Handcrafted Outlier Detection Revisited

Luca Cavalli[1], Viktor Larsson[1], Martin Ralf Oswald[1], Torsten Sattler[2], and Marc Pollefeys[1,3]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Chalmers University of Technology, Gothenburg, Sweden
[3] Microsoft Mixed Reality & AI Zurich Lab

Abstract. Local feature matching is a critical part of many computer vision pipelines, including among others Structure-from-Motion, SLAM, and Visual Localization. However, due to limitations in the descriptors, raw matches are often contaminated by a majority of outliers. As a result, outlier detection is a fundamental problem in computer vision and a wide range of approaches, from simple checks based on descriptor similarity to geometric verification, have been proposed over the last decades. In recent years, deep learning-based approaches to outlier detection have become popular. Unfortunately, the corresponding works rarely compare with strong classical baselines. In this paper we revisit handcrafted approaches to outlier filtering. Based on best practices, we propose a hierarchical pipeline for effective outlier detection as well as integrate novel ideas which in sum lead to an efficient and competitive approach to outlier rejection. We show that our approach, although not relying on learning, is more than competitive to both recent learned works as well as handcrafted approaches, both in terms of efficiency and effectiveness. The code is available at https://github.com/cavalli1234/AdaLAM.

Keywords: low-level vision, matching, spatial matching, spatial consistency, spatial verification

1 Introduction

Image matching is a key component in any image processing pipeline based on correspondences between images, such as Structure from Motion (SfM) [15, 41, 42, 49, 52], Simultaneous Localization and Mapping (SLAM) [3, 13, 29] and Visual Localization [8, 24, 36, 39]. Classically, the problem is tackled by computing high dimensional descriptors for keypoints which are robust to a set of transformations, then a keypoint is matched with its most similar counterpart in the other image, i.e. the nearest neighbor in descriptor space. Due to limitations in the descriptors, the set of nearest neighbor matches usually contains a great majority of outliers as many features in one image often have no corresponding feature in the other image. Consequently, outlier detection and filtering is an important problem in these applications. Several methods have been proposed for this task, from simple low-level filters based only on descriptors such as the ratio-test [27], to local spatial consistency checks [1, 6, 8, 18, 19, 22, 26, 28, 31, 38, 43, 47, 54, 55, 58] and global

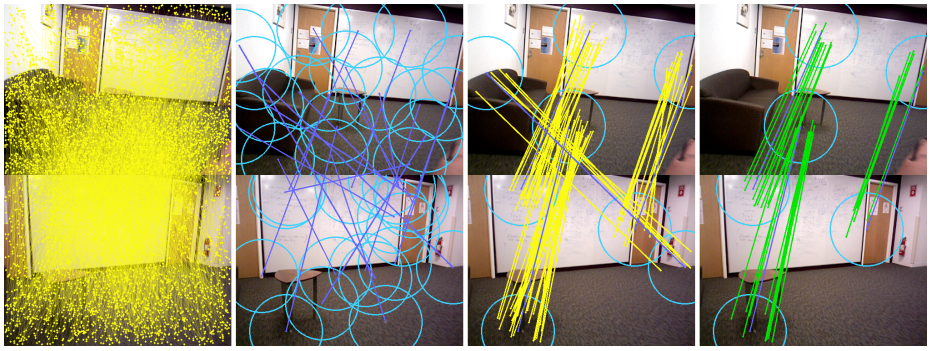


Fig. 1: **Main steps in our method**, from left to right: *1.* we take as input a wide set of putative matches (in yellow), *2.* we select well spread hypotheses of rough region correspondences (blue circles), *3.* for each region we consider the set of all putative matches consistent with the same region correspondence hypothesis, *4.* we only keep the correspondences which are locally consistent with an affine transform with sufficient support (in green). Note that for visualization purposes we do *not* show all the hypotheses *nor* all the matches.

geometric verification methods, either exact [4, 5, 9–11, 14, 18, 21, 31, 32, 47, 48] or approximate [2, 17, 23, 40, 53, 54]. In the last years, many methods have been proposed to learn either local neighborhood consistency [34,59] or global geometric verification [7, 12, 30, 33, 37, 57]. Yet, this line of research usually overlooks prior classical methods, and rarely compares with strong classical baselines.

In this paper we revisit handcrafted approaches to outlier filtering. Based on best practices, we propose a hierarchical pipeline for efficient and effective outlier filtering. We show that even though this approach does not involve learning, it achieves competitive performance to learned approaches, greatly outperforming the current state of the art in outdoor scenes and being superior or on par in indoor scenarios. Our results indicate that more research is needed in this area, including properly understanding the performance of learned methods.

Thus, we can summarize our contributions in the following: (**1**) We propose a novel framework that builds up from several past ideas in spatial matching into a coherent, robust, and highly parallel algorithm for fast spatial verification of image correspondences. (**2**) As our framework is based on geometrical assumptions that can have different discriminative power in different scenarios, we propose a novel method that adaptively relaxes our assumptions, to achieve better generalization to different domains while still mining as much information as available from each image region. (**3**) We experimentally show that our adaptive relaxation improves generalization, and that our method can greatly outperform current learned and non-learned state-of-the-art methods on favorable domains, while being on par in unfavorable domains as well. (**4**) We demonstrate that handcrafted methods still have considerable potential and can perform comparably to or better than current state-of-the-art learned methods, showing that there is still much research

to be done in this area. (**5**) We provide a publicly available implementation of our method at https://github.com/cavalli1234/AdaLAM.

2 Related Work

Outlier rejection is a long-standing problem which has been studied in many contexts, producing many diverse approaches that act at different levels, with different complexity and different objectives.

Simple filters are widely used as a straightforward heuristic that already greatly improves the inlier ratio of available correspondences based on very low-level descriptor checks. In this category we include the classical ratio-test [27] and mutual nearest neighbor check, that filter out ambiguous matches, as well as (Hamming) distance thresholding to prune obvious outliers. These heuristics are extremely efficient and easy to implement, though they are not always sufficient as they can easily leave many outliers or filter out inliers present in the initial putative matches set.

