-stopping strategy in practice.

RANSAC chooses a model hypothesis as the final estimate  $\hat{\mathbf{h}}$  according to a scoring function  $s$ :

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathcal{H}}{\operatorname{argmax}} s(\mathbf{h}, \mathcal{Y}). \quad (1)$$

The scoring function measures the consensus of an hypothesis w.r.t. all observations, and is traditionally implemented as inlier counting [12].

**Neural Guidance.** RANSAC chooses observations uniformly random to create the hypothesis pool  $\mathcal{H}$ . We aim at sampling observations according to a learned distribution instead that is parametrized by a neural network with parameters  $\mathbf{w}$ . That is, we select observations according to  $\mathbf{y} \sim p(\mathbf{y}; \mathbf{w})$ . Note that  $p(\mathbf{y}; \mathbf{w})$  is a categorical distribution over the discrete set of observations  $\mathcal{Y}$ , *not* a continuous distribution in observation space. We wish to learn parameters  $\mathbf{w}$  in a way that increases the chance of selecting outlier-free minimal sets, which will result in accurate estimates  $\hat{\mathbf{h}}$ . We sample a hypothesis pool  $\mathcal{H}$  according to  $p(\mathcal{H}; \mathbf{w})$  by sampling observations and minimal sets independently, *i.e.*

$$p(\mathcal{H}; \mathbf{w}) = \prod_{j=1}^M p(\mathbf{h}_j; \mathbf{w}), \text{ with } p(\mathbf{h}; \mathbf{w}) = \prod_{i=1}^N p(\mathbf{y}_i; \mathbf{w}). \quad (2)$$

From a pool  $\mathcal{H}$ , we estimate model parameters  $\hat{\mathbf{h}}$  with RANSAC according to Eq. 1. For training, we assume that we can measure the quality of the estimate with a task loss function  $\ell(\hat{\mathbf{h}})$ . The task loss can be calculated w.r.t. a ground truth model  $\mathbf{h}^*$ , or self-supervised, *e.g.* by using the inlier count of the final estimate:  $\ell(\hat{\mathbf{h}}) = -s(\hat{\mathbf{h}}, \mathcal{Y})$ . We wish to learn the distribution  $p(\mathcal{H}; \mathbf{w})$  in a way that we receive a small task loss with high probability. Inspired by DSAC [4], we define our training objective as the minimization of the expected task loss:

$$\mathcal{L}(\mathbf{w}) = \mathbb{E}_{\mathcal{H} \sim p(\mathcal{H}; \mathbf{w})} [\ell(\hat{\mathbf{h}})]. \quad (3)$$

We compute the gradients of the expected task loss w.r.t. the network parameters as

$$\frac{\partial}{\partial \mathbf{w}} \mathcal{L}(\mathbf{w}) = \mathbb{E}_{\mathcal{H}} \left[ \ell(\hat{\mathbf{h}}) \frac{\partial}{\partial \mathbf{w}} \log p(\mathcal{H}; \mathbf{w}) \right]. \quad (4)$$

Integrating over all possible hypothesis pools to calculate the expectation is infeasible. Therefore, we approximate the gradients by drawing  $K$  samples  $\mathcal{H}_k \sim p(\mathcal{H}; \mathbf{w})$ :

$$\frac{\partial}{\partial \mathbf{w}} \mathcal{L}(\mathbf{w}) \approx \frac{1}{K} \sum_{k=1}^K \left[ \ell(\hat{\mathbf{h}}) \frac{\partial}{\partial \mathbf{w}} \log p(\mathcal{H}_k; \mathbf{w}) \right]. \quad (5)$$

Note that gradients of the task loss function  $\ell$  do *not* appear in the expression above. Therefore, differentiability of the

task loss  $\ell$ , the robust solver  $\hat{\mathbf{h}}$  (*i.e.* RANSAC) or the minimal solver  $f$  is not required. These components merely generate a training signal for steering the sampling probability  $p(\mathcal{H}; \mathbf{w})$  in a good direction. Due to the approximation by sampling, the gradient variance of Eq. 5 can be high. We apply a standard variance reduction technique from reinforcement learning by subtracting a baseline  $b$  [45]:

$$\frac{\partial}{\partial \mathbf{w}} \mathcal{L}(\mathbf{w}) \approx \frac{1}{K} \sum_{k=1}^K \left[ [\ell(\hat{\mathbf{h}}) - b] \frac{\partial}{\partial \mathbf{w}} \log p(\mathcal{H}_k; \mathbf{w}) \right]. \quad (6)$$

We found a simple baseline in the form of the average loss per image sufficient, *i.e.*  $b = \bar{\ell}$ . Subtracting the baseline will move the probability distribution towards hypothesis pools with lower-than-average loss for each training example.

**Combination with DSAC.** Brachmann *et al.* [4] proposed a RANSAC-based pipeline where a neural network with parameters  $\mathbf{w}$  predicts observations  $\mathbf{y}(\mathbf{w}) \in \mathcal{Y}(\mathbf{w})$ . End-to-end training of the pipeline, and therefore learning the observations  $\mathbf{y}(\mathbf{w})$ , is possible by turning the argmax hypothesis selection of RANSAC (cf. Eq. 1) into a probabilistic action:

$$\hat{\mathbf{h}}_{\text{DSAC}} = \mathbf{h}_j \sim p(j|\mathcal{H}) = \frac{\exp s(\mathbf{h}_j, \mathcal{Y}(\mathbf{w}))}{\sum_{k=1}^M \exp s(\mathbf{h}_k, \mathcal{Y}(\mathbf{w}))}. \quad (7)$$

This differentiable variant of RANSAC (DSAC) chooses a hypothesis randomly according to a distribution calculated from hypothesis scores. The training objective aims at learning network parameters such that hypotheses with low task loss are chosen with high probability:

$$\mathcal{L}_{\text{DSAC}}(\mathbf{w}) = \mathbb{E}_{j \sim p(j)} [\ell(\mathbf{h}_j)]. \quad (8)$$

In the following, we extend the formulation of DSAC with neural guidance (NG-DSAC). We let the neural network predict observations  $\mathbf{y}(\mathbf{w})$  and, additionally, a probability associated with each observation  $p(\mathbf{y}; \mathbf{w})$ . Intuitively, the neural network can express a confidence in its own predictions through this probability. This can be useful if a certain input for the neural network contains no information about the desired model  $\mathbf{h}$ . In this case, the observation prediction  $\mathbf{y}(\mathbf{w})$  is necessarily an outlier, and the best the neural network can do is to label it as such by assigning a low probability. We combine the training objectives of NG-RANSAC (Eq. 3) and DSAC (Eq. 8) which yields:

$$\mathcal{L}_{\text{NG-DSAC}}(\mathbf{w}) = \mathbb{E}_{\mathcal{H} \sim p(\mathcal{H}; \mathbf{w})} \mathbb{E}_{j \sim p(j|\mathcal{H})} [\ell(\mathbf{h}_j)], \quad (9)$$

where we again construct  $p(\mathcal{H}; \mathbf{w})$  from individual  $p(\mathbf{y}; \mathbf{w})$ 's according to Eq. 2. The training objective of NG-DSAC consists of two expectations. Firstly, the expectation w.r.t. sampling a hypothesis pool according to the probabilities predicted by the neural network. Secondly, the expectation w.r.t. sampling a final estimate from the pool according

to the scoring function. As in NG-RANSAC, we approximate the first expectation via sampling, as integrating over all possible hypothesis pools is infeasible. For the second expectation, we can calculate it analytically, as in DSAC, since it integrates over the discrete set of hypotheses  $\mathbf{h}_j$  in a given pool  $\mathcal{H}$ . Similar to Eq. 6, we give the approximate gradients  $\frac{\partial}{\partial \mathbf{w}} \mathcal{L}(\mathbf{w})$  of NG-DSAC as:

$$\frac{1}{K} \sum_{k=1}^K \left[ [\mathbb{E}_j [\ell] - b] \frac{\partial}{\partial \mathbf{w}} \log p(\mathcal{H}_k; \mathbf{w}) + \frac{\partial}{\partial \mathbf{w}} \mathbb{E}_j [\ell] \right], \quad (10)$$

