StereoGlue: Robust Estimation with Single-Point Solvers

Daniel Barath[1], Dmytro Mishkin[2,4], Luca Cavalli[1], Paul-Edouard Sarlin[1], Petr Hruby[1], and Marc Pollefeys[1,3]

[1]ETH Zurich [2]VRG, Faculty of Electrical Engineering, CTU in Prague, Czech Republic [3]Microsoft [4]HOVER Inc.

Abstract. We propose StereoGlue, a method designed for joint feature matching and robust estimation that effectively reduces the combinatorial complexity of these tasks using single-point minimal solvers. StereoGlue is applicable to a range of problems, including but not limited to relative pose and homography estimation, determining absolute pose with 2D-3D correspondences, and estimating 3D rigid transformations between point clouds. StereoGlue starts with a set of one-to-many tentative correspondences, iteratively forms tentative matches, and estimates the minimal sample model. This model then facilitates guided matching, leading to consistent one-to-one matches, whose number serves as the model score. StereoGlue is superior to the state-of-the-art robust estimators on real-world datasets on multiple problems, improving upon a number of recent feature detectors and matchers. Additionally, it shows improvements in point cloud matching and absolute camera pose estimation. The code is at: `https://github.com/danini/stereoglue`.

Keywords: robust estimation · RANSAC · feature matching

1 Introduction

Matching multiple observations (*e.g.*, image-to-image, image-to-point cloud, point cloud-to-point cloud) of the same scene is a fundamental problem in computer vision and robotics with a wide range of applications. These include image retrieval [2, 61, 75, 80, 101], Structure-from-Motion [1, 10, 50, 94, 119], localization [63, 76, 89, 91], SLAM [31, 32, 37, 72], multi-view stereo [24, 40, 41, 54], and point cloud mosaicking [22, 42, 107, 111].

Conventionally, the matching process adheres to a three-stage framework: local feature detection, feature matching, and geometric robust estimation. Its sequential nature poses a significant challenge, as failures in any stage lead to an overall failure, undermining the reliability of the entire process. While recent algorithms [23, 79, 98, 108] perform feature detection and matching jointly, at the cost of significantly increased run-time for all-pair 3D reconstruction, a gap remains in the literature of methods for simultaneous matching and robust estimation. To address this deficiency, we introduce *StereoGlue*, a novel method

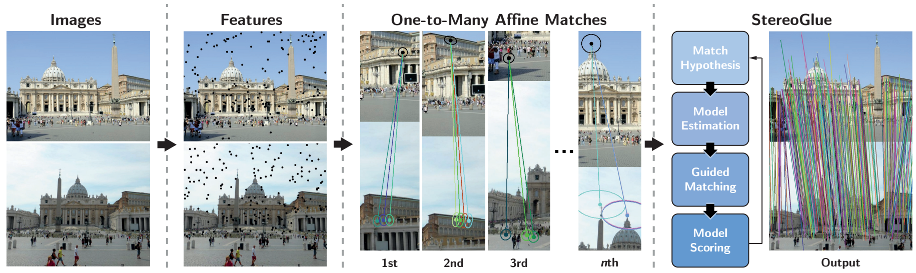

Fig. 1: For two-view estimation, the steps of the proposed **StereoGlue** are: (1) features with affine shapes are detected in the input images, *e.g.*, by SuperPoint [30] combined with AffNet [68]. (2) For each feature in the source image, the matching by, *e.g.* SuperGlue [88], is often ambiguous, especially at repeated patterns. Thus, we form *one-to-many* matches for each point in the source image. (3) StereoGlue iteratively selects a candidate one-to-one correspondence and estimates the model (*e.g.*, relative pose) by a single-point solver. Guided sampling then forms one-to-one correspondences consistent with the estimated model to calculate its score and select its inliers.

performing joint matching and robust estimation by iteratively selecting potential matches, estimating the model, and performing guided matching to calculate the model score and select its inliers. While most methods must commit to one-to-one matches to keep the problem tractable, we relax this to one-to-k matches, making matching more robust and accurate in real-world scenes.

Feature detection and matching. Local image features are the main workhorse in 3D reconstruction. Traditionally, such features encompass three main steps: (scale-covariant) keypoint detection, orientation estimation, and descriptor extraction. Keypoint detection typically operates on a scale pyramid, using handcrafted response functions such as Hessian [17,66], Harris [46,66], Difference of Gaussians [61], or learned alternatives like FAST [86] or Key.Net [15]. Keypoint detection provides a triplet (x, y, scale) that defines a square or circular patch. Subsequently, the patch orientation is obtained using handcrafted approaches, such as the dominant gradient orientation [61] or center of mass [87], or learned ones like [59,68,112]. Optionally, the affine-covariant shape [16,68] might be determined. Finally, the patch is geometrically rectified and described using local patch descriptors such as SIFT [61], HardNet [67], SOSNet [100], and others.

Recent advances in deep learning have led to feature detection and description methods that do not rely on patch extraction. Methods like SuperPoint [30], R2D2 [84], D2Net [34] and DISK [103] employ feedforward Convolutional Neural Networks and assume up-is-up image orientation. Some recent methods have proposed learning matching directly, such as SuperGlue [88] or LightGlue [60], while others skip the detection step entirely [23,98,108]. While operating in a different domain, state-of-the-art pairwise point cloud registration algorithms [49,79,114] perform similar steps to find corresponding 3D points.

Robust Estimation. Feature matching often leads to several outliers inconsistent with the scene geometry. This holds especially in wide-baseline cases, where

Algorithm 1 StereoGlue

Input: $\mathcal{P}_1$, $\mathcal{P}_2$ – two sets of data points
Output: $\mathcal{M}^*$ – correspondences, θ – model params.
$\theta^* \leftarrow \mathbf{0}, q^* \leftarrow 0$, $\mathcal{M}^* \leftarrow \varnothing$ ▷ Initialization
while $\neg$Terminate() **do**
  $\mathcal{S} \leftarrow$ NextBestMatch($\mathcal{P}_1, \mathcal{P}_2$) ▷ Generate a match
  $\theta \leftarrow$ EstimateModel($\mathcal{S}$) ▷ A one-point solver
  $\mathcal{M} \leftarrow$ GuidedMatching($\theta, \mathcal{P}_1, \mathcal{P}_2$)
  $q \leftarrow$ GetScore($\theta, \mathcal{M}$)
  if $q > q^*$ **then** ▷ Update the best model
    $q', \theta', \mathcal{M}' \leftarrow$ LocalOptimization($\theta, \mathcal{P}_1, \mathcal{P}_2$)
    $\theta^* \leftarrow \theta', q^* \leftarrow q'$, $\mathcal{M}^* \leftarrow \mathcal{M}'$

the inlier ratio often falls below 10%. Robust estimation is thus crucial to find the sought model (*e.g.*, relative pose) and the matches consistent with it. Classical approaches employ a RANSAC-like [38] hypothesize-and-verify strategy, iteratively applying minimal solvers [38, 47, 48, 56, 57, 97] to random subsets of the input data until an all-inlier sample is found. To improve upon RANSAC, various techniques have been developed, such as local optimization methods (LO-RANSAC, LO$^+$-RANSAC, and GC-RANSAC) [8,27,58], advanced scoring functions (MLESAC, MSAC, MAGSAC, and MAGSAC++) [4,9,11,102], speed-ups using probabilistic sampling (PROSAC, NAPSAC, and P-NAPSAC) [11,25,73], preemptive verification (SPRT and SP-RANSAC) [13, 26], degeneracy checks (DEGENSAC, QDEGSAC, and NeFSAC) [21, 28, 39], and methods for auto-tuning of the inlier threshold (MINPRAN and a contrario RANSAC) [69,85,96].

Recently, several learning-based algorithms have been proposed for robust relative pose estimation. Such methods generally fall into two main categories: ones aiming to learn correspondence weights for an iteratively re-weighted least-squares approach [82, 99, 113, 115] or for outlier pre-filtering [117]. Other ones learn importance scores to condition the random sampling process [18,109,110].

Motivation. Despite the recent progress, feature matchers still have to commit to one-to-one matches even if such a decision is ambiguous (*e.g.*, due to repetitive structures) without knowing the underlying scene geometry. On the other hand, jointly performing feature matching and robust model estimation is a prohibitively complex problem, making it impractical in the general case. For example, when matching n features, the complexity is n^2. Injecting this into the complexity of robust estimation, we get $\binom{n^2}{m}$, where m is the sample size to fit a minimal model, such as $m = 5$ for essential matrix estimation. This makes the probability of selecting an all-inlier sample that leads to an accurate model extremely low. Having 1000 features and estimating an essential matrix requires trying more than 10^{26} minimal sample combinations.