Local neighborhoods methods filter correspondences based on the observation that correct matches should be consistent with other correct matches in their vicinity, while wrong matches are normally inconsistent with their neighbors. Consistency can be formulated as a co-neighboring constraint [6,8,28,31,38,43,47], or enforcing a local transformation between neighboring correspondences [19, 26, 54,55,58], or as a graph of mutual pairwise agreements of local transformations [1, 18, 22]. Methods acting at this level can also be very efficient, and represent a more informative selection compared to simple filters.

Geometric verification approaches filter matches based on a global transformation on which correct correspondences must agree. This can be achieved by robustly fitting a global transformation (be it similarity, affinity, homography or fundamental) to the set of all the matches, with sampling methods, including RANSAC [14] and its numerous later improvements, either biasing the sampling probabilities towards more likely inliers [9, 18, 31, 47], making iterations more efficient with a sequential probability ratio test [10] or adding local optimization [4, 11, 21, 48], combining all of the previous [32], or marginalizing over the inlier decision threshold [5]. A different line of research in the context of image retrieval uses fast approximate spatial verification to determine whether two images have the same content. They only approximately fit a geometric transformation to efficiently prune the majority of outliers, using the local affine or similarity transformation encoded by each individual match [27]. The space of all transforms is quantized and the set of accepted correspondences is determined by majority voting with a Hough scheme in linear time [2, 17, 23, 40, 53, 54].

Learned methods extract an implicit consistency model directly from data. Several works have been proposed in the last years, acting on different levels, either learning a local neighborhood consistency model [34, 59], or a global consistency model [7, 12, 30, 33, 37, 57]. Many of these target learning epipolar geometry constraints explicitly, formulating the problem either as outlier classification [12, 30],

or as an iteratively reweighted least-squares problem [33], or biasing RANSAC's sampling distribution towards matches more likely to be correct [7].

The line of research of learned methods in this field, however, usually gives little consideration to the vast literature of classical methods that have been proposed for outlier rejection, and rarely compares against strong classical baselines. As a result, the performance of these methods is not yet well understood. In this paper, we take inspiration from prior work on outlier filtering and compare our classical pipeline with the learning-based ones, showing that we can achieve comparable to superior results on the same datasets they trained on, while offering a comparable runtime on the same hardware.

3 Method

Given the sets of keypoints $\mathcal{K}_1$ and $\mathcal{K}_2$ respectively in images I_1 and I_2, generally the set of all putative matches $\mathcal{M}$ is taken as the set of nearest neighbor matches from $\mathcal{K}_1$ to $\mathcal{K}_2$, where nearest neighbors are defined in descriptor space. In practice, due to limitations in the descriptors, $\mathcal{M}$ is contaminated by a great majority of incorrect correspondences, thus our objective is to produce a subset $\mathcal{M}' \subseteq \mathcal{M}$ that is the nearest possible approximation of the set of all and only correct inlier matches $\mathcal{M}^* \subseteq \mathcal{M}$.

Our method builds on classical spatial matching approaches used both in the field of matching and image retrieval. To keep computational costs down, we limit our search of matches to a subset of a fixed set of initial putative matches $\mathcal{M}$, which we take as the nearest neighbors in descriptor space, and employ classical filters on orientation and scale to efficiently prune confidently wrong matches. The main steps in our algorithm are reported in Figure 1: (**1**) We select a limited number of confident and well distributed matches, which we call seed points. (**2**) For each seed point we select neighboring compatible correspondences. (**3**) We verify local affine consistency in the neighborhood of each seed point via highly parallel RANSACs [14] with multiple inlier thresholds. For each seed point, we select the best threshold a posteriori, and we accept it if enough inliers agree on the fitted affinity. We output $\mathcal{M}'$ as the union of all the set of inliers of the accepted seed points within the chosen inlier threshold.

3.1 Preliminaries and core assumptions

The 3D plane tangent to a point induces an homography between two views, which can be well approximated locally by an affine transformation A in image space [20]. This affine transformation strongly constraints geometrical cross consistency of correct keypoint correspondences, acting as a very reliable filter. However, the underlying assumptions of planarity, locality and correct projections can break in multiple ways in real images: (**1**) The surface on which 3D points lie may not be planar. The offset between the 3D tangent plane at a point and the real surface produces a non-linear deviation in the projections of all the 3D points not lying on the tangent plane, which is more and more significant with the

curvature of the surface. (**2**) The detected points may not be near to each other, adding distortion to the affine model which is no longer a good approximation of the induced homography. This error increases with the relative distance of keypoints and with the tilt of the tangent plane. (**3**) Matching keypoints may not represent the projection of exactly the same 3D point. This is a very common problem with wide baseline viewpoint changes, as slight changes in illumination and self occlusions can easily move the peak in saliency for keypoint localization.

To address these problems we propose an adaptive relaxation on our core assumption, that we describe in Section 3.4

3.2 Seed points selection

As affine transforms A are a good approximation of local transformations around a 3D point P, we use available nearest neighbor correspondences to guide the search for candidate 3D surface points. More specifically we want to select a restricted set of confident and well spread correspondences to be used as hypotheses for P, around which consistent point correspondences are to be searched, as in [19]. We call such hypotheses *seed points*. As a confidence score we use the classical descriptor ratio test between the nearest neighbor and the second nearest neighbor, while we require a correspondence to have the highest score within its neighborhood with radius R to be selected as a seed point. This way we ensure both distinctiveness and coverage of seed points without causing grid artifacts, while keeping the selection completely parallel for efficient computation on GPU, as each correspondence can be scored and compared to neighbors for seed point selection independently of the final selection of the others.

3.3 Local neighborhood selection, filtering and validation

The assignment of correspondences to seed points is a crucial step in the algorithm as it builds the search space around each hypothesis of P to find the affine transform A. Wider neighborhoods can more easily include correct correspondences to fit A, while at the same time they implicitly loosen the affine constraints as they violate the assumption on locality.