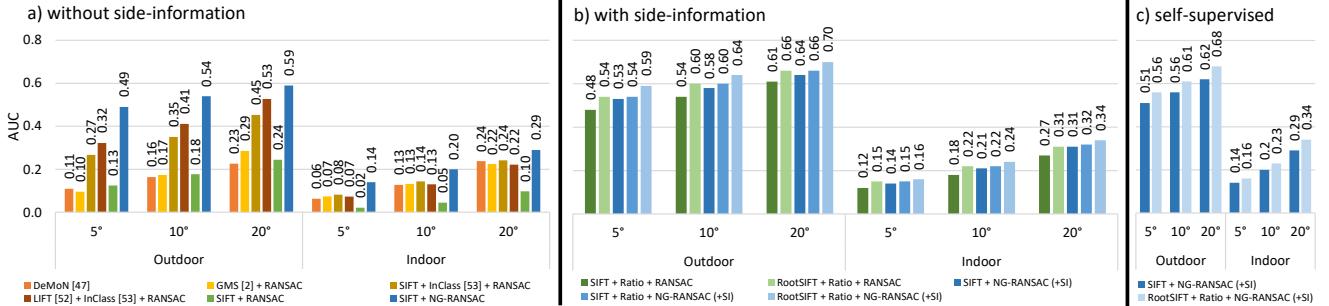
where we use  $\mathbb{E}_j [\ell]$  as a stand-in for  $\mathbb{E}_{j \sim p(j|\mathcal{H})} [\ell(\mathbf{h}_j)]$ . The calculation of gradients for NG-DSAC requires the derivative of the task loss (note the last part of Eq. 10) because  $\mathbb{E}_j [\ell]$  depends on parameters  $\mathbf{w}$  via observations  $\mathbf{y}(\mathbf{w})$ . Therefore, training NG-DSAC requires a differentiable task loss function  $\ell$ , a differentiable scoring function  $s$ , and a differentiable minimal solver  $f$ . Note that we inherit these restrictions from DSAC. In return, NG-DSAC allows for learning observations and observation confidences, at the same time.

## 4. Experiments

We evaluate neural guidance on multiple, classic computer vision tasks. Firstly, we apply NG-RANSAC to estimating epipolar geometry of image pairs in the form of essential matrices and fundamental matrices. Secondly, we apply NG-DSAC to horizon line estimation and camera relocalization. We present the main experimental results here, and refer to the appendix for details about network architectures, hyper-parameters and more qualitative results. Our implementation is based on PyTorch [32], and we will make the code publicly available for all the tasks discussed below.

### 4.1. Essential Matrix Estimation

Epipolar geometry describes the geometry of two images that observe the same scene [16]. In particular, two image points  $\mathbf{x}$  and  $\mathbf{x}'$  in the left and right image corresponding to the same 3D point satisfy  $\mathbf{x}'^\top F \mathbf{x} = 0$ , where the  $3 \times 3$  matrix  $F$  denotes the fundamental matrix. We can estimate  $F$  uniquely (but only up to scale) from 8 correspondences, or from 7 correspondences with multiple solutions [16]. The essential matrix  $E$  is a special case of the fundamental matrix when the calibration parameters  $K$  and  $K'$  of both cameras are known:  $E = K'^\top F K$ . The essential matrix can be estimated from 5 correspondences [31]. Decomposing the essential matrix allows to recover the relative pose between the observing cameras, and is a central step in image-based 3D reconstruction [40]. As such, estimating the fundamental or essential matrices of image pairs is a classic and well-researched problem in computer vision.



**Figure 2. Essential Matrix Estimation.** We calculate the relative pose between outdoor and indoor image pairs via the essential matrix. We measure the AUC of the cumulative angular error up to a threshold of  $5^\circ$ ,  $10^\circ$  or  $20^\circ$ . **a)** We use no side-information about the sparse correspondences. **b)** We use side-information in the form of descriptor distance ratios between the best and second best match. We use it to filter correspondences with a threshold of 0.8 (*+Ratio*), and as an additional input for our network (*+SI*). **c)** We train NG-RANSAC in a self-supervised fashion by using the inlier count as training objective.

In the following, we firstly evaluate NG-RANSAC for the calibrated case and estimate essential matrices from SIFT correspondences [27]. For the sake of comparability with the recent, learned robust estimator of Yi *et al.* [56] we adhere closely to their evaluation setup, and compare to their results.

**Datasets.** Yi *et al.* [56] evaluate their approach in outdoor as well as indoor settings. For the outdoor datasets, they select five scenes from the structure-from-motion (SfM) dataset of [19]: *Buckingham*, *Notredame*, *Sacre Coeur*, *St. Peter’s* and *Reichstag*. They pick two additional scenes from [44]: *Fountain* and *Herzjesu*. They reconstruct each scene using a SfM tool [53] to obtain ‘ground truth’ camera poses, and co-visibility constraints for selecting image pairs. For indoor scenes Yi *et al.* choose 16 sequences from the SUN3D dataset [54] which readily comes with ground truth poses captured by KinectFusion [30]. See Appendix A for a listing of all scenes. Indoor scenarios are typically very challenging for sparse feature-based approaches because of texture-less surfaces and repetitive elements (see Fig. 1 for an example). Yi *et al.* train their best model using one outdoor scene (*St. Peter’s*) and one indoor scene (*Brown I*), and test on all remaining sequences (6 outdoor, 15 indoor). Yi *et al.* kindly provided us with their exact data splits, and we will use their setup. Note that training and test is performed on completely separate scenes, *i.e.* the neural network has to generalize to unknown environments.

**Evaluation Metric.** Via the essential matrix, we recover the relative camera pose up to scale, and compare to the ground truth pose as follows. We measure the angular error between the pose rotations, as well as the angular error between the pose translation vectors in degrees. We take the maximum of the two values as the final angular error. We calculate the cumulative error curve for each test sequence, and compute the area under the curve (AUC) up to a threshold of  $5^\circ$ ,  $10^\circ$  or  $20^\circ$ . Finally, we report the average AUC over all test sequences (but separately for the indoor and outdoor setting).

**Implementation.** Yi *et al.* train a neural network to classify a set of sparse correspondences in inliers and outliers. They represent each correspondence as a 4D vector combining the 2D coordinate in the left and right image. Their network is inspired by PointNet [33], and processes each correspondence independently by a series of multilayer perceptrons (MLPs). Global context is infused by using instance normalization [49] in-between layers. We re-build this architecture in PyTorch, and train it according to NG-RANSAC (Eq. 3). That is, the network predicts weights to guide RANSAC sampling instead of inlier class labels. We use the angular error between the estimated relative pose, and the ground truth pose as task loss  $\ell$ . As minimal solver  $f$ , we use the 5-point algorithm [31]. To speed up training, we initialize the network by learning to predict the distance of each correspondence to the ground truth epipolar line, see Appendix A for details. We initialize for 75k iterations, and train according to Eq. 3 for 25k iterations. We optimize using Adam [23] with a learning rate of  $10^{-5}$ . For each training image, we extract 2000 SIFT correspondences, and sample  $K = 4$  hypothesis pools with  $M = 16$  hypotheses. We use a low number of hypotheses during training to obtain variation when sampling pools. For testing, we increase the number of hypotheses to  $M = 1000$ . We use an inlier threshold of  $10^{-3}$  assuming normalized image coordinates using camera calibration parameters.

**Results.** We compare NG-RANSAC to the inlier classification (*InClass*) of Yi *et al.* [56]. They use their approach with SIFT as well as LIFT [55] features. The former works better for outdoor scenes, the latter works better for indoor scenes. We also include results for DeMoN [50], a learned SfM pipeline, and GMS [2], a semi-dense approach using ORB features [38]. See Fig. 2 a) for results. RANSAC achieves poor results for indoor and outdoor scenes across all thresholds, scoring as the weakest method among all competitors. Coupling it with neural guidance (NG-RANSAC) elevates it to the leading position with a comfortable margin. See also Fig. 3 for qualitative results.

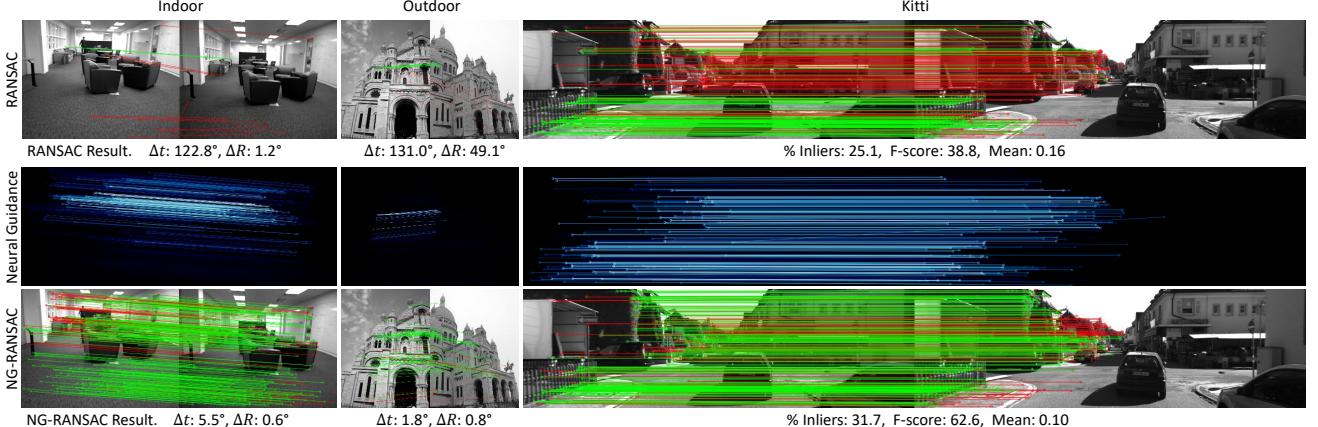


Figure 3. **Qualitative Results.** We compare fitted models for RANSAC and NG-RANSAC. For the indoor and outdoor image pairs, we fit essential matrices, and for the Kitti image pair we fit the fundamental matrix. We draw final model inliers in green if they adhere to the ground truth model, and red otherwise. We also measure the quality of each estimate, see the main text for details on the metrics.