Here, we recognize that the problem complexity can be tamed by employing single-point solvers [5,36,43–45,93]. This reduces the complexity of the joint procedure to that of the matching $\mathcal{O}(n^2)$, as $m = 1$ in this special case. As the main

Algorithm 2 Model Scoring and Guided matching

Input: $\mathcal{P}_1$ - points, θ - model, H - hashing fn.
K - k best match, ϵ - thr., W - weight fn., Q - scoring
Output: $\mathcal{M}$ - correspondences, q - model score

```
M ← ∅                                              ▷ Initialization to empty set
for each p_1 ∈ P_1 do                              ▷ Each point in the 1st domain
    r* ← ε, p_2* ← 0                               ▷ Best residual and match
    for each p_2 ∈ (K(p_1) ∩ H(p_1, θ)) do
        if φ((p_1, p_2), θ) < r* then
            r* ← φ((p_1, p_2), θ), p_2* ← p_2
    if r* < ε then
        M ← M ∪ {(p_1, p_2*)}
        q ← q + W(K(p_1))Q(θ)
```

contribution, we propose *StereoGlue*, a joint matching and robust estimation pipeline that is general and improves upon the state-of-the-art robust estimators. *StereoGlue* uses an off-the-shelf feature matcher to obtain a soft matching, efficiently forming one-to-many correspondence pools, which are leveraged to simultaneously estimate the sought model and form consistent one-to-one matches. Additionally, we explore various minimal solvers for relative [5, 6, 36] and absolute camera pose estimation [105], for pairwise point cloud registration [51], and we propose one for homographies. *StereoGlue* outperforms state-of-the-art estimators by a significant margin on various real-world and large-scale datasets.

2 Joint Matching and Estimation

StereoGlue is proposed in this section to robustly estimate the parameters of the sought model while simultaneously performing feature matching. See Fig. 1. The pseudo-code of the algorithm is in Alg. 1. Similar to RANSAC, we formalize the problem as iterative sampling and model estimation. However, we assume to have a solver that estimates the model from one match. This allows formalizing function `NextBestMatch` that selects sample $\mathcal{S}$ in each iteration, comprising a single match. Model θ is estimated from $\mathcal{S}$.

After estimating the model, we perform guided matching [10, 64, 95] using model θ to find a set $\mathcal{M}$ of correspondences consistent with the model parameters. The model quality q is calculated from $\mathcal{M}$, *e.g.*, as its support (*i.e.*, $|\mathcal{M}|$), or by any existing scoring technique. If a new best model is found, we apply local optimization to improve its accuracy. The algorithm runs until the termination criterion is triggered. Next, we will describe each step in depth.

Next Best Match Selection. Suppose that we are given $n_1, n_2 \in \mathbb{N}^+$ features in the first and second domains (*e.g.*, image), respectively. Forming correspondences has quadratic complexity $\mathcal{O}(n_1 n_2)$. Thus, iterating through all potential matches severely affects the run-time. To alleviate this computational burden, we employ an off-the-shelf matcher to obtain the k best matches for each feature in

Table 1: Relative pose estimation on PhotoTourism [52] on a total of 9900 image pairs. We report the avg. and median pose errors (in degrees; max. of the translation and rotation errors), their AUC scores, and the inlier numbers. We use the 3PC+uG [33] and the 1AC+uG [43] solvers with *upright* gravity, the 1AC+mD solver [36] on depth from MiDaS-v3 [81,83], and the five point method (5PC) [74]. Upright gravity means that the solvers do not need gravity measurements – they assume it is $[0, -1, 0]$. For solvers requiring more than a single match, we apply the state-of-the-art MAGSAC++ [11]. Levenberg-Marquardt method [71] minimizes pose error on all inliers. The best values are bold in each group. The absolute best ones are underlined.

Features	Estimator	Solver	AVG ↓	MED ↓	AUC@1° ↑	@2.5° ↑	@5° ↑	@10° ↑	@20° ↑	# inliers
SuperPoint + SuperGlue	**StereoGlue**	1AC+uG	**2.6**	**0.7**	**34.5**	55.9	70.3	81.3	**89.2**	394
	StereoGlue	1AC+mD	**2.6**	0.8	**34.5**	**56.0**	**70.4**	**81.4**	**89.2**	395
	MAGSAC++	5PC	4.1	1.3	23.0	43.5	59.9	74.1	84.6	276
	MAGSAC++	1PC+uG	4.0	1.3	23.0	43.4	59.6	74.0	84.7	276
ALIKED + LightGlue	**StereoGlue**	1AC+uG	**3.0**	$\underline{\mathbf{0.5}}$	**41.4**	**62.0**	**74.9**	**83.9**	**89.9**	510
	StereoGlue	1AC+mD	3.6	0.6	38.7	58.5	71.2	80.5	87.2	532
	MAGSAC++	5PC	3.4	0.6	39.0	60.7	74.1	83.4	89.4	547
	MAGSAC++	1PC+uG	4.9	0.6	37.8	59.2	72.3	81.2	87.1	548
DeDoDe + LightGlue	**StereoGlue**	1AC+uG	$\underline{\mathbf{2.3}}$	$\underline{\mathbf{0.5}}$	$\underline{\mathbf{43.5}}$	$\underline{\mathbf{64.3}}$	$\underline{\mathbf{76.7}}$	$\underline{\mathbf{85.4}}$	$\underline{\mathbf{91.2}}$	361
	StereoGlue	1AC+mD	3.7	$\underline{\mathbf{0.5}}$	41.6	60.7	72.8	81.7	88.1	361
	MAGSAC++	5PC	3.2	0.7	38.1	58.0	71.6	81.7	88.7	273
	MAGSAC++	1PC+uG	4.3	0.7	36.8	56.3	69.5	79.3	86.1	273
DoG-8k + HardNet + AffNet	**StereoGlue**	1AC+uG	**3.4**	**0.7**	**38.7**	**57.4**	**70.0**	**79.9**	**87.4**	286
	StereoGlue	1AC+mD	5.2	0.9	22.2	50.6	62.6	73.0	81.7	202
	MAGSAC++	5PC	6.3	1.4	27.7	42.7	54.3	66.2	77.2	210
	MAGSAC++	1AC+uG	5.1	0.9	33.3	50.5	62.5	72.9	81.6	257
DoG-8k + HardNet + Adalam	MAGSAC++	5PC	8.8	**0.8**	**34.3**	**52.5**	**65.0**	**74.8**	82.4	307
LoFTR	MAGSAC++	5PC	**3.6**	1.3	22.5	43.4	59.6	73.7	**84.5**	866
LoFTR	MAGSAC++	3PC+uG	4.1	1.4	21.0	40.9	56.7	71.1	82.6	878
DISK	MAGSAC++	5PC	4.7	0.9	27.9	44.3	55.7	64.5	71.2	474
DISK	MAGSAC++	3PC+uG	4.5	**0.8**	29.1	45.8	57.1	66.1	72.9	617
R2D2 + NN	MAGSAC++	5PC	13.0	2.7	13.6	28.8	42.9	57.9	70.3	169
R2D2 + NN	MAGSAC++	3PC+uG	12.9	2.7	13.9	28.8	42.8	57.5	70.2	169
DoG-8k + SOSNet + NN	MAGSAC++	5PC	40.4	5.9	12.8	23.9	33.5	43.3	52.9	55
DoG-8k + SOSNet + NN	MAGSAC++	3PC+uG	40.4	5.9	12.9	23.8	33.4	43.3	52.9	55

the source domain, where $k \ll n_2$, $k \in \mathbb{N}^+$. For nearest-neighbors-based descriptor matching, like in SIFT [61], we can simply obtain the k-nearest-neighbors (kNN) to get the one-to-many pool. For algorithms like SuperGlue [88], LightGlue [60] or GeoTransformer [79] that solve the optimal transport problem, we can obtain the k best matches from the matching score matrix as the ones with the k highest scores. This allows *StereoGlue* to explore the k best matches and, thus, reduce the matching ambiguity during robust estimation. For example, see Fig. 1, where the potential matches are on the windows, and SuperGlue struggles to find the correct correspondence due to the repetitive nature of the features.

As the objective is to find a good correspondence that leads to an accurate model early, we employ a PROSAC-like [25] procedure where the potential matches are ordered by a quality prior. For matchers performing nearest neighbors search, we use the SNN ratio [62]. For other matchers, we utilize the matching score. Note that learning-based techniques [18, 21] can also be used to predict importance scores that can be used quality prior.

Table 2: Relative pose estimation on ScanNet [29] on the 1500 image pairs from [88, 98]. We report the avg. and median pose errors (in degrees; max. of the translation and rotation errors), their AUC scores and the inlier numbers. We use the 3PC+uG [33] and 1AC+uG [43] solvers with upright gravity, the 1AC+mD solver [36] on depth from MiDaS-v3 [81, 83], and the five point method (5PC) [74]. For solvers requiring more than a single match, we apply the state-of-the-art MAGSAC++ [11]. Finally, the Levenberg-Marquardt method [71] minimizes the pose error on all inliers. The best values are bold in each group. The absolute best ones are underlined.

