Pose-graph via Adaptive Image Re-ordering

Daniel Barath[1]
https://people.inf.ethz.ch/dbarath/
Jana Noskova[2]
https://mat.fsv.cvut.cz/noskova/
Ivan Eichhardt[3]
http://cv.inf.elte.hu/
Jiri Matas[2]
https://cmp.felk.cvut.cz/~matas/

[1] Computer Vision and Geometry Group
ETH Zurich, Switzerland

[2] Visual Recognition Group,
FEE, CTU in Prague, Czech Republic

[3] Algorithms and Applications
Department, Eotvos Lorand University,
Budapest, Hungary

Abstract

We introduce novel methods that speed up the pose-graph generation for global Structure-from-Motion algorithms. We replace the widely used "accept-or-reject" strategy for image pairs, where often thousands of RANSAC iterations are wasted on pairs with low inlier ratio or on non-matchable ones. The new algorithm exploits the fact that every unsuccessful RANSAC iteration reduces the probability of an image pair being matchable, *i.e.*, it reduces its inlier ratio expectation. The method always selects the most promising pair for matching. While running RANSAC on the pair, it updates the distribution of its inlier ratio probability in a principled way via a Bayesian approach. Once the expected inlier ratio drops below an adaptive threshold, the method puts back the pair in the processing queue ordered by the updated inlier ratio expectations. The algorithms are tested on more than 600k real image pairs. They accelerate the pose-graph generation by an order-of-magnitude on average. The source code is available at `https://github.com/danini/pose-graph-creation`

1 Introduction

Structure-from-Motion (SfM) has been intensively researched in computer vision for decades. Most of the early methods adopt an incremental strategy, where the reconstruction is built progressively and the images are carefully added one-by-one in the procedure [1, 37, 38, 43, 44, 53]. Recent studies [4, 10, 13, 15, 19, 21, 22, 23, 31, 35, 54] show that global approaches that consider all images simultaneously when reconstructing the scene geometry, lead to comparable or better accuracy than incremental ones while being significantly more efficient. Moreover, global methods are less dependent on local decisions or image ordering.

Typically, global Structure-from-Motion pipelines consist of the following main steps. First, feature points are extracted in all $n \in \mathbb{N}$ images. Such step is easily parallelizable and has $\mathcal{O}(n)$ time complexity. These features are then used to order the image pairs from the most probable to match to the most difficult ones, *e.g.*, via bag-of-visual-words [46]. Next, tentative correspondences are generated between all image pairs by matching the high-dimensional (*e.g.*, 128 for SIFT [30]) descriptors of the detected features. Then, the detected correspondences are filtered and relative poses are estimated by applying RANSAC [20] or one of its state-of-the-art variants, *e.g.*, [8, 25, 41]. The feature matching and geometric estimation steps are by far the slowest parts, both having quadratic complexity in the number of images. Finally, a global pose graph is obtained from the pair-wise poses via rotation and translation averaging, optimized by additional bundle adjustment. Interestingly, this step takes almost negligible time, *i.e.*, a few minutes in our experiments, compared to the initial pose-graph generation that often runs for hours.

Accelerating the pose graph generation is a long-standing problem in SfM. One of the most studied approaches is the efficient selection of potentially overlapping image pairs. Traditionally, this problem is solved by generating compact image descriptors constructed by an aggregation of local features, like Fisher vectors [36], VLAD [26] and other alternatives [3, 39, 45, 50]. To recent works, learning-based alternatives, like GeM descriptors [40] or [11, 42], dominate the image search task by significantly outperforming state-of-the-art traditional methods. All these works focus on finding potential image matches prior to running an SfM reconstruction which then processes the selected pairs one by one.

Another group of algorithms focuses on speeding up independent RANSAC runs by advanced sampling strategies, pre-emptive model verification, or early rejection. In order to find an all-inlier sample early, NAPSAC [51], GroupSAC [33] and Progressive NAPSAC [6] assume that the inliers of a model have "similar" properties and, therefore, can be separated into groups prior to the estimation. The points are sampled from randomly selected groups. The PROSAC [17] algorithm exploits a predicted inlier probability rank of each point. Preemptive model verification strategies [12, 16, 18, 32] have been proposed to recognize incorrect models early during the model quality calculation. Recently, papers also focus on rejecting likely ill-conditioned or degenerate minimal samples early [9, 14] to avoid estimating the model parameter unnecessarily. All of these algorithms focus on accelerating a single RANSAC run and do not exploit the fact that, in many cases, we run robust estimation on a large number of image pairs sequentially.

Recently, Barath *et al.* [7] studied the problem that RANSAC-like robust estimation is often time-consuming, especially, when the images do not match or the inlier ratio is low. In these cases, the iteration number of the applied randomized robust estimator inevitably reaches the maximum iteration number set by the user. This means doing thousands of unnecessary iterations. They introduce an approach running the A* algorithm to find walks in a partially built pose-graph when estimating the relative pose of two images. The pose is then recovered by chaining the transformations along the found walk and improved by an iteratively re-weighted least-squares approach re-selecting the inliers in every iteration. This strategy helps to avoid running RANSAC if a found walk leads to a non-random number of inliers. Due to the skipped RANSAC runs, the pose estimation is significantly accelerated.

