Improving Structure from Motion with Reliable Resectioning

Rajbir Kataria[1] Joseph DeGol[2] Derek Hoiem[1]

[1]University of Illinois Urbana-Champaign [2]Microsoft
{rk2,dhoiem}@illinois.edu jodegol@microsoft.com

Abstract

A common cause of failure in structure-from-motion (SfM) is misregistration of images due to visual patterns that occur in more than one scene location. Most work to solve this problem ignores image matches that are inconsistent according to the statistics of the tracks graph, but these methods often need to be tuned for each dataset and can lead to reduced completeness of normally good reconstructions when valid matches are removed. Our key idea is to address ambiguity directly in the reconstruction process by using only a subset of reliable matches to determine resectioning order and the initial pose. We also introduce a new measure of similarity that adjusts the influence of feature matches based on their track length. We show this improves reconstruction robustness for two state-of-the-art SfM algorithms on many diverse datasets.

1. Introduction

The modern incremental Structure from Motion (SfM) approach was developed to meet the challenge of reconstructing landmarks from Internet photos [33, 32, 28], where the goal is to correctly register most of the photos. Now, SfM is widely used to model buildings, bridges, and cities for inspection and maintenance [1, 3, 4]. SfM algorithms face new challenges in these applications due to more stringent completeness and correctness requirements and the prevalence of incorrect feature matches on repeated or symmetric structures, such as signs, windows, and architectural patterns, that lead to unregistered or misregistered photographs (Fig. 1). Failure in SfM is costly, requiring laborious manual corrections or traveling back to the site to take more photographs.

In this paper, we propose techniques to improve robustness to incorrect matches in incremental SfM. Consider Fig. 2. Robust fitting methods [12] and outlier checks fail to discard the many incorrect matches on two different Oats containers in images 1 and 6 because *the matches agree on an incorrect relative pose*. Existing approaches to deal with this problem, such as trying to prune bad matches in the tracks graph, are not widely used due to their complexity or brittleness. Our approach is to instead keep all matches and focus on fixing the reconstruction steps that are corrupted by bad matches: (1) choosing the next image to add ("resectioning"); and (2) initializing the pose estimate based on currently reconstructed points. Suppose the reconstruction in Fig. 2 initially contains images 1-3. Given the 458 robust matches (that passed RANSAC verification), image 6 looks like a good candidate to resection next, but doing so will cause a misregistration. *Our first key insight* is that "long tracks" (features that match across many images) are more likely to be due to repeated structures than short tracks. Long tracks can be good for precise triangulation, but they are less trustworthy than short tracks *for resectioning* for a simple reason: features on duplicate structures match across more images than features on unique structures because, by definition, duplicate structures appear more often. Therefore, we give shorter tracks more weight when determining the reliability of matches between two images, and resection the image that has the most reliable matches with a reconstructed image. Even if image 6 is resectioned last, the many consistent but incorrect matches with images 1-

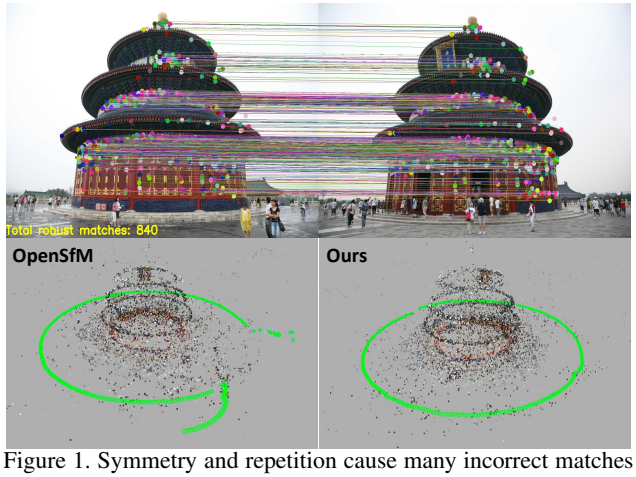

Figure 1. Symmetry and repetition cause many incorrect matches in this image pair, which causes misregistrations in current SfM systems (e.g. [2]). Our method successfully reconstructs the scene by scoring reliability of image matches and basing resectioning order and pose initialization on the most reliably matching images.