Let $S_i = (\mathbf{x}_1^{S_i}, \mathbf{x}_2^{S_i})$ be a seed point correspondence, which induces a similarity transformation $(\alpha^{S_i} = \alpha_2^{S_i} - \alpha_1^{S_i}, \sigma^{S_i} = \sigma_2^{S_i}/\sigma_1^{S_i})$ from its local feature frame, decomposed in the orientation component α^{S_i} and scale component σ^{S_i}, and $\mathcal{N}_i \subseteq \mathcal{M}$ be the set of correspondences that are assigned to S_i to verify affine consistence. Let t_α and t_σ be thresholds for orientation and scale agreement between a candidate correspondence and the seed correspondence S_i. In analogy to [38], correspondence $(p_1, p_2) = ((\mathbf{x}_1, \mathbf{d}_1, \sigma_1, \alpha_1), (\mathbf{x}_2, \mathbf{d}_2, \sigma_2, \alpha_2)) \in \mathcal{M}$, which induces a transformation $(\alpha^p = \alpha_2 - \alpha_1, \sigma^p = \sigma_2/\sigma_1)$ is assigned to $\mathcal{N}_i$ if all the following constraints are satisfied:

$$\left\|\mathbf{x}_1^{S_i} - \mathbf{x}_1\right\| \leq \lambda R_1, \quad \left\|\mathbf{x}_2^{S_i} - \mathbf{x}_2\right\| \leq \lambda R_2, \quad \left|\alpha^{S_i} - \alpha^p\right| \leq t_\alpha, \quad \left|\ln\left(\frac{\sigma^{S_i}}{\sigma^p}\right)\right| \leq t_\sigma \quad (1)$$

where R_1 and R_2 are the radii used to spread seed points respectively in image I_1 and I_2, and λ is a hyperparameter that regulates the overlap between inclusion neighborhoods. Note that we consider angles α in modulo 2π lying within the interval $(-\pi, \pi]$. Different radii R_1 and R_2 are chosen proportionally to the image area to be invariant to image rescaling.

As from Eq. (1), we include in $\mathcal{N}_i$ all the correspondences in $\mathcal{M}$ that are locally consistent in both images and that induce a similarity transform (α^p, σ^p) which is consistent with $(\alpha^{S_i}, \sigma^{S_i})$ within independent thresholds t_α and t_σ. The independent thresholds encode a confidence over the reliability of the orientation and scale information provided by keypoints. The idea of verification using orientation and scale consistency has been repeatedly proposed for template matching [1, 27] and image retrieval [2, 17, 40, 54] as a coarse but powerful indication for outlier pruning.

For each set $\mathcal{N}_i$ corresponding to a seed point S_i, we translate the keypoint coordinates to have their origin in S_i, and we robustly fit an affine transformation A_i using RANSAC with a fixed number m of iterations to run efficiently on highly parallel hardware. At each iteration j, we uniformly sample two correspondences in $\mathcal{N}_i$ and fit the affine transform hypothesis A_i^j centered in S_i with this minimal set of constraints. As the best inlier threshold for RANSAC depends on the amount of noise on the inliers, we score each hypothesis based on multiple thresholds $t_1 \dots t_n$ and select the best one a posteriori, as explained in the next section (Section 3.4). For a correspondence $(\mathbf{x}_1, \mathbf{x}_2)$ we can compute the residuals with respect to A_i^j and the corresponding inlier set $\mathcal{P}$ as follows:

$$R(A_i^j, \mathbf{x}_1, \mathbf{x}_2) = \left\| A_i^j \mathbf{x}_1 - \mathbf{x}_2 \right\| \tag{2}$$

$$\mathcal{P}_i^j(t_k) = \left\{ (\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{N}_i \;\middle|\; R(A_i^j, \mathbf{x}_1, \mathbf{x}_2) \leq t_k \sqrt{|\det(A_i^j)|} \right\} \tag{3}$$

leading to the hypothesis scoring function C:

$$C(A_i^j, t_k) = \begin{cases} 0 & \text{if } \det(A_i^j) \geq t_\sigma^2 \vee \det(A_i^j) \leq \frac{1}{t_\sigma^2} \\ \left| \mathcal{P}_i^j(t_k) \right| & \text{otherwise} \end{cases} \tag{4}$$

where $|.|$ over a set is the count of its elements. Notice that in Eq. (3) we rescale our tolerance threshold t_k with $\sqrt{|\det(A_i^j)|}$, so that t_k is expressed in pixels of error tolerated in image I_1, and it is rescaled in image I_2 according to the scale change encoded in A_i^j. Moreover, we do not accept affine hypotheses with extreme scale changes above t_σ to filter out degenerate cases, as evident in Eq. (4). We do not include any prior from α_i and σ_i from the seed correspondence S_i as they encode the local transformation as a similarity, which may not agree in orientation and scale with the same parameters in the fit affinity when skew is not negligible.

We finally select the affine model that maximizes the score C for each seed point S_i and for each threshold $t_k \in \{t_1 \dots t_n\}$, we fit A_{i,t_k}^* as the least squares solution that minimizes the residuals on the highest scoring inlier set $\mathcal{P}_i^*(t_k)$.

3.4 Adaptive assumption relaxation

Threshold t_k directly determines the tolerance for deviations from the affine model that would hold when all assumptions discussed in Section 3.1 are valid. Increasing values for t_k thus relaxes the assumptions, while reducing the reliability of the scoring function C as more outliers pass the checks. The RANSAC inlier count consistently increases in the presence of only outliers with increasing t_k, while it increases strongly for lower thresholds in the presence of noisy inliers. In the following we will make very similar assumptions to [48] about the inlier and outlier distributions.