In this paper, we propose to revisit the traditional "accept-or-reject" strategy used in state-of-the-art large-scale pose estimation algorithms [44]. In brief, the "accept-or-reject" approach is an iteration of two steps. *First*, an image pair with the highest probability of being matchable is selected by, *e.g.*, bag-of-visual-words. *Second*, RANSAC is applied with its maximum iteration number parameter set to a reasonably large value, *e.g.* 5000 or 10 000. This maximum iteration number is a hard-constraint on the number of iterations that is controlled by the manually set confidence parameter (typically, set to 0.99). If RANSAC fails to find a pose with a large number of inliers, the image pair is rejected and is never used again. In this case, the iteration number *always* reaches its maximum. In terms of run-time, this approach is sub-optimal since the steps are applied consecutively. In the first step, the image pair with the highest probability of being matchable is selected. Theoretically, this probabil-

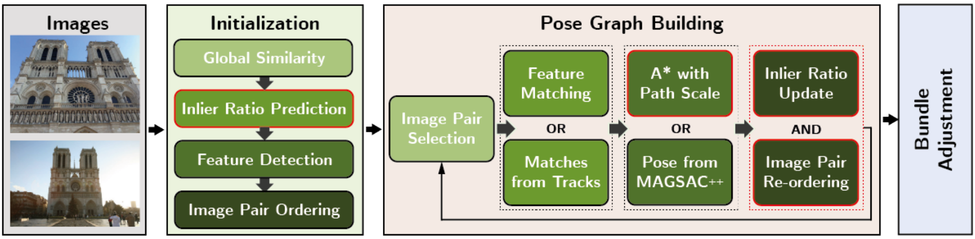


Figure 1. The proposed pipeline creating an initial pose-graph to be then improved, *e.g.*, by bundle adjustment of global Structure-from-Motion algorithms. The new steps compared to [7] are the ones with red outline, *i.e.*, the inlier ratio prediction (Section 2), the path scale estimation for the A*-based path finding (Section 4), and the adaptive image pair re-ordering (Section 3).

ity decreases monotonically during the robust estimation in the next step. Eventually, it falls below the probability of the second best image pair. From that point, the RANSAC iterations are done on a pair which does not have the highest probability of being matchable anymore.

We propose, instead, to alternate between the two steps by always running only a "few" RANSAC iterations on each image pair, possibly, multiple times. This "few" is controlled in a principled way via a Bayesian approach which updates the inlier ratio expectations after every unsuccessful RANSAC iteration. If the expectation falls below an adaptively set threshold, the image pair is put back in the processing queue with an updated probability of being matchable. Again, the next most likely pair is selected and robust estimation is applied to recover the pose of the pair with the highest probability of being matchable. Note that this approach is analogous to playing puzzle, where, first, the most promising piece is selected and the player attempts to locate its place. In the case of failure, the piece is put to the side and the next most promising one comes. This approach can be straightforwardly combined with any image retrieval technique, *e.g.*, GeM descriptors [40].

In order to provide a prior for the inlier ratio expectations, we train a fully connected network on GeM [40] descriptors in a self-supervised manner. Even though GeM descriptors can be used directly to estimate an image similarity score, *e.g.* by the inner product as in [7], we found that it can be further improved by the proposed algorithm. The proposed network predicts the inlier ratio of an image pair expected after running the robust estimation. These predicted inlier ratios are used in the Bayesian approach as prior knowledge.

Moreover, we found that the pose chaining proposed for the A* algorithm in [7] implicitly assumes that the translations are unit-length. Thus, it yields only approximations of the poses. As a technical contribution to the A*-based pipeline, we recover the translation scales along the path found by A*, improving the accuracy of the estimated relative poses and, also, the success rate of the A* algorithm. The proposed pipeline is summarized in Fig. 1.

2 Inlier Ratio Learning

Deciding a priori about which image pairs are matchable (*i.e.*, have a common field-of-view) and, thus, should be processed is an extremely important task for SfM algorithms. The geometric verification procedure, including feature detection, matching and robust pose estimation, is expensive. Applying them to all the $\binom{n}{2}$ combinatorically possible image pairs is unnecessary and impossible in practice. The objective, in this section, is to predict an

Algorithm 1 Pose-Graph by Adaptive Re-ordering.

```
Input: $p_1, \dots, p_r$ – image pairs
Output: $G$ – pose graph (initialization: $G \leftarrow$ EmptyGraph())
$\mu_1, \dots, \mu_r \leftarrow f_\gamma(p_1, \dots, p_r)$    ▷ Inlier ratio prediction
$Q \leftarrow$ Sort$((p_1, \mu_1), \dots, (p_r, \mu_r))$    ▷ Processing queue sorted by predicted inlier ratio
for $(p, \mu) \leftarrow$ NextPair$(Q)$ do    ▷ Selecting the next most likely image pair
    $(\mathbf{R}, \mathbf{t}) \leftarrow$ RANSAC$(p, k(\mu))$    ▷ Run $k(\mu)$ iterations
    if $(\mathbf{R}, \mathbf{t}) \neq 0$ then    ▷ Pose $(\mathbf{R}, \mathbf{t}) \in$ SE(3) found
        $G \leftarrow$ AddEdge$(G, p, \mathbf{R}, \mathbf{t})$    ▷ Pose is accepted, image pair is considered done
    else
        $\mu' \leftarrow$ Update$(p, k(\mu))$    ▷ Bayesian inlier ratio expectation update
        $Q \leftarrow$ Sort$(Q \cup \{(p, \mu')\})$    ▷ Putting the pair back in the queue
```

inlier ratio for each image pair that can be later used for filtering and ordering image pairs.

