# Practical solutions to the relative pose of three calibrated cameras
# -
# Supplementary Material

Charalambos Tzamos[1,*] Viktor Kocur[2,*] Yaqing Ding[1] Daniel Barath[3] Zuzana Berger Haladová[2]
Torsten Sattler[4] Zuzana Kukelova[1]

[1]Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague
[2]Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava
[3]ETH Zürich (Zurich), HUN-REN SZTAKI (Budapest)
[4]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague

This supplementary material provides additional details and experimental results promised in the main paper: Sec. 1 provides the proof on the bound of the epipolar error of the mean point correspondence mentioned in Sec. 3.1 of the main paper, additional synthetic and noise experiments that were discussed in Sec. 4 of the main paper, accuracy-speed trade-off results for two-view approximate geometry inside RANSAC (see Sec. 4 of the main paper), and additional details and plots on more scenes for the mean point correspondence accuracy (see Sec. 4 of the main paper). Sec. 2 contains ablation studies to validate our choices regarding the modifications discussed in Sec. 3.2 of the main paper. Sec. 3 provides results for the accuracy-speed trade-off for individual scenes from the PhotoTourism [7] and Cambridge Landmarks [8] datasets, evaluations of the three-view solvers for different thresholds inside RANSAC (see Sec. 4 of the main paper), runtimes of the proposed and state-of-the-art solvers for the 4p3v problem, semi-synthetic experiments for increasing outliers ratios on a scene from PhotoTourism [7], details and results on an alternative evaluation measure (see Sec. 4 of the main paper), and results with GC-RANSAC (see Sec. 4 of the main paper).

## 1. Approximate camera geometry

In this section, in addition to the experiments presented in Sec. 4 of the main paper (paragraph "Approximate camera geometry"), we present additional experiments and results to support our idea of estimating approximate geometry in the first two views. We start with the proof on the bound of the epipolar error of the mean-point correspondence used in the proposed `4p3v(M)`-based solvers (see Sec 3.1 of the main paper). Next, Sec. 1.2 discusses why the mean-point correspondence provides an additional constraint (compared to the original point correspondences) that can be used to estimate the essential matrix. Then, to further assess the accuracy of the two-view variants of our approximate solvers (outside of RANSAC), *i.e.*, `4p(A)`, `4p(M)`, and `4p(M`$\pm\delta$`)`, we design two additional synthetic experiments (see Sec. 1.3). The goal of these synthetic experiments is to study how the accuracy of approximate solutions varies with varying properties of the scene and the cameras. Moreover, Sec. 1.3 provides a synthetic noise experiment on data extracted from a real scene from the ETH3D dataset [13], outside of RANSAC for the three-view solvers. Lastly, Sec. 1.4 contains a speed-accuracy evaluation of the two-view variants of the proposed solvers, inside Poselib RANSAC, and in Sec. 1.5 we study the accuracy of the mean point correspondence (as in Fig. 3 of the main paper).

### 1.1. Proof of the bounds on the epipolar error

While the mean point correspondence $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ used in the `4p3v(M)`-based solvers can provide a good approximation of a correct correspondence, such a correspondence can be noisy. Note that all state-of-the-art 4p3v solvers (including `4p3v(HC)` [6] and the solver from [12]) rely on certain approximations without establishing theoretical proofs to quantify their accuracy. In the `4p3v(HC)` solver [6], the failures that appear quite often are usually the results of tracking a geometrically incorrect solution inside the homotopy continuation method.[1] Thus, this solution can be arbitrarily far from the correct solution. The solver from [12] requires sampling epipoles from a $10^{th}$-degree curve on

[*] Equal contribution

[1]The solver is tracking only one from 272 solutions of the relaxed version of the 4p3v problem and this solution does not need to be a geometrically correct one.

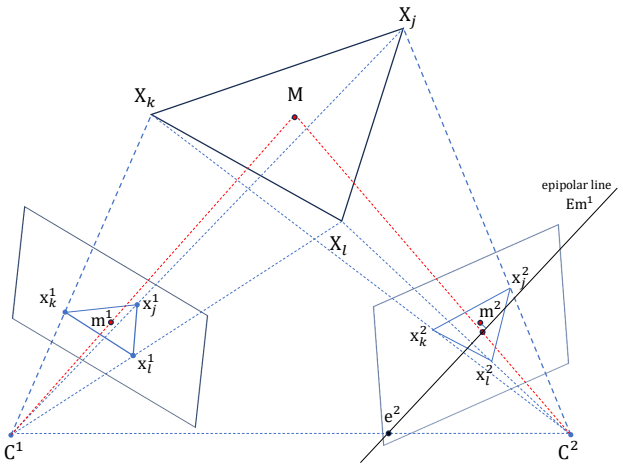



Figure 1. Illustration of the geometric configuration considered in the proof of Lemma 1.

which the true epipole must lie. For any selected point on the curve of epipoles, the error of the sampled epipole is not bounded, since the true epipole can lie anywhere on the unbounded curve. The curve is unbounded since the epipole can be located arbitrarily far away from the image center based on the relative pose of the two cameras (up to infinity for sideways motion).

In contrast, the error of our mean point correspondence is bounded. The $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ correspondence can be seen as a correspondence of points that are projections of the mean point of three 3D points. In this case, the error of both projections $\mathbf{m}^1$ and $\mathbf{m}^2$ can be computed as an error that is introduced by approximating the perspective projection using the para-perspective projection. The approximation error introduced by the para-perspective projection is studied in the literature and can be found *e.g.* in [5, 15].

However, we can also look at the $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ correspondence from a different point of view. We can consider this correspondence as a correspondence in which we fix a point in one view, *e.g.*, $\mathbf{m}^1$, and generate a corresponding point in the second view. In this case, as mentioned in the main paper, it can be proven that the epipolar error of the mean point correspondence $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ is bounded by the maximum distance of the mean point $\mathbf{m}^2$ from the vertices of the triangle $\left\{\mathbf{x}_i^2, \mathbf{x}_j^2, \mathbf{x}_k^2\right\}$. Here we provide a simple proof.

**Lemma 1.** *Let us assume two cameras with camera centers $\mathbf{C}^1$ and $\mathbf{C}^2$ that observe 3D points $X_i, X_j$, and $X_k$ (see Figure 1 for an illustration). Let $\left\{\mathbf{x}_i^1, \mathbf{x}_j^1, \mathbf{x}_k^1\right\}$ and $\left\{\mathbf{x}_i^2, \mathbf{x}_j^2, \mathbf{x}_k^2\right\}$ be the projections of these 3D points in camera 1 and camera 2, respectively. Let $\mathbf{m}^1$ be the mean point of the points $\left\{\mathbf{x}_i^1, \mathbf{x}_j^1, \mathbf{x}_k^1\right\}$ and let $\mathbf{E}$ be the essential matrix between these two cameras,* i.e.*, a matrix that satisfies $\mathbf{x}_l^{2^\top} \mathbf{E} \mathbf{x}_l^1 = 0,\ l \in \{i, j, k\}$. Then the epipolar line $\mathbf{Em}^1$ passes through the triangle $\left\{\mathbf{x}_i^2, \mathbf{x}_j^2, \mathbf{x}_k^2\right\}$.*

*Proof.* The camera center $\mathbf{C}^1$ and the 3D points $\mathbf{X}_i, \mathbf{X}_j$, and $\mathbf{X}_k$ form a tetrahedron $T^1$ (see Figure 1). The projections $\left\{\mathbf{x}_i^1, \mathbf{x}_j^1, \mathbf{x}_k^1\right\}$ in the first camera lie at the edges of this tetrahedron $T^1$. The ray from the camera center $\mathbf{C}^1$ through the mean point $\mathbf{m}^1$ thus lies inside the tetrahedron $T^1$ and intersects the plane defined by 3D points $\mathbf{X}_i, \mathbf{X}_j$, and $\mathbf{X}_k$ in a point $\mathbf{M}$ that lies inside the triangle defined by $\{\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k\}$.

The camera center $\mathbf{C}^2$ and the 3D points $\mathbf{X}_i, \mathbf{X}_j$, and $\mathbf{X}_k$ form a tetrahedron $T^2$. Again, the projections $\left\{\mathbf{x}_i^2, \mathbf{x}_j^2, \mathbf{x}_k^2\right\}$ lie at the edges of the tetrahedron $T^2$. The ray passing through the camera center $\mathbf{C}^2$ and the 3D point $\mathbf{M}$ lies inside the tetrahedron $T^2$ and thus intersects the image plane of the second camera at a point that lies inside the triangle defined by the points $\left\{\mathbf{x}_i^2, \mathbf{x}_j^2, \mathbf{x}_k^2\right\}$. By construction, the projection of $\mathbf{M}$ into the second camera lies on the epipolar line $\mathbf{Em}^1$. Therefore, the epipolar line $\mathbf{Em}^1$ which is a line connecting this point and the epipole $\mathbf{e}^2$, passes through the triangle $\left\{\mathbf{x}_i^2, \mathbf{x}_j^2, \mathbf{x}_k^2\right\}$. □

It follows from Lemma 1 that since the epipolar line $\mathbf{Em}^1$ passes through the triangle $\left\{\mathbf{x}_i^2, \mathbf{x}_j^2, \mathbf{x}_k^2\right\}$, the maximum distance of the mean point $\mathbf{m}^2$ to the epipolar line $\mathbf{Em}^1$ is equal to the maximum distance of $\mathbf{m}^2$ to the vertices of the triangle.

### 1.2. Mean-point constraint

As already mentioned in the main paper, under the assumption of a para-perspective projection, *i.e.*, of affine geometry, the mean point of three 3D points is projected to the mean points of the 3D points' projections in both images [15]. Thus, the mean point correspondence $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ does not add a new constraint if used to estimate an affine camera. This can be easily shown. In the case of affine cameras, the essential matrix $\mathtt{E}_A$ has the form

$$\mathtt{E}_A = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & f \end{bmatrix} . \tag{1}$$

Thus the epipolar constraint for the mean point correspondence $\mathbf{m}^1 \leftrightarrow \mathbf{m}^2$ with the homogeneous coordinates $\mathbf{m}^1 = (\mathbf{x}_i^1/3 + \mathbf{x}_j^1/3 + \mathbf{x}_k^1/3)$, and $\mathbf{m}^2 = (\mathbf{x}_i^2/3 + \mathbf{x}_j^2/3 + \mathbf{x}_k^2/3)$:

$$(\mathbf{m}^2)^\top \mathtt{E}_A \mathbf{m}^1 = 0 \tag{2}$$

can be written as a linear combination of the epipolar constraint for the three input points, *i.e.* $(x_l^2)^\top \mathtt{E}_A \mathbf{x}_l^1 = 0,\ l \in \{i, j, k\}$.