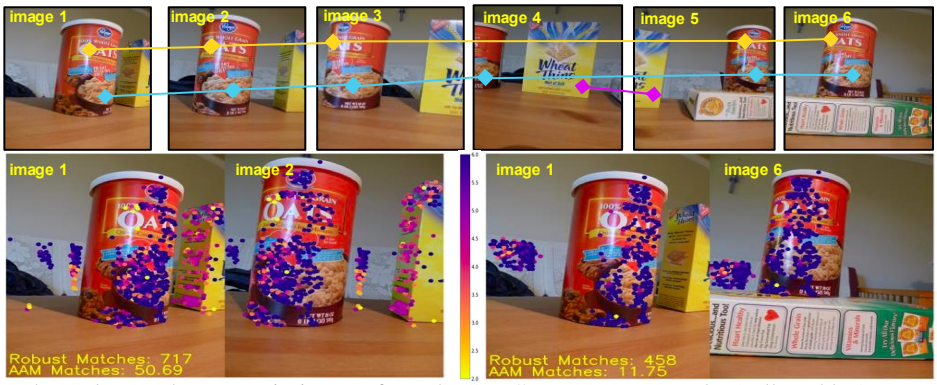

Figure 2. **Top**: Example tracks are shown on six images from the "oats" scene. Long tracks (yellow, blue) are more likely to contain incorrect matches to duplicate structures than short ones (purple) because duplicate structures appear more often. Thus, we give shorter tracks more weight in determining match reliability and resection order, and initialize pose using reconstructed points from the most reliably matching images. **Bottom**: Tracks are colored by length. Both pairs of images have many robust matches (RANSAC inliers) on the duplicate Oats container, but the correctly matching pair on left also has many short tracks on the unique Wheat Thins box, causing its AAM score to be much higher than the incorrect pair on right.

3 could cause pose to be incorrectly initialized in robust PnP [19], causing outlier checks to discard the true matches to images 4 and 5. *Our second key insight* is to initialize pose based only on the reconstructed points that are observed by the most reliable image matches. After pose is initialized, all reconstructed points that pass outlier checks are used to refine the pose. Together, these two key ideas – to give short tracks more weight in determining resection order and to initialize pose using only the most reliable matches – lead to correct reconstruction of this scene. To be clear, we still employ long tracks in pose estimation and triangulation, we simply weigh them less when determining match reliability between images.

Our experiments show that our proposed techniques lead to dramatically better robustness in dozens of tabletop [25] and hallway [10] scenes selected for their challenges of duplicate structures. To show that our method works on amorphous capture patterns, we also show promising results on internet scenes [15] (see supplemental material). We also demonstrate improvements on a standard multiview stereo (MVS) dataset [17] in terms of SfM and MVS outputs. The generality of our ideas is further demonstrated by showing that our modified resection order and pose estimation steps improve both the OpenSfM [2] and COLMAP [28] systems, which are two of the most widely used SfM systems. [1]

In summary, the **contributions** of this paper are: (1) Improved resection ordering that gives more weight to matches that are part of shorter tracks; (2) Improved initial pose estimation that uses only points reconstructed from reliably matching images to initialize pose before using all reconstructed points to refine pose; and (3) Experiments with two state-of-the-art SfM systems on more than 30 image sets that quantitatively and qualitatively validate the effectiveness and generality of our approach.

[1] https://github.com/rajkataria/ReliableResectioning

2. Related Work

Incremental Structure from Motion: The pipeline of feature extraction, feature matching, and incremental SfM for wide baseline, unordered images was first established by Schaffalitzky and Zisserman [27] and Snavely et al. [33]. Images are iteratively added to the reconstruction using the correspondences to triangulated points to estimate each new pose (called "resectioning"). In this paper, we address two aspects of resectioning: (1) the initial pose estimation of each new image; and (2) the order in which images are resectioned. Although many works (e.g. Agarwal et al. [5, 6], Frahm et al. [13], and Wu et al. [38]) focus on improving bundle adjustment or forms of RANSAC (e.g. Wetzel et al. [35] and Raguram et al. [24]) to improve camera localization, no works to our knowledge consider the initial pose estimation step in resectioning, which is typically performed by solving the Perspective-n-Point(PnP) [19] problem with RANSAC based on all observations of reconstructed points. Our approach modifies this by relying only on reconstructed points from images that reliably match the resectioned image, which we show in Sec. 4 leads to significantly improved results.