We use the following approach to generate a fully connected graph as a preliminary step, where the vertices are the images and the edges represent the inlier ratios. We extract GeM [40] descriptors with ResNet-50 [24] CNN, pre-trained on GLD-v1 dataset [34]. In [7], the inner product of the GeM descriptors is used to predict the similarity score. We, instead, propose a *self-supervised* learning approach, where the input of the network is a pair (d_i, d_j) of GeM descriptors of an image pair (I_i, I_j), where $i, j \in \{1, \dots, n\}$, and the output is the expected inlier ratio $\mu_{ij}(1) \in [0, 1]$.

Data Generation. In order to generate training and validation data, we first calculate the GeM descriptors of all images from the current scene. For each of them, we detect RootSIFT keypoints in the way as proposed in [27]. We then iterate through all image pairs.

Since the objective is to predict the inlier ratio, we apply MAGSAC++ [8] to each image pair with its inlier threshold set to 0.75 px (as recommended in [5]), confidence to 0.99 and max. iteration number to 10 000. As shown in [5], more iterations do not improve the accuracy noticeably. The inlier ratio of the found model is used as target for learning.

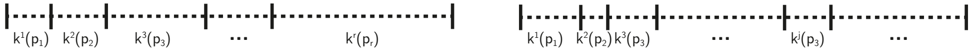
Network Training. We can consider the problem as binary classification, where 0 is a not and 1 is a matchable image pair. For the probability conditioned on (d_i, d_j) of being matchable $p(d_i, d_j)$, logistic regression is used as follows:

$$\ln \frac{p(d_i, d_j)}{1 - p(d_i, d_j)} = f_\gamma(d_i, d_j), \quad \text{with} \quad p(d_i, d_j) = \frac{1}{1 + \exp\{-f_\gamma(d_i, d_j)\}},$$

where $f_\gamma(d_i, d_j)$ is a point estimate, γ are the trained network parameters, (d_i, d_j) is the input image descriptor pair, and we are given a set $\mathcal{D} = \{((d_i, d_j), \mu_{ij})\}_{i,j=1}^{M}$ of training data where M is the number of training images and μ_{ij} are inlier ratios of the image pairs (I_i, I_j).

We train network f_γ to $\mu_{ij} = 1/(1 + \exp\{-f_\gamma(d_i, d_j)\})$ return $p(d_i, d_i) = 1$ if the descriptors are the same, and be permutation invariant, $f_\gamma(d_i, d_j) = f_\gamma(d_j, d_i)$. It is not important to have an accurate ranking of the non-matching images with ≈0 inlier ratio, as long as they are all recognized, contrary to the matching ones, among which we want to discriminate accurately. However, it is crucial to cover the wide distribution of possible negative examples to prevent out-of-distribution negatives from ending up among the image pairs with high inlier ratio predictions. Our solution is a highly unbalanced dataset covering the full distribution of negatives, but with a focal loss [28] that promotes accuracy on positive samples.

We use a small network that allows the proposed scoring technique to be fast. To do so,



(a) Traditional "accept-or-reject" strategy.

(b) Proposed strategy.

Figure 2. Typical RANSAC iteration numbers $k^t(p_i)$ when processing the tth image pair p_i, $i \in \{1,\dots,r\}$. In (a), t equals to i and, usually, $k^t(p_i) \leq k^{t+1}(p_{i+1})$ since the image pairs are ordered, prior to the estimation, by their probability of being matchable. In (b), only a *few* iterations are done on the tth processed image pair controlled by the expected inlier ratio. A pair may be processed multiple times with an updated expected inlier ratio.

we use $n_l \in \mathbb{N}$ linear layers of size s_l, each followed by a 1D batch normalization, a leaky ReLU and a dropout layer. To our experiments, the proposed network works the best with $n_l = 5$ layers of linear, batch normalization and drop-out layers.

3 Adaptive Image Pair Re-ordering

We propose a principled method to adaptively re-order image pairs when processing large-scale datasets to speed up the pose-graph generation by always running on the most likely to match image pairs. In the traditional approach, the pairs are ordered prior to the estimation. A relative pose is estimated from each image pair by running RANSAC. However, such randomized robust estimators tend to be extremely time-consuming on non-matching image pairs or on ones with low inlier ratio. This is caused by the adaptive termination criterion where the implied iteration number depends exponentially on the inlier ratio. In the non-matching or low inlier ratio cases, only a manually set maximum iteration number $k_{\max}$ prevents RANSAC from doing millions of iterations unnecessarily. In such cases the iteration number is always $k_{\max}$ which is still slow in practice.