Let us assume that outlier correspondences $(\mathbf{x}_1^o, \mathbf{x}_2^o)$ are independent and uniformly distributed around S_i within radius λR_1 in I_1, as we only consider correspondences lying within such radius from S_i. Given an affine transform A_i^j then their images in I_2 are uniformly distributed in an area of size $|\det(A_i^j)|\pi\lambda^2 R_1^2$. Given threshold t_k, the acceptance region in I_2 is a circle of radius $t_k\sqrt{|\det(A_i^j)|}$ centered in $A_i^j\mathbf{x}_1^o$, thus having area $\pi t_k^2|\det(A_i^j)|$. The probability of a single outlier correspondence to be counted as inlier in a RANSAC iteration is thus:

$$p_o = \frac{\pi t_k^2|\det(A_i^j)|}{|\det(A_i^j)|\pi\lambda^2 R_1^2} = \frac{t_k^2}{\lambda^2 R_1^2} \quad . \tag{5}$$

And the number of positively counted outliers in a single RANSAC iteration follows the distribution $\mathcal{B}(n_o, p_o)$ where $\mathcal{B}$ is binomial and n_o is the number of outliers included in $\mathcal{N}_i$. Let $MAX(n, \mathcal{Y})$ be the distribution obtained by taking the maximum value of n independent random variables following the distribution $\mathcal{Y}$, then the m-iteration RANSAC score $C_{i,k}^*$ of an outlier seed correspondence S_i with all outlier correspondences in $\mathcal{N}_i$ follows the distribution:

$$C_{i,t_k}^* \sim MAX\left(m, \mathcal{B}\left(\|\mathcal{N}_i\|, p_o\right)\right) \quad . \tag{6}$$

Let us now assume that inlier correspondences $(\mathbf{x}_1^i, \mathbf{x}_2^i)$ have dependent distributions of $\mathbf{x}_1^i$ and $\mathbf{x}_2^i$ such that $\mathbf{x}_2^i \sim \mathcal{N}(A_i^*\mathbf{x}_1^i, \sigma^2 I)$. Thus, squared residuals follow a chi-square distribution with two degrees of fredom $R(A_i^*, \mathbf{x}_1^i, \mathbf{x}_2^i)^2 \sim \sigma^2\chi_2^2$, and the RANSAC score distribution for only n_i inlier correspondences follows the binomial $C_{i,t_k}^* \sim \mathcal{B}(n_i, P_k^i)$ where $P_k^i = P(\sigma^2\chi_2^2 \leq t_k^2\det(A_i^*))$ is the probability of an inlier to meet threshold t_k. Considering n_i inlier and n_o outlier correspondences in the same set, assuming independence and assuming that the RANSAC iterations m are enough to find A_i^*, we can approximate the final score distribution as:

$$C_{i,t_k}^* \sim \mathcal{B}(n_i, P_k^i) + MAX\left(\left\lceil m\frac{n_i^2}{(n_o+n_i)^2}\right\rceil, \mathcal{B}(\|\mathcal{N}_i\|, p_o)\right) \quad . \tag{7}$$

We correct the number of RANSAC iterations in the outlier counts distribution to consider only the RANSAC iterations that actually sample two inliers.

As we intend to compensate the influence of outliers in RANSAC inlier counts, we subtract from all scores $C_{i,k}^*$ the expected score of the inlier-free case as an

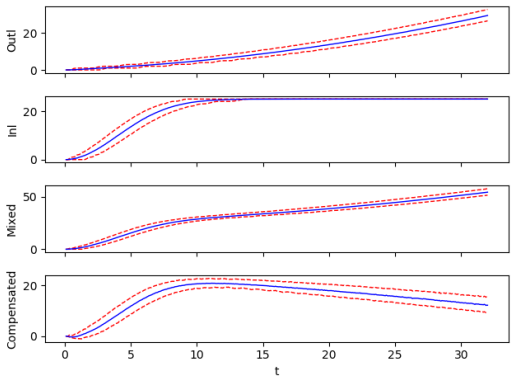


Fig. 2: **Simulated inlier count distributions** with varying threshold t_k according to our model, with 90% upper and lower confidence intervals (red dashed) and expected value (blue solid). The parameters for this simulation are: $n_i = 25,\ n_o = 80,\ \sigma = 16,\ \lambda R_1 = 60,\ m = 128$. From top to bottom: count component only from outliers, count component only from inliers, overall count distribution, and compensated count.

upper bound of the actual influence of the outliers. As shown in Figure 2, this allows to clean the inlier count signal from RANSAC to highlight the threshold range where most inliers are included without exceeding with outlier inclusion.

A perfectly compensated inlier count signal has constant expected value after all inliers are included in the counts. However, outliers still represent a zero-mean noise that can make the optimal threshold unclear. We robustify this approach by overcompensation: the overestimation of the outlier compensation causes their component to have degreasing negative mean. As a result, the best range of thresholds is more robustly highlighted as a peak in the overcompensated inlier counts, as in the last plot of Figure 2.

Let $\mathbb{E}_o$ be the expectation assuming all outliers, then for each seed correspondence S_i we select the threshold $t^i_* = \text{argmax}_{t_k}\, C^*_{i,t_k} - \mathbb{E}_o\left[C^*_{i,t_k}\right]$ that maximizes the compensated inlier count of the best fit model. We then output all inliers in $\mathcal{N}_i$ included in the set of inliers $\mathcal{P}^*_i(t^i_*)$ for the best threshold, if and only if $C^*_{i,t^i_*} - \mathbb{E}_o\left[C^*_{i,t^i_*}\right] \geq 3$ to ensure that we have a minimal number of inliers and suppress noise from outliers.

As a final robustness step, if only $s < 20$ seed points passed the inlier count test, we also output the top $20 - s$ correspondences in $\mathcal{M}$ based on the ratio-test score. This is to ensure that, when we detect a failure of our procedure, we can still output a set of confident matches. However, in our experiments we observed no significant variation in performance due to this option, which triggers only in extreme cases.

More implementation details and our hyperparameter setup are available in the supplementary material.

Table 1: **Comparative experiments** with the state of the art in indoor and outdoor scenes. All methods fit the essential matrix with LO-RANSAC with maximum 10^4 iterations, except *Ratio test (100k)* that uses 10^5 LO-RANSAC iterations and MAGSAC which runs Ratio test + 100k iterations of MAGSAC. All numbers are in percentages.