This is not the case for perspective cameras. For perspective cameras, *i.e.*, when estimating the full essential matrix

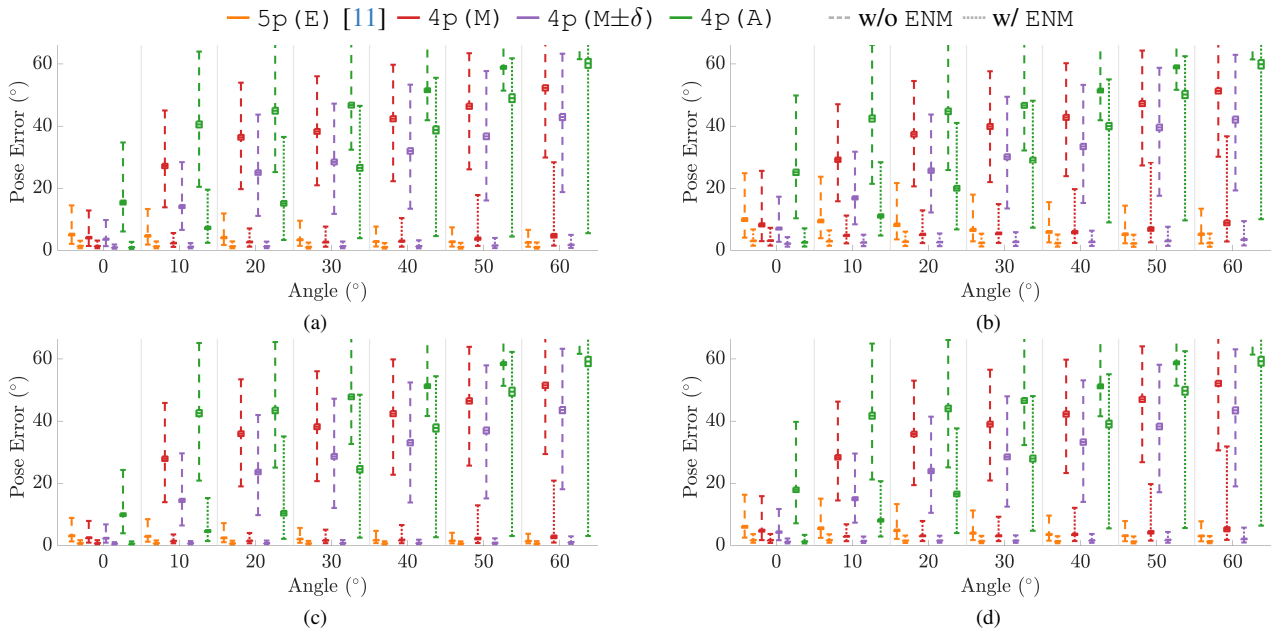



Figure 2. Results from a synthetic experiment evaluating the accuracy of two-view variants of our solvers as a function of the angle between the principal axes of the cameras are presented. The top row, comprising Subfigures (a) and (b), shows results for Gaussian noise with standard deviations of 2px and 4px, respectively. The bottom row, consisting of Subfigures (c) and (d), presents results for uniform noise with 2px and 4px deviations, respectively. The outlier ratio is set to 20% in all cases. From the solutions for each solver and sample we select the one with the lowest error.

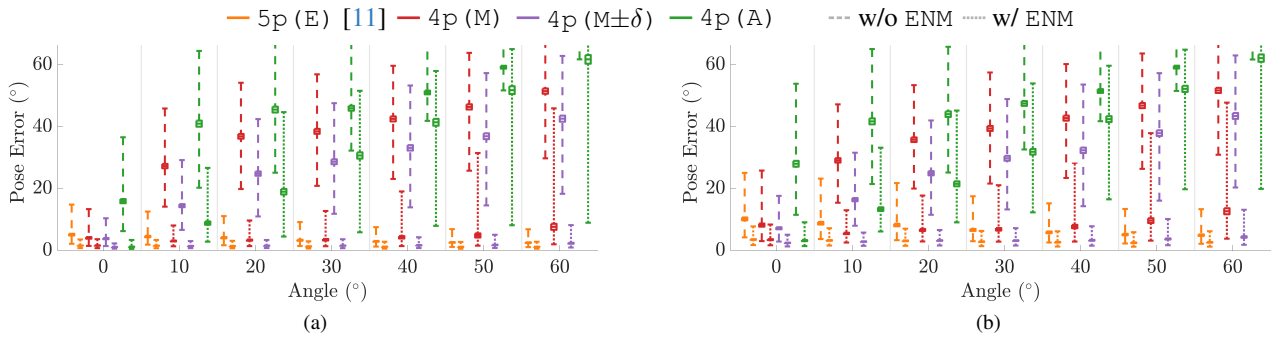



Figure 3. Results from a synthetic experiment evaluating the accuracy of two-view variants of our solvers as a function of the angle between the principal axes of the cameras are presented. We present results for Gaussian noise with standard deviations of 2px and 4px, respectively. The outlier ratio is set to 40% in all cases. From the solutions for each solver and sample we select the one with the lowest error.

$\mathtt{E}$, the mean point correspondence introduces an additional constraint. In this case the epipolar constraint

$$(\mathbf{m}^2)^\top \mathtt{E} \mathbf{m}^1 = 0 \ , \tag{3}$$

after expansion, contains terms $x_a^1 x_b^2, a \neq b, a, b \in \{i, j, k\}$. Thus, the epipolar constraint (3) is not a linear combination of the individual epipolar constraints $(\mathbf{x}_l^2)^\top \mathtt{E} \mathbf{x}_l^1 = 0$, $l = i, j, k$. Therefore, the mean point correspondence provides an independent constraint when used to estimate the epipolar geometry of perspective cameras.

## 1.3. Synthetic Experiments

The error of the relative poses estimated with the proposed approximate `4p3v (M)` -based and `4p3v (A)` -based solvers depends on many aspects, *e.g.*, the baseline and the view angles of the cameras w.r.t. the three points used to compute the mean point correspondence, the depth of these points, the size and shape of the triangles defined by the three points, the type of motion of the cameras, the depth of the scene and the distance of cameras from the scene, the level of noise in the correspondences, *etc*. Isolating the impact

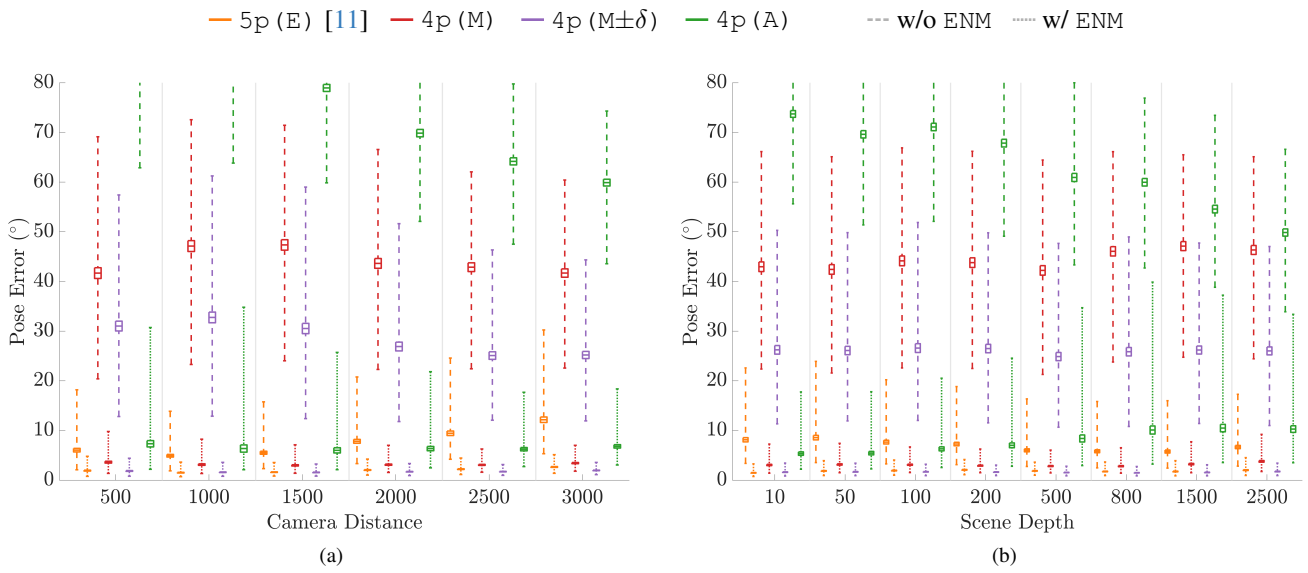



Figure 4. Synthetic experiments for two-view solvers, measuring the pose error under varying camera distance from the scene and scene depths, with added 1px noise. In **(a)**, points are uniformly sampled within a 2000x2000x100 cube, and projected to cameras positioned at distances (in units) from the scene as indicated by the x-axis, looking towards the scene. In **(b)**, the depth of the scene is varied, with points sampled inside a 2000x2000x*depth* cube as specified by the x-axis. Cameras are randomly placed at distances between 1000 and 1200 units from the scene, looking towards the scene. On the y-axis of both figures is the pose error measured as $\max(\mathtt{R}_{err}, \mathtt{t}_{err})$. From the solutions for each solver and sample we select the one with the lowest error. The results are displayed by boxplots which shows the 25% to 75% quantiles as a box with a horizontal line at the median.

of the individual aspects, *e.g.*, through experiments on synthetic data, is highly non-trivial (*e.g.*, how to generate realistic synthetic scenarios that allow conclusions to generalize to real-world scenarios) and analysing the co-dependencies between different aspects on the overall performance seems to need a paper on its own. Moreover, the effect of approximation introduced by using para-perspective projection was already studied in the literature [5, 15].

In the main paper, we thus presented mostly results on real-world scenes, without trying to isolate individual factors (see Figure 3 and Table 1 in the main paper). However, we also tested interesting camera and scene setups using synthetically generated data.

We extend the synthetic experiment for increasing angles between the projection rays of the cameras in the main paper (Fig. 2 of the main paper), by investigating the impact of increased noise, alternative noise models, and higher outlier ratio. Similar to the experiment of increasing angles, we evaluate the two-view solvers (outside of RANSAC) in two additional interesting scenarios. The goal is to study the effect of the proposed approximations on the relative pose estimation under varying properties of the scene and the cameras.

Additionally, we test the performance (outside of RANSAC) of our proposed approximate solvers and the state-of-the-art solvers for the 4p3v problem w.r.t. increasing image noise added to ground-truth correspondences extracted from a scene from the ETH3D dataset [13]. As an evaluation metric for the two-view geometry, we use the pose error measured as $\max(\mathtt{R}_{err}, \mathtt{t}_{err})$ [7].

**Increasing angle between principal axes of cameras.** In Fig. 2, we present results analogous to Fig. 2 of the main paper, but for $\sigma = 2$px and $\sigma = 4$px noise levels. We also include results for the uniform noise model (noise is evenly distributed in range $[-\sigma, \sigma]$). As in Fig. 2 of the main paper, the synthetic data contain 20% outliers. As in the results presented in Fig. 2 of the main paper, the accuracy of both approximate solvers decreases as the angle increases, where `4p(A)` demonstrates notably lower accuracy than `4p(M)` and `4p(M±δ)`. However, ENM singnificantly improves the accuracy, with `4p(M±δ)+ENM` achieving the same or slightly better accuracy than `5pt+ENM` for angle $\leq 30°$. In Fig. 3 we present results for Gaussian noise and higher outlier ratio, in particular 40% outliers. The results are consistent with those of Fig. 2, though the increased outlier ratio leads to a higher error when using ENM.