Features	Estimator	Solver	AVG ↓	MED ↓	AUC@1° ↑	@2.5° ↑	@5° ↑	@10° ↑	@20° ↑	# inliers
SuperPoint + SuperGlue	StereoGlue	1AC+uG	**12.9**	5.8	**0.8**	**7.1**	20.6	**39.7**	**58.4**	119
		1AC+mD	14.0	**5.5**	**0.8**	7.0	**20.7**	39.8	58.1	110
	MAGSAC++	5PC	21.4	6.5	0.7	5.9	17.3	33.9	50.9	89
		3PC+uG	32.4	21.0	0.5	4.2	11.5	21.9	33.1	84
ALIKED + LightGlue	StereoGlue	1AC+uG	23.0	**6.8**	**0.7**	**6.6**	18.7	**35.1**	**50.7**	138
		1AC+mD	24.3	6.9	0.6	**6.6**	**18.8**	34.8	49.8	138
	MAGSAC++	5PC	18.0	7.1	**0.7**	6.3	17.7	33.0	48.0	176
		1AC+uG	**16.9**	7.2	0.6	5.6	16.9	32.6	48.5	186
DeDoDe + LightGlue	StereoGlue	1AC+uG	26.6	9.7	0.5	5.3	**15.6**	**29.6**	**43.8**	102
		1AC+mD	27.2	10.3	**0.8**	5.5	15.2	28.9	43.0	101
	MAGSAC++	5PC	**14.7**	**6.8**	0.6	**5.6**	**15.6**	28.7	42.0	88
		1AC+uG	15.2	7.4	0.7	5.2	14.5	27.7	41.3	88
DoG-8k + HardNet + AffNet	StereoGlue	1AC+uG	26.8	15.0	**0.7**	**5.0**	**13.0**	24.2	37.2	146
		1AC+mD	**24.7**	**12.4**	0.6	4.5	12.6	**25.3**	**39.6**	120
	MAGSAC++	5PC	33.7	29.9	0.3	2.3	6.6	13.6	22.9	81
		1AC+uG	25.3	13.0	0.3	3.1	9.0	18.4	29.4	64
DoG-8k + HardNet + Adalam	MAGSAC++	5PC	54.1	17.8	0.5	3.7	11.1	22.3	34.9	101
LoFTR		5PC	30.3	**6.6**	**1.1**	**8.3**	**22.5**	**41.2**	**57.7**	468
R2D2 + NN		5PC	32.9	13.6	0.6	4.2	12.0	24.6	38.1	190
R2D2 + NN		3PC+uG	**18.9**	10.6	0.4	2.8	8.2	16.8	27.4	137
DoG-8k + SOSNet + NN		5PC	33.3	29.7	0.4	2.6	6.6	13.6	23.4	78
DoG-8k + SOSNet + NN		3PC+uG	60.8	36.4	0.3	1.6	5.3	12.4	22.5	38

Scoring and Guided Matching. Assume that we are given a model $\theta \in \mathbb{R}^{d_\theta}$ estimated from a single correspondence ($d_\theta \in \mathbb{N}$ is the dimensionality of the model manifold), point sets $\mathcal{P}_1$ and $\mathcal{P}_2$ in the two domains, and a point-to-model residual function $\phi : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_p} \to \mathbb{R}$, where $d_p \in \mathbb{N}$ is the data dimension. Model θ can be, for example, an essential matrix and ϕ the Sampson distance or symmetric epipolar error. In short, we iterate through all potential matches and select the pair with the lowest point-to-model residual for each point in the first domain. Finally, the number of consistent correspondences serves as the model score. The pseudo-code for the guided sampling is in Alg. 2. The inputs of the algorithm are the points $\mathcal{P}_1$ in the first domain; model θ; a function $K : \mathcal{P}_1 \to \mathcal{P}_2^k$ assigning the k best match in the second domain to a point in the first one; the inlier-outlier threshold $\epsilon \in \mathbb{R}^+$; a weighting $W : \mathbb{R} \to \mathbb{R}$, a model scoring $Q : \mathbb{R}^d \to \mathbb{R}$, and a hashing function $H : \mathcal{P}_1 \times \mathbb{R}^d \to \mathcal{P}_2^*$. We use MAGSAC++ [11] as scoring function Q to calculate the model score.

Given point $\mathbf{p}_1$ and model θ, the purpose of the hashing function H is to efficiently select matches from $\mathcal{P}_2$ that are consistent with θ when paired $\mathbf{p}_1$, *i.e.*, $\forall \mathbf{p}_2 \in H(\mathbf{p}_1, \theta) : \phi(\mathbf{p}_1, \mathbf{p}_2) \leq \epsilon$. Such H can be constructed for all popular $f : \mathbb{R}^n \to \mathbb{R}^m$ mappings, such as homography or epipolar geometry, using regular grids [14]. We adapt the method proposed in [14] for all tested problems.

Table 3: Homography estimation on HPatches [3]. The AUC scores and avg. times are reported. StereoGlue is applied with the proposed 1AC+uG-H solver assuming upright gravity. We also run MAGSAC++ [11] with the 4PC [47] and 1AC+uG-H solvers. The best values are bold in each group, the absolute bests are underlined.

Features	Estimator	Solver	AUC@1px ↑	@2.5px ↑	@5px ↑	@10px ↑	Time (secs) ↓
SuperPoint + SuperGlue	StereoGlue	1AC+uG-H	**50.5**	**73.9**	**84.9**	**91.1**	**0.04**
	MAGSAC++	1AC+uG-H	45.6	71.7	83.9	90.9	0.66
		4PC	37.9	65.6	79.0	90.1	0.60
DoG-2k + HardNet + AffNet	StereoGlue	1AC+uG-H	40.1	68.0	81.4	88.8	0.29
	MAGSAC++	1AC+uG-H	40.3	68.8	82.3	89.8	0.11
		4PC	**40.9**	**69.3**	**82.7**	**90.4**	**0.01**
ALIKED + LightGlue	StereoGlue	1AC+uG-H	**68.5**	**81.9**	**89.6**	**93.4**	0.22
	MAGSAC++	1AC+uG-H	68.4	81.4	88.8	92.5	0.07
		4PC	67.8	81.2	89.1	93.0	**0.02**
DeDoDe + LightGlue	StereoGlue	1AC+uG-H	**66.5**	**79.6**	**87.3**	**91.1**	0.03
	MAGSAC++	1AC+uG-H	65.4	78.1	85.9	89.9	0.05
		4PC	65.6	78.7	86.6	90.7	**0.01**
LoFTR	MAGSAC++	4PC	**41.8**	**68.6**	**81.2**	**87.9**	0.40
DoG-2k + SOSNet + NN		1AC+uG-H	38.3	65.5	79.5	87.4	0.47
DoG-2k + SOSNet + NN		4PC	36.9	63.3	77.0	85.1	0.25
R2D2 + NN		1AC+uG-H	27.6	51.5	65.9	75.1	0.20
R2D2 + NN		4PC	27.4	51.0	65.5	75.4	**0.09**
DISK + NN		1AC+uG-H	25.1	51.8	68.5	77.8	0.29
DISK + NN		4PC	25.0	51.5	68.1	78.7	0.20

We found it important to use a weighting W in the score calculation, especially when estimating relative pose, *i.e.*, fundamental or essential matrix. The reason is that the point-to-model residual (*e.g.*, Sampson distance) being zero does not necessarily mean it is a correct correspondence. We are unable to measure the translation along the epipolar lines [47]. Without accounting for this, the process hallucinates many incorrect matches consistent with the found model. The model has lots of inliers, while being incorrect. Therefore, for cases with such residual functions, we introduce an additional parameter $\mu \in [0, 1]$ that will act similarly to the Lowe ratio threshold [61] or Wald criterion [106]. For each point $\mathbf{p}_1$, we are given $K(\mathbf{p}_1) = \{\mathbf{p}_2^1, ..., \mathbf{p}_2^k\}$ with matching scores $S(\mathbf{p}_1) = \{s_{12}^1, ...s_{12}^k\}$ from the feature matcher. We only keep those potential matches from $K(\mathbf{p}_1)$, where the matching score $s_{12}^i \geq \mu\,(\max S(\mathbf{p}_1))$. Thus, $K'(\mathbf{p}_1) = \{\mathbf{p}_2^i \mid \mathbf{p}_2^i \in K(\mathbf{p}_1) \wedge s_{12}^i \geq \mu\,(\max S(\mathbf{p}_1))\}$. Weight $W(\mathbf{p}_1) = |K'(\mathbf{p}_1)|^{-1}$ in the proposed algorithm. Therefore, the weight is inversely proportional to the number of matches that have similar matching scores.