The termination criterion requiring the confidence in the solution to exceed a manually set threshold works well when estimating the pose of a single view pair. However, its "accept-or-reject" strategy makes large-scale estimation more time-consuming than necessary. Each image pair is processed until we are sure if it is matchable or not, without considering that the probability of being matchable drops monotonically during the estimation and, thus, several other image pairs might become more likely to match. In case of success, the pair is added to the pose-graph. Otherwise, it is rejected and *never* processed again.

In this paper, we propose a new approach, replacing the "accept-or-reject" strategy. This is done by alternating between the selection of the most likely to be matchable image pair and the robust estimation with keeping the estimation light-weight. Under light-weight, we mean that the estimation runs only for a reasonably short time that is enough to solve the easy cases without wasting time on the hard or impossible ones (see Alg. 1).

To do so, we start by predicting the expected inlier ratio $\mu_{ij}(1) = p(d_i, d_j) \in [0,1]$ for each image pair (I_i, I_j) as described in Section 2. Note that this prediction is done once and, in our tests, takes only a few seconds. Using this a priori predicted inlier ratio, we calculate the implied number of RANSAC iterations via formula $k(\eta,\mu,m) = \log(1-\eta)\,/\log(1-\mu^m)$, where $\eta \in [0,1]$ is a manually set confidence (typically, set to 0.99), μ is the inlier ratio, and $m \in \mathbb{N}_{>0}$ is the sample size required for the model estimation, *e.g.*, $m = 5$ in case of estimating essential matrices. For the sake of simplicity, we will write $k(\mu)$ instead of $k(\eta,\mu,m)$.

After calculating iteration number $k(\mu)$ from the predicted inlier ratios, we sort the image pairs in an increasing order as $k(\mu_{i_1^t,j_1^t}(t)) \leq k(\mu_{i_2^t,j_2^t}(t)) \leq \dots \leq k(\mu_{i_p^t,j_p^t}(t))$, where $p = \binom{n}{2}$

is the total number of image pairs, pair (i^t_k, j^t_k) $(k \in \{1,\dots,p\};\ i^t_k, j^t_k \in \{1,\dots,n\};\ i^t_k \neq j^t_k)$ are the indices of the images in an image pair in the t-th iteration (denoted by the upper index; $t \in \mathbb{N}_{>0}$), and n is the number of images. Note that we use $p' \ll p$, as done in [7], by a priori rejecting all image pairs with $\mu < \mu_{\min}$ inlier ratio, where $\mu_{\min}$ is a manually set lower bound for the acceptable inlier ratios.

The robust estimation starts with the image pair with indices (i^t_1, j^t_1) and its maximum iteration number set to $k(\mu_{i^t_1,j^t_1}(t))$. This maximum iteration number is an upper bound. Thus, the estimation finishes if its termination criterion is triggered by finding a model with a non-random number of inliers [25], or if the iteration number exceeds $k(\mu_{i^t_1,j^t_1}(t))$. In case the estimation is successful, *i.e.* the found model has enough inliers, the image pair is removed from the queue and added to the pose-graph. Otherwise, inlier ratio $\mu_{i^t_1,j^t_1}(t)$ is updated (as explained in the next section), image pair with indices (i^t_1, j^t_1) is put back to the queue and the pairs are re-ordered according to the updated inlier ratio. Parameter t becomes $(t+1)$. In practice, this insertion of single value $k(\mu_{i^t_1,j^t_1}(t+1))$ takes $\log p$ time when using a heap. Since the algorithm selects the most probable pair in every iteration, where the iteration number k is likely low, RANSAC always runs only for a short time. Then, the next best pair is selected after updating the expected inlier ratio.

Expected Inlier Ratio Update. We describe an algorithm to update the expected inlier ratio $\mu_{i^t_1,j^t_1}(t)$ of an image pair with indices (i^t_1, j^t_1) after running the robust estimation for $k(\mu_{i^t_1,j^t_1}(t))$ iterations without finding an accurate model. For simplification, let us use notation $\mu(t) = \mu_{i^t_1,j^t_1}(t)$. Since we did not gather additional information about other image pairs, their estimated inlier ratio in the $(t+1)$-th iteration remains unchanged.

To estimate the expected inlier ratio of the processed pair, we use the Bayesian approach. This procedure is applied if the robust estimation, after $k(\mu(t))$ iterations, has not found a model with a reasonable number of inliers and, thus, the pair should be processed again later. Adopting the RANSAC assumption, we consider that this case happens only if there was no all-inlier sample among all the $k(\mu(t))$ tested ones. The random number of all-inlier samples N_{all} in $k(\mu(t))$ samples follows the binomial distribution with parameters $\mu^m(t)$ and $k(\mu(t))$. The usual conjugate prior for a binomial distribution is a beta distribution with prior hyper-parameters $a(t)$ and $b(t)$, having the expectation and variance, respectively, as

$$\frac{a(t)}{a(t)+b(t)}, \qquad v = \frac{a(t)b(t)}{(a(t)+b(t))^2(a(t)+b(t)+1)},$$

and posterior hyper-parameters $(a(t)+N_{all})$ and $(b(t)+k(\mu(t))-N_{all})$. When RANSAC is unsuccessful, *i.e.* $N_{all}=0$, the posterior distribution parameters are $a(t+1)=a(t)$ and $b(t+1)=b(t)+k(\mu(t))$. The best estimator for $\mu^m(t+1)$ using a quadratic loss function is an expectation of the posterior distribution. Consequently, we set

$$\mu^m(t+1) = \frac{a(t+1)}{a(t+1)+b(t+1)}$$

and the next maximum iteration number to $k(\mu(t+1))$.