**Increasing distance of cameras to the 3D scene.** It is known that the quality of the affine approximation, *i.e.*, the approximation of the perspective projection using the para-perspective projection, depends on the distance of the points from the camera [15]. Thus in the first experiment, we evaluate the performance of all solvers w.r.t. increasing distance

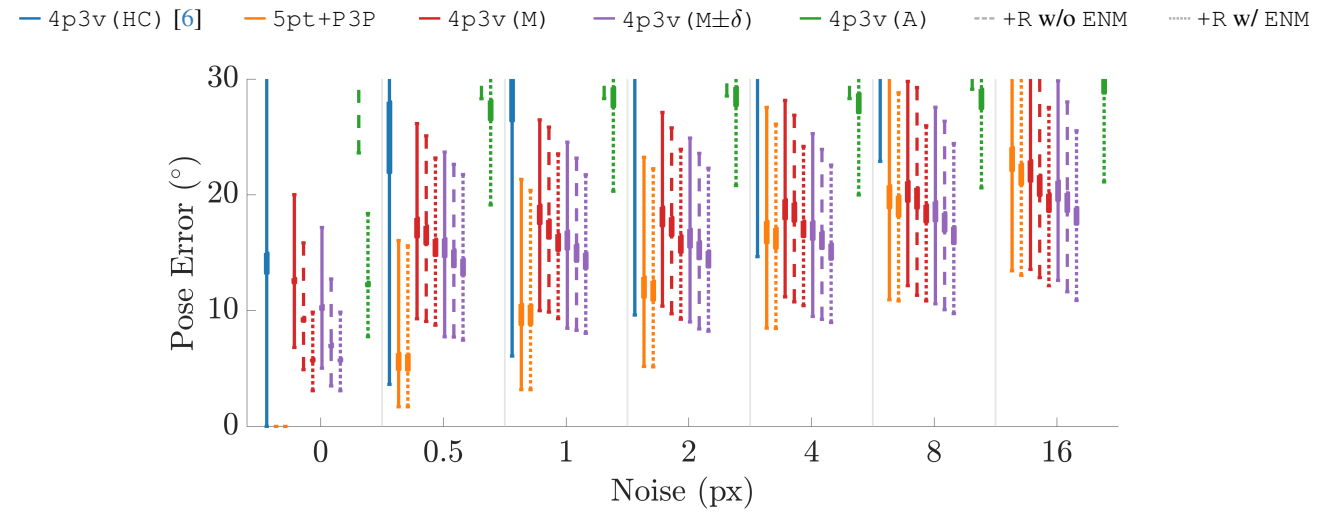



Figure 5. Noise experiment showing the pose error measured as max $\left(0.5(\mathtt{R}^{12}_{err} + \mathtt{R}^{13}_{err}), 0.5(\mathbf{t}^{12}_{err} + \mathbf{t}^{13}_{err})\right)$, as a function of the noise scale in pixels. Here, $\mathtt{R}_{ij}$ and $\mathtt{t}_{ij}$ are the relative rotation and translation of the $i^{\text{th}}$ and $j^{\text{th}}$ views, respectively. From the solutions for each solver and sample we select the one with the lowest error. Note that the errors observed for the pure `4p(A)` solver (without `ENM` and without refinement) are outside of the error range shown in this plot.

of the cameras to the 3D scene.

We perform this experiment on 10k synthetically generated instances. For each of the 10k instances, we uniformly sample 3D points inside a 2000x2000x100 unit cube, and the camera centers are placed at random points with the same distance from the scene. The distances tested (in units) are $\{500, 1000, 1500, 2000, 2500, 3000\}$. The cameras are generated such that they look towards the scene. We add 1px noise to the projected points. Note that the scene is generated such that the projections of the points cover a large portion of the image for cameras at all distances.

Fig. 4a shows the results of this experiment represented by the boxplot function, which shows values between the 25% and 75% quantiles as a box with a horizontal line at the median. As expected, the errors of the `4p(A)` solver decrease as the distance of the cameras from the scene increases, since affine geometry can be better satisfied with larger distances from the scene. Without considering `ENM`, `5p(E)` is the best performing solver. The errors of the `5p(E)` solver are increasing with increasing distance of the cameras from the scene (due to the fact that fixed image noise is generating larger errors for points that are farther from cameras). This effect is less visible for the proposed `4p(M)` and `4p(M±δ)` solvers since for these solvers the error is originally more dominated by the error in the mean point correspondence. When considering `ENM`, `5p(E)`, `4p(M)`, and `4p(M±δ)` solvers perform similarly, with `4p(M±δ)` being the most accurate for distances $\geq 1500$. The `4p(A)` solver is also greatly improved when using `ENM`, reaching similar or even better accuracy than `5p(E)` (w/o `ENM`), for distances $\geq 1000$.

**Increasing depth of the 3D scene.** In Fig. 4b, instead of increasing the distance of the cameras to the 3D scene, we place the cameras randomly at distances between 1000 and 1200 units away from the scene, looking towards the scene, while changing the depth of the scene. In particular, the 3D points are generated uniformly at random inside a 2000x2000x*depth* unit cube, where the depth of the scene is specified by the values on the x-axis. The tested depths are $\{10, 50, 100, 200, 500, 800, 1500, 2500\}$. We add 1px noise to the projected points. Without using `ENM`, the pose errors of `4p(A)` are visibly decreasing as the depth of the scenes increases. This is to be expected since increasing the scene depth increases the chance of sampling four points that are more consistent with the para-perspective / affine camera model (since the points are more likely to be farther away from the cameras). The remaining solvers, that is, `5p(E)`, `4p(M)`, and `4p(M±δ)`, are not significantly affected by the changes in the depth of the scene. When using `ENM`, all tested solvers are improved significantly, with `4p(M±δ)` being the most accurate in terms of pose error. When using `ENM`, the errors in the estimated poses increase with increasing scene depths, which is particularly visible for the `4p(A)` solver. This behavior is due to the fact that for points farther away from the camera, the same amount of image noise (1px) has a larger impact, thus leading to the non-minimal samples being more affected by noise. Still, using `ENM` clearly leads to significantly smaller errors for all solvers.

**Noise experiments.** We test the performance of our solvers and the state-of-the-art solvers w.r.t. increasing im-

age noise. We used the SfM model of the botanical garden scene (randomly selected from all scenes) from the ETH3D dataset [13] to obtain instances of 5 points in three views by identifying images in the scene that share 3D points. Perfect noise-free image correspondences are generated by projecting the 3D points into the images. We then add increasing amounts of normally distributed noise to these correspondences. We generated more than 1k instances. Note that the `4p3v(HC)` solver was trained on the ETH3D dataset [13].

The results for increasing noise in the image points are shown in Fig. 5. The figure shows boxplots of the pose errors measured in the same way as in our experiments in the main paper (*cf*. Sec. 4 in the main paper), *i.e*., as $\max\left(0.5(\mathtt{R}^{12}_{err} + \mathtt{R}^{13}_{err}), 0.5(\mathbf{t}^{12}_{err} + \mathbf{t}^{13}_{err})\right)$.[2] The errors are zoomed into an interesting interval and are shown as functions of varying Gaussian noise from 0px to 16px.

Due to the approximations used in our proposed `4p3v(M)`-based and `4p3v(A)`-based solvers, these solvers exhibit non-zero errors for zero noise. However, at noise levels $\geq$ 4px, our `4p3v(M±δ)+R` solvers return comparable or even better (w/ `ENM`) results than the `5pt+P3P` solver. For noise $\geq$ 8px, also the `4p3v(M)+R` solver with `ENM` returns slightly more accurate poses than the `5pt+P3P` solver with `ENM`. In general, the effect of increasing image noise is less visible for approximate `4p3v(M)`-based and `4p3v(A)`-based solvers. In this case, the error of the approximation is dominating the error introduced by the noise in the image correspondences. While for `4p3v(A)`-based solvers the approximation error is dominant at all noise levels, for `4p3v(M)`-based solvers, at noise $\geq$ 4px, the error introduced by the approximate mean point correspondence (and points in their vicinity in $\delta$-based solvers) is suppressed by the error introduced by noise in the remaining point correspondences.[3] Note that, although the pose errors for the `4p3V(A)+R+ENM` solver are higher than those of the rest of the solvers, as shown in our real experiments, this solver still returns reasonably low errors to provide local optimization (LO) within RANSAC with a good initialization in real-world settings. Further, note that the `5pt+P3P` solver samples one more point (real correspondence) in the first two cameras, and these points are affected only by the considered noise. This shows that the mean point correspondence used in the `4p3v(M)`-based solver is a good approximation to a real correspondence. The recent state-of-the-art `4p3v(HC)` solver [6] is failing in about 50% of the instances for noiseless data, even though the solver was trained on the ETH3D dataset. Thus, the median errors are significantly larger than the median errors of the remaining solvers.

[2]Here $\mathtt{R}^{ij}_{err}$ is the error of the estimated relative rotation between cameras $i$ and $j$, computed as the angle in the axis-angle representation of $\mathtt{R}^{-1}_{ij}\mathtt{R}^{\mathtt{GT}}_{ij}$, and $\mathbf{t}^{ij}_{err}$ is the error of the estimated translation computed as the angle between the two unit vectors corresponding to the translations [7].

[3]This can be seen from the comparable pose accuracy of the `5pt+P3P` and `4p3v(M)` solvers for $\geq$ 8px noise, and the comparable accuracy of the `5pt+P3P` and `4p3v(M±δ)` solvers for $\geq$ 4px noise.

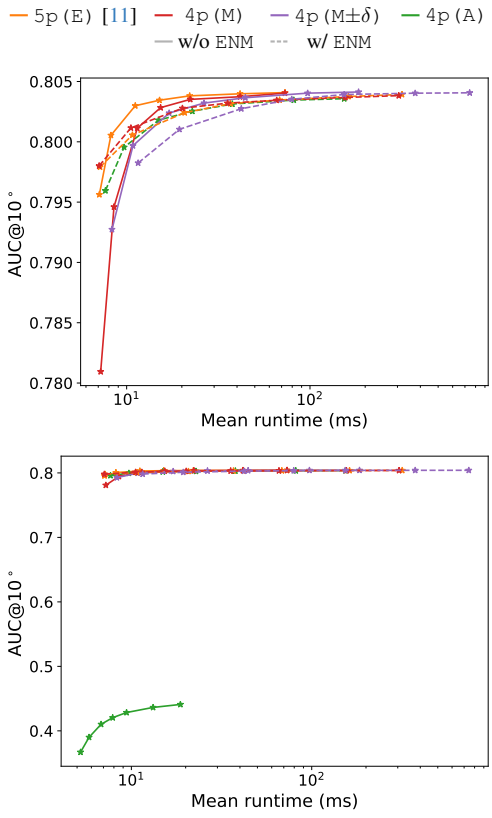



Figure 6. Speed-accuracy evaluation of various solvers for two view relative pose estimation, evaluated using PoseLib [9] on 12 scenes from the Phototourism dataset [7] (excluding *St. Peter's Square*). We report the AUC@10° of the pose error and vary the number of Poselib RANSAC iterations ($\{10, 20, 50, 100, 200, 500, 1000\}$) for a fixed 5 px epipolar threshold. The top plot is a zoomed-in version of the bottom plot.

## 1.4. Two-view approximate solutions inside RANSAC

In the next experiment, we evaluate the discussed two-view solvers, *i.e*., the `5p(E)`, `4p(M)`, `4p(M±δ)`, and `4p(A)` solvers, inside RANSAC as well. This experiment indicates how the proposed approximate solvers would have behaved if used as two-view solvers. Note that in the two-view case, the proposed filtering (`+F`) and refinement (`+R`) using the $4^{th}$ point correspondence in the third view are not applicable.

Fig. 6 shows the speed-accuracy evaluation of the solvers for the problem of two view relative pose estimation evaluated using PoseLib RANSAC [9] on 12 scenes from the Phototourism dataset [7] (excluding *St. Peter's Square*) with pairwise point correspondences obtained us-

ing [3, 10]. The statistic reported is the AUC@10° of the pose error for a varied number of RANSAC iterations ($\{10, 20, 50, 100, 200, 500, 1000\}$) and a fixed epipolar threshold of 5px. The upper figure is a zoom-in of the lower figure to an interesting interval, where differences between the solvers are more visible. Although all proposed solvers (except the `4p(A)` solver without `ENM`) have a performance comparable to that of the state-of-the-art two view `5pt` solver, the `5pt` solver is the best performing one for the two-view scenario. This result is not surprising, given the well-known good performance (in terms of speed and accuracy) of the `5pt` solver and not very high outlier contamination of the data (for which sampling one point less would have potentially had a more visible effect). It also indicates that the proposed modifications, *i.e.*, the filtering `+F` the and refinement `+R`, for the three-view scenario are important and are making the proposed 4p3v approximate solvers practical and more precise than the `5pt+P3P` sovler.