Local Optimization. In state-of-the-art robust estimators [8, 11, 27], local optimization is crucial to achieve high accuracy. Thus, when a new best model is found, we apply a few iterations of inner RANSAC only on the selected matches as proposed in [58]. In practice, the LO runs only $\log t$ times [27], where t is the total iteration number of the outer loop. The iteration number spent inside the local optimization is set to a small value, *e.g.*, 20.

3 Solvers from a Single Correspondence

This section discusses minimal solvers for various problems capable of estimating a model from a single match. Such solvers can be designed by making assumptions about the model manifold or leveraging rich features. Under assumptions, we mean prior constraints that allow for reducing the degrees of freedom. For example, we can assume that the camera is mounted to a moving vehicle and, thus, the relative rotation between two frames acts only around the vertical axis, and the y component of the translation is zero. Under rich features, we mean ones that provide more constraints than solely the point locations. Such features include affine correspondences (AC) [12], oriented 3D points, or surface patches.

Relative Pose can be estimated from a single AC accompanied with either monocular depth predictions [36] or gravity direction [43]. Assuming a known direction is not restricting. Consumer devices are usually equipped with Inertial Measurement Units (IMUs) that provide accurate gravity direction *by default*. In case of unknown gravity, it is often safe to assume upright orientation [33], especially when the estimator runs LO that alleviates the impact of a noisy prior.

Absolute Pose can be estimated from a single AC by the recent P1AC solver [105]. While the method requires the 3D points to be oriented, such information can be easily obtained from the point cloud of the stored 3D map.

Rigid Transformation. Given a 3D-3D correspondence predicted by, *e.g.*, GeoTransformer [79], the Q-REG algorithm [51] fits a quadratic surface to each point, considering their neighbors in the point cloud. The principle curvatures of this local quadratic surface serve as a local coordinate system. In case of having a match, the pair of local coordinate systems provide the relative rotation. The point locations give the translation between the point clouds.

Homography. As we are unaware of homography solvers that do not assume special camera motions, we propose a novel one leveraging ACs and known gravity directions. The design and equations of the solver are detailed in Appendix A.

4 Experiments

StereoGlue is evaluated on real-world datasets for relative pose, homography, absolute pose, and rigid transformation estimation. All experiments were implemented in C++ and run on an Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz.

4.1 Relative Pose Estimation

Affine Features. As existing single-point solvers require ACs, we need to obtain them from images. The standard way is to use a local feature detector, like DoG [61] or Key.Net [15], estimate keypoint locations and scales, and use the patch-based AffNet [68] to get affine shapes. Finally, a patch-based descriptor, like HardNet [67] or SOSNet [100], runs. This approach is among leaders in the IMC 2020 benchmark [52]. The second way is to use handcrafted AC detectors,

Table 4: Results of different affine correspondence detectors.

Detector	Desc.	+AffNet	AUC@1°	2.5°	5°	10°	20°
DoG-8k [61]	HardNet+NN	✓	**38.7**	**57.4**	70.0	79.9	87.4
Key.Net [15]	HardNet+NN	✓	22.6	38.8	51.1	62.7	73.6
DISK [103]	HardNet+NN	✓	16.4	27.7	37.9	49.6	63.0
MSER [65]	HardNet+NN	✗	13.6	24.3	34.4	46.2	58.6
SP [30]	HardNet+NN	✓	11.5	22.0	31.6	42.9	55.4
WαSH [104]	HardNet+NN	✗	0.0	0.1	0.8	4.0	13.6
SP [30]	+NN	✓	8.7	17.5	26.4	37.0	48.7
SP [30]	+SG	✓	34.5	55.9	**70.3**	**81.3**	**89.2**
DISK [103]	+NN	✓	30.1	47.3	59.5	69.6	77.7

(a) Affine features on PhotoTourism [52] used inside *StereoGlue* on a total of 9900 image pairs.

Detector	Desc.	+AffNet	AUC@1°	2.5°	5°	10°	20°
DoG-8k [61]	HardNet+NN	✓	0.5	4.5	12.6	25.3	39.6
SP [30]	HardNet+NN	✓	0.4	2.6	7.7	16.3	26.9
DISK [103]	HardNet+NN	✓	0.3	2.2	6.3	13.4	21.3
Key.Net [15]	HardNet+NN	✓	0.3	1.8	5.3	10.7	17.4
MSER [65]	HardNet+NN	✗	0.1	1.2	3.5	7.2	12.5
WαSH [104]	HardNet+NN	✗	0.0	0.1	0.5	1.9	5.7
SP [30]	+NN	✓	0.6	4.2	11.7	23.1	36.1
SP [30]	+SG	✓	**0.8**	**7.0**	**20.7**	**39.8**	**58.1**
DISK [103]	+NN	✓	0.3	2.4	7.2	14.7	25.1

(b) Affine features on ScanNet [29] used inside *StereoGlue* on a total of 1500 image pairs.

such as MSER [65] and WαSH [104]. On top of such features, we can detect any patch-based descriptors, *e.g.*, HardNet [67] or SOSNet [100].

We also experimented with joint detector-descriptor models, such as SuperPoint [30], DISK [103], DeDoDe [35], and ALIKED [118], that output keypoints and descriptors. We run Self-Scale-Ori [59] to get the scale and orientation and then AffNet to upgrade point features to affine ones.

In the main experiments, we run the proposed *StereoGlue* on DoG + HardNet + AffNet + NN (NN – nearest neighbor matching) and SuperPoint / ALIKED / DeDoDe with Self-Scale-Ori, AffNet, and SuperGlue / LightGlue. Obtaining a pool of potential matches is straightforward when using NN on HardNet descriptors. To get a similar pool for SuperGlue, we directly access the matching score matrix that is obtained when solving the optimal transport problem. This allows selecting the k best matches for each point. Additionally, we will show other methods, those that achieve reasonable performance on particular datasets.

Minimal Solvers. We compare three solvers. 5PC [97] is the widely-used algorithm estimating the pose from five point correspondences. The 1AC+mD solver is proposed in [36]. It estimates the pose from a single AC and predicted monocular depth. To allow running this solver, we obtain relative depth by MiDaS-v3 [81,83]. We also compare solver 1AC+G [43] that requires a single AC and a known direction in the images. To demonstrate the robustness of the proposed *StereoGlue*, we *always* run 1AC+G assuming that the gravity points downwards it is of upright direction $[0, -1, 0]^{\mathrm{T}}$. Thus, we call the solver 1AC+uG. This way, we *do not need* to know the gravity direction prior to running the algorithm. This is based on two assumptions that proved true on the tested datasets: (i) people tend to roughly align their cameras with the gravity direction [52,77]; (ii) *StereoGlue* is robust enough due to the employed local optimization procedure. We also test the 3PC+G [33] solver that requires three PCs and gravity.

PhotoTourism. For testing the methods, we use the data from the CVPR IMC 2020 PhotoTourism challenge [52]. It consists of 25 scenes (2 – validation; 12 – training; 11 – test sets) of landmarks with photos of varying sizes collected from the internet. The algorithms are tested on the two scenes for validation – a total of 9900 pairs. For robust estimation, we chose MAGSAC++ [11] as

Table 5: Rigid transformation estimation on the 3DLoMatch dataset [49] with matches from GeoTr [79]. The compared methods are RANSAC with 50K iterations and Q-REG [51] (results copied from [51]). Metrics are registration recall at 0.2m (RR), mean rotation (RRE) and translation (RTE) errors, and RMSE. The best values are bold.

Model	RR (%) ↑	RRE (cm) ↓	RTE (cm) ↓	RMSE (cm) ↓
GeoTransformer	74.1	23.15	58.3	57.8
GeoTr + 50K	75.0	22.69	57.8	57.3
GeoTr + Q-REG	77.1	16.70	46.0	44.6
GeoTr + StereoGlue	**80.7**	**16.04**	**43.9**	**36.3**

the main competitor. We compare the following detectors: SuperPoint [30] with SuperGlue [88], DeDoDe [35] and ALIKED [118] with LightGlue [60], DoG [61] with HardNet [67] descriptors, DoG with HardNet followed by Adalam [20], DoG with SOSNet [100] descriptors, DISK [103], and R2D2 [84]. Also, we show the results of LoFTR [98]. The average error of the gravity prior $[0, -1, 0]^{\mathrm{T}}$ is 10.8°.