Method	TUM [45]			SUN3D [56]			YFCC100M [46]		
	AUC5	AUC10	AUC20	AUC5	AUC10	AUC20	AUC5	AUC10	AUC20
Ours	**24.7**	**37.2**	**48.4**	**7.6**	**18.3**	**33.2**	**57.8**	**71.1**	**81.7**
OA-Net [57]	20.9	32.2	43.3	6.9	16.3	29.4	53.5	66.0	76.7
NGRANSAC [7]	19.4	29.6	38.7	6.2	15.0	27.3	53.8	66.7	77.7
GMS [6]	19.6	30.5	41.3	6.8	15.9	29.1	52.3	65.0	76.0
Ratio test [27] (10k)	16.1	24.8	33.6	5.9	14.1	25.6	51.9	64.9	76.3
Ratio test [27] (100k)	17.3	26.6	36.2	6.1	14.5	26.3	53.2	66.3	77.5
MAGSAC [5]	17.5	27.2	36.5	5.9	14.6	27.0	47.2	58.9	70.6

4 Experiments

Our experiments aim at comparing our method with existing state-of-the-art methods in Section 4.3, and to understand the influence of each component of our method with ablation studies in Section 4.4. All experiments measure relative pose estimation performance under the same pipeline and on the same datasets. We evaluate on the same test sets as OA-Net [57], NGRANSAC [7] and GMS [6]: the same four scenes from YFCC100M [46], two from Strecha [44] and fifteen from SUN3D [56] as [7, 57], and the same six sequences from TUM [45] as [6].

4.1 Evaluation Pipeline

Our evaluation pipeline aims at measuring relative pose estimation performance within the same settings. More specifically, all methods receive exactly the same keypoints as input and need to output a set of matches that will be used to robustly fit an essential matrix, which is decomposed to rotation and translation. We then measure the rotation and translation errors in degrees and take the maximum of the two, and report the exact Area Under the Curve (AUC) with thresholds of 5, 10, and 20 degrees.

The keypoints are all extracted with OpenCV SIFT [27] with the same parameters as in the code provided by OA-Net [57] and NGRANSAC [7], with a maximum number of 8000 keypoints per image. Keypoints with locations, descriptors, orientation and scale are provided to the matching methods, and matches are produced. For fitting the essential matrix we use the LO-RANSAC [11] implementation in COLMAP [41, 42] with minimum 10^3 iterations and maximum 10^4, unless differently specified. The intrinsic camera calibration is assumed to be known and is taken from ground truth.

4.2 Datasets

We evaluate our method on large and diverse indoor and outdoor datasets, using the same scenes as the methods we compare with. For outdoor scenes we use the YFCC100M [46] internet photos, that were later organized into 72 scenes [16] reconstructed with the Structure from Motion software VisualSfM [51, 52], providing bundle adjusted camera poses, intrinsics and triangulated point clouds. We select scenes and image couples as to reproduce the test set used by [7,30,57], thus we used the same six scenes, including the two from Strecha [44], with the same sampling procedure. From now on when we refer to YFCC100M, we are referring to the four scenes actually coming from YFCC100M *and* the two coming from Strecha. All images are used with the original resolution.

For indoor scenes we use six sequences from the TUM [45] visual odometry benchmark and the SUN3D [56] dataset, both of which provide ground truth poses together with the RGB images. In particular, for TUM we select the same sequences as the authors of GMS [6], but we use a different subsampling scheme to provide a wider range of image transformations. We take one keyframe every 150 frames, and match it with other 9 images sampled at 15 frames intervals from it. This ensures a sufficient image overlap while gradually increasing the difficulty of the image pair, differentiating the break-down point of the competing alternatives. On SUN3D we use the same fifteen scenes and sampling procedure as [7, 30, 57]. All images are used with the original resolution.

4.3 Comparison with the State of the Art

We compare our method against sample representatives of the current state of the art. GMS (Grid-based Motion Statistics) [6] is a non-learned method that models the statistics of having locally consistent matches and filters matches based on a statistical significance test over large groups. Designed with the objective of being fast, the authors use 10 000 ORB features [35]. However, we found that with appropriate tuning the performance is higher using our SIFT setup with a ratio-test filtering beforehand, as suggested by the authors. Thus, we report these results using the public OpenCV implementation of the method with rotation and scale invariance. NGRANSAC (Neural Guided RANSAC) [7] uses a neural network to predict sampling probabilities for RANSAC from keypoint locations and ratio-test scores. We use the pre-trained models provided by the authors for essential matrix estimation with SIFT keypoints pre-filtered with a ratio-test of 0.8 (SIFT+Ratio+NG-RANSAC(+SI) label in [7]), which have been trained on both YFCC100M [46] and SUN3D [56]. We experimentally found that, although the method outputs an essential matrix, better performance is achieved by using LO-RANSAC only on the inlier set found by NGRANSAC. Thus, after running both versions we report these results. OA-Net (Order Aware Network) [57] learns to infer confidence scores on nearest neighbor matches looking at the global keypoint spatial consistency. They propose a soft assignment to latent clusters in canonical order, and an order-aware upsampling operation that restores the original size of the input to infer confidences. The authors provide a model