### 1.5. Accuracy of the mean point correspondence

Fig. 3 in the main paper showed results obtained by establishing correspondences between the mean of the triangle in one image and various points in the triangle in the second image. We expressed points in the second triangle via their barycentric coordinates and uniformly sample $19 \times 19$ barycentric coordinates $(a, b) \in [0, 1]^2$, such that $a + b \leq 1$ (ensuring points inside the triangle). The 3rd coordinate is given as $c = 1 - a - b$. For each correspondence, we measured the symmetric epipolar error w.r.t. the ground truth pose, translation and rotation errors, and the percentage of inliers. Fig. 3 in the main paper showed the rotation error and percentage of inliers, as observed for the *St. Peter's Square* scene from the PhotoTourism dataset [7]. Here, Fig. 7 shows the same statistics, including translation and symmetric epipolar errors, for the *St. Peter's Square* scene already used in the main paper (Fig. 7 (top row)), and two more scenes from the PhotoTourism dataset: *Sacre Coeur* (Fig. 7 (middle row)), and *Temple Nara Japan* (Fig. 7 (bottom row)).

As with Fig. 3 in the main paper, to suppress the effect of discrete sampling, for each metric, we fit a 2D Gaussian distribution and report the mean value (in barycentric coordinates) as numbers in brackets in the caption of the figure. As can be seen, the same conclusion can be drawn from Fig. 7 as from Fig. 3 in the main paper: The optima of the studied metrics are reached very close to the mean point of the triangles, which has barycentric coordinates $(0.\bar{3}, 0.\bar{3})$. This validates our approach of using the mean point correspondence as an approximate correspondence in our `4p3v(M)`-based solvers.

| Estimator | $\delta$ | AVG (°) ↓ | MED (°) ↓ | AUC@5 ↑ | @10 ↑ | @20 ↑ |
|---|---|---|---|---|---|---|
| `4p3v(M±δ)` | 0.2 | 4.03 | 2.09 | 57.81 | 73.48 | 84.67 |
| | 0.1 | 3.99 | 2.03 | 58.29 | 73.71 | 84.78 |
| | 0.09 | 3.99 | 2.02 | 58.52 | 73.95 | 84.93 |
| | 0.08 | **3.95** | 2.04 | 58.66 | 74.11 | **85.04** |
| | 0.07 | 4.02 | 2.03 | 58.63 | 74.01 | 84.98 |
| | 0.06 | 4.01 | 2.01 | 58.62 | 73.92 | 84.90 |
| | 0.05 | 3.98 | **1.98** | **58.94** | **74.19** | **85.04** |
| | 0.01 | 4.01 | 2.07 | 57.90 | 73.60 | 84.75 |
| | 0.005 | 4.12 | 2.07 | 57.54 | 73.23 | 84.55 |
| | 0.001 | 4.28 | 2.14 | 56.65 | 72.51 | 84.08 |
| `4p3v(M±δ)+R` | 0.2 | 3.75 | 1.87 | 61.36 | 75.83 | 86.07 |
| | 0.1 | 3.75 | 1.87 | 61.33 | 75.78 | 86.03 |
| | 0.09 | 3.74 | 1.87 | 61.39 | 75.81 | 86.02 |
| | 0.08 | **3.71** | 1.87 | **61.45** | **75.90** | **86.16** |
| | 0.07 | 3.76 | 1.87 | 61.41 | 75.82 | 86.06 |
| | 0.06 | 3.79 | 1.87 | 61.40 | 75.81 | 86.04 |
| | 0.05 | 3.75 | **1.86** | 61.38 | 75.79 | 86.03 |
| | 0.01 | 3.75 | 1.87 | 61.25 | 75.70 | 85.95 |
| | 0.005 | 3.73 | 1.87 | 61.22 | 75.75 | 86.03 |
| | 0.001 | 3.86 | 1.90 | 60.82 | 75.46 | 85.80 |
| `4p3v(M±δ)+R+F` | 0.2 | 3.78 | 1.88 | 61.12 | 75.70 | 85.96 |
| | 0.1 | 3.77 | **1.87** | 61.16 | 75.68 | 85.95 |
| | 0.09 | 3.76 | **1.87** | 61.24 | 75.70 | 85.92 |
| | 0.08 | **3.73** | **1.87** | **61.30** | **75.78** | **86.07** |
| | 0.07 | 3.77 | **1.87** | **61.30** | 75.71 | 85.99 |
| | 0.06 | 3.80 | **1.87** | 61.27 | 75.75 | 85.99 |
| | 0.05 | 3.75 | **1.87** | 61.28 | 75.73 | 85.98 |
| | 0.01 | 3.78 | 1.88 | 61.11 | 75.62 | 85.89 |
| | 0.005 | 3.76 | 1.89 | 61.09 | 75.65 | 85.95 |
| | 0.001 | 3.90 | 1.91 | 60.53 | 75.26 | 85.67 |

Table 1. Evaluation of the effects of the scale of the $\delta$ shift on the *St. Peter's Square* scene from PhotoTourism [7].

## 2. Ablation studies

This section contains ablation studies to validate our choices in modifications discussed in Sec. 3.2 of the main paper.

**Validation of $\delta$.** We tested our $\delta$-based solvers for different values of $\delta$ and measured their performance. In general, there is no common value of the $\delta$ shift that leads to the best results on all datasets. This is expected since the precision of the mean-point correspondence depends on many different factors, *e.g.*, the viewing angles of the cameras, the type of the motion, the depth and spatial distributions of the 3D points, *etc*. We set the value for $\delta$ by evaluating their effects on the *St. Peter's Square* scene from the PhotoTourism dataset [7], which we used for validation only and did not include it in the other results for PhotoTourism [7] in the paper. Tab. 1 shows how the different settings of the scale of the $\delta$ shift affect the accuracy of the $\delta$-based solvers. Based on these experiments, we use $\delta = 0.08$ as it typically provides the best or the second best results for all variants of the `4p3v(M±δ)`-based solvers. However, note that `4p3v(M±δ)`-based solvers achieve a very similar accuracy even with different settings of $\delta$. Thus, we can conclude that the choice of $\delta$ is not critical. In some scenarios, the choice of the optimal $\delta$ parameter may be more scene-dependent and could potentially be set using learning-based approaches.

**Refinement validation.** We also perform validation of the

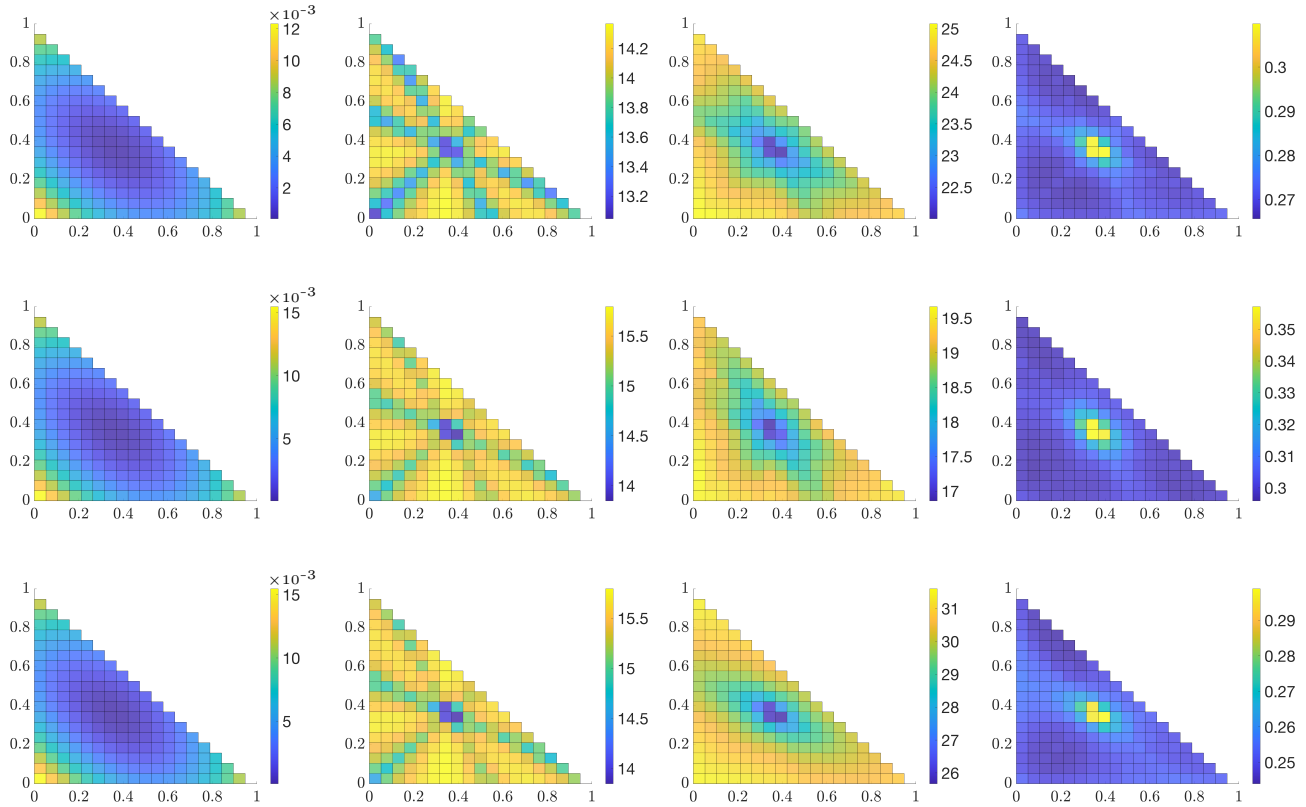


Figure 7. Left to right: Distribution of the average symmetric epipolar error (top: 0.3337, 0.3327) (middle: 0.3319, 0.3308) (bottom: 0.3355, 0.3290); rotation error (top: 0.3373, 0.3349) (middle: 0.3373, 0.3347) (bottom: 0.3261, 0.3496); translation error (top: 0.3336, 0.3417) (middle: 0.3325, 0.3382) (bottom: 0.3213, 0.3515); and percentage of inliers gathered (top: 0.3266, 0.3434) (middle: 0.3377, 0.3354) (bottom: 0.3198, 0.3552), as a function of the barycentric coordinates of the triangle in the second image w.r.t. the mean point of the corresponding triangle in the first image on 485k four-tuples of correspondences from scenes (top) *St. Peter's Square*, (middle) *Sacre Coeur*, and (bottom) *Temple Nara Japan* from the PhotoTourism dataset [7]. For each metric, we fit a 2D Gaussian distribution and report the mean of the distribution in brackets.

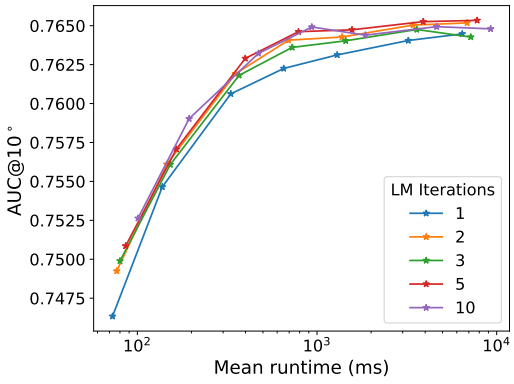



Figure 8. Evaluation of the effects of the number of inner refinement (+R) iterations within the 4p3v(M±$\delta$)+R+F solver on the *St. Peter's Square* scene from the PhotoTourism [7] dataset. Shown is the speed-accuracy evaluation, where the curves are obtained by varying the number of Poselib RANSAC iterations.

total number of LM steps in the refinement (+R). Again, we used the *St. Peter's Square* scene from the PhotoTourism dataset [7] for validation. The results of this experiment are shown in Fig. 8. We chose the value of 2 for other experiments as it provides good speed-accuracy trade-off across a range of RANSAC iterations. However, we note that other settings provide very similar performance.