The results are in Table 1. We report the average and median pose errors (*i.e.*, the max. of the rotation and translation errors) in degrees, the AUC scores at 1°, 2.5°, 5°, 10°, and 20°, and the average inlier number. Note that the inlier number is not informative when different detectors and matchers are compared. We show it to highlight that the proposed method increases the inlier number compared to MAGSAC++ with 5PC on the same features.

DeDoDe + LightGlue, in conjunction with the proposed *StereoGlue*, leads to the highest accuracy across all detectors and robust estimator combinations. It is important to note that the proposed *StereoGlue* improves all methods in all accuracy metrics. Interestingly, the solver, AC+uG, assuming upright gravity performs better than the one with monodepth predictions. The 3PC+uG [33] solver only marginally improves the results of MAGSAC++.

ScanNet. The ScanNet dataset [29] contains 1613 monocular sequences with ground truth poses and depth. We evaluate our method on the 1500 pairs used in [88,98]. These pairs contain wide baselines and extensive texture-less regions. The avg. error of the gravity prior is 24.8°.

The results are shown in Table 2. Here, ALIKED and DeDoDe are significantly less accurate than SuperPoint features with SuperGlue matcher. *StereoGlue* with DoG or SuperPoint+SuperGlue key points improves the performance by a large margin. It makes SuperPoint+SuperGlue comparable to the detector-free LoFTR [98] with achieving even smaller avg. and med. errors and higher AUC@20°. With *StereoGlue*, DoG+HardNet is among the top-performing methods, with not much worse results than the recent ALIKED and DeDoDe. Both 1AC+uG and 1AC+mD lead to similar accuracy.

Feature Ablation. We compared a number of affine detectors to choose the best ones. The AUC scores on PhotoTourism are shown in Table 4a and on ScanNet in Table 4b. On PhotoTourism, we used the 1AC+uG solver. On ScanNet, we

Table 6: Absolute pose estimation on the Cambridge Landmarks [55] and Aachen Day-Night [92] datasets compared with P3P [78] and P1AC [105] inside GC-RANSAC [8]. For Cambridge L., we report the recall at 5cm/1°, 0.1m/1°, 0.2m/1°; for Aachen at 0.25m/2°, 0.5m/5°, 5m/10°. The best values are bold.

		P3P + GC-RSC	P1AC + GC-RSC	P1AC + StereoGlue
Cambridge L.	5cm/1°	52.6	53.4	**62.4**
	0.1m/1°	54.6	65.1	**77.9**
	0.2m/1°	73.1	80.7	**82.9**
Aachen Day	0.25m/2°	62.0	62.0	**64.8**
	0.5m/5°	83.4	84.6	**85.6**
	5m/10°	**96.0**	95.9	**96.0**
Aachen Night	0.25m/2°	47.1	51.3	**53.9**
	0.5m/5°	60.2	66.0	**67.5**
	5m/10°	74.3	82.2	**80.1**

used 1AC+mD. All methods use *StereoGlue*. DoG with HardNet and AffNet is on par with SuperPoint with SuperGlue on PhotoTourism. On ScanNet, SP+SG is the best. Interestingly, SuperPoint works better with HardNet descriptors than its own when NN matching is used. As expected, classical affine shape detectors, *i.e.* MSER and WαSH, are inaccurate even with HardNet descriptors.

4.2 Homography Estimation

The **HPatches** [3] dataset contains 52 sequences under significant illumination changes and 56 sequences that exhibit large viewpoint variation. Since the intrinsic matrices are not provided in HPatches, we calibrate the cameras of the 56 sequences with viewpoint changes by the RealityCapture software [19]. We use these sequences in the evaluation.

The results are reported in Table 3. *StereoGlue* improves on all recent detector and matcher combinations. It leads to the best performance in all accuracy metrics when combined with ALIKED + LightGlue.

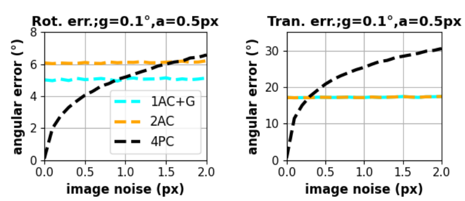

Fig. 2: Image noise study. The average (over 100k runs) angular errors of the rotations and translation estimated by the 4PC [47], 2AC [5], and proposed 1AC+G(H) homography solvers plotted as a function of the image noise in pixels.

Run-time. As reported in Table 3, the avg. run-time of *StereoGlue* on **H** estimation runs for at most a few tens of milliseconds. The avg. time of pose estimation on PhotoTourism is 0.09, and on ScanNet is 0.03 seconds. For comparison, MAGSAC++ with the 5PC solver runs for 0.01 secs on ScanNet and for 0.04 secs on PhotoTourism. Even though *StereoGlue* is slower, it still runs in real-time while achieving SOTA accuracy.

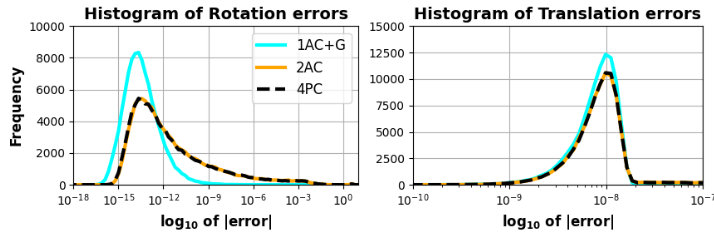

Fig. 3: Stability study. Frequencies (100k runs) of $\log_{10}$ rot. and trans. errors (°) in homographies estimated by the 4PC [47], 2AC [5], and proposed 1AC+G(H) solvers.

Synthetic Experiments. To create a synthetic scene, we generate two cameras with random rotations and translations and focal length set to 1000. A randomly oriented 3D point is generated and projected into both cameras. The affine transformation is calculated from the point orientation. We generated 100k random problem instances and ran the solvers on noiseless samples. Fig. 3 shows histograms of the $\log_{10}$ rotation and translation errors. The plots show that all solvers are stable – there is no peak close to 10^0. In Fig. 2, the average errors in degrees are shown as a function of the image noise. We use a fixed gravity (0.1°) and affine noise (0.5 px). It is important to note that the realistic affine noise is unclear in practice, with no work analyzing it. These plots only intend to demonstrate that the solvers act reasonably w.r.t. increasing noise levels, which they do.

4.3 Absolute Pose Estimation

To evaluate our method on image-based localization, we use the Cambridge Landmarks [55] and Aachen Day-Night v1.1 [90, 92, 116] datasets. For Cambridge Landmarks, we report the recall at 0.05m/1°, 0.1m/1° and 0.2m/1° of the pose errors; for Aachen at 0.25m/2°, 0.5m/5°, 5m/10°. We compare with P3P [78] and P1AC [105] combined with GC-RANSAC (results copied from [105]) on DoG+HardNet+AffNet features. The results are shown in Table 6. *StereoGlue* with P1AC [105] improves significantly on all datasets.

4.4 Rigid Transformation Estimation

To evaluate *StereoGlue* on this task, we use the 3DLoMatch [49] dataset. It contains 62 scenes, with 46 used for training, 8 for validation, and 8 for testing. The point cloud pairs in 3DLoMatch exhibit particularly low overlap, thus making the dataset complicated. We calculate the correspondence RMSE; Registration Recall (RR), which measures the fraction of successfully registered pairs, defined as having a correspondence RMSE below 0.2 m; the average relative rotation (RRE), and translation errors (RTE).

The results, using GeoTransformer [79] to obtain potential one-to-many matches, are reported in Table 5. The values of the competitors are copied from [51]. The proposed *StereoGlue* substantially improves in all metrics.

5 Conclusion

We propose *StereoGlue* to jointly perform feature matching and robust estimation by leveraging a pool of one-to-many correspondences. It is substantially less sensitive to matching ambiguities than using traditional top-1 matches. *StereoGlue* improves performance in various applications when applied on top of popular and state-of-the-art feature detectors. Although the used solvers for image matching assume that the gravity direction is known, *StereoGlue* is so robust that the upright $[0, -1, 0]^\mathrm{T}$ prior works even on ScanNet, where it is only a rough approximation with an avg. error of 24.8° compared to the actual direction.