NG-RANSAC outperforms *InClass* of Yi *et al.* [56] despite some similarities. Both use the same network architecture, are based on SIFT correspondences, and both use RANSAC at test time. Yi *et al.* [56] train using a hybrid classification-regression loss based on the 8-point algorithm, and ultimately compare essential matrices using squared error. Therefore, their training objective is very different from the evaluation procedure. During evaluation, they use RANSAC with the 5-point algorithm on top of their inlier predictions, and measure the angular error. NG-RANSAC incorporates all these components in its training procedure, and therefore optimizes the correct objective.

**Using Side-Information.** The evaluation procedure defined by Yi *et al.* [56] is well suited to test a robust estimator in high-outlier domains. However, it underestimates what classical approaches can achieve on these datasets. Lowe’s ratio criterion [27] is often deployed to drastically reduce the amount of outlier correspondences. It is based on the distance ratio of the best and second-best SIFT match. Matches with a ratio above a threshold (we use 0.8) are removed before running RANSAC. We denote the ratio criterion as *+Ratio* in Fig. 2 b), and observe a drastic improvement when combined with SIFT and RANSAC. This classic approach outperforms all learned methods of Fig. 2 a) and is competitive to NG-RANSAC (when not using any side-information). We can push performance even further by applying RootSIFT normalization to the SIFT descriptors [1]. By training NG-RANSAC on ratio-filtered RootSIFT correspondences, using distance ratios as additional network input (denoted by *+SI*), we achieve best accuracy.

**Self-supervised Learning.** We can train NG-RANSAC self-supervised when we define a task loss  $\ell$  to assess the quality of an estimate independent of a ground truth model  $\mathbf{h}^*$ . A natural choice is the inlier count of the final estimate. We found the inlier count to be a very stable training signal, even in the beginning of training such that we require

no special initialization of the network. We report results of self-supervised NG-RANSAC in Fig. 2 c). It outperforms all competitors, but achieves slightly worse accuracy than NG-RANSAC trained with supervision. A supervised task loss allows NG-RANSAC to adapt more precisely to the evaluation measure used at test time. For the datasets used so far, the process of image pairing uses co-visibility information, and therefore a form of supervision. In the next section, we learn NG-RANSAC fully self-supervised by using the ordering of sequential data to assemble image pairs.

## 4.2. Fundamental Matrix Estimation

We apply NG-RANSAC to fundamental matrix estimation, comparing it to the learned robust estimator of Ranftl and Koltun [36], denoted *Deep F-Mat*. They propose an iterative procedure where a neural network estimates observation weights for a robust model fit. The residuals of the last iteration are an additional input to the network in the next iteration. The network architecture is similar to the one used in [56]. Correspondences are represented as 4D vectors, and they use the descriptor matching ratio as an additional input. Each observation is processed by a series of MLPs with instance normalization interleaved. *Deep F-Mat* was published very recently, and the code is not yet available. We therefore follow the evaluation procedure described in [36] and compare to their results.

**Datasets.** Ranftl and Koltun [36] evaluate their method on various datasets that involve custom reconstructions which are not publicly available. Therefore, we compare to their method on the Kitti dataset [14], which is online. Ranftl and Koltun [36] train their method on sequences 00-05 of the Kitti odometry benchmark, and test on sequences 06-10. They form image pairs by taking subsequent images within a sequence. For each pair, they extract SIFT correspondences and apply Lowe’s ratio criterion [27] with a threshold of 0.8.

	Training Objective	% Inliers	F-score	Mean	Median
RANSAC	-	21.85	13.84	0.35	0.32
USAC [31]	-	21.43	13.90	0.35	0.32
Deep F-Mat [33]	Mean	24.61	14.65	<b>0.32</b>	<b>0.29</b>
NG-RANSAC	Mean	25.05	<b>14.76</b>	<b>0.32</b>	<b>0.29</b>
NG-RANSAC	F-score	24.13	14.72	0.33	0.31
NG-RANSAC	%Inliers	<b>25.12</b>	14.74	<b>0.32</b>	<b>0.29</b>

Figure 4. **Fundamental Matrix Estimation.** We measure the average percentage of inliers of the estimated model, the alignment of estimated inliers and ground truth inliers (*F-score*), and the mean and median distance of estimated inliers to ground truth epilines. For NG-RANSAC, we compare the performance after training with different objectives. Note that *%Inliers* is a self-supervised training objective.

**Evaluation Metric.** Ranftl and Koltun [36] evaluate using multiple metrics. They measure the percentage of inlier correspondences of the final model. They calculate the F-score over correspondences where true positives are inliers of both the ground truth model and the estimated model. The F-score measures the alignment of estimated and true fundamental matrix in image space. Both metrics use an inlier threshold of 0.1px. Finally, they calculate the mean and median epipolar error of inlier correspondences w.r.t. the ground truth model, using an inlier threshold of 1px.

**Implementation.** We cannot use the architecture of *Deep F-Mat* which is designed for iterative application. Therefore, we re-use the architecture of Yi *et al.* [56] from the previous section for NG-RANSAC (also see Appendix B for details). We adhere to the training setup described in Sec. 4.1 with the following changes. We observed faster training convergence on Kitti, so we omit the initialization stage, and directly optimize the expected task loss (Eq. 3) for 300k iterations. Since Ranftl and Koltun [36] evaluate using multiple metrics, the choice of the task loss function  $\ell$  is not clear. Hence, we train multiple variants with different objectives (*%Inliers*, *F-score* and *Mean* error) and report the corresponding results. As minimal solver  $f$ , we use the 7-point algorithm, a RANSAC threshold of 0.1px, and we draw  $K = 8$  hypothesis pools per training image with  $M = 16$  hypotheses each.

**Results.** We report results in Fig. 4 where we compare NG-RANSAC with RANSAC, USAC [34] and *Deep F-Mat*. Note that USAC also uses guided sampling based on matching ratios according to the PROSAC strategy [9]. NG-RANSAC outperforms the classical approaches RANSAC and USAC. NG-RANSAC also performs slightly superior to *Deep F-Mat*. We observe that the choice of the training objective has small but significant influence on the evaluation. All metrics are highly correlated, and optimizing a metric in training generally also achieves good (but not necessarily best) accuracy using this metric at test time. Interestingly, optimizing the inlier count during training performs competitively, although being a self-supervised objective. Fig. 3 shows a qualitative result on Kitti.

	AUC (%)
Simon et al. [39]	54.4
Kluger et al. [22]	57.3
Zhai et al. [54]	58.2
Workman et al. [49]	71.2
DSAC	74.1
NG-DSAC	75.2
SLNet [23]	82.3

Figure 5. **Horizon Line Estimation.** **Left:** AUC on the HLW dataset. **Right:** Qualitative results. We draw the ground truth horizon in green and the estimate in blue. Dots mark the observations predicted by NG-DSAC, and the dot colors mark their confidence (dark = low). Note that the horizon can be outside the image.

### 4.3. Horizon Lines

We fit a parametric model, the horizon line, to a single image. The horizon can serve as a cue in image understanding [52] or for image editing [25]. Traditionally, this task is solved via vanishing point detection and geometric reasoning [37, 24, 57, 42], often assuming a Manhattan or Atlanta world. We take a simpler approach and use a general purpose CNN that predicts a set of 64 2D points based on the image to which we fit a line with RANSAC, see Fig. 5. The network has two output branches (see Appendix C for details) predicting (i) the 2D points  $\mathbf{y}(\mathbf{w}) \in \mathcal{Y}(\mathbf{w})$ , and (ii) probabilities  $p(\mathbf{y}; \mathbf{w})$  for guided sampling.

**Dataset.** We evaluate on the HLW dataset [52] which is a collection of SfM datasets with annotated horizon line. Test and training images partly show the same scenes, and the horizon line can be outside the image area.

**Evaluation Metric.** As is common practice on HLW, we measure the maximum distance between the estimated horizon and ground truth within the image, normalized by image height. We calculate the AUC of the cumulative error curve up to a threshold of 0.25.