**Validation of +F/+R/+ENM.** Fig. 9 ablates the impact of individual modifications (+R/+C/+ENM) proposed in Sec. 3.2 in the main paper on the speed-accuracy trade-off. It especially highlights the importance of the refinement using the 4th point in the third view (+R). Fig. 9 also shows the performance of the top-performing solvers when different maximum epipolar thresholds are used within RANSAC. Compared to 4p3v(A)+R+F+ENM, the proposed 4p3v(M±$\delta$) -based solvers are not as sensitive to the selection of the epipolar threshold. The results presented in Fig. 9 were obtained on the PhotoTourism dataset. Fig. 10

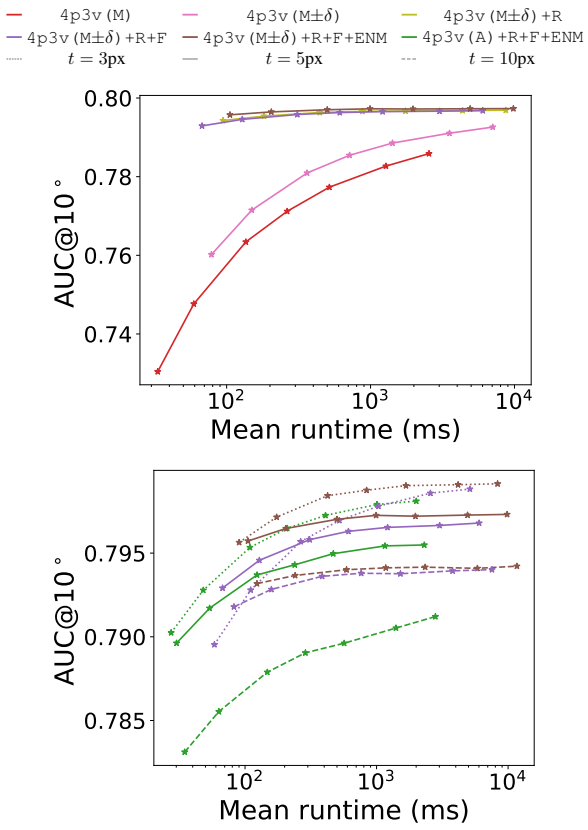



Figure 9. Speed-accuracy trade-off on 12 scenes of Phototourism [7]. We show the impact of (**Top:**) different modifications presented in Sec. 3.2 in the main paper on the performance of the solvers and (**Bottom:**) the maximum epipolar threshold used in RANSAC on the performance of the three best-performing methods.

shows results of the same ablation study, focused on the `4p3v(M)`-based solvers, on the Cambridge Landmarks and Aachen Day-Night v1.1 datasets.

# 3. Experiments on real data

In this section, we aim to further study the performance of the proposed methods, supplementing Sec. 4 (paragraph "Experiments on real data") of main paper with more detailed evaluations. Section 3.1 presents results on individual scenes, extending the analysis in Fig. 4 of the main paper. Section 3.2 investigates the impact of varying the RANSAC epipolar threshold on solver performance. These experiments extend Fig. 9 (bottom) by comparing additional solvers across all three datasets. Section 3.3 evaluates and compares the run-times of each of the proposed and state-of-the-art solvers. Section 3.4 explores the robustness of the solvers under varying inlier ratios using semi-synthetic data. In Section 3.5, we provide results using Poselib RANSAC [9] for an alternative pose error metric that considers errors across all three camera pairs. Section 3.7 evaluates the solvers within the GC-RANSAC [1] framework for all three datasets.

## 3.1. Results on individual scenes

Fig. 4 in the main paper showed results jointly on all PhotoTourism [7] scenes (except *St. Peter's Square*), jointly on the 5 Cambridge Landmarks [8] scenes (except the Street scene, which is commonly not used due to issues with its ground truth), and Aachen Day-Night v1.1 [16]. It also showed results on one individual scene from [8], *i.e.*, the *St. Mary's Church* scene. In Fig. 11, we provide results for the accuracy-speed trade-off evaluation for all remaining individual scenes of PhotoTourism [7] and Cambridge Landmarks [8].

As discussed in the main paper, the accuracy of the proposed approximate solvers is scene-dependent. This also applies to the state-of-the-art `4p3v(HC)` solver [6], since in this solver the scene needs to be similar enough to the training scenes for the MLP-based classifier to work well. The proposed `ENM` refitting suppresses to some extent the scene dependency of the proposed `4p3v(M)`-based and `4p3v(A)`-based solvers. It can be seen that the proposed `4p3v(M±δ)+R+F` solver consistently provides the best speed-accuracy trade-off both with and without `ENM` across all scenes. `4p3v(A)+R+F+ENM` provides a similar performance, typically beating `4p3v(HC)` [6]. However, it may perform worse for some specific scenes, *e.g.*, *Shop Facade* and *Palace of Westminster*. In general, the results on individual scenes are consistent with the results from the main paper.

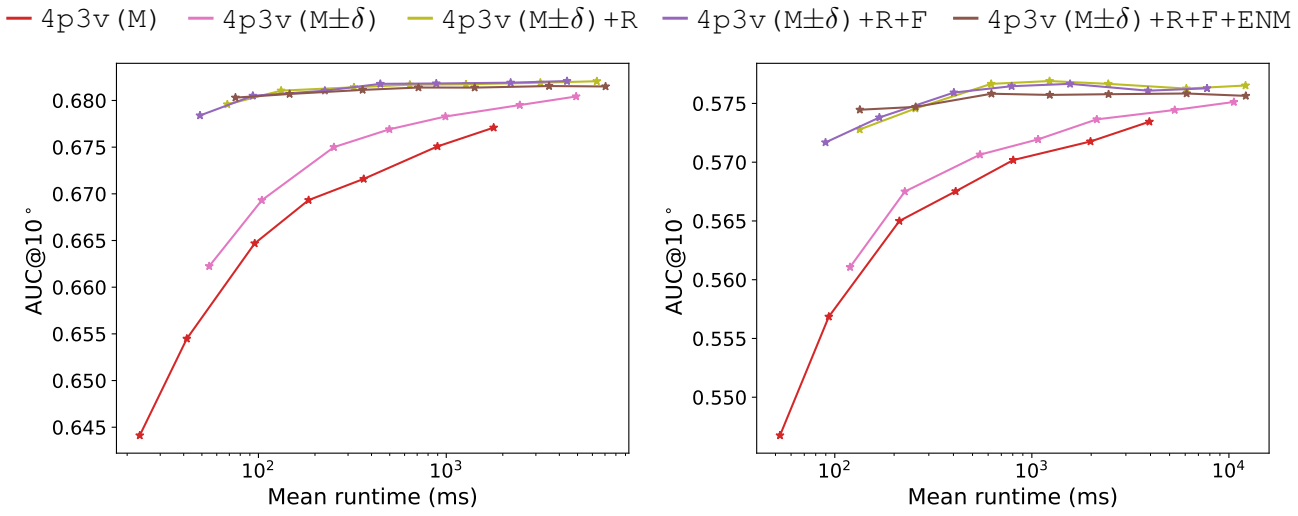
## 3.2. RANSAC threshold sensitivity

In Fig. 9 (bottom), we provided experiments showing how the performance of the selected methods changes when we vary the RANSAC epipolar threshold. In Fig. 12 we provide more extensive results comparing the methods with different thresholds on all three datsets.

Similar to the results presented in the main paper, `4p3v(M±δ)+R+F` in both variants (with and without `ENM`) shows consistently good performance even when using a different threshold in RANSAC. In contrast, `4p3v(A)+R+F+ENM` performs worse than `4p3v(M±δ)+R+F` when considering a higher epipolar threshold in RANSAC.

## 3.3. Solver run-times

In this section, we present run-times of the proposed solvers as well as the state-of-the-art solvers for the relative pose problem of three calibrated cameras. While the main paper reports run-time results for full RANSAC-based estimation, we now report the run-times of the individual solvers out-




(a) Cambridge [8]

(b) Aachen [16]

Figure 10. We show the impact of the different strategies (+F/+R/+ENM) introduced in Sec. 3.2 of the main paper on the performance of our 4p3v(M)-based solvers on Cambridge Landmarks [8] and Aachen Day-Night v1.1 [16]. We report the AUC@10°. We vary the number of Poselib RANSAC iterations ($\{100, 200, 500, 1000, 2000, 5000, 10000\}$). We use an epipolar threshold of 5px in RANSAC. Runtimes are averaged over all image triplets.

| | 5pt+P3P | 4p3v (HC) | 4p3v (M) | 4p3v (M±$\delta$) | 4p3v (A) |
|---|---|---|---|---|---|
| Time ($\mu$s) | 77.90 | 66.06 | 83.92 | 218.71 | 61.12 |

Table 2. The average run-time, averaged over more than 50k instances of the *Sacre Coeur* scene of the PhotoTourism dataset [14], of the solvers for the 4p3v problem.

side of RANSAC. To measure the run-times of the solvers[4], we calculated the average run-time of each solver on more than 50k instances of the *Sacre Coeur* scene of the PhotoTourism dataset [14]. The run-times are reported in Table 2. The experiments were performed on an Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz. In general, the implementations of all proposed solvers are not optimized for speed, and we still see room for speeding them up.

### 3.4. Outlier experiments

To show how the different solvers perform even under varying inlier ratios, we perform a semi-synthetic experiment. We use the *Notre Dame* scene from PhotoTourism [7]. We keep all inlier triplets w.r.t. a 5px epipolar threshold using the ground truth poses. We add additional synthetic outlier correspondences by generating random points in all three views. This allows us to study how the different methods perform when the inlier ratio changes. The results are shown in Fig. 13. As expected, the performance of all solvers decreases with lower inlier ratios. We also observe that 4p3v(M±$\delta$)+R+F performs well even with a low inlier ratio. In contrast, the relative performance of 4p3v(A)+R+F+ENM worsens with a decreased inlier ratio. However, we note that even with an inlier ratio of 40%, it still results in performance comparable to the baseline 5pt+P3P solver. This suggests that a high inlier ratio is not necessary for the ENM to work well in conjunction with the solver 4p3v(A).