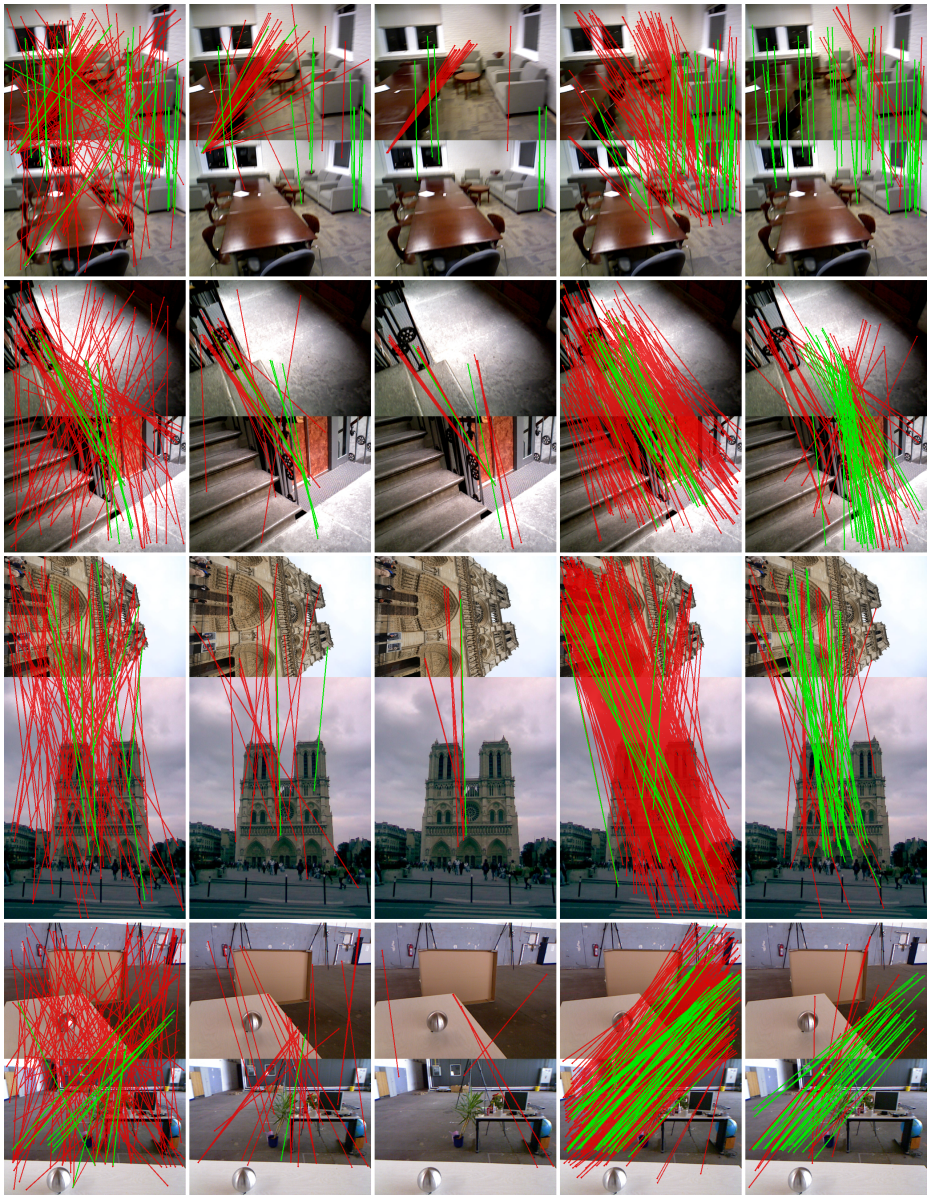
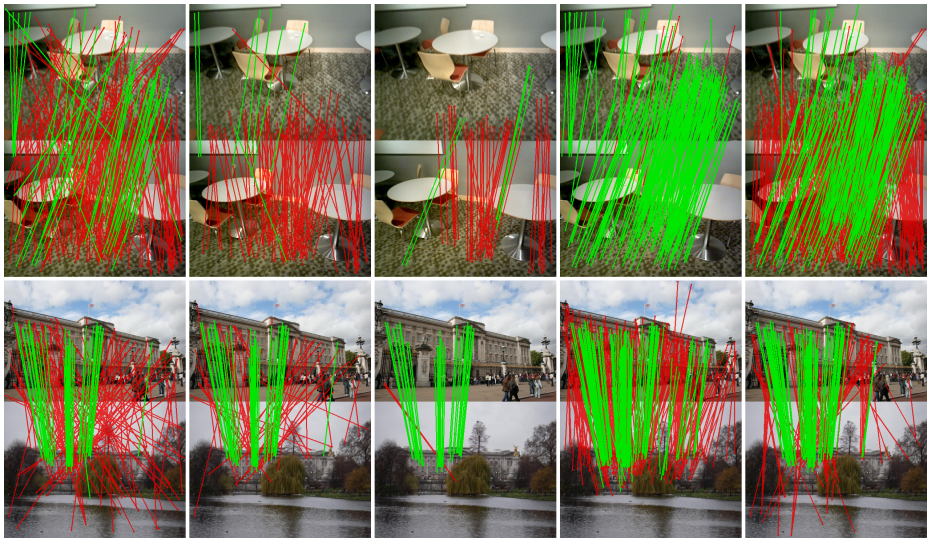


Fig. 3: **Success cases** from our experiments. Matches agreeing with ground truth epipolar geometry are shown in green, others are in red. Examples include cases with very sparse correspondences, local repeated structures, weak texture, strong rotations and perspective deformations.



Ratio-test [27] NGRANSAC [7] GMS [6] OA-Net [57] Ours

Fig. 4: **Failure cases** from our experiments. Matches agreeing with ground truth epipolar geometry are shown in green, others are in red. The main failure case for our method is wide repeated structures along the image, which can locally mimic the correspondence distribution of the correct region.

pre-trained on both YFCC100M [46] and SUN3D [56]. Our SIFT parameters are taken from the public implementation provided by the authors with the pre-trained model. MAGSAC [5] is a modern RANSAC variant based on the idea of marginalizing over a range of possible inlier thresholds for the purpose of model scoring. In our experiments we run MAGSAC with 100k iterations on correspondences filtered by the ratio-test with a 0.8 threshold. Finally we include a simple baseline using the standard ratio test with a 0.8 threshold, as the default in SiftGPU [50] used in COLMAP [41]. We also try the performance of this simple baseline with ten times more LO-RANSAC iterations, going from the 10^4 used for all methods to 10^5 iterations.

Table 1 reports the results of our experiments on both indoor and outdoor scenes. For comparability and deeper insights we report additional metrics in the supplementary material, including inlier statistics and an upper bound approximation of the AUC used by some of the methods. All the competitor methods outperformed their original paper scores in our setup when comparable, where the main difference is the use of LO-RANSAC rather than OpenCV's RANSAC implementation. We found that local optimization can refine the solution by some degrees, improving the scores for low errors. Results show that our method can drastically outperform current state of the art in outdoor scenarios by exploiting the planarity of most scenes and buildings, while still being very competitive in indoor scenarios where our assumptions are violated

more often. While TUM is a completely new dataset for all learned methods, both OA-Net and NGRANSAC are trained on YFCC100M and SUN3D. However, we make sure not to have overlaps between our test set and their training set.

Figures 3 and 4 show qualitative results that represent success cases and failure cases for our method with respect to others. Figure 3 shows how our method captures consistent global motion even when available correct matches are sparse, and is fully invariant to strong rotation and scale changes. As affine coherence in keypoint patterns can give confidence to matches even when descriptors are ambiguous, our method is able to mine correspondences even from almost textureless surfaces or in the presence of locally repeating structures. However, this is not always the case for widely repeating regular structures, as illustrated in Figure 4. In such cases, there is one or more independent clusters of wrong correspondences that locally mimic the distribution of the correct correspondences. Global approaches in this case have a chance to disambiguate the right cluster, and learned approaches can give priority to the cluster compatible with more likely motions, as OA-Net is doing.