For resection order, the original approach of [27, 33] has remained popular (e.g., [38, 2]) — to choose the next image that observes the maximum number of reconstructed 3D points in the current reconstruction. Haner et al. [14] modify this method to account for the uncertainties of 3D point positions. More recently, Schönberger and Frahm [28], for the COLMAP system, use a pyramid-weighting scheme to give higher preference for spatially distributed observations. These approaches, despite being robust to some incorrect matches, fail in the presence of large duplicate structures mainly due to their inability to disambiguate. DeGol et al. [10] show using fiducial markers for disambiguation to improve the resectioning order (without other changes) leads to better reconstructions. Rather than relying on fidu-

cial markers, which are often not available, our method determines whether two images are likely to match based on our proposed ambiguity-adjusted match score. Our ablation study shows that our proposed change to resection order improves performance when integrated into both COLMAP and OpenSfM systems.

There is extensive research in next best view planning (e.g. Chen et al. [8], Dunn et al. [11], Bircher et al. [7], Kriegel et al. [18], and Mendoza et al. [22]), but most of this work focuses on robotics applications where the robot can move to capture the next image. In resectioning, we choose from a collection of already captured images. Our method does not apply to GlobalSfM approaches, which do not have a resectioning step. Our method also does not apply to the progressive pipeline proposed by Locher et al. [20] which addresses the problem of online reconstruction.

Disambiguation: Work in disambiguation specifically addresses incorrect matches and registration due to repeated structures in scenes. One approach tries to catch geometric inconsistencies due to confusion of repeated structures during reconstruction. For example, Zach et al. [40] enforce loop consistency (i.e. chained transformations along a cycle should yield an identity transformation). Shen et al. [31] extend this approach by checking first, second, and third order triplet loops for consistency. Roberts et al. [25] take a different approach by using an expectation-maximization framework to cluster matches that are inconsistent with the geometric constraints defined by the majority of matches. Heinly et al. [15] post-process the reconstruction looking for 3D points that conflict in their spatial location when projected into pairs of registered images. Cohen et al. [9] use appearance and geometric cues to detect symmetries and impose them as constraints during bundle adjustment. Other approaches analyze the tracks graph for inconsistencies. For example, Wilson et al. [36] prune out bad tracks using the bipartite local clustering coefficient as an indicator of noisy tracks. Yan et al. [39] attempt to tease out the capture path by exploiting the geodesic relationship of photo collections. Shah et al. [30] prune observations from the tracks graph using a min-cost network flow problem.

Most of the disambiguation approaches [31, 36, 39, 30] use a set of heuristics and carefully tuned parameters, which are usually adjusted based on the characteristics of the scene and, as we show in Section 4, do not generalize well to scenes with low levels of ambiguities. For example, Shah et al. [30] requires a different set of features to model a general scene while Yan et al. [39] and Wilson et al. [36] require parameters to be tuned based on the size and statistics of the scene. Shen et al. [31] rejects outliers based on a preset threshold and would also require tuning. In contrast, our approach addresses the problem of repeated structures directly in the reconstruction process, which prevents discarding potential matches too early and does not require dataset-specific parameter tuning.

Image Retrieval: Visual burstiness, a phenomenon in which visual words co-occur in the same spatial configuration, is analogous to disambiguation in the SfM literature. Jegou et al. [16] and Sattler et al. [26] address this by weighting the features using *idf* and feature descriptor similarities. These weighting schemes relate to our similarity measure, AAM, which downweights features that are observed in many images (long track lengths).

Algorithm 1: Structure-from-Motion Overview

```
   Input : Set of Images
1  Extract feature points on each image
2  Match pairs of images using RANSAC to obtain pairs of
     corresponding features
3  Create tracks graph using sets of corresponding features
4  Reconstruct one pair of images to initialize reconstruction
5  do
6      Select next resection candidate I_next
7      Estimate initial pose for I_next
8      if resection is successful then
9          Add I_next to the reconstruction
10         Triangulate tracks in reconstruction
11         Bundle adjust to optimize camera params and 3D points
12     end
13 until no images can be added;
   Output: Camera parameters for each image and 3D points
```

3. Method

Section 3.1 provides context for our contributions with an overview of incremental SfM. In Sec. 3.2, we describe our method to determine resection order including our ambiguity-adjusted match score. In Sec. 3.3, we describe our method to intialize pose when an image is resectioned.