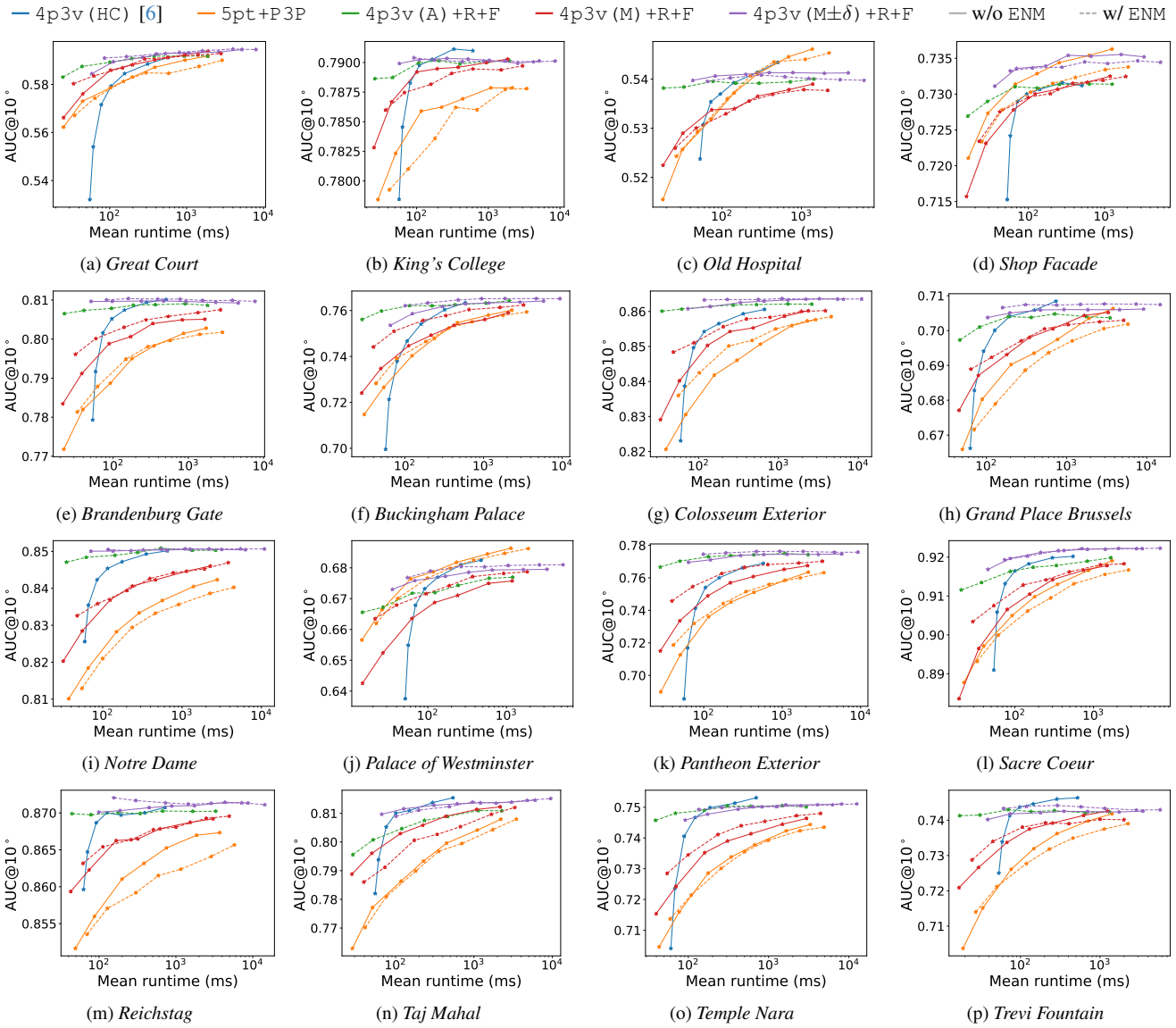
[4]Note that for 4p3v(HC) solver, in Tab 2, the time needed to load the weights of the network (or any other required data) is not added to the runtime of the solver. The data are loaded once per RANSAC, and the loading takes on average 45ms.

### 3.5. Alternative evaluation measure

For the evaluation in the main paper, we defined the pose error as $\max\left(0.5(\mathtt{R}^{12}_{err} + \mathtt{R}^{13}_{err}), 0.5(\mathbf{t}^{12}_{err} + \mathbf{t}^{13}_{err})\right)$, where $\mathtt{R}^{ij}_{err}$ and $\mathbf{t}^{ij}_{err}$ are the angular errors of rotation and translation (both in degrees) for camera pair $ij$. The 4p3v problem also includes the estimation of $\mathtt{R}_{23}$ and $\mathbf{t}_{23}$ since the relative scale of $\mathbf{t}_{12}$ and $\mathbf{t}_{13}$ is recovered. We therefore also present results for the pose error defined as

$$P_{err} = \max\left(\mathtt{R}^{12}_{err}, \mathtt{R}^{13}_{err}, \mathtt{R}^{23}_{err}, \mathbf{t}^{12}_{err}, \mathbf{t}^{13}_{err}, \mathbf{t}^{23}_{err}\right) \quad . \tag{4}$$

The results equivalent to Tab. 2 from the main paper using this pose error definition are presented in Tab. 3. A speed-accuracy comparison equivalent to Fig. 4 in the main paper is presented in Fig. 14. The overall ranking of the methods remains the same under both the metric used in the main paper and the alternative described in this section.




(a) *Great Court* (b) *King's College* (c) *Old Hospital* (d) *Shop Facade*

(e) *Brandenburg Gate* (f) *Buckingham Palace* (g) *Colosseum Exterior* (h) *Grand Place Brussels*

(i) *Notre Dame* (j) *Palace of Westminster* (k) *Pantheon Exterior* (l) *Sacre Coeur*

(m) *Reichstag* (n) *Taj Mahal* (o) *Temple Nara* (p) *Trevi Fountain*

Figure 11. Results for individual scenes from the Cambridge Landmarks [8] (a-d) and Phototourism [7] (e-p) scenes which were not presented in the main paper. We report the AUC@10° of the pose error and vary the number of Poselib RANSAC iterations ($\{100, 200, 500, 1000, 2000, 5000, 10000\}$). We use an epipolar threshold of 5px inside RANSAC. Runtimes are averaged over all image triplets.

### 3.6. Comparison with [12]

The authors of [12] kindly shared their source code with us. Unfortunately, we were not able to run the part of the code that samples epipole candidates from a 10-degree polynomial curve (appropriately sampling the curve is hard as the epipole can be arbitrarily far from the image center). At the same time, the authors were also not able to run it.

Based on the working parts of the code, we tested an oracle version of [12], where instead of sampling the 10-degree polynomial curve, the oracle gives us the correct epipole. Given a sample close to the correct epipole, [12] performs comparable to our M-based solvers. In practice it is hard to find good samples (the epipole can be arbitrarily far from the image center). [12] report using 40-1,000 samples with additional local optimization for robust estimation. Even then, [12] show that this approach performs worse than 5pt+P3P on synthetic data. In contrast, two additional samples in our $\delta$-based solvers already lead to better accuracy than 5pt+P3P.

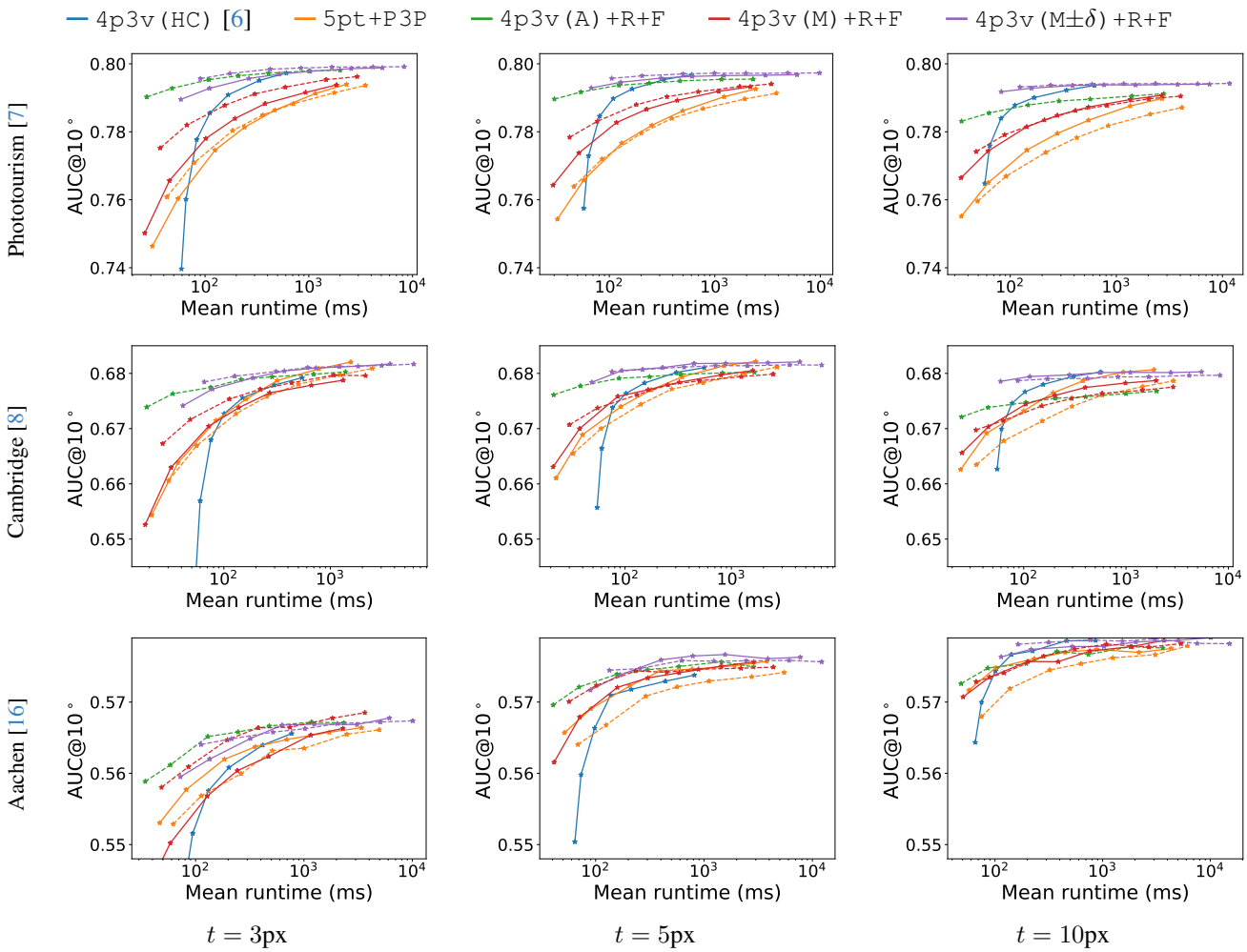
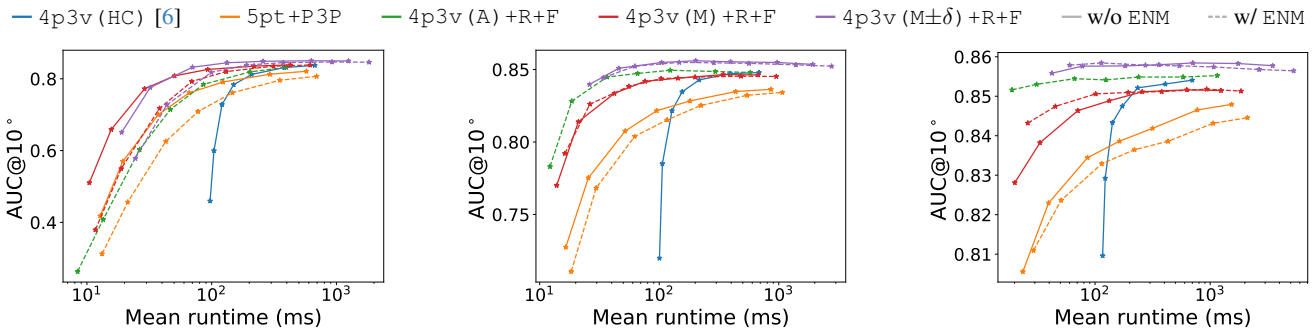



Figure 12. Speed-accuracy trade-off on all scenes from Phototourism [7], except *St. Peter's Square*, 5 scenes from Cambridge Landmarks [8], and the Aachen Day-Night v1.1 dataset [16]. We report the AUC@10° of the pose error and vary the number of Poselib RANSAC iterations ($\{100, 200, 500, 1000, 2000, 5000, 10000\}$) for different maximum epipolar thresholds ($t$). Runtimes are averaged over all image triplets.