4.4 Ablation studies

We aim at understanding the contribution of each element we introduce in our method, thus we extensively evaluate different versions of our method subtracting one element at a time. For comparability with other methods, we run the same experiments in the same setting as in Section 4.3 on TUM and YFCC100M.

We target three optional steps in our pipeline and re-evaluate removing one or multiple of them. We report as *Full* the complete method, denoted as "Ours" in Section 3.1. We remove the filtering with side information in Eq. (1) for the *No-Side* method, we skip refitting the estimated affinities on the final set of inliers for the *No-Refit* method, and we drop adaptive thresholding in the *No-Adaptive* method. We run this last ablation with all the evaluated thresholds of the full method and choose only the one scoring best with respect to ground truth.

We report the results of our ablation in Table 2. On the full method, we measure a runtime of 20-40ms on image pairs with 4000-8000 extracted keypoints, running on an RTX2080Ti. Since most of the methods we compare with in our experiments provide CPU implementations, or important CPU preprocessing steps, their runtimes are usually higher but not directly comparable with ours; however we found that the public implementation of OA-Net [57] also performs all operations on PyTorch as we do. We measure runtimes of 20-40ms on the same hardware and keypoint collections. For comparability with the ablations, we also report the performance of OA-Net [57] in Table 2.

The full adaptive method generally outperforms the best fixed threshold, showing that it can make a positive decision on which threshold to use case by case. In general, the adaptive thresholding increases the generalization performance of the method, allowing it to operate effectively in diverse settings without the need to decide for a single fixed threshold. Moreover, as refitting and running multiple thresholds is overall a significant component of our runtime, the ablated versions, particularly the *No-Refit-No-Adaptive*, are straightforward solutions to tune the

Table 2: **Ablation tests** with varying setups of our method. The numbers are comparable with Table 1. Areas under the curve (AUC) are in percentage; times in milliseconds include nearest neighbor search and outlier rejection. Results and timings for OA-Net [57] are additionally reported for better comparability.

Method	TUM [45]				YFCC100M [46]			
	AUC5	AUC10	AUC20	time	AUC5	AUC10	AUC20	time
Full (Ours)	**24.7**	**37.2**	**48.4**	26ms	**57.8**	71.1	81.7	40ms
No-Side	22.4	33.8	44.2	42ms	54.5	67.4	78.4	64ms
No-Adaptive	22.4	33.7	43.8	17ms	57.5	70.8	81.4	28ms
No-Refit-No-Adaptive	24.4	36.7	47.8	**16ms**	**57.8**	**71.2**	**81.8**	**26ms**
No-Refit	23.8	35.5	45.5	20ms	57.0	70.25	80.9	33ms
OA-Net [57]	20.9	32.2	43.3	21ms	53.5	66.0	76.7	41ms

trade-off between quality and runtime, especially for a fixed domain in which generalization of performance is not a real concern. We finally highlight that smart classical filters can increase both runtime and quality as they reduce the size of the problem by pruning grossly incorrect correspondences at the beginning, and at the same time reduce the number of outliers, providing a more stable inlier count signal.

5 Conclusions

In this paper we proposed a method for outlier rejection of an initial set of putative correspondences inspired by local consistency constraints which have been re-discovered repeatedly in the last years [6, 17, 19, 25–27, 38, 58]. We show that, by proposing an adaptive relaxation of the underlying assumptions for local consistency, we improve the generalization of this approach to make it competitive in diverse and challenging scenarios. Our method can greatly outperform the current state of the art in favorable settings, where the planarity assumption can be more discriminative, while being on par on unfavorable, less structured ones. At the same time, we formulate our approach as a highly parallel algorithm to be run on modern GPUs in the order of the tens of milliseconds.

Acknowledgements: This work was supported by a Google Focused Research Award, by the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots), the Chalmers AI Research Centre (CHAIR) (VisLocLearn) and Innosuisse funding (Grant No. 34475.1 IP-ICT). Viktor Larsson was supported by an ETH Zurich Postdoctoral Fellowship.

References

1. Albarelli, A., Rodola, E., Torsello, A.: Robust game-theoretic inlier selection for bundle adjustment. In: International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT2010) (2010)
2. Avrithis, Y., Tolias, G.: Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. International Journal of Computer Vision (IJCV) **107**(1), 1–19 (2014)
3. Bailey, T., Durrant-Whyte, H.: Simultaneous localization and mapping (slam): Part ii. IEEE robotics & automation magazine **13**(3), 108–117 (2006)
4. Barath, D., Matas, J.: Graph-cut ransac. In: Computer Vision and Pattern Recognition (CVPR) (2018)
5. Barath, D., Matas, J., Noskova, J.: Magsac: marginalizing sample consensus. In: Computer Vision and Pattern Recognition (CVPR) (2019)
6. Bian, J., Lin, W.Y., Matsushita, Y., Yeung, S.K., Nguyen, T.D., Cheng, M.M.: Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: Computer Vision and Pattern Recognition (CVPR) (2017)
7. Brachmann, E., Rother, C.: Neural-guided ransac: Learning where to sample model hypotheses. In: International Conference on Computer Vision (ICCV) (2019)
8. Cech, J., Matas, J., Perdoch, M.: Efficient sequential correspondence selection by cosegmentation. Trans. Pattern Analysis and Machine Intelligence (PAMI) **32**(9), 1568–1581 (2010)
9. Chum, O., Matas, J.: Matching with prosac-progressive sample consensus. In: Computer Vision and Pattern Recognition (CVPR) (2005)
10. Chum, O., Matas, J.: Optimal randomized ransac. Trans. Pattern Analysis and Machine Intelligence (PAMI) **30**(8), 1472–1482 (2008)
11. Chum, O., Matas, J., Kittler, J.: Locally optimized ransac. In: Joint Pattern Recognition Symposium. pp. 236–243. Springer (2003)
12. Dang, Z., Moo Yi, K., Hu, Y., Wang, F., Fua, P., Salzmann, M.: Eigendecomposition-free training of deep networks with zero eigenvalue-based losses. In: European Conference on Computer Vision (ECCV) (2018)
13. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part i. IEEE robotics & automation magazine **13**(2), 99–110 (2006)
14. 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)
15. Hartley, R.I., Sturm, P.: Triangulation. Computer Vision and Image Understanding (CVIU) **68**(2), 146–157 (1997)
16. Heinly, J., Schönberger, J.L., Dunn, E., Frahm, J.M.: Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In: Computer Vision and Pattern Recognition (CVPR) (2015)
17. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: European Conference on Computer Vision (ECCV) (2008)
18. Johns, E., Yang, G.Z.: Ransac with 2d geometric cliques for image retrieval and place recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW) (2015)
19. Jung, I.K., Lacroix, S.: A robust interest points matching algorithm. In: International Conference on Computer Vision (ICCV) (2001)