3.1. Incremental Structure from Motion Overview

Algorithm 1 outlines the steps of an incremental Structure from Motion (SfM) algorithm. The blue text in the algorithm box highlights the two steps our paper addresses: (1) selecting the next resection candidate I_{next} (often referred to as resectioning order), and (2) estimating the initial pose of the resection candidate I_{next}.

Incremental SfM takes a set of images as input. For each image, feature points (i.e. points and feature descriptors from SIFT [21]) are detected (line 1). Then these feature points are matched for each image pair (line 2). These image pairs can be exhaustive or filtered using Vocab Trees [34], GPS, etc. Only matches that are inliers according to a model, such as the fundamental or essential matrix, are retained, and RANSAC [12] is used to jointly estimate the model and inlier matches. We call these inlier matches "robust matches". Next, the feature matches across image pairs are connected to form tracks (line 3). A "track" is a set of features that are transitively matched and thought to be generated by a single 3D scene point. For example, features matched across (image 1, image 2) and (image 2, image 3) can be connected into a feature track (image 1, image 2, image 3), as illustrated in Fig. 2. The tracks graph

encodes which tracks are observed by which features and images. We call an "observation" a feature in an image that is part of a track. At this point, the reconstruction process begins. First, an initial pair is chosen to begin the reconstruction (line 4). Matches between this initial pair are used to estimate the relative positions of these two images, and an initial set of 3D points is triangulated.

The iterative process of adding images to the reconstruction now begins (line 5-13). Within each iteration, the first step is to choose which image should be added next (line 6). One common approach is to select the next image to resection I_{next} as the image that observes the most 3D points in the current reconstruction (e.g., [2]). An alternative is the approach of COLMAP [28] which weights the images by the spatial distribution of feature tracks (giving preference to more spread out distributions) to encourage more stable pose estimates. We introduce another alternative in Section 3.2 that provides robustness to structured outliers.

Once the image is chosen, the next step is to estimate an initial pose for I_{next} (line 7). This is done by solving the PnP [19] problem in a RANSAC loop and refining the solution by minimizing the reprojection error using the Levenberg-Marquardt algorithm. Typically, this step uses all common tracks between I_{next} and the reconstruction. We propose an alternative approach in Section 3.3 that can improve robustness in cases where there are strong matches to images from different parts of the scene. If the initial pose estimation succeeds, the image is added to the reconstruction (line 9), and new 3D points are triangulated from tracks with two reconstructed images (line 10). *An incorrect registration can occur* if I_{next} has matches to two different parts of the scene due to a repeated structure. In this case, the pose of I_{next} may be initialized based on a large number of incorrect matches to the wrong part of the scene, and the correct matches are later discarded in outlier checks because they are inconsistent with the initial pose. This leads to inaccurate and/or incomplete reconstructions.

Finally, bundle adjustment is run (line 11) to refine the poses of all cameras and 3D points by minimizing the reprojection error of all the observations in all the images of the reconstruction. This process continues until no further images can be added. The final output is a set of camera poses and 3D points.

3.2. Resectioning Order

To determine resection order, we need to estimate whether the pose of an image is likely to be correctly estimated based on reconstructed points. It is common to choose the image that has the most tracks in common with the reconstruction, but this can cause problems due to repeating structures, as shown in Fig. 2. Long tracks are more likely to be due to duplicate structures because more images observe a repeated pattern than a unique pattern. Fig. 3 (top) quantifies this phenomenon, in terms of the probability of track length for good versus bad image matches.

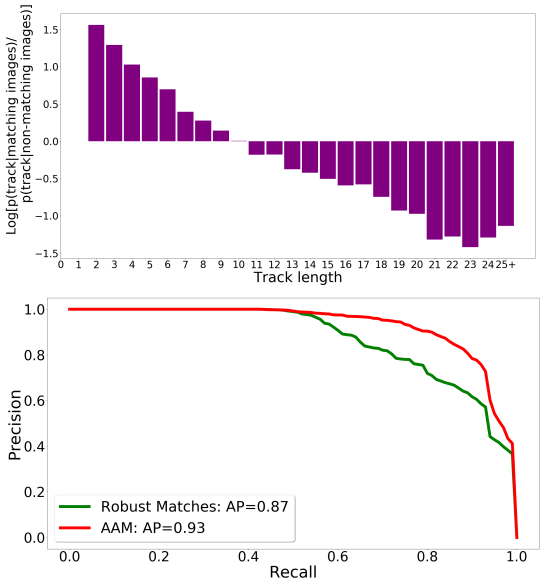

Figure 3. **Top:** The log odds ratio of the track distributions for matching images over non-matching images for all 6 scenes in the Duplicate Structures dataset[25]. Ground truth image matches were labeled manually. Matching images have a significantly higher distribution of short tracks while non-matching images have a higher distribution of long tracks. **Bottom:** This plot shows the mean Precision-Recall curves for the 6 scenes in the Duplicate Structures[25] dataset. Our proposed similarity measure, ambiguity-adjusted matches, is better at discriminating false image matches compared to the number of robust matches.