A Homography Solver

In this section, we describe the proposed single-match-based homography solver.
Affine correspondence $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{A})$ is a triplet, where $\mathbf{p}_1 = [u_1 \; v_1 \; 1]^\mathrm{T}$ and $\mathbf{p}_2 = [u_2 \; v_2 \; 1]^\mathrm{T}$ are a homogeneous point pair in two images and $\mathbf{A}$ is a 2×2 linear transformation called *local affine transformation*. For $\mathbf{A}$, we use the definition provided in [70] as it is given as the first-order Taylor approximation of the 3D → 2D projection function.
Fundamental matrix $(\mathbf{F}) \in \mathbb{R}^{3\times3}$ is rank-2 matrix relating points $\mathbf{p}_1$, $\mathbf{p}_2$ as:

$$\mathbf{p}_2^\mathrm{T} \mathbf{F} \mathbf{p}_1 = 0. \tag{1}$$

Essential matrix $(\mathbf{E})$ is related to $\mathbf{F}$ as $\mathbf{K}'^{-\mathrm{T}} \mathbf{E} \mathbf{K}^{-1} = \mathbf{F}$, where $\mathbf{K}$, $\mathbf{K}'$ are the intrinsic parameters of the cameras [47]. (1) can be written as $\mathbf{p}_2^\mathrm{T} \mathbf{K}'^{-\mathrm{T}} \mathbf{E} \mathbf{K}^{-1} \mathbf{p}_1 = 0$. From now on, we assume that corresponding points $\mathbf{p}_1$, $\mathbf{p}_2$ have been premultiplied by $\mathbf{K}$, $\mathbf{K}'$. This simplifies (1) to

$$\mathbf{p}_2^\mathrm{T} \mathbf{E} \mathbf{p}_1 = 0. \tag{2}$$

Essential matrix $\mathbf{E}$ is decomposed as $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$, where $\mathbf{R} \in \mathrm{SO}(3)$, $\mathbf{t} \in \mathbb{R}^3$ is the relative pose of the two views. The relationship of an affine correspondence (AC) and essential matrix $\mathbf{E}$ was first defined in [7] as

$$\mathbf{A}^{-\mathrm{T}} \mathbf{n}_1 = -\mathbf{n}_2, \tag{3}$$

where $\mathbf{n}_1$, $\mathbf{n}_2$ are the normals to the epipolar lines in the images. In summary, an affine correspondence imposes three independent constraints on the essential matrix. One is given by (2), and two others by (3).
Homography $\mathbf{H} \in \mathbb{R}^3$ is defined as $\mathbf{H} = \mathbf{R} - \frac{1}{d} \mathbf{t} \mathbf{n}^\mathrm{T}$, where $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ are the relative camera rotation and translation, respectively, $d \in \mathbb{R}$ is the plane intercept and $\mathbf{n} \in \mathbb{R}^3$ is its normal. To solve for $\mathbf{H}$, we derive the

constraints for relative pose $\mathbf{R}, \mathbf{t}$ from a single AC $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{A})$, and the gravity directions $\mathbf{v}_1 = [x_{v_1}, y_{v_1}, z_{v_1}]^\mathrm{T}, \mathbf{v}_2 = [x_{v_2}, y_{v_2}, z_{v_2}]^\mathrm{T}$ known in both images. The relative pose with a known vertical direction has three degrees of freedom (DoF), and the AC imposes three constraints on it.

To [53], we can express the rotation as $\mathbf{R} = \mathbf{R}_2^\mathrm{T} \mathbf{R}_y \mathbf{R}_1$, where $\mathbf{R}_y$ is a rotation around y-axis, $\mathbf{R}_1$ transforms $\mathbf{v}_1$ to y-axis, $\mathbf{R}_2$ transforms $\mathbf{v}_2$ to y-axis. Let $\mathbf{y} = [0,\ 1,\ 0]^\mathrm{T}$ be the y-axis. The axis of $\mathbf{R}_1$ is computed as $\mathbf{v}_1 \times \mathbf{y} = [-z_{v_1}/d, \quad 0, \quad x_{v_1}/d]^\mathrm{T}$, where $d = x_{v_1}^2 + z_{v_1}^2$, the angle is obtained as $\arccos{(\mathbf{v}_1^\mathrm{T}\mathbf{y})} = \arccos{(y_{v_1})}$. Rotation $\mathbf{R}_1$ is computed using the Rodrigues formula, rotation $\mathbf{R}_2$ is obtained similarly. Matrix $\mathbf{R}_y$ is expressed elementwise as

$$\mathbf{R}_y = \frac{1}{1+x^2} \begin{bmatrix} 1-x^2 & 0 & -2x \\ 0 & 1+x^2 & 0 \\ 2x & 0 & 1-x^2 \end{bmatrix}, \tag{4}$$

where $x = \tan \phi/2$. Now, we can express the essential matrix $\mathbf{E}$ as $\mathbf{E} = \mathbf{R}_2^\mathrm{T} [\mathbf{t}']_\times \mathbf{R}_y \mathbf{R}_1$, where $\mathbf{t}' = \mathbf{R}_2 \mathbf{t}$. Let $\mathbf{q}_1 = \mathbf{R}_1 \mathbf{p}_1$ and $\mathbf{q}_2 = \mathbf{R}_2 \mathbf{p}_2$. Eq. (2) becomes

$$\mathbf{q}_2^\mathrm{T} [\mathbf{t}']_\times \mathbf{R}_y \mathbf{q}_1 = 0, \tag{5}$$

To modify constraints (3) in a similar way, we define $\mathbf{B} = \mathbf{A}^{-\mathrm{T}} [\mathbf{r}_1^1\, \mathbf{r}_1^2]^\mathrm{T}$, $\mathbf{C} = [\mathbf{r}_2^1\, \mathbf{r}_2^2]^\mathrm{T}$, where $\mathbf{r}_i^1$, $\mathbf{r}_i^2$ $\mathbf{r}_i^3$ are the column vectors of $\mathbf{R}_i$, $i \in \{1, 2\}$. The elements of $\mathbf{B}$ are written in row-major order as $b_1, ..., b_6$, and the elements of $\mathbf{C}$ as $c_1, ..., c_6$. We can rewrite the constraints (3) as

$$\begin{aligned} \mathbf{A}^{-\mathrm{T}} \mathbf{n}_1 - \mathbf{n}_2 &= \mathbf{A}^{-\mathrm{T}} \mathbf{l}_{1[1:2]} - \mathbf{l}_{2[1:2]} \\ &= \mathbf{A}^{-\mathrm{T}} [\mathbf{r}_1^1\ \mathbf{r}_1^2]^\mathrm{T} \mathbf{R}_y^\mathrm{T} [\mathbf{t}']_\times^\mathrm{T} \mathbf{q}_2 - [\mathbf{r}_2^1\ \mathbf{r}_2^2]^\mathrm{T} [\mathbf{t}']_\times \mathbf{R}_y \mathbf{q}_1 = 0. \end{aligned} \tag{6}$$

Constraints (5), (6) give 3 equations in variables $x \in \mathbb{R}$ and $\mathbf{t}' \in \mathbb{R}^3$. After multiplying the equations with $1 + x^2$, we get three equations that are linear in the elements of translation $\mathbf{t}'$. We can, therefore, use the *hidden variable approach* to rewrite the equations in the form $\mathbf{M}(x)\mathbf{t}' = 0$, where $\mathbf{M}(x)$ is a 3×3 matrix whose elements depend on x. If $(x, \mathbf{t}')$ is a solution to the linear system, then matrix $\mathbf{M}(x)$ must be singular. Consequently, $\det \mathbf{M}(x) = 0$ holds. This is a univariate polynomial of degree 6. We find its roots as the eigenvalues of its *companion matrix*. After finding x, we calculate $\mathbf{t}'$ as the kernel of matrix $\mathbf{M}(x)$ and the rotation $\mathbf{R}_y$ according to (4). Finally, we compute the relative pose $(\mathbf{R}, \mathbf{t})$ as $\mathbf{R} = \mathbf{R}_2^\mathrm{T} \mathbf{R}_y \mathbf{R}_1$, $\mathbf{t} = \mathbf{R}_2^\mathrm{T} \mathbf{t}'$.

Next, we will solve for the unknown plane parameters using the estimated relative pose. We can set $\mathbf{n}' = \frac{1}{d}\mathbf{n}$ and simplify the expression as follows:

$$\mathbf{H} = \mathbf{R} - \mathbf{t}\mathbf{n}'^\mathrm{T}. \tag{7}$$

To find homography $\mathbf{H}$ consistent with $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{A})$ and vertical directions $\mathbf{v}_1$ and $\mathbf{v}_2$, we substitute $(\mathbf{R}, \mathbf{t})$ into (7). Then, we only need to find $\mathbf{n}' \in \mathbb{R}^3$. We substitute (7) into the constraints from [7] connecting ACs and homography $\mathbf{H}$. We obtain 6 linear equations in 3 unknowns. The LS method obtains vector $\mathbf{n}'$ from the above system. Finally, we compute the homography $\mathbf{H}$ from $\mathbf{R}$, $\mathbf{t}$, $\mathbf{n}'$ using the equation (7).