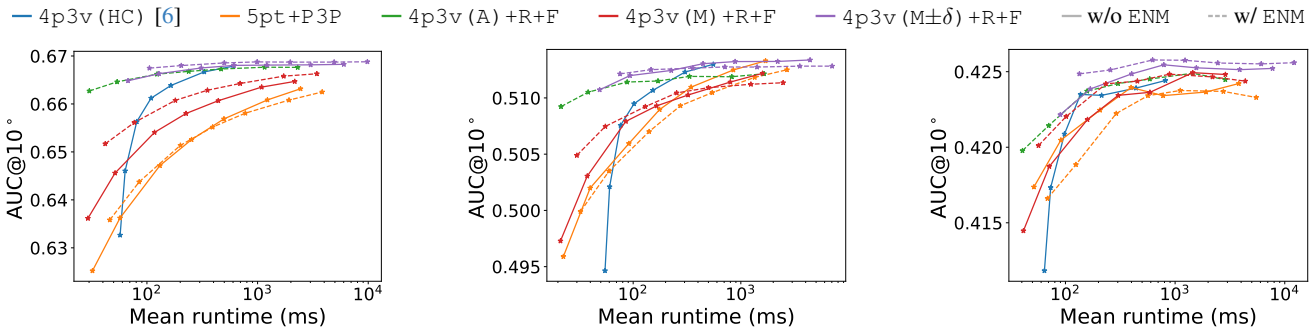



(a) 40% Inlier Ratio (b) 60% Inlier Ratio (c) 80% Inlier Ratio

Figure 13. Speed-accuracy trade-off on the *Notre Dame* scene from Phototourism [7]. We perform semi-synthetic experiments where we add random outlier correspondences to modify the inlier ratios.

| | Phototourism [7] | | | | | | Cambridge Landmarks [8] | | | | | | Aachen Day-Night v1.1 [16] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | AVG (°) ↓ | MED (°) ↓ | AUC@5 ↑ | @10 ↑ | @20 ↑ | Runtime (ms) ↓ | AVG (°) ↓ | MED (°) ↓ | AUC@5 ↑ | @10 ↑ | @20 ↑ | Runtime (ms) ↓ | AVG (°) ↓ | MED (°) ↓ | AUC@5 ↑ | @10 ↑ | @20 ↑ | Runtime (ms) ↓ |
| 4p3v (HC) [6] | 10.28 | 2.81 | 46.97 | 61.37 | 73.16 | 64.10 | 13.95 | 4.58 | 29.54 | 48.68 | 65.12 | 60.11 | 23.46 | 6.58 | 28.56 | 41.20 | 53.12 | 67.26 |
| 5pt+P3P | 9.02 | 2.81 | 47.06 | 61.89 | 74.12 | 33.77 | 12.39 | 4.57 | 29.56 | 49.03 | 65.94 | 24.04 | 21.17 | 6.33 | 29.07 | 41.90 | 54.13 | 53.34 |
| 5pt+P3P+ENM | 8.77 | 2.73 | 47.93 | 62.68 | 74.73 | 48.79 | 12.00 | 4.52 | 29.76 | 49.35 | 66.25 | 34.82 | 21.39 | 6.42 | 28.91 | 41.73 | 53.78 | 71.61 |
| 4p3v (A) | 58.51 | 46.32 | 16.16 | 21.93 | 27.71 | 16.58 | 59.75 | 43.50 | 13.59 | 22.93 | 31.44 | 13.33 | 53.53 | 41.38 | 17.05 | 23.62 | 30.10 | 32.04 |
| 4p3v (A) +ENM | 8.44 | 2.60 | 49.46 | 64.10 | 75.85 | 40.45 | 11.86 | 4.46 | 30.17 | 49.89 | 66.74 | 28.35 | 21.00 | 6.23 | 29.08 | 42.05 | 54.32 | 62.44 |
| 4p3v (A) +R | 53.59 | 38.24 | 18.31 | 24.91 | 31.34 | 16.32 | 54.17 | 24.26 | 16.60 | 27.32 | 36.63 | 12.48 | 51.81 | 38.69 | 17.55 | 24.35 | 31.05 | 29.56 |
| 4p3v (A) +R+F | 56.31 | 43.55 | 16.89 | 23.07 | 29.15 | **11.10** | 55.99 | 30.64 | 15.71 | 26.07 | 35.10 | **9.35** | 54.12 | 42.69 | 16.64 | 23.04 | 29.50 | **19.75** |
| 4p3v (A) +R+F+ENM | 8.45 | 2.59 | 49.59 | 64.22 | 75.92 | 32.37 | 11.78 | 4.41 | 30.41 | 50.12 | 66.93 | 23.43 | 21.11 | 6.29 | 29.02 | 41.97 | 54.18 | 42.36 |
| 4p3v (M) | 9.98 | 3.01 | 45.16 | 60.02 | 72.55 | 35.18 | 13.63 | 4.73 | 28.66 | 47.83 | 64.70 | 25.15 | 23.44 | 6.82 | 27.90 | 40.53 | 52.69 | 54.89 |
| 4p3v (M) +ENM | 8.86 | 2.75 | 47.68 | 62.54 | 74.67 | 48.83 | 12.10 | 4.52 | 29.80 | 49.31 | 66.17 | 34.93 | 21.51 | 6.40 | 28.91 | 41.73 | 53.83 | 70.87 |
| 4p3v (M) +R | 9.24 | 2.74 | 47.74 | 62.44 | 74.47 | 41.52 | 12.76 | 4.53 | 29.77 | 49.27 | 66.12 | 30.76 | 22.06 | 6.39 | 28.92 | 41.69 | 53.70 | 60.40 |
| 4p3v (M) +R+F | 9.19 | 2.71 | 48.10 | 62.73 | 74.69 | 30.88 | 12.84 | 4.53 | 29.74 | 49.29 | 66.11 | 22.72 | 22.06 | 6.42 | 28.70 | 41.53 | 53.62 | 42.38 |
| 4p3v (M) +R+F+ENM | 8.52 | 2.62 | 49.16 | 63.82 | 75.60 | 44.51 | 11.79 | 4.44 | 30.27 | 49.91 | 66.72 | 32.48 | 20.93 | 6.21 | 29.16 | 42.16 | 54.46 | 58.08 |
| 4p3v (M±δ) | 9.28 | 2.94 | 45.92 | 61.05 | 73.66 | 83.70 | 12.68 | 4.59 | 29.35 | 48.88 | 65.89 | 59.23 | 22.09 | 6.42 | 28.60 | 41.55 | 53.76 | 125.02 |
| 4p3v (M±δ) +ENM | 8.44 | 2.73 | 48.01 | 63.02 | 75.20 | 125.66 | 11.61 | 4.47 | 30.04 | 49.75 | 66.68 | 89.05 | 20.80 | 6.22 | 29.26 | 42.16 | 54.32 | 175.54 |
| 4p3v (M±δ) +R | 8.32 | 2.58 | 49.79 | 64.53 | 76.30 | 100.61 | 11.93 | 4.40 | 30.47 | 50.35 | 67.22 | 73.94 | 21.15 | 6.12 | 29.35 | 42.31 | 54.53 | 138.53 |
| 4p3v (M±δ) +R+F | 8.39 | 2.58 | 49.69 | 64.41 | 76.18 | 71.73 | 12.06 | 4.41 | 30.42 | 50.25 | 67.08 | 52.84 | 21.38 | 6.15 | 29.24 | 42.22 | 54.43 | 92.89 |
| 4p3v (M±δ) +R+F+ENM | **7.99** | **2.56** | **49.93** | **64.67** | **76.45** | 112.60 | **11.36** | **4.39** | **30.54** | **50.40** | **67.30** | 81.98 | **20.64** | **6.08** | **29.40** | **42.48** | **54.67** | 139.19 |

Table 3. Experiments with the alternative evaluation measure described in Sec. 3.5. Results for different solvers implemented in the PoseLib framework [9] on all scenes from the PhotoTourism [7], 5 scenes from the Cambridge Landmarks [8], and the Aachen Day-Night v1.1 [16] datasets. We mark the **best** and second best results. Reported runtimes are for the whole RANSAC.




(a) Phototourism [7] (b) Cambridge Landmarks [8] (c) *Aachen Day-Night v1.1* [16]

Figure 14. Experiments with the alternative evaluation measure described in Sec. 3.5: Speed-accuracy trade-off on (a) all scenes from PhotoTourism [7], except *St. Peter's Square*, (b) 5 Cambridge Landmarks [8] scenes, and (c) the Aachen Day-Night v1.1 [16] dataset. We report the AUC@10° using the alternative definition of the pose error (4). We vary the number of Poselib RANSAC iterations ($\{100, 200, 500, 1000, 2000, 5000, 10000\}$). We use an epipolar threshold of 5px in RANSAC. Runtimes are averaged over all image triplets.

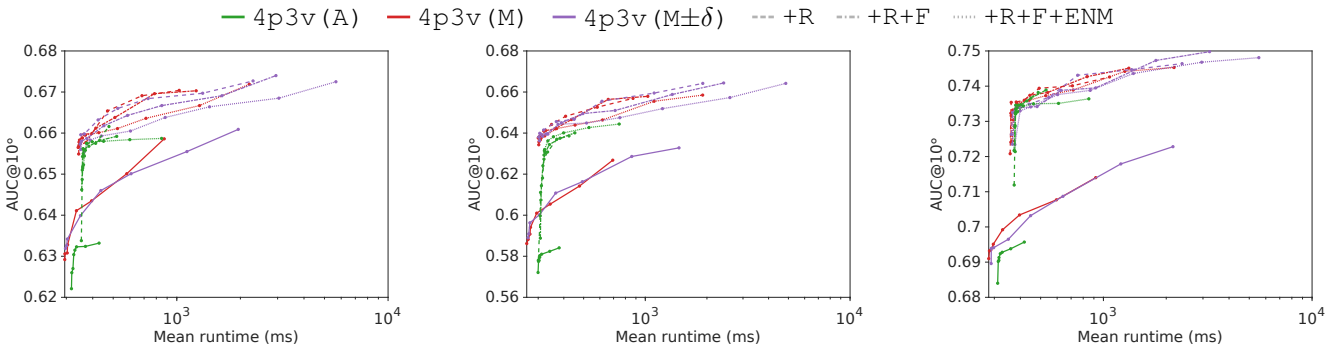



Figure 15. Speed-accuracy evaluation of various solvers for three view relative pose estimation, evaluated using GC-RANSAC [1] on the (left) *St. Mary's Church*, (middle) *Shop Facade*, and (right) *King's College* scenes from the Cambridge Landmarks dataset [8]. We report the AUC@10° of the pose error and vary the number of RANSAC iterations ($\{5, 10, 20, 50, 100, 200, 500, 1000\}$) with fixed 5px epipolar threshold.