Therefore, we propose an **ambiguity-adjusted match** (AAM) similarity measure that assigns more weight to matches in short tracks, corresponding to visual patterns observed by few images:

$$S_{AAM}(i,j) = \sum_{t_k \in T_{ij}} \gamma^{|t_k|-2} \tag{1}$$

where $|t_k|$ is the length of a track that includes observations from the i^{th} and j^{th} images. In all experiments, we set the discount factor γ to 0.5. Fig. 3 (bottom) shows that our AAM score better predicts whether two images correctly match than the unweighted number of robust matches.

We determine the next image to resection as the image that is the most similar to any image in the reconstruction:

$$I_{next} = \arg\max_{c} \left(\max_{r} S_{AAM}(c \in C, r \in R) \right) \tag{2}$$

where C is the set of all candidate images, R is the set of all reconstructed images, and $S_{AAM}(c,r)$ is the ambiguity-adjusted match similarity between images c and r. We confirmed experimentally that resectioning based on the most similar image tends to outperform aggregating similarity to all reconstructed images.

3.3. Local Pose Initialization

Typically, SfM approaches estimate pose of the selected candidate image using all triangulated points that are ob-

served by the candidate image. To reduce the influence of incorrect track matches, we use a local pose initialization approach that uses reconstructed points from only reliable images to estimate pose. First, we obtain the similarity value (s_{max}) between the resectioned image and the most similar reconstructed image (similarity criteria can be robust matches or S_{AAM}) where $s_{max} = \max_r \mathrm{S}(I_{next}, r \in R)$. Second, we use s_{max} to select additional similar reconstructed images. Points generated using images that have a similarity value greater than $\tau * s_{max}$ will be used for pose initialization as shown in Equation 3.

$$R_{reliable} = \{r \in R : \mathrm{S}(I_{next}, r) > s_{max} \cdot \tau\} \quad (3)$$

R is the set of all images in the reconstruction and $R_{reliable}$ is the set of reliable images for I_{next}. We use $\tau = 0.5$ in all our experiments.

The pose for I_{next} is initialized using the reconstructed points in common with $R_{reliable}$ using RANSAC and PnP. Next, *all* reconstructed points observed by I_{next} are projected onto I_{next}, and observations with high reprojection error are pruned. Finally, the pose of I_{next} is refined using gradient descent to minimize reprojection error using all reconstructed points that correspond to inlier observations.

4. Experiments

We describe the datasets, performance measures, base SfM systems, and implementation details in Section 4.1. We compare results with our full approach to the base systems in Sec. 4.2, evaluate on the downstream task of multiview stereo in Sec. 4.3, compare to other disambiguation approaches in Sec. 4.4, and perform ablations in Sec. 4.5.

4.1. Experimental Setup

Our goal is to evaluate whether: (1) our proposed ideas to modify resection order and initial pose estimation lead to better performance in image sets that depict repeated structures; and (2) the improvements are general enough to provide good performance on varied image sets for different SfM systems. Using 29 image sets, we provide qualitative and quantitative evaluation of reconstructions, evaluation of downstream task performance, and controlled experiments with two SfM systems.

Datasets: Our main experiments are performed on three datasets: the Duplicate Structure dataset [25], the UIUCTag dataset [10], and the Tanks and Temples dataset [17]. The first two contain a total of 22 image sets that test robustness to repeated structures, such as nearly identical consumer goods in [25] and exit signs, posters, and architectural details in [10]. We also use the seven training image sets from Tanks and Temples to evaluate more general scenes. We do not train or tune parameters on any of these datasets.

In the supplemental material, we show the efficacy of our system on several challenging unstructured internet datasets of Heinly et al. [15]. Since these datasets have no ground truth or a discernable capture pattern, we provide qualitative results and identify misregistrations for comparison.