For each image pair (I_i, I_j), the initial parameters of the beta distribution $a(1)$ and $b(1)$ are set using the predicted inlier ratio $\mu(1) = \mu_{ij}(1) = p(d_i, d_j)$. We assume that the procedure described in Section 2 provides the expectation of the prior beta distribution and with the same mean precision for all image pairs. Therefore, the v variances of all these initial beta distributions are equal and can be learned in advance. Given the learned variance, equations

$$\mu^m(1) = \frac{a(1)}{a(1)+b(1)}, \quad v = \frac{a(1)b(1)}{(a(1)+b(1))^2(a(1)+b(1)+1)}$$

lead to

$$a(1) = \frac{(\mu^m(1))^2(1-\mu^m(1))}{v} - \mu^m(1), \quad b(1) = a(1)\frac{1-\mu^m(1)}{\mu^m(1)}.$$

We compute v from the predicted and GT inlier ratios on the validation set (Section 2).

We add a safe-guard to the iteration number as requiring $\sum_{i=1}^{t} k(\mu(i)) \leq k_{\max}$ to hold for all image pairs. This means that the total iteration number spent on a particular image pair should be lower or equal than a maximum iteration number parameter set by the user. This acts exactly in the same way as in RANSAC to prevent it from running millions of iterations on non-matching pairs. The typical iteration numbers using the traditional and proposed approaches are visualized in Fig. 2.

4 Path Scale Recovery

The algorithm ϕ proposed by Barath *et al.* [7] for recovering the relative pose between views v_s and v_d (s – source; d – destination) from a walk in the pose-graph, *i.e.* a chain of relative poses, provides only an approximation of the translation.

$$\phi(\mathcal{W}) = \phi(f_{w_1}, f_{w_2}, \dots, f_{w_{n-1}}) = \dots = \phi(f_{w_1})\phi(f_{w_2})\dots\phi(f_{w_{n-1}}), \tag{1}$$

where $\mathcal{W} = (f_{w_1}, f_{w_2}, \dots, f_{w_{n-1}})$ is a finite walk in the graph.

The approximative nature of (1) comes from the fact that we are given unit-length translations that renders the final pose from (1) an approximation. In case the absolute scales of the edges are similar, the implied error is marginal. Otherwise, it leads to inaccurate solutions with a low number of inliers. In [7], this approximative nature is not crucial. It can only make A* fail more often and, thus, RANSAC-based robust estimation applied at the cost of only a few milliseconds. However, A* can be made successful more often than in [7] when the path scales are recovered along the found walks. Using these recovered scales, (1) is not an approximation anymore. To find the scales, we apply a similar strategy as proposed in [19]. We select consecutive image triplets and estimate the relative scales along the path. The procedure is shown in depth in the supplementary material.

5 Experiments

We tested the proposed algorithms on the 1DSfM dataset [52]. It consists of 13 scenes of landmarks with photos of varying sizes collected from the internet. 1DSfM provides 2-view matches with epipolar geometries and a reference reconstruction from incremental SfM (computed with Bundler [47, 48]) for measuring error. Since Bundler was published more than ten years ago, we reconstructed the scenes with COLMAP [44] to get a more accurate reconstruction that can be considered ground truth. We use scenes Piccadilly and Madrid Metropolis for training, thus, we do not show results on them.

To get point correspondences in each image pair, we used the RootSIFT [2] algorithm with mutual nearest neighbor check and SNN ratio test [29], as recommended in [27]. We

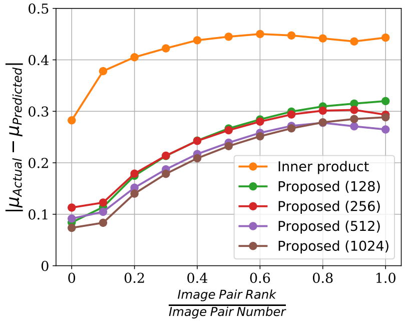


Figure 3. The absolute inliers ratio errors, of the predicted and actual ones, plotted as a function of the image pair rank. The pairs are ranked by their actual inlier ratio. Thus, the left side of the plot shows the error on image pairs with high inlier ratio. The hidden layer size is in brackets.

	Mean error	Median error
Inner product	0.418	0.419
Proposed (128)	0.235	0.250
Proposed (256)	0.234	0.228
Proposed (512)	0.209	0.220
Proposed (1024)	0.205	0.230

Figure 4. The mean and median absolute error of the predicted and ground truth (found by MAGSAC++ after 10 000 iterations) inlier ratios. The numbers in the brackets are the sizes of the used hidden layers. The layer size used in the other experiments is highlighted in gray.

matched all image pairs with global similarity higher than 0.4 with which we got accurate reconstructions in reasonable time. This leads to using only 0.33% (614 366 in total) of all pairs that $\sum_{s \in \text{scenes}} \binom{n_s}{2}$ would imply, where n_s is the number of images in a particular scene. All algorithms are run on the same image pairs.