**Implementation.** We train using the NG-DSAC objective (Eq. 9) from scratch for 300k iterations. As task loss  $\ell$ , we use the normalized maximum distance between estimated and true horizon. For hypothesis scoring  $s$ , we use a soft inlier count [6]. We train using Adam [23] with a learning rate of  $10^{-4}$ . For each training image, we draw  $K = 2$  hypothesis pools with  $M = 16$  hypotheses. We also draw 16 hypotheses at test time. We compare to DSAC which we train similarly but disable the probability branch.

**Results.** We report results in Fig. 5. DSAC and NG-DSAC achieve competitive accuracy on this dataset, ranking among the top methods. NG-DSAC has a small but significant advantage over DSAC alone. Our method is only surpassed by SLNet [25], an architecture designed to find semantic lines in images. SLNet generates a large number of random candidate lines, selects a candidate via classification, and refines it with a predicted offset. We could couple SLNet with neural guidance for informed candidate sampling. Unfortunately, the code of SLNet is not online and the authors did not respond to inquiries.

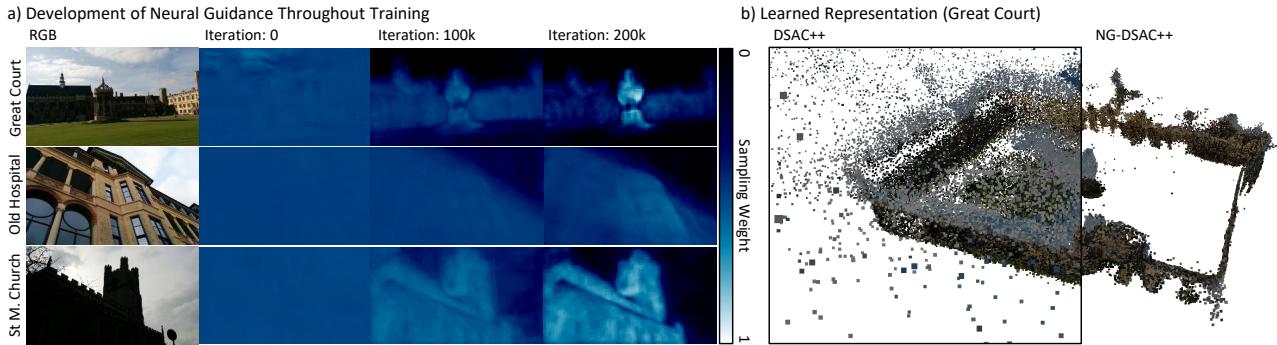


Figure 6. **Neural Guidance for Camera Re-localization.** **a)** We show how the predicted sampling probabilities of NG-DSAC++ develop during end-to-end training. **b)** We visualize the internal representation of the neural network. We predict scene coordinates for each training image, plotting them with their RGB color. For DSAC++ we choose training pixels randomly, for NG-DSAC++ we choose randomly among the top 1000 pixels per training image according to the predicted distribution.

	DSAC++ [6] (VGGNet)	DSAC++ (ResNet)	NG-DSAC++ (ResNet)
Great Court	40.3cm	40.3cm	<b>35.0cm</b>
Kings College	17.7cm	13.0cm	<b>12.6cm</b>
Old Hospital	<b>19.6cm</b>	22.4cm	21.9cm
Shop Facade	5.7cm	5.7cm	<b>5.6cm</b>
St M. Church	12.5cm	9.9cm	<b>9.8cm</b>

Figure 7. **Camera Re-Localization.** We report median position error for Cambridge Landmarks [22]. DSAC++ (ResNet) is our re-implementation of [6] with an improved network architecture.

#### 4.4. Camera Re-Localization

We estimate the absolute 6D camera pose (position and orientation) w.r.t. a known scene from a single RGB image.

**Dataset.** We evaluate on the Cambridge Landmarks [22] dataset. It is comprised of RGB images depicting five landmark buildings<sup>1</sup> in Cambridge, UK. Ground truth poses were generated by running a SfM pipeline.

**Evaluation Metric.** We measure the median translational error of estimated poses for each scene<sup>2</sup>.

**Implementation.** We build on the publicly available DSAC++ pipeline [6] which is a scene coordinate regression method [41]. A neural network predicts for each image pixel a 3D coordinate in scene space. We recover the pose from the 2D-3D correspondences using a perspective-n-point solver [13] within a RANSAC loop. The DSAC++ pipeline implements geometric pose optimization in a fully differentiable way which facilitates end-to-end training. We re-implement the neural network integration of DSAC++ with PyTorch (the original uses LUA/Torch). We also update the network architecture of DSAC++ by using a ResNet [18] instead of a VGGNet [43]. As with horizon

<sup>1</sup>We omitted one additional scene (*Street*). Like other learning-based methods before [6] we failed to achieve sensible results for this scene. By visual inspection, the corresponding SfM reconstruction seems to be of poor quality, which potentially harms the training process.

<sup>2</sup>The median rotational accuracies are between  $0.2^\circ$  to  $0.3^\circ$  for all scenes, and do hardly vary between methods.

line estimation, we add a second output branch to the network for estimating a probability distribution over scene coordinate predictions for guided RANSAC sampling. We denote this extended architecture *NG-DSAC++*. We adhere to the training procedure and hyperparameters of DSAC++ (see Appendix D) but optimize the NG-DSAC objective (Eq. 9) during end-to-end training. As task loss  $\ell$ , we use the average of the rotational and translational error w.r.t. the ground truth pose. We sample  $K = 2$  hypothesis pools with  $M = 16$  hypotheses per training image, and increase the number of hypotheses to  $M = 256$  for testing.

**Results.** We report our quantitative results in Fig. 7. Firstly, we observe a significant improvement for most scenes when using DSAC++ with a ResNet architecture. Secondly, comparing DSAC++ with NG-DSAC++, we notice a small to moderate, but consistent, improvement in accuracy. The advantage of using neural guidance is largest for the *Great Court* scene, which features large ambiguous grass areas, and large areas of sky visible in many images. NG-DSAC++ learns to ignore such areas, see the visualization in Fig. 6 a). The network learns to mask these areas solely guided by the task loss during training, as the network fails to predict accurate scene coordinates for them. In Fig. 6 b), we visualize the internal representation learned by DSAC++ and NG-DSAC++ for one scene. The representation of DSAC++ is very noisy, as it tries to optimize geometric constraints for sky and grass pixels. NG-DSAC++ learns a cleaner representation by focusing entirely on buildings.

## 5. Conclusion

We have presented NG-RANSAC, a robust estimator using guided hypothesis sampling according to learned probabilities. For training we can incorporate non-differentiable task loss functions and non-differentiable minimal solvers. Using the inlier count as training objective allows us to also train NG-RANSAC self-supervised. We applied NG-RANSAC to multiple classic computer vision tasks and observe a consistent improvement w.r.t. RANSAC alone.

**Acknowledgements:** This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 647769). The computations were performed on an HPC Cluster at the Center for Information Services and High Performance Computing (ZIH) at TU Dresden.

## A. Essential Matrix Estimation

### List of Scenes Used for Training and Testing.

Training:

- Staint Peter’s (Outdoor)
- brown\_bm\_3 - brown\_bm\_3 (Indoor)

Testing (Outdoor):

- Buckingham
- Notre Dame
- Sacre Coeur
- Reichstag
- Fountain
- HerzJesu

Testing (Indoor):

- brown\_cogsci\_2 - brown\_cogsci\_2
- brown\_cogsci\_6 - brown\_cogsci\_6
- brown\_cogsci\_8 - brown\_cogsci\_8
- brown\_cs\_3 - brown\_cs3
- brown\_cs\_7 - brown\_cs7
- harvard\_c4 - hv\_c4\_1
- harvard\_c10 - hv\_c10\_2
- harvard\_corridor\_lounge - hv\_lounge1\_2
- harvard\_robots.lab - hv\_s1\_2
- hotel\_florence\_jx - florence\_hotel\_stair\_room\_all
- mit\_32\_g725 - g725\_1
- mit\_46\_6conf - bcs\_floor6.conf\_1
- mit\_46\_6lounge - bcs\_floor6\_long
- mit\_w85g - g\_0
- mit\_w85h - h2\_1

**Network Architecture.** As mentioned in the main paper, we replicated the architecture of Yi *et al.* [56] for our experiments on epipolar geometry (estimating essential and fundamental matrices). For a schematic overview see Fig. 8. The network takes a set of feature correspondences as input, and predicts as output a weight for each correspondence which we use to guide RANSAC hypothesis sampling. The network consists of a series of multilayer perceptrons (MLPs) that process each correspondence independently. We implement the MLPs with  $1 \times 1$  convolutions. The network infuses global context via instance normalization layers [49], and it accelerate training via batch normalization [20]. The main body of the network is comprised of 12 blocks with skip connections [18]. Each block con-

sists of two linear layers followed by instance normalization, batch normalization and a ReLU activation [17] each. We apply a Sigmoid activation to the last layer, and normalize by dividing by the sum of outputs.<sup>3</sup>