### 3.7. GC-RANSAC

Besides PoseLib's RANSAC implementation, we also evaluated and compared our proposed solvers with the state-of-

| Estimator | AVG (°) ↓ | MED (°) ↓ | AUC@5 ↑ | @10 ↑ | @20 ↑ | Runtime (s) ↓ |
|---|---|---|---|---|---|---|
| **Phototourism [7]** | | | | | | |
| 4p3v(HC) [6] | 5.34 | 1.89 | 59.19 | 72.25 | 82.10 | 2.95 |
| 5pt+P3P | 5.18 | 1.86 | 59.48 | 72.35 | 82.16 | **1.99** |
| 5pt+P3P+ENM | 5.15 | 1.86 | 59.53 | 72.39 | 82.20 | 2.05 |
| 4p3v(A) | 5.75 | 1.90 | 59.02 | 71.87 | 81.61 | 2.34 |
| 4p3v(A)+R | 5.69 | 1.72 | 61.59 | 73.83 | 82.88 | 2.98 |
| 4p3v(A)+R+F | 5.76 | 1.73 | 61.53 | 73.79 | 82.83 | 3.00 |
| 4p3v(A)+R+F+ENM | 5.34 | 1.72 | 61.71 | 74.00 | 83.09 | 2.87 |
| 4p3v(M) | 5.21 | 1.88 | 59.35 | 72.29 | 82.15 | **1.99** |
| 4p3v(M)+R | 4.94 | 1.72 | 61.91 | 74.21 | 83.36 | 2.71 |
| 4p3v(M)+R+F | 4.94 | 1.72 | 61.88 | 74.22 | 83.35 | 2.71 |
| 4p3v(M)+R+F+ENM | 4.93 | 1.72 | 61.84 | 74.21 | 83.38 | 2.73 |
| 4p3v(M±$\delta$) | 5.09 | 1.89 | 59.41 | 72.50 | 82.38 | 2.01 |
| 4p3v(M±$\delta$)+R | 4.90 | **1.71** | 61.90 | 74.26 | 83.42 | 2.76 |
| 4p3v(M±$\delta$)+R+F | 4.88 | **1.71** | **61.95** | **74.31** | 83.47 | 2.75 |
| 4p3v(M±$\delta$)+R+F+ENM | **4.86** | 1.72 | 61.90 | 74.29 | **83.48** | 2.84 |
| **Cambridge Landmarks [8]** | | | | | | |
| 4p3v(HC) [6] | 8.13 | 3.05 | 43.75 | 60.73 | 73.93 | 2.37 |
| 5pt+P3P | 8.01 | 3.09 | 43.17 | 60.17 | 73.67 | **2.31** |
| 5pt+P3P+ENM | 8.09 | 3.11 | 43.15 | 60.02 | 73.50 | 2.48 |
| 4p3v(A) | 8.59 | 3.11 | 43.16 | 59.98 | 73.15 | 2.62 |
| 4p3v(A)+R | 7.95 | **2.80** | 46.27 | 63.20 | 75.86 | 2.96 |
| 4p3v(A)+R+F | 8.05 | 2.81 | 46.28 | 63.17 | 75.75 | 2.98 |
| 4p3v(A)+R+F+ENM | 7.75 | 2.81 | 46.38 | 63.20 | 75.86 | 2.98 |
| 4p3v(M) | 7.95 | 3.08 | 43.32 | 60.37 | 73.75 | 2.34 |
| 4p3v(M)+R | 7.22 | **2.80** | 46.48 | **63.52** | 76.30 | 2.86 |
| 4p3v(M)+R+F | 8.05 | 2.81 | 46.28 | 63.17 | 75.75 | 2.98 |
| 4p3v(M)+R+F+ENM | 7.75 | 2.81 | 46.38 | 63.20 | 75.86 | 2.98 |
| 4p3v(M±$\delta$) | 7.82 | 3.06 | 43.58 | 60.60 | 73.99 | 2.41 |
| 4p3v(M±$\delta$)+R | 7.16 | **2.80** | **46.52** | 63.51 | 76.25 | 3.01 |
| 4p3v(M±$\delta$)+R+F | **7.12** | 2.81 | 46.33 | 63.47 | **76.35** | 2.95 |
| 4p3v(M±$\delta$)+R+F+ENM | 7.19 | 2.80 | 46.42 | 63.44 | 76.24 | 3.29 |
| **Aachen Day-Night v1.1 [16]** | | | | | | |
| 4p3v(HC) [6] | 10.73 | 3.84 | 39.94 | 53.01 | 64.77 | 1.90 |
| 5pt+P3P | 10.76 | 3.91 | 39.54 | 52.67 | 64.56 | **1.77** |
| 5pt+P3P+ENM | 10.78 | 3.79 | 39.86 | 53.06 | 64.79 | 1.89 |
| 4p3v(A) | 11.06 | 3.88 | 39.54 | 52.63 | 64.19 | 2.24 |
| 4p3v(A)+R | 10.09 | 3.50 | 42.64 | 55.75 | 67.01 | 3.32 |
| 4p3v(A)+R+F | 10.22 | 3.49 | 42.58 | 55.75 | 66.87 | 3.38 |
| 4p3v(A)+R+F+ENM | 10.23 | 3.48 | 42.58 | 55.75 | 66.88 | 3.32 |
| 4p3v(M) | 10.62 | 3.83 | 39.62 | 52.92 | 64.74 | 1.87 |
| 4p3v(M)+R | 10.11 | 3.46 | 42.69 | 55.90 | 67.11 | 3.17 |
| 4p3v(M)+R+F | 10.22 | 3.53 | 42.58 | 55.62 | 66.71 | 3.18 |
| 4p3v(M)+R+F+ENM | 10.06 | 3.52 | 42.55 | 55.76 | 67.00 | 3.20 |
| 4p3v(M±$\delta$) | 10.55 | 3.84 | 39.86 | 53.30 | 65.15 | 1.89 |
| 4p3v(M±$\delta$)+R | **10.04** | **3.45** | **42.91** | **56.02** | **67.15** | 3.24 |
| 4p3v(M±$\delta$)+R+F | 10.15 | 3.47 | 42.65 | 55.76 | 66.94 | 3.17 |
| 4p3v(M±$\delta$)+R+F+ENM | **10.04** | 3.50 | 42.67 | 55.82 | 66.95 | 3.33 |

Table 4. Results for different solvers and strategies implemented in the GC-RANSAC framework [1] for all scenes from the PhotoTourism [7], the Cambridge Landmarks [8] and Aachen Day-Night v1.1 [16] datasets. We mark the **best** and second best results. Runtimes are reported in seconds for the whole RANSAC with early termination (0.9999 confidence, minimum 100 iterations) and the epipolar threshold set to 5px.

the-art solvers inside the GC-RANSAC [1] framework.

In GC-RANSAC, local optimization (LO) is performed using non-minimal solvers that fit models to larger-than-minimal samples. We use the non-minimal version[5] of the 5pt solver [11] and the non-minimal absolute pose DLSPnP [4] solver.[6] In contrast to LO used in Poselib RANSAC, where the estimated model is used as an initialization of the Levenberg–Marquardt algorithm, in GC-RANSAC, the estimated model is used only to score inliers.

Tab. 4 shows the results for GC-RANSAC with PROSAC sampling [2] and a 5px epipolar threshold for all scenes from the PhotoTourism [7] dataset, 5 scenes from the Cambridge Landmarks [8] dataset, and the Aachen Day-Night v1.1 [16] dataset. Similarly to what was observed for Poselib RANSAC (see Table 2 in the main paper), with the suggested modifications, all the proposed solvers outperform the state-of-the-art 4p3v(HC) solver [6] and the baseline 5pt+P3P solver in terms of pose accuracy with comparable runtimes. Again, the $\delta$-based solvers provide, in general, the best speed-accuracy trade-off.

Due to a different LO, there are several differences compared to the results from Poselib RANSAC. Since in GC-RANSAC the estimated model is used only to score inliers, it does not need to be as precise as in Poselib RANSAC. Thus, even 4p3v(A) solver without any modification provides reasonably precise results.[7] In contrast to this, in Poselib RANSAC, the 4p3v(A) solver without any modification results in large errors (see Table 2 in the main paper). Without refitting using ENM, the affine model estimated for the first two views in the 4p3v(A) solver is not sufficiently precise to provide a good initialization for Levenberg–Marquardt-based optimization in Poselib's LO. Still, even for GC-RANSAC, the pure 4p3v(A) solver performs worse than the remaining variants of the proposed 4p3v(A)-based and 4p3v(M)-based solvers.

Another difference is in refitting using ENM. For GC-RANSAC, the effect of ENM is not as significant as for Poselib RANSAC. When applied without refinement (+R), the early non-minimal refitting (ENM), in general, increases the precision of solvers. When combined with +R, the improvement is not very visible. The reason is that the model returned after refining the initial, approximate model estimated by the 4p3v(M)-based and 4p3v(A)-based solvers on the $4^{th}$ correspondence is usually sufficiently accurate to score inliers. Moreover, in the LO of GC-RANSAC, this approximate model is refitted using the non-minimal 5pt solver (which is similar to the refitting that is used in ENM), and the non-minimal DLSPnP solver [4]. Similarly to ENM, the filtering (+F) that uses the $4^{th}$ correspondence does not bring an improvement that is as visible as for Poselib RANSAC. This is because for GC-RANSAC, the speedup obtained using the filtering +F is not as significant, compared to the longer running times of the LO part of GC-RANSAC.

On the other hand, the remaining two suggested modifications, *i.e.*, the $\delta$-based solvers and the refinement (+R) using the $4^{th}$ correspondence bring visible improvements. This behavior is also visible in Figure 15. Here we present

[5] The non-minimal version of the 5pt solver [11] uses the last four vectors from the SVD/QR decomposition of a $n \times 9$ matrix instead of the 4-dim null space of a $5 \times 9$ matrix to parameterize the unknown essential matrix.

[6] This may not be the most efficient way how to perform non-minimal refitting. However, since all methods use the same LO, it is sufficient for a fair comparison.

[7] Note that the model of 4p3v(A) is refitted in the LO step with the non-minimal 5pt solver. This is similar to the refitting used in ENM, *i.e.*, GC-RANSAC's local optimization includes some form of ENM, explaining why the 4p3v(A) performs quite well.

an ablation study on the effects of the various modifications ($\delta$ and `+F/+R/+ENM`, which were introduced in Sec. 3.2 of the main paper) on the `4p3v(M)`-based and `4p3v(A)`-based solvers. The results are reported on the *St. Mary's Church*, *Shop Facade*, and *King's College* scenes from the Cambridge Landmarks dataset [8]. These results especially highlight the importance of the refinement using the $4^{th}$ point in the third view (`+R`). On the other hand, the benefits of `4p3v(M`$\pm\delta$`)`-based solvers over the `4p3v(M)`-based solvers that are also visible in Table 4 are not so significant as in Poselib RANSAC. Still, `4p3v(M`$\pm\delta$`)`-based solvers lead to improved pose accuracy.

## References


[1] D. Barath and J. Matas. Graph-Cut RANSAC. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6733–6741, 2018. 9, 13, 14

[2] Ondrej Chum and Jiri Matas. Matching with prosac-progressive sample consensus. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 220–226. IEEE, 2005. 14

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 7

[4] Joel A. Hesch and Stergios I. Roumeliotis. A direct least-squares (dls) method for pnp. In *2011 International Conference on Computer Vision*, pages 383–390, 2011. 14

[5] Radu Horaud, Fadi Dornaika, and Bart Lamiroy. Object pose: The link between weak perspective,paraperspective, and full perspective. *Int. J. Comput. Vision*, 22(2):173–189, 1997. 2, 4

[6] Petr Hruby, Timothy Duff, Anton Leykin, and Tomas Pajdla. Learning to solve hard minimal problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5532–5542, 2022. 1, 5, 6, 9, 11, 12, 13, 14

[7] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 1, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14

[8] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 1, 9, 10, 11, 12, 13, 14, 15

[9] Viktor Larsson. PoseLib - Minimal Solvers for Camera Pose Estimation, 2020. 6, 9, 13

[10] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *International Conference on Computer Vision (ICCV)*, pages 17627–17638, 2023. 7

[11] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004. 3, 4, 6, 14

[12] D. Nistér and F. Schaffalitzky. Four points in two or three calibrated views: Theory and practice. *International Journal of Computer Vision*, 67(2):211–231, 2006. 1, 11

[13] Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark With High-Resolution Images and Multi-Camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 4, 6

[14] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 10

[15] Zhengyou Zhang. *Weak Perspective Projection*, pages 877–883. Springer US, Boston, MA, 2014. 2, 4

[16] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *International Journal of Computer Vision*, 129(4):821–844, 2021. 9, 10, 12, 13, 14