All methods are implemented in C++. For the A*-based graph traversal algorithm [7], we used the implementation provided by the authors. For robust estimation, we used the state-of-the-art MAGSAC++ algorithm [7] implemented in OpenCV.

Inlier Ratio Prediction. To train the inlier ratio prediction network, we used a total of 80 715 465 image pairs from scenes Piccadilly and Madrid Metropolis. We used the 10% of the image pairs as validation set. For testing the inlier ratio prediction techniques, we used scene Alamo on which no training was performed. Note that we did not filter with the global similarity when training the network, thus, providing a wide range of negative samples.

For Fig. 3, we ranked the image pairs according to their actual inlier ratios (horizontal axis) found by MAGSAC++. The absolute inlier ratio error is shown on the vertical axis. Thus, the left side of the plot shows the errors on pairs with high inlier ratio. On the right side, the error on pairs with low inlier ratio is plotted. The predicted inlier ratios are significantly more accurate than by using the inner product of GeM descriptors, especially, for image pairs with high overlap, *i.e.*, high inlier ratio cases. Moreover, as expected due to using the focal loss, the errors in the predicted inlier ratios are the lowest for the image pairs with the highest overlap (*i.e.*, left part of the figure).

Figure 4 reports the average and median absolute errors of the predicted and actual inlier ratios. We compared the proposed method to using the inner product of the GeM [40] descriptors as similarity score as it was proposed in [7]. Also, we compare the results using different hidden layer sizes shown in brackets. The proposed technique leads to significantly lower errors than using the inner products. The improvement from the layer size stops over 512, where the averages of the mean and median values are the minimal. Therefore, we use 512 as layer size in the other experiments in the rest of the paper.

Image Pair Re-ordering. In order to test the proposed adaptive image pair re-ordering and path scale recovery, we ran the following algorithms on each scene of the 1DSfM dataset:

1. (*Baseline*) Matching the image pairs, which survived the filtering by the inlier ratio

			total time	MAGSAC++			A* pose estimation			overhead
	# edges	# inliers	(hours)	t_{total}	t_{avg}	# runs	t_{total}	t_{avg}	# runs	(hours)
Baseline	417 572	56 148 287	55.11	50.93	4.63	614 366		–		4.18
A* w/o scale [7]	524 221	65 176 541	16.69	13.99	1.27	217 109	**0.042**	**0.004**	525 831	2.66
A* + scale + re-ord.	**554 182**	68 442 654	**2.06**	**1.06**	**0.10**	301 105	0.306	0.028	593 712	0.69
Ablation study										
A* + no scale + re-ord.	532 947	61 900 737	5.94	1.23	0.11	348 344	0.057	0.005	634 217	4.65
A* + scale + no re-ord.	538 119	**70 991 857**	10.96	10.12	0.92	183 904	0.281	0.026	529 280	**0.56**
Baseline + re-ord.	392 779	48 070 653	6.82	2.14	0.19	1 174 609		–		4.68

Table 1. The results of pose-graph creation by the baseline algorithm, A* w/o scale [7] and the proposed A* plus scale recovery and adaptive image re-ordering techniques on the scenes from the 1DSfM dataset with a total of 614 366 image pairs. The bottom part of the table contains results with different algorithm combinations. The reported properties are the total number of edges in the pose-graph (# edges); the number of inlier correspondences (# inliers); the total run-time (in hours); the total, average (over the scenes) run-time of the MAGSAC-based pose estimation and the A* algorithm (all in hours); the number of MAGSAC and A* runs; and the overhead time coming from other parts of the pipeline. The total time is the sum of the MAGSAC++, A* and overhead times. The reported times are projected to a single CPU core, while all of the method can be straightforwardly parallelised.

predicted from GeM descriptors, by MAGSAC++.

2. (*A* w/o scale*) The A*-based technique proposed in [7] combined with MAGSAC++.
3. (*A* + scale + re-ord.*) A* with all the proposed algorithms and MAGSAC++.

We used the proposed inlier ratio prediction for all methods as a preliminary step. Thus, the image pairs are processed in a descending order according to the predicted inlier ratios. While this does not change the results of the baseline, it is beneficial both for [7] and for the proposed method. As ablation study, we also tested combinations *A* + no scale + re-ord.*, *A* + scale + no re-ord.*, and *Baseline + re-ord.* In all combinations, the maximum iteration number and the confidence of MAGSAC++ are set, respectively, to 5000 and 0.99.

Table 1 reports the total number of pose-graph edges and inlier correspondences found in the 11 scenes; the total run-time in hours; the total and average time spent on pose estimation with MAGSAC++ in hours and the number of MAGSAC++ runs; the total and average time spent on A*-based pose estimation in hours and the number of A* runs. It can be seen that the proposed technique combining A*, path scale estimation and adaptive image pair re-ordering is the fastest method. It is an *order-of-magnitude* faster than the baseline algorithm while returning more pose graph edges. Moreover, it is 8–9 times faster than the original A*-based method. Additional results are shown in the supplementary material.