**Initialization Procedure.** We initialize our network in the following way. We define a target sampling distribution  $g(\mathbf{y}; E^*)$  using the ground truth essential matrix  $E^*$  given for each training pair. Intuitively, the target distribution should return a high probability when a correspondence  $\mathbf{y}$  is aligned with the ground truth essential matrix  $E^*$ , and a low probability otherwise. We assume that correspondence  $\mathbf{y}$  is a 4D vector containing two 2D image coordinates  $\mathbf{x}$  and  $\mathbf{x}'$  (3D in homogeneous coordinates). We define the epipolar error of a correspondence w.r.t. essential matrix  $E$ :

$$d(\mathbf{y}, E) = \frac{(\mathbf{x}'^\top E \mathbf{x})^2}{[\mathbf{E}\mathbf{x}]_0^2 + [\mathbf{E}\mathbf{x}]_1^2 + [\mathbf{E}^\top \mathbf{x}']_0^2 + [\mathbf{E}^\top \mathbf{x}']_1^2}, \quad (11)$$

where  $[\cdot]_i$  returns the  $i$ th entry of a vector. Using the epipolar error, we define the target sampling distribution:

$$g(\mathbf{y}; E^*) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{d(\mathbf{y}, E^*)}{2\sigma^2}\right). \quad (12)$$

Parameter  $\sigma$  controls the softness of the target distribution, and we use  $\sigma = 10^{-3}$  which corresponds to the inlier threshold we use for RANSAC. To initialize our network, we minimize the KL divergence between the network prediction  $p(\mathbf{y}; w)$  and the target distribution  $g(\mathbf{y}; E^*)$ . We initialize for 75k iterations using Adam [23] with a learning rate of  $10^{-3}$  and a batch size of 32.

**Implementation Details.** For the following components we rely on the implementations provided by OpenCV [7]: the 5-point algorithm [31], epipolar error, SIFT features [27], feature matching, and essential matrix decomposition. We extract 2000 features per input image which yields 2000 correspondences for image pairs after matching. When applying Lowe’s ratio criterion [27] for filtering and hence reducing the number of correspondences, we randomly duplicate correspondences to restore the number of 2000. We minimize the expected task loss using Adam [23] with a learning rate of  $10^{-5}$  and a batch size of 32. We choose hyperparameters based on validation error of the *Reichstag* scene. We observe that the magnitude of the validation error corresponds well to the magnitude of the training error, *i.e.* a validation set would not be strictly required.

**Qualitative Results.** We present additional qualitative results for indoor and outdoor scenarios in Fig. 9. We compare results of RANSAC and NG-RANSAC, also visualizing neural guidance as predicted by our network. We obtain these results in the high-outlier setup, *i.e.* without using

---

<sup>3</sup>The original architecture of Yi *et al.* [56] uses a slightly different output processing due to using the output as weights for a robust model fit. They use a ReLU activation followed by a tanh activation.

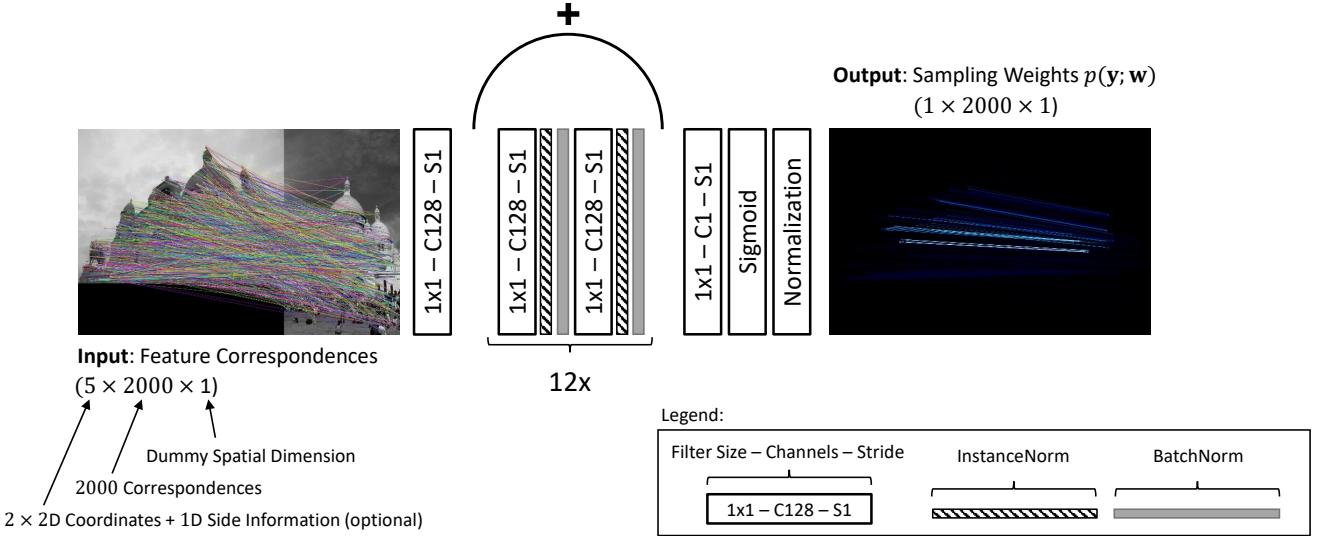


Figure 8. **NG-RANSAC Network Architecture for F/E-matrix Estimation.** The network takes a set of feature correspondences as input and predicts as output a weight for each correspondence. The network consists of linear layers interleaved by instance normalization [49], batch normalization [20] and ReLUs [17]. The arc with a plus marks a skip connection for each of the twelve blocks [18]. This architecture was proposed by Yi *et al.* [56].

Lowe’s ratio criterion and without using side-information as additional network input.

## B. Fundamental Matrix Estimation

**Implementation Details.** We reuse the architecture of Fig. 8. We normalize image coordinates of feature matches before passing them to the network. We subtract the mean coordinate and divide by the coordinate standard deviation, where we calculate mean and standard deviation over the training set. Ranftl and Koltun [36] fit the final fundamental matrix to the top 20 weighted correspondences as predicted by their network. Similarly, we re-fit the final fundamental matrix to the largest inlier set found by NG-RANSAC. This refinement step results in a small but noticeable increase in accuracy. For the following components we rely on the implementations provided by OpenCV [7]: the 7-point algorithm, epipolar error, SIFT features [27] and feature matching.

**Qualitative Results.** We present additional qualitative results for the Kitti dataset [14] in Fig. 10. We compare results of RANSAC and NG-RANSAC, also visualizing neural guidance as predicted by our network.

## C. Horizon Lines

**Network Architecture.** We provide a schematic of our network architecture for horizon line estimation in Fig. 11. The network takes a  $256 \times 256$ px RGB image as input. We re-scale images of arbitrary aspect ratio such that the long side is 256px. We symmetrically zero-pad the short side to 256px. The network has two output branches. The first

branch predicts a set of  $8 \times 8 = 64$  2D points, our observations  $y(w)$ , to which we fit the horizon line. We apply a Sigmoid and re-scale output points to  $[-1.5, 1.5]$  in relative image coordinates to support horizon lines outside the image area. We implement the network in a fully convolutional way [26], *i.e.* each output point is predicted for a patch, or restricted receptive field, of the input image. Therefore, we shift the coordinate of each output point to the center of its associated patch.

The second output branch predicts sampling probabilities  $p(y; w)$  for each output point. We apply a Sigmoid to the output of the second branch, and normalize by dividing by the sum of outputs. During training, we block the gradients of the second output branch when back propagating to the base network. The sampling gradients have larger variance and magnitude than the observation gradients of the first branch, especially in the beginning of training. This has a negative effect on convergence of the network as a whole. Intuitively, we want to give priority to the observation prediction because they determine the accuracy of the final model parameters. The sampling prediction should address deficiencies in the observation predictions without influencing them too much. The gradient blockade ensures these properties.