Acknowledgements

This work was partially funded by the Hasler Stiftung Research Grant via the ETH Zurich Foundation and an ETH Zurich Career Seed Award. Dmytro Mishkin was supported by the 13162/122/1222100C000_1ND funds.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011) 1
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: CVPR. pp. 5297–5307 (2016) 1
3. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: CVPR. pp. 5173–5182 (2017) 7, 11
4. Barath, D., Cavalli, L., Pollefeys, M.: Learning to find good models in RANSAC. In: CVPR. pp. 15744–15753 (2022) 3
5. Barath, D., Hajder, L.: A theory of point-wise homography estimation. Pattern Recognition Letters **94**, 7–14 (2017) 3, 4, 11, 12
6. Barath, D., Hajder, L.: Efficient recovery of essential matrix from two affine correspondences. Transactions on Image Processing **27**(11), 5328–5337 (2018) 4
7. Barath, D., Hajder, L.: Efficient recovery of essential matrix from two affine correspondences. IEEE Trans. Image Process. **27**(11), 5328–5337 (2018). https://doi.org/10.1109/TIP.2018.2849866, https://doi.org/10.1109/TIP.2018.2849866 13, 14
8. Barath, D., Matas, J.: Graph-cut ransac: local optimization on spatially coherent structures. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4961–4974 (2021) 3, 7, 11
9. Barath, D., Matas, J., Noskova, J.: Magsac: marginalizing sample consensus. In: CVPR. pp. 10197–10205 (2019) 3
10. Barath, D., Mishkin, D., Eichhardt, I., Shipachev, I., Matas, J.: Efficient initial pose-graph generation for global sfm. In: Computer Vision and Pattern Recognition. pp. 14546–14555 (2021) 1, 4
11. Barath, D., Noskova, J., Ivashechkin, M., Matas, J.: Magsac++, a fast, reliable and accurate robust estimator. In: CVPR. pp. 1304–1312 (2020) 3, 5, 6, 7, 9
12. Barath, D., Polic, M., Förstner, W., Sattler, T., Pajdla, T., Kukelova, Z.: Making affine correspondences work in camera geometry computation. In: ECCV. pp. 723–740. Springer (2020) 8
13. Barath, D., Valasek, G.: Space-partitioning RANSAC. European Conference on Computer Vision (2022) 3
14. Barath, D., Valasek, G.: Space-partitioning ransac. In: European Conference on Computer Vision. pp. 721–737. Springer (2022) 6
15. Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K.: Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In: International Conference on Computer Vision (2019) 2, 8, 9
16. Baumberg, A.: Reliable feature matching across widely separated views. In: Computer Vision and Pattern Recognition. pp. 1774–1781. IEEE Computer Society (2000) 2

17. Beaudet, P.R.: Rotationally invariant image operators. In: Proceedings of the 4th International Joint Conference on Pattern Recognition (1978) 2
18. Brachmann, E., Rother, C.: Neural-guided ransac: Learning where to sample model hypotheses. In: International Conference on Computer Vision. pp. 4322–4331 (2019) 3, 5
19. Capturing Reality: Realitycapture, https://www.capturingreality.com/ 11
20. Cavalli, L., Larsson, V., Oswald, M.R., Sattler, T., Pollefeys, M.: Handcrafted outlier detection revisited. In: European Conference on Computer Vision (2020) 10
21. Cavalli, L., Pollefeys, M., Barath, D.: Nefsac: Neurally filtered minimal samples. European Conference on Computer Vision (2022) 3, 5
22. Chen, C., Liu, X., Li, Y., Ding, L., Feng, C.: Deepmapping2: Self-supervised large-scale lidar map optimization. In: CVPR. pp. 9306–9316 (2023) 1
23. Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., McKinnon, D., Tsin, Y., Quan, L.: Aspanformer: Detector-free image matching with adaptive span transformer. In: European Conference on Computer Vision. pp. 20–36. Springer (2022) 1, 2
24. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: International Conference on Computer Vision (October 2019) 1
25. Chum, O., Matas, J.: Matching with prosac-progressive sample consensus. In: CVPR. vol. 1, pp. 220–226. IEEE (2005) 3, 5
26. Chum, O., Matas, J.: Optimal randomized RANSAC. Transactions on Pattern Analysis and Machine Intelligence **30**(8), 1472–1482 (2008) 3
27. Chum, O., Matas, J., Kittler, J.: Locally optimized ransac. In: Joint Pattern Recognition Symposium. pp. 236–243. Springer (2003) 3, 7
28. Chum, O., Werner, T., Matas, J.: Two-view geometry estimation unaffected by a dominant plane. In: CVPR. IEEE (2005) 3
29. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017) 6, 9, 10
30. Detone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-Supervised Interest Point Detection and Description. CVPRW Deep Learning for Visual SLAM (2018) 2, 9, 10
31. DeTone, D., Malisiewicz, T., Rabinovich, A.: Toward geometric deep slam. arXiv preprint arXiv:1707.07410 (2017) 1
32. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: CVPR workshops. pp. 224–236 (2018) 1
33. Ding, Y., Yang, J., Kong, H.: An efficient solution to the relative pose estimation with a common direction. In: International Conference on Robotics and Automation. pp. 11053–11059. IEEE (2020) 5, 6, 8, 9, 10
34. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In: CVPR (2019) 2
35. Edstedt, J., Bökman, G., Wadenbäck, M., Felsberg, M.: DeDoDe: Detect, Don't Describe – Describe, Don't Detect for Local Feature Matching (2023) 9, 10
36. Eichhardt, I., Barath, D.: Relative pose from deep learned depth and a single affine correspondence. In: European Conference on Computer Vision. pp. 627–644. Springer (2020) 3, 4, 5, 6, 8, 9
37. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: European conference on computer vision. pp. 834–849. Springer (2014) 1

38. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981) 3
39. Frahm, J.M., Pollefeys, M.: Ransac for (quasi-) degenerate data (qdegsac). In: CVPR. vol. 1, pp. 453–460. IEEE (2006) 3
40. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: CVPR. pp. 1434–1441. IEEE (2010) 1
41. Furukawa, Y., Hernández, C., et al.: Multi-view stereo: A tutorial. Foundations and Trends® in Computer Graphics and Vision **9**(1-2), 1–148 (2015) 1
42. Gojcic, Z., Zhou, C., Wegner, J.D., Guibas, L.J., Birdal, T.: Learning multiview 3d point cloud registration. In: CVPR. pp. 1759–1769 (2020) 1
43. Guan, B., Su, A., Li, Z., Fraundorfer, F.: Rotational alignment of imu-camera systems with 1-point ransac. In: Chinese Conference on Pattern Recognition and Computer Vision. pp. 172–183. Springer (2019) 3, 5, 6, 8, 9
44. Guan, B., Zhao, J., Li, Z., Sun, F., Fraundorfer, F.: Relative pose estimation with a single affine correspondence. Transactions on Cybernetics (2021) 3
45. Hajder, L., Barath, D.: Relative planar motion for vehicle-mounted cameras from a single affine correspondence. In: International Conference on Robotics and Automation. pp. 8651–8657. IEEE (2020) 3
46. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference. pp. 147–151 (1988) 2
47. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003) 3, 7, 11, 12, 13
48. Hartley, R.I.: In defense of the eight-point algorithm. IEEE Transactions on pattern analysis and machine intelligence **19**(6), 580–593 (1997) 3
49. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: CVPR. pp. 4267–4276 (2021) 2, 10, 12
50. Jared, H., Schonberger, J.L., Dunn, E., Frahm, J.M.: Reconstructing the world in six days. In: CVPR (2015) 1
51. Jin, S., Barath, D., Pollefeys, M., Armeni, I.: Q-REG: End-to-end trainable point cloud registration with surface curvature. 3DV (2024) 4, 8, 10, 13
52. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image matching across wide baselines: From paper to practice. International Journal of Computer Vision (2020) 5, 8, 9
53. Kalantari, M., Hashemi, A., Jung, F., Guédon, J.: A new solution to the relative orientation problem using only 3 points and the vertical direction. J. Math. Imaging Vis. **39**(3), 259–268 (2011). https://doi.org/10.1007/s10851-010-0234-2, https://doi.org/10.1007/s10851-010-0234-2 14
54. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. Advances in neural information processing systems **30** (2017) 1
55. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In: ICCV. pp. 2938–2946 (2015) 11, 12
56. Kukelova, Z., Bujnak, M., Pajdla, T.: Automatic generator of minimal problem solvers. In: European Conference on Computer Vision. pp. 302–315. Springer (2008) 3
57. Kukelova, Z., Kileel, J., Sturmfels, B., Pajdla, T.: A clever elimination strategy for efficient minimal solvers. In: CVPR. pp. 4912–4921 (2017) 3
58. Lebeda, K., Matas, J., Chum, O.: Fixing the locally optimized ransac–full experimental evaluation. In: British machine vision conference. vol. 2. Citeseer (2012) 3, 7