Besides the time spent on MAGSAC++ and the A*-based pose estimation, the overhead is also visible in Table 1. This overhead stems from, *e.g.*, reading correspondences multiple times from the disk or from visibility checks in the pose graph. This additional time is calculated by subtracting the run-times of MAGSAC and A* from the total time. For example, the time overhead of the Baseline method is 4.18 hours (= total time - MAGSAC - A*). The overhead of Baseline + re-ordering is 4.68 hours. Even though the overhead is slightly higher when using the proposed re-ordering, the caused speed-up is more significant, *i.e.*, the total time is reduced by 48.29 hours.

Global Structure-from-Motion. Once relative poses are estimated they are fed to the Theia library [49] that performs global SfM [15, 52]. That is, image matching and relative pose estimation were performed by our code. The key steps of global SfM are robust orientation estimation, proposed by Chatterjee *et al.* [15], followed by robust nonlinear position opti-

	# views	# points	AVG $\epsilon_{\mathbf{R}}$ (°)	MED $\epsilon_{\mathbf{R}}$ (°)	AVG $\epsilon_{\mathbf{p}}$ (m)	MED $\epsilon_{\mathbf{p}}$ (m)
Baseline	820	108 161	9.83	7.41	3.14	2.19
A* w/o scale [7]	815	106 336	9.80	7.41	3.18	2.25
A* + scale + re-ordering	**821**	**106810**	9.61	7.27	**3.05**	**2.04**
A* + no scale + re-ordering	816	106 408	**9.45**	**7.02**	3.17	2.28
A* + scale + no re-ordering	**821**	107 827	9.95	7.62	3.20	2.27
Baseline + re-ordering	819	107 750	9.53	7.11	3.13	2.14

Table 2. The results of a global SfM [49] averaged over the scenes from the 1DSfM dataset [52]. The SfM is initialized with pose-graphs generated by the methods shown in the first column. The reported properties are: number of views (# views) and 3D points (# points) reconstructed by the SfM given an initial pose-graph; the average and median rotation ($\epsilon_{\mathbf{R}}$; degrees) and position errors ($\epsilon_{\mathbf{p}}$; meters).

mization by [52]. The estimation of global rotations and positions enables triangulating 3D points, and the reconstruction is finalized by the bundle adjustment of camera parameters and point coordinates. Since the reconstruction always failed on scene Gendarmenmarkt, we did not consider that scene when calculating the errors.

Table 2 reports the results of Theia initialized by pose-graphs generated by the tested algorithms. All methods lead to similar number of views and 3D points reconstructed and similar accuracy compared to the COLMAP reconstruction considered as ground truth. The proposed technique with path scale recovery and image pair re-ordering leads to the most reconstructed views and the best position accuracy by a small margin. Consequently, while the pose-graph estimation is sped up by an order-of-magnitude, there is no deterioration in the accuracy of the applied structure-from-motion algorithm when using the proposed algorithms. The average run-time of the global SfM bundle adjustment by [49] is 1-5 minutes. Note that the proposed algorithms can also be applied as a prior step for incremental SfMs. However, incremental SfMs often runs for hours or days. Thus, the speed-up from the proposed pipeline is not as big as for global SfMs.

6 Conclusion

The new strategy of estimating relative poses from a large-scale dataset allows for building a pose-graph an 20 times faster than by the traditional approach. The method is 8-9 times faster than the recently proposed [7]. The proposed image pair re-ordering, on its own, has a large impact on the run-time (A* time drops to its 12%; baseline time drops to its 11%). The proposed network predicts the inlier ratio of image pairs significantly more accurately than the method used in [27] with an average error reduced to its half from 0.418 to 0.209. Moreover, we improved the A*-based pose estimation of [7] by estimating the scale of the found paths.

Acknowledgments: This work was supported by the ETH Zurich Postdoctoral Fellowship and by the Research Center for Informatics (project CZ.02.1.01/0.0/0.0/16_019/0000765 funded by OP VVV.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building Rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[2] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012.

[3] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.

[4] Mica Arie-Nachimson, Shahar Z Kovalsky, Ira Kemelmacher-Shlizerman, Amit Singer, and Ronen Basri. Global motion estimation from point matches. In *Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 81–88. IEEE, 2012.

[5] Daniel Barath, Jiří Matas, Dmytro Mishkin, Rene Ranftl, and Tat-Jun Chin. RANSAC in 2020 tutorial. In *CVPR*, 2020. URL http://cmp.felk.cvut.cz/cvpr2020-ransac-tutorial/.

[6] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. MAGSAC++, a fast, reliable and accurate robust estimator. In *CVPR*, pages 1304–1312, 2020.

[7] Daniel Barath, Dmytro Mishkin, Ivan Eichhardt, Ilia Shipachev, and Jiri Matas. Efficient initial pose-graph generation for global sfm. In *CVPR*, pages 14546–14555, 2021.

[8] Daniel Barath, Jana Noskova, and Jiri Matas. Marginalizing sample consensus. *TPAMI*, 2021.

[9] Daniel Barath, Luca Cavalli, and Marc Pollefeys. Learning to find good models in RANSAC. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15744–15753, 2022.

[10] Matthew Brand, Matthew Antone, and Seth Teller. Spectral solution of large-scale extrinsic camera calibration as a graph embedding problem. In *ECCV*, pages 262–273. Springer, 2004.

[11] Song Cao and Noah Snavely. Learning to match images in large-scale collections. In *European Conference on Computer Vision*, pages 259–270. Springer, 2012.