**Implementation Details.** We use a differentiable soft inlier count [6] as scoring function, *i.e.*:

$$s(\mathbf{h}, \mathcal{Y}) = \alpha \sum_{\mathbf{y} \in \mathcal{Y}} 1 - \text{sig}[\beta d(\mathbf{y}, \mathbf{h}) - \beta \tau], \quad (13)$$

where  $d(\mathbf{y}, \mathbf{h})$  denotes the point-line distance between observation  $\mathbf{y}$  and line hypothesis  $\mathbf{h}$ . Hyperparameter  $\alpha$  deter-

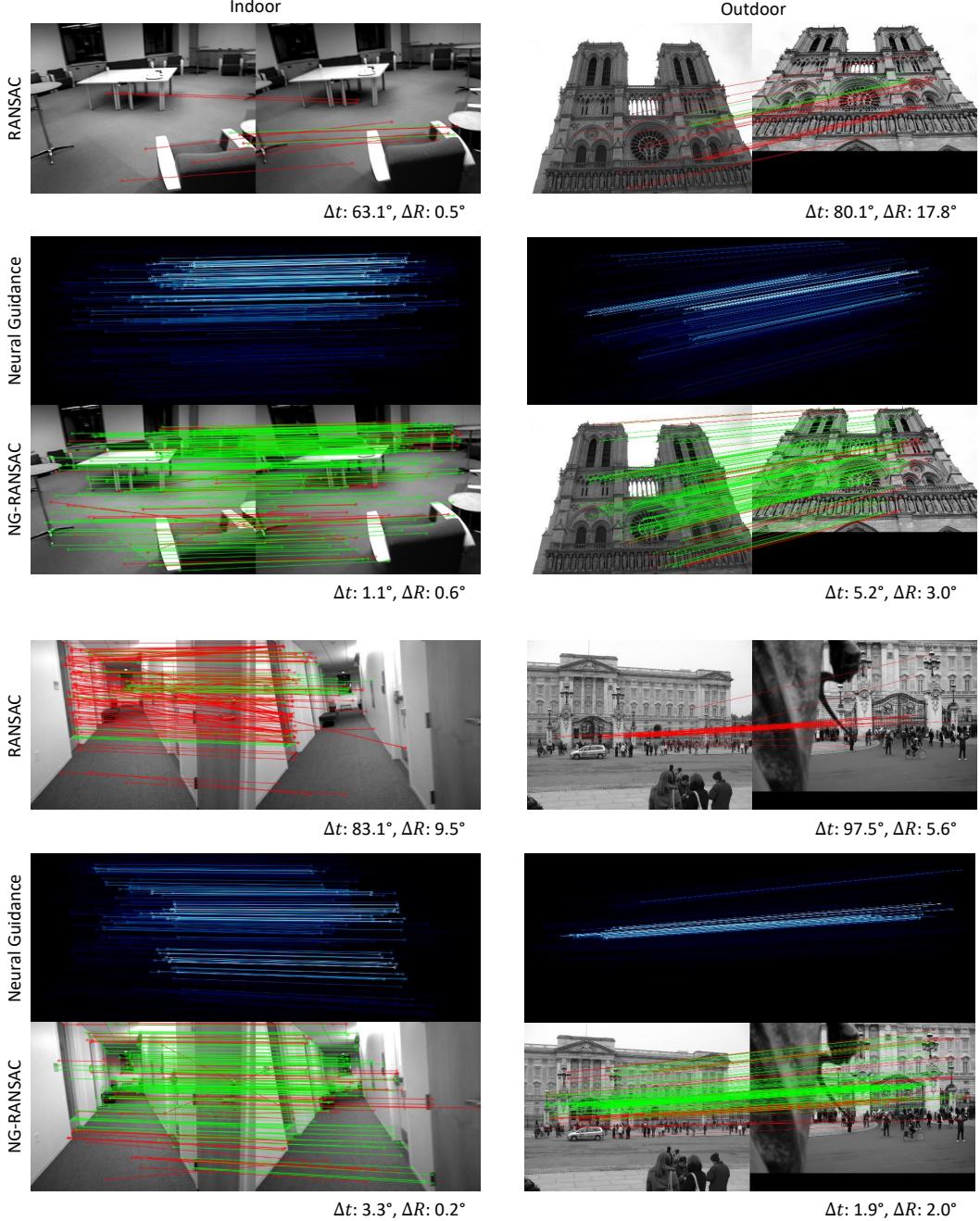


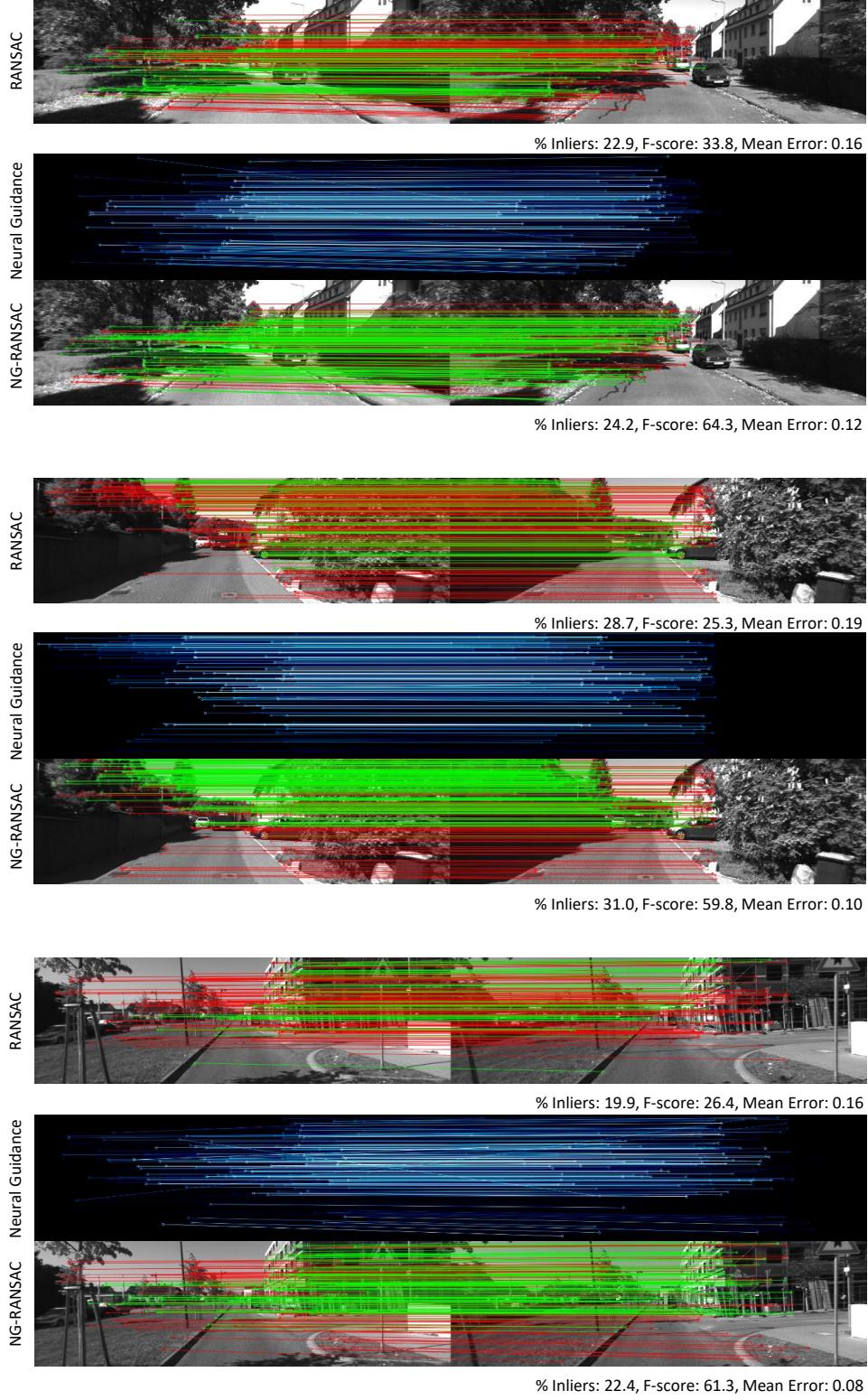
Figure 9. **Qualitative Results for Essential Matrix Estimation.** We compare results of RANSAC and NG-RANSAC. Below each result, we give the angular error between estimated and true translation vectors, and estimated and true rotation matrices. We draw correspondences in green if they adhere to the ground truth essential matrix with an inlier threshold of  $10^{-3}$ , and red otherwise.

mines the softness of the scoring distribution in DSAC,  $\beta$  determines the softness of the Sigmoid, and  $\tau$  is the inlier threshold. We use  $\alpha = 0.1$ ,  $\beta = 100$  and  $\tau = 0.05$ . We choose hyperparameters based on grid search for the minimal training loss.

As discussed in the main paper, we use the normalized maximum distance between a line hypothesis and the

ground truth horizon in the image as task loss  $\ell$ . This can lead to stability issues when we sample line hypotheses with very steep slope. Therefore, we clamp the task loss to a maximum of 1, *i.e.* the normalized image height.

As mentioned before, some images in the HLW dataset [52] have their horizon outside the image. Some of these images contain virtually no visual cue where the horizon



**Figure 10. Qualitative Results for Fundamental Matrix Estimation.** We compare results of RANSAC and NG-RANSAC. Below each result, we give the percentage of inliers of the final model, the F-score which measures the alignment of estimated and true fundamental matrix, and the mean epipolar error of estimated inlier correspondences w.r.t. the ground truth fundamental matrix. We draw correspondences in green if they adhere to the ground truth fundamental matrix with an inlier threshold of 0.1px, and red otherwise.