20. Köser, K.: Geometric estimation with local affine frames and free-form surfaces. Ph.D. thesis, University of Kiel (2009), http://d-nb.info/994782322
21. Lebeda, K., Matas, J., Chum, O.: Fixing the locally optimized ransac–full experimental evaluation. In: British Machine Vision Conference (BMVC) (2012)
22. Leordeanu, M., Hebert, M.: A spectral technique for correspondence problems using pairwise constraints. In: International Conference on Computer Vision (ICCV) (2005)
23. Li, X., Larson, M., Hanjalic, A.: Pairwise geometric matching for large-scale object retrieval. In: Computer Vision and Pattern Recognition (CVPR) (2015)
24. Li, Y., Snavely, N., Huttenlocher, D.P.: Location recognition using prioritized feature matching. In: European Conference on Computer Vision (ECCV) (2010)
25. Lin, W.Y., Liu, S., Jiang, N., Do, M.N., Tan, P., Lu, J.: Repmatch: Robust feature matching and pose for reconstructing modern cities. In: European Conference on Computer Vision (ECCV) (2016)
26. Lin, W.Y., Wang, F., Cheng, M.M., Yeung, S.K., Torr, P.H., Do, M.N., Lu, J.: Code: Coherence based decision boundaries for feature correspondence. Trans. Pattern Analysis and Machine Intelligence (PAMI) **40**(1), 34–47 (2017)
27. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60**(2), 91–110 (2004)
28. Ma, J., Zhao, J., Jiang, J., Zhou, H., Guo, X.: Locality preserving matching. International Journal of Computer Vision (IJCV) **127**(5), 512–531 (2019)
29. Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B., et al.: Fastslam: A factored solution to the simultaneous localization and mapping problem. Conference on Artificial Intelligence (AAAI) (2002)
30. Moo Yi, K., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: Computer Vision and Pattern Recognition (CVPR) (2018)
31. Ni, K., Jin, H., Dellaert, F.: Groupsac: Efficient consensus in the presence of groupings. In: International Conference on Computer Vision (ICCV) (2009)
32. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: Usac: a universal framework for random sample consensus. Trans. Pattern Analysis and Machine Intelligence (PAMI) **35**(8), 2022–2038 (2012)
33. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: European Conference on Computer Vision (ECCV) (2018)
34. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: Neural Information Processing Systems (NeurIPS) (2018)
35. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: International Conference on Computer Vision (ICCV) (2011)
36. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Computer Vision and Pattern Recognition (CVPR) (2019)
37. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Computer Vision and Pattern Recognition (CVPR). pp. 4938–4947 (2020)
38. Sattler, T., Leibe, B., Kobbelt, L.: Scramsac: Improving ransac's efficiency with a spatial consistency filter. In: International Conference on Computer Vision (ICCV) (2009)
39. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8601–8610 (2018)
40. Schönberger, J.L., Price, T., Sattler, T., Frahm, J.M., Pollefeys, M.: A vote-and-verify strategy for fast spatial verification in image retrieval. In: Asian Conference on Computer Vision (ACCV) (2016)
41. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR) (2016)
42. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
43. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. Trans. Pattern Analysis and Machine Intelligence (PAMI) **31**(4), 591–606 (2008)
44. Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Computer Vision and Pattern Recognition (CVPR) (2008)
45. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: International Conference on Intelligent Robots and Systems (IROS) (2012)
46. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
47. Torr, P.H., Nasuto, S.J., Bishop, J.M.: Napsac: High noise, high dimensional robust estimation-it's in the bag. British Machine Vision Conference (BMVC) (2002)
48. Torr, P.H., Zisserman, A.: Mlesac: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding (CVIU) **78**(1), 138–156 (2000)
49. Ullman, S.: The interpretation of structure from motion. Proceedings of the Royal Society of London. Series B. Biological Sciences **203**(1153), 405–426 (1979)
50. Wu, C.: Siftgpu: A gpu implementation of scale invariant feature transform (sift). `http://cs.unc.edu/~{}ccwu/siftgpu` (2011)
51. Wu, C., Agarwal, S., Curless, B., Seitz, S.M.: Multicore bundle adjustment. In: Computer Vision and Pattern Recognition (CVPR) (2011)
52. Wu, C., et al.: Visualsfm: A visual structure from motion system (2011)
53. Wu, X., Kashino, K.: Adaptive dither voting for robust spatial verification. In: International Conference on Computer Vision (ICCV) (2015)
54. Wu, X., Kashino, K.: Robust spatial matching as ensemble of weak geometric relations. In: British Machine Vision Conference (BMVC) (2015)
55. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: Computer Vision and Pattern Recognition (CVPR) (2009)
56. Xiao, J., Owens, A., Torralba, A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. In: International Conference on Computer Vision (ICCV) (2013)
57. 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 (ICCV) (2019)
58. Zhang, Z., Deriche, R., Faugeras, O., Luong, Q.T.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artificial Intelligence **78**(1-2), 87–119 (1995)
59. Zhao, C., Cao, Z., Li, C., Li, X., Yang, J.: Nm-net: Mining reliable neighbors for robust feature correspondences. In: Computer Vision and Pattern Recognition (CVPR) (2019)