[12] D. P. Capel. An effective bail-out test for RANSAC consensus scoring. In *BMVC*, 2005.

[13] Luca Carlone, Roberto Tron, Kostas Daniilidis, and Frank Dellaert. Initialization techniques for 3d slam: a survey on rotation estimation and its use in pose graph optimization. In *Int. Conf. on Robotics and Automation*, pages 4597–4604. IEEE, 2015.

[14] Luca Cavalli, Marc Pollefeys, and Daniel Barath. NeFSAC: Neurally filtered minimal samples. In *The European Conference on Computer Vision*, 2022.

[15] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In *ICCV*, pages 521–528, 2013.

[16] O. Chum and J. Matas. Randomized RANSAC with tdd test. In *BMVC*, volume 2, pages 448–457, 2002.

[17] Ondrej Chum and Jiri Matas. Matching with PROSAC-progressive sample consensus. In *CVPR*, volume 1, pages 220–226. IEEE, 2005.

[18] Ondřej Chum and Jiří Matas. Optimal randomized RANSAC. *TPAMI*, 30(8):1472–1482, 2008.

[19] Zhaopeng Cui and Ping Tan. Global structure-from-motion by similarity averaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 864–872, 2015.

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

[21] Venu Madhav Govindu. Combining two-view constraints for motion estimation. In *CVPR*, volume 2, pages II–II. IEEE, 2001.

[22] Venu Madhav Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *CVPR*, volume 1, pages I–I. IEEE, 2004.

[23] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *IJCV*, 103(3):267–305, 2013.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[25] Maksym Ivashechkin, Daniel Barath, and Jiri Matas. VSAC: Efficient and accurate estimator for H and F. 2021.

[26] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1704–1716, 2011.

[27] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 2020.

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[29] David Lowe. Object recognition from local scale-invariant features. In *ICCV*. IEEE, 1999.

[30] David Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2): 91–110, 2004.

[31] Daniel Martinec and Tomas Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *CVPR*, pages 1–8. IEEE, 2007.

[32] J. Matas and O. Chum. Randomized RANSAC with sequential probability ratio test. In *ICCV*, volume 2, pages 1727–1732. IEEE, 2005.

[33] K. Ni, H. Jin, and F. Dellaert. GroupSAC: Efficient consensus in the presence of groupings. In *ICCV*, pages 2193–2200. IEEE, 2009.

[34] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *CVPR*, pages 3456–3465, 2017.

[35] Onur Ozyesil and Amit Singer. Robust camera location estimation by convex programming. In *CVPR*, pages 2674–2683, 2015.

[36] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3384–3391. IEEE, 2010.

[37] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *IJCV*, 59(3): 207–232, 2004.

[38] Marc Pollefeys, David Nistér, Jan-Michael Frahm, Amir Akbarzadeh, Philippos Mordohai, Brian Clipp, Chris Engels, David Gallup, S-J Kim, Paul Merrell, et al. Detailed real-time urban 3d reconstruction from video. *IJCV*, 78(2-3):143–167, 2008.

[39] Filip Radenović, Hervé Jégou, and Ondrej Chum. Multiple measurements and joint dimensionality reduction for large scale image search with short vectors. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 587–590, 2015.

[40] Filip Radenović, Giorgos Tolias, and Ondrej Chum. Fine-tuning CNN image retrieval with no human annotation. *TPAMI*, 2018.

[41] Rahul Raguram, Ondrej Chum, Marc Pollefeys, Jiri Matas, and Jan-Michael Frahm. USAC: A universal framework for random sample consensus. *TPAMI*, 35(8):2022–2038, 2012.

[42] Anita Rau, Guillermo Garcia-Hernando, Danail Stoyanov, Gabriel J Brostow, and Daniyar Turmukhambetov. Predicting visual overlap of images through interpretable non-metric box embeddings. In *European Conference on Computer Vision*, pages 629–646. Springer, 2020.

[43] Carsten Rother. *Multi-view reconstruction and camera recovery using a real or virtual reference plane*. PhD thesis, Numerisk analys och datalogi, 2003.

[44] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016.

[45] Johannes L. Schonberger, Alexander C. Berg, and Jan-Michael Frahm. PAIGE: Pairwise image geometry encoding for improved efficiency in structure-from-motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[46] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003.

[47] Noah Snavely, Steve Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ToG*, volume 25, pages 835–846. ACM, 2006.

[48] Noah Snavely, Steve Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008.

[49] Chris Sweeney. Theia multiview geometry library. `http://theia-sfm.org`.

[50] Giorgos Tolias, Teddy Furon, and Hervé Jégou. Orientation covariant aggregation of local descriptors with embeddings. In *European Conference on Computer Vision*, pages 382–397. Springer, 2014.

[51] P. H. Torr, S. J Nasuto, and J. M. Bishop. NAPSAC: High noise, high dimensional robust estimation-it's in the bag. 2002.

[52] K. Wilson and N. Snavely. Robust Global Translations with 1DSfM. In *ECCV*, pages 61–75, 2014.

[53] Changchang Wu. Towards linear-time incremental structure from motion. In *Int. Conf. on 3D Vision*, pages 127–134. IEEE, 2013.

[54] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *CVPR*, pages 4568–4577, 2018.