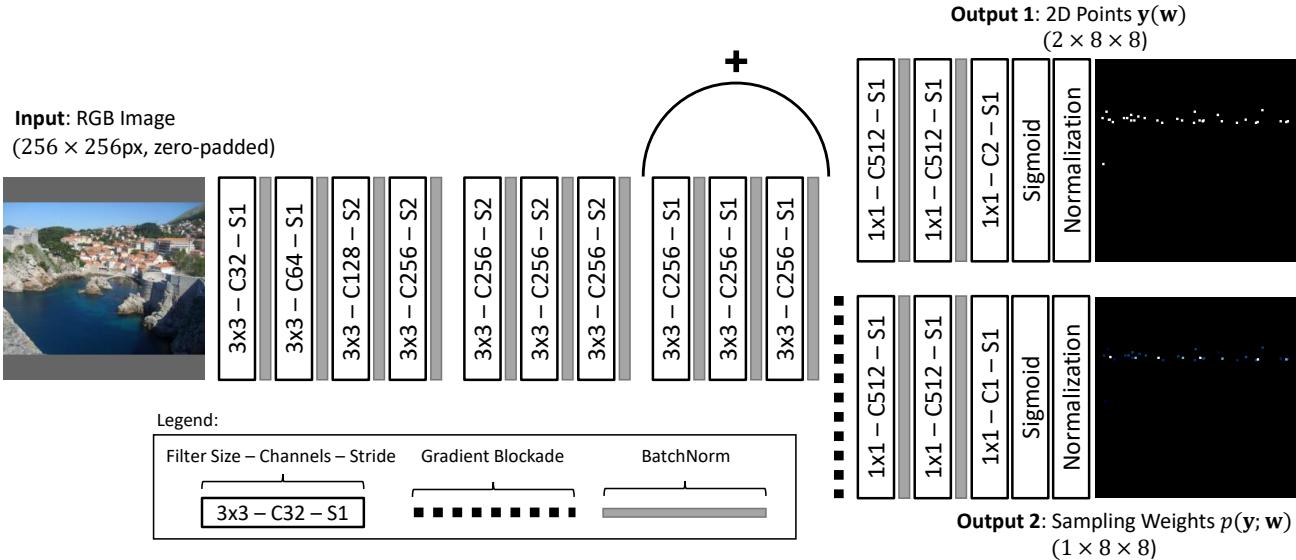


Figure 11. **NG-DSAC Network Architecture for Horizon Line Estimation.** The network takes a RGB image as input and predicts as output a set of 2D points and corresponding sampling weights. The network consists of convolution layers interleaved by batch normalization [20] and ReLUs [17]. The arc with a plus marks a skip connection [18]. We use the gradient blockade during training to prevent direct influence of the sampling prediction (second branch) to learning the observations (first branch).

exactly lies. Therefore, we find it beneficial to use a robust variant of the task loss  $\ell'$  that limits the influence of such outliers. We use:

$$\ell' = \begin{cases} \ell & \ell < 0.25 \\ 0.25\sqrt{\ell} & \text{otherwise} \end{cases}, \quad (14)$$

i.e. we use the square root of the task loss after a magnitude of 0.25, which is the magnitude up to which the AUC is calculated when evaluating on HLW [52].

**Qualitative Results.** We present additional qualitative results for the HLW dataset [52] in Fig. 12.

## D. Camera Re-Localization

**Network Architecture.** We provide a schematic of our network architecture for camera re-localization in Fig. 13. The network is a FCN [26] that takes an RGB image as input, and predicts dense outputs, sub-sampled by a factor of 8. The network has two output branches. The first branch predicts 3D scene coordinates [41], our observations  $y(w)$ , to which we fit the 6D camera pose. The second output branch predicts sampling probabilities  $p(y; w)$  for the scene coordinates. We apply a Sigmoid to the output of the second branch, and normalize by dividing by the sum of outputs. During training, we block the gradients of the second output branch when back propagating to the base network. The sampling gradients have larger variance and magnitude than the observation gradients of the first branch, especially in the beginning of training. This has a negative effect on convergence of the network as a whole. Intuitively, we want to

give priority to the scene coordinate prediction because they determine the accuracy of the pose estimate. The sampling prediction should address deficiencies in the scene coordinate predictions without influencing them too much. The gradient blockade ensures these properties.

**Implementation details.** We follow the three-stage training procedure proposed by Brachmann and Rother for DSAC++ [6].

Firstly, we optimize the distance between predicted and ground truth scene coordinates. We obtain ground truth scene coordinates by rendering the sparse reconstructions given in the Cambridge Landmarks dataset [22]. We ignore pixels with no corresponding 3D point in the reconstruction. Since the reconstructions contain outlier 3D points, we use the following robust distance:

$$(y, y^*) = \begin{cases} \|y - y^*\|_2 & \|y - y^*\|_2 < 10 \\ 10\sqrt{\|y - y^*\|_2} & \text{otherwise} \end{cases}, \quad (15)$$

i.e. we use the Euclidean distance up to a threshold of 10m after which we use the square root of the Euclidean distance. We train the first stage for 500k iterations using Adam [23] with a learning rate of  $10^{-4}$  and a batch size of 1 image.

Secondly, we optimize the reprojection error of the scene coordinate predictions w.r.t. to the ground truth camera pose. Similar to the first stage, we use a robust distance function with a threshold of 10px after which we use the square root of the reprojection error. We train the second stage for 300k iterations using Adam [23] with a learning rate of  $10^{-4}$  and a batch size of 1 image.

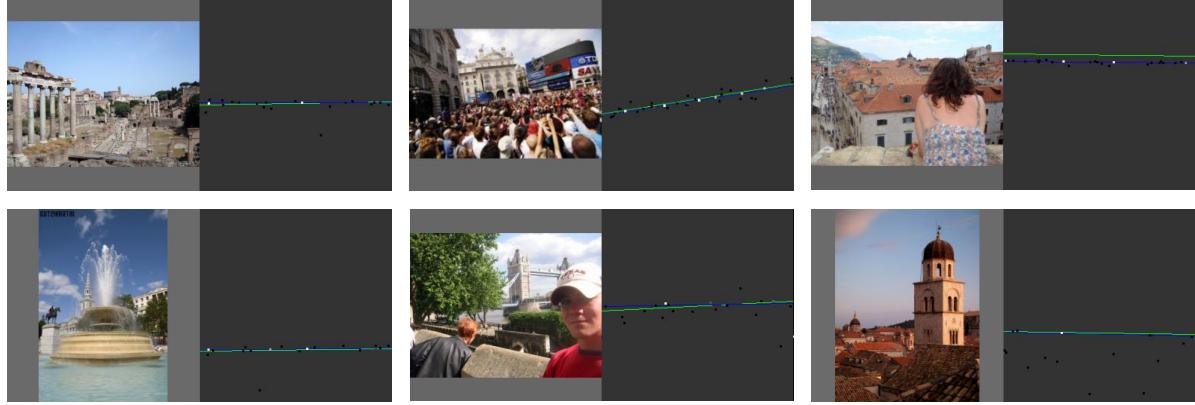


Figure 12. **Qualitative Results for Horizon Line Estimation.** Next to each input image, we show the estimated horizon line in blue and the true horizon line in green. We also show the observation points predicted by our network, colored by their sampling weight (dark = low).