59. Lee, J., Jeong, Y., Cho, M.: Self-supervised learning of image scale and orientation. In: 31st British Machine Vision Conference 2021, BMVC 2021, Virtual Event, UK. BMVA Press (2021) 2, 9
60. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. Internation Conference on Computer Vision (2023) 2, 5, 10
61. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60**(2), 91–110 (2004) 1, 2, 5, 7, 8, 9, 10
62. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004) 5
63. Lynen, S., Zeisl, B., Aiger, D., Bosse, M., Hesch, J., Pollefeys, M., Siegwart, R., Sattler, T.: Large-scale, real-time visual–inertial localization revisited. The International Journal of Robotics Research **39**(9), 1061–1084 (2020) 1
64. Ma, J., Jiang, J., Zhou, H., Zhao, J., Guo, X.: Guided locality preserving feature matching for remote sensing image registration. IEEE transactions on geoscience and remote sensing **56**(8), 4435–4447 (2018) 4
65. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extrema regions. In: BMVC. pp. 384–393 (2002) 9
66. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision (IJCV) **60**(1), 63–86 (2004) 2
67. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In: NeurIPS (2017) 2, 8, 9, 10
68. Mishkin, D., Radenovic, F., Matas, J.: Repeatability is Not Enough: Learning Affine Regions via Discriminability. In: European Conference on Computer Vision (2018) 2, 8
69. Moisan, L., Moulon, P., Monasse, P.: Automatic homographic registration of a pair of images, with a contrario elimination of outliers. Image Processing On Line **2**, 56–73 (2012) 3
70. Molnár, J., Chetverikov, D.: Quadratic transformation for planar mapping of implicit surfaces. Journal of Mathematical Imaging and Vision (2014) 13
71. Moré, J.J.: The levenberg-marquardt algorithm: implementation and theory. In: Numerical analysis, pp. 105–116. Springer (1978) 5, 6
72. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE transactions on robotics **31**(5), 1147–1163 (2015) 1
73. Myatt, D., Torr, P., Nasuto, S., Bishop, J., Craddock, R.: NAPSAC: High noise, high dimensional robust estimation-it's in the bag. In: Proceedings of the British Machine Vision Conference. pp. 44.1–44.10. BMVA Press (2002) 3
74. Nister, D.: An efficient solution to the five-point relative pose problem. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(6), 756–770 (2004). https://doi.org/10.1109/TPAMI.2004.17 5, 6
75. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: ICCV. pp. 3456–3465 (2017) 1
76. Panek, V., Kukelova, Z., Sattler, T.: Meshloc: Mesh-based visual localization (2022) 1
77. Perdoch, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. In: CVPR. pp. 9–16 (2009) 9
78. Persson, M., Nordberg, K.: Lambda Twist: An accurate fast robust perspective three point (P3P) solver. In: ECCV. pp. 318–332 (2018) 11, 12

79. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Ilic, S., Hu, D., Xu, K.: Geotransformer: Fast and robust point cloud registration with geometric transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 1, 2, 5, 8, 10, 13
80. Radenović, F., Tolias, G., Chum, O.: Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In: European conference on computer vision. pp. 3–20. Springer (2016) 1
81. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021) 5, 6, 9
82. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: European Conference on Computer Vision (2018) 3
83. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022) 5, 6, 9
84. Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., Humenberger, M.: R2d2: Repeatable and reliable detector and descriptor. In: NeurIPS (2019) 2, 10
85. Riu, C., Nozick, V., Monasse, P., Dehais, J.: Classification performance of ransac algorithms with automatic threshold estimation. In: VISIGRAPP. vol. 5, pp. 723–733. Scitepress (2022) 3
86. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision. pp. 430–443. ECCV'06, Springer-Verlag, Berlin, Heidelberg (2006). https://doi.org/10.1007/11744023_34, http://dx.doi.org/10.1007/11744023_34 2
87. Rublee, E., Rabaud, V., Konolidge, K., Bradski, G.: ORB: An Efficient Alternative to SIFT or SURF. In: International Conference on Computer Vision (2011) 2
88. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR. pp. 4938–4947 (2020) 2, 5, 6, 10
89. Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: European conference on computer vision. pp. 752–765. Springer (2012) 1
90. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6DOF urban visual localization in changing conditions. In: CVPR. pp. 8601–8610 (2018) 12
91. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., et al.: Benchmarking 6dof outdoor visual localization in changing conditions. In: CVPR. pp. 8601–8610 (2018) 1
92. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: Proceedings of the British Machine Vision Conference (BMVC). vol. 1, p. 4 (2012) 11, 12
93. Scaramuzza, D.: 1-point-RANSAC structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints. International journal of computer vision **95**(1), 74–85 (2011) 3
94. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. pp. 4104–4113 (2016) 1
95. Shah, R., Srivastava, V., Narayanan, P.: Geometry-aware feature matching for structure from motion applications. In: Winter Conference on Applications of Computer Vision. pp. 278–285. IEEE (2015) 4

96. Stewart, C.V.: Minpran: A new robust estimator for computer vision. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(10), 925–938 (1995) 3
97. Stewenius, H., Engels, C., Nistér, D.: Recent developments on direct relative orientation. ISPRS Journal of Photogrammetry and Remote Sensing **60**(4), 284–294 (2006) 3, 9
98. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR. pp. 8922–8931 (2021) 1, 2, 6, 10
99. Sun, W., Jiang, W., Tagliasacchi, A., Trulls, E., Yi, K.M.: Attentive context normalization for robust permutation-equivariant learning. In: CVPR (2020) 3
100. Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., Balntas, V.: Sosnet: Second order similarity regularization for local descriptor learning. In: CVPR (2019) 2, 8, 9, 10
101. Tolias, G., Avrithis, Y., Jégou, H.: Image search with selective match kernels: aggregation across single and multiple images. International Journal of Computer Vision **116**(3), 247–261 (2016) 1
102. Torr, P.H.S., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding (CVIU) (2000) 3
103. Tyszkiewicz, M.J., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. In: NeurIPS (2020) 2, 9, 10
104. Varytimidis, C., Rapantzikos, K., Avrithis, Y.: Wash: Weighted α-shapes for local feature detection. In: European Conference on Computer Vision 2012 (2012) 9
105. Ventura, J., Kukelova, Z., Sattler, T., Baráth, D.: P1AC: Revisiting absolute pose from a single affine correspondence. In: ICCV. pp. 19751–19761 (2023) 4, 8, 11, 12
106. Wald, A.: Sequential Analysis. John Wiley and Sons, 1st edn. (1947) 7
107. Wang, H., Liu, Y., Dong, Z., Guo, Y., Liu, Y.S., Wang, W., Yang, B.: Robust multiview point cloud registration with reliable pose graph initialization and history reweighting. In: CVPR. pp. 9506–9515 (2023) 1
108. Wang, Q., Zhang, J., Yang, K., Peng, K., Stiefelhagen, R.: Matchformer: Interleaving attention in transformers for feature matching. In: Asian Conference on Computer Vision (2022) 1, 2
109. Wei, T., Matas, J., Barath, D.: Adaptive reordering sampler with neurally guided magsac. In: ICCV. pp. 18163–18173 (2023) 3
110. Wei, T., Patel, Y., Shekhovtsov, A., Matas, J., Barath, D.: Generalized differentiable ransac. In: ICCV. pp. 17649–17660 (2023) 3
111. Yew, Z.J., Lee, G.H.: Learning iterative robust transformation synchronization. In: 2021 International Conference on 3D Vision (3DV). pp. 1206–1215. IEEE (2021) 1
112. Yi, K.M., Verdie, Y., Fua, P., Lepetit, V.: Learning to Assign Orientations to Feature Points. In: CVPR (2016) 2
113. Yi*, K.M., Trulls*, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: CVPR (2018) 3
114. Yu, J., Ren, L., Zhang, Y., Zhou, W., Lin, L., Dai, G.: Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration. In: CVPR. pp. 17702–17711 (2023) 2
115. Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H.: Learning two-view correspondences and geometry using order-aware network. International Conference on Computer Vision (2019) 3
116. Zhang, Z., Sattler, T., Scaramuzza, D.: Reference pose generation for long-term visual localization via learned features and view synthesis. International Journal of Computer Vision **129**, 821–844 (2021) 12

117. Zhao, C., Ge, Y., Zhu, F., Zhao, R., Li, H., Salzmann, M.: Progressive correspondence pruning by consensus learning. In: International Conference on Computer Vision (2021) 3
118. Zhao, X., Wu, X., Chen, W., Chen, P.C., Xu, Q., Li, Z.: Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. IEEE Transactions on Instrumentation and Measurement (2023) 9, 10
119. Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L.: Very large-scale global sfm by distributed motion averaging. In: CVPR. pp. 4568–4577 (2018) 1