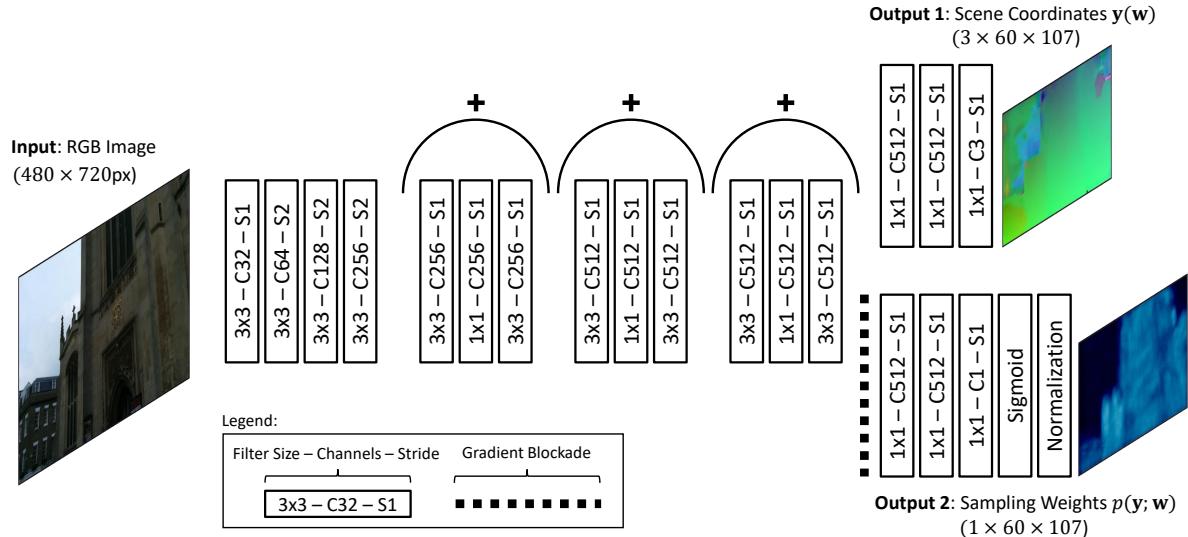
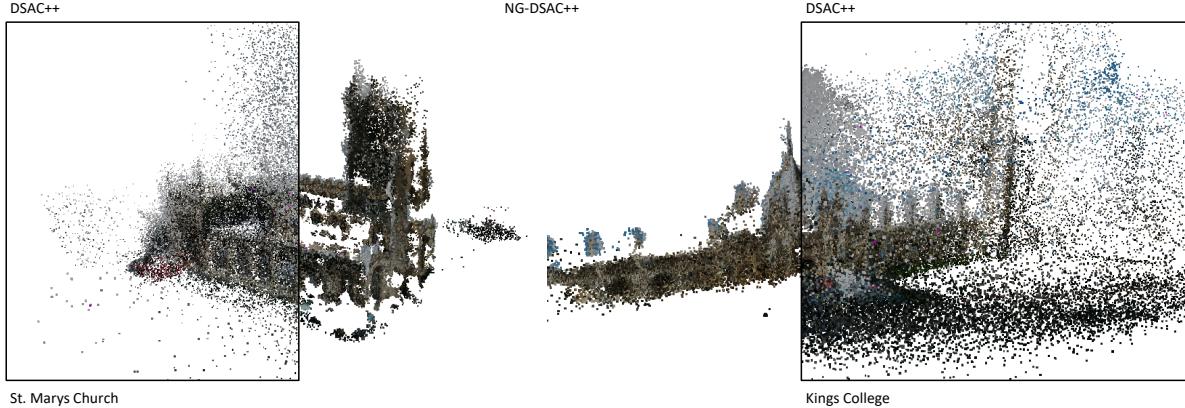


Figure 13. **NG-DSAC++ Network Architecture for Camera Re-Localization.** The network takes an RGB image as input and predicts as output dense scene coordinates and corresponding sampling weights. The network consists of convolution layers followed by ReLUs [17]. An arc with a plus marks a skip connection [18]. We use the gradient blockage during training to prevent direct influence of the sampling prediction (second branch) to learning the scene coordinates (first branch).

Thirdly, we optimize the expected task loss according to the NG-DSAC objective as explained in the main paper. As task loss we use  $\ell = \angle(\theta, \theta^*) + \|\mathbf{t} - \mathbf{t}^*\|_2$ . We measure the angle between estimated camera rotation  $\theta$  and ground truth rotation  $\theta^*$  in degree. We measure the distance between the estimated camera position  $\mathbf{t}$  and ground truth position  $\mathbf{t}^*$  in meters. As with horizon line estimation (see previous section), we use a soft inlier count as hypothesis scoring function with hyperparameters  $\alpha = 10$ ,  $\beta = 0.5$  and  $\tau = 10$ . We train the third stage for 200k iterations using Adam [23] with a learning rate of  $10^{-6}$  and a batch size of 1 image. **Learned 3D Representations.** We visualize the internal 3D scene representations learned by DSAC++ and NG-DSAC++ in Fig. 14 for two more scenes.

## References

- [1] R. Arandjelovic. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 6
- [2] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T. D. Nguyen, and M.-M. Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *CVPR*, 2017. 5
- [3] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6D object pose estimation using 3D object coordinates. In *ECCV*, 2014. 2
- [4] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. DSAC-Differentiable RANSAC for camera localization. In *CVPR*, 2017. 2, 3, 4
- [5] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold,



**Figure 14. Learned 3D Representations.** We visualize the internal representation of the neural network. We predict scene coordinates for each training image, plotting them with their RGB color. For DSAC++ we choose training pixels randomly, for NG-DSAC++ we choose randomly among the top 1000 pixels per training image according to the predicted distribution.

- and C. Rother. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In *CVPR*, 2016. 2
- [6] E. Brachmann and C. Rother. Learning less is more-6D camera localization via 3D surface regression. In *CVPR*, 2018. 1, 2, 7, 8, 10, 13
- [7] G. Bradski. OpenCV. *Dr. Dobb's Journal of Software Tools*, 2000. 1, 3, 9, 10
- [8] T. Cavallari, S. Golodetz, N. A. Lord, J. Valentin, L. Di Stefano, and P. H. Torr. On-the-fly adaptation of regression forests for online camera relocalisation. In *CVPR*, 2017. 2
- [9] O. Chum and J. Matas. Matching with PROSAC - Progressive sample consensus. In *CVPR*, 2005. 2, 7
- [10] O. Chum and J. Matas. Optimal randomized ransac. *TPAMI*, 2008. 2
- [11] O. Chum, J. Matas, and J. Kittler. Locally optimized RANSAC. In *DAGM*, 2003. 2
- [12] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981. 1, 2, 3
- [13] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng. Complete solution classification for the perspective-three-point problem. *TPAMI*, 2003. 8
- [14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 6, 10
- [15] R. I. Hartley. In defense of the eight-point algorithm. *TPAMI*, 1997. 3
- [16] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 3, 4
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *ICCV*, 2015. 9, 10, 13, 14
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8, 9, 10, 13, 14
- [19] J. Heinly, J. L. Schönberger, E. Dunn, and J.-M. Frahm. Reconstructing the World\* in Six Days \*(As Captured by the Yahoo 100 Million Image Dataset). In *CVPR*, 2015. 5
- [20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 9, 10, 13
- [21] O. H. Jafari, S. K. Mustikovela, K. Pertsch, E. Brachmann, and C. Rother. ipose: Instance-aware 6d pose estimation of partly occluded objects. 2018. 2
- [22] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-DoF camera relocalization. In *ICCV*, 2015. 8, 13
- [23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5, 7, 9, 13, 14
- [24] F. Kluger, H. Ackermann, M. Y. Yang, and B. Rosenhahn. Deep learning for vanishing point detection using an inverse gnomonic projection. In *GCPR*, 2017. 7
- [25] J.-T. Lee, H.-U. Kim, C. Lee, and C.-S. Kim. Semantic line detection and its applications. In *ICCV*, 2017. 7
- [26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 10, 13
- [27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3, 5, 6, 9, 10
- [28] D. Massiceti, A. Krull, E. Brachmann, C. Rother, and P. H. S. Torr. Random forests versus neural networks - what's best for camera localization? In *ICRA*, 2017. 2
- [29] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. *T-RO*, 2017. 1
- [30] R. Newcombe, S. Izadi, O. Hilliges, D. Molynieux, D. Kim, A. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *Proc. ISMAR*, 2011. 5
- [31] D. Nistér. An efficient solution to the five-point relative pose problem. *TPAMI*, 2004. 4, 5, 9
- [32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 4
- [33] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 5

- [34] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm. Usac: A universal framework for random sample consensus. *TPAMI*, 2013. 2, 7
- [35] R. Raguram, J.-M. Frahm, and M. Pollefeys. A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In *ECCV*, 2008. 2
- [36] R. Ranftl and V. Koltun. Deep fundamental matrix estimation. In *ECCV*, 2018. 2, 3, 6, 7, 10
- [37] C. Rother. A new approach for vanishing point detection in architectural environments. In *BMVC*, 2002. 7
- [38] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, 2011. 5
- [39] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *TPAMI*, 2016. 1
- [40] J. L. Schönberger and J.-M. Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 1, 4
- [41] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013. 2, 8, 13
- [42] G. Simon, A. Fond, and M.-O. Berger. A-contrario horizon-first vanishing point detection using second-order grouping laws. In *ECCV*, 2018. 7
- [43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014. 8
- [44] C. Strecha, W. von Hansen, L. J. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008. 5
- [45] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998. 4
- [46] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 2
- [47] B. Tordoff and D. W. Murray. Guided sampling and consensus for motion estimation. In *ECCV*, 2002. 2
- [48] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *CVIU*, 2000. 2
- [49] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, 2016. 5, 9, 10
- [50] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 5
- [51] J. Valentin, M. Nießner, J. Shotton, A. Fitzgibbon, S. Izadi, and P. H. S. Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *CVPR*, 2015. 2
- [52] S. Workman, M. Zhai, and N. Jacobs. Horizon lines in the wild. In *BMVC*, 2016. 7, 11, 13
- [53] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013. 5
- [54] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013. 5
- [55] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned invariant feature transform. In *ECCV*, 2016. 5
- [56] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua. Learning to find good correspondences. In *CVPR*, 2018. 2, 3, 5, 6, 7, 9, 10
- [57] M. Zhai, S. Workman, and N. Jacobs. Detecting vanishing points using global image context in a non-manhattan world. In *CVPR*, 2016. 7