Measures: Typically, SfM papers [15, 36, 39] report only the number of registered images and points reconstructed and reprojection error, but these metrics are easily gamed by manipulating outlier checks or sampling features more densely. We visually verify whether a reconstruction has large misregistrations, which is easy for our selected datasets due to the regular capture patterns (e.g., see Fig. 5). For correct reconstructions, we evaluate completeness by the percent of images and of observations that are reconstructed. The percent of observations is the percent of detected/matched features that have low reprojection error after reconstruction. We use percent observations rather than number of reconstructed points to avoid overcounting due to splitting tracks into multiple points. As a summary of performance, we categorize the reconstruction outcomes into: "success" if it is complete (at least 90% of images registered) and has no major misregistrations; "partial" if 30% to 90% of images are registered and there are no major misregistrations; and "failure" if there are misregistrations or fewer than 30% of images registered. Additionally, for the Tanks and Temples dataset we quantitatively evaluate the precision-recall performance of a multiview stereo (MVS) reconstruction using COLMAP MVS [29].

Base SfM Systems: Our method addresses the resectioning part of the SfM pipeline and requires a complete system for testing. We choose OpenSfM (v0.2.0) [2] and COLMAP [28] (v3.4) as base systems because they are actively developed, open source, state-of-the-art systems. Except where otherwise noted, we use default parameters. We perform most ablations and comparisons with OpenSfM, and use COLMAP to demonstrate general applicability. Our method substantially improves both systems for scenes with high levels of ambiguities while performing at least as well as base systems for more general scenes.

Implementation Details: For speed, we use vocabulary trees [34, 23] to get match candidates for both systems. For OpenSfM, we employ the efficient bundle adjustment strategy of VisualSfM [38]. Since OpenSfM uses the entire tracks graph for refining the initial pose, we remove observations with high reprojection error after initial pose estimation. This modification is not needed for COLMAP because it already uses only inliers for pose refinement. Our method has two parameters ($\gamma = 0.5$ and $\tau = 0.5$) which *were set prior to the experiments and never tuned for our test datasets*. γ is chosen based on probability of a match being erroneous in the presence of duplicate structures, while τ is chosen to yield a small but sufficient neighborhood for local pose estimation (ideally 2-4 strongly matching images). The computation of image similarities (AAM), resection order, and selection of images for pose estimation have negligible compute cost, and the resulting more reliable initial pose estimates sometimes reduce the time spent in bundle adjustment. The mean speed increase of reconstruction (not

	Images	OpenSfM[2]		OOS		COLMAP [28]		OCM	
		%R	%O	%R	%O	%R	%O	%R	%O
Books	21	✗	✗	100	85	✗	✗	100	80
Cereal	25	✗	✗	100	82	✗	✗	100	68
Cup	64	✗	✗	100	74	✗	✗	✗	✗
Desk	31	✗	✗	100	91	✗	✗	100	78
Oats	23	✗	✗	100	74	✗	✗	✗	✗
Street	19	✗	✗	100	63	✗	✗	100	55
ece_floor2_hall	74	✗	✗	96	68	96	37	95	36
ece_floor3_loop	362	✗	✗	100	72	✗	✗	83	35
ece_floor3_loop_ccw	192	✗	✗	99	75	✗	✗	✗	✗
ece_floor3_loop_cw	170	✗	✗	100	77	✗	✗	100	51
ece_floor5	239	✗	✗	✗	✗	90	35	87	39
ece_floor5_stairs	328	✗	✗	94	64	79	33	80	33
ece_floor5_wall	39	44	37	97	90	✗	✗	✗	✗
ece_stairs	89	✗	✗	100	90	73	31	100	52
yeh_day_all	252	98	63	100	82	94	48	✗	✗
yeh_day_atrium	37	100	71	100	69	100	43	97	44
yeh_day_backward	120	✗	✗	100	88	92	55	90	55
yeh_day_forward	63	75	55	98	86	✗	✗	27	20
yeh_night_all	170	✗	✗	✗	✗	✗	✗	✗	✗
yeh_night_atrium	41	100	81	100	85	98	50	93	47
yeh_night_backward	79	100	66	✗	✗	91	47	90	46
yeh_night_forward	96	73	51	100	88	✗	✗	100	56
Barn	410	100	72	100	84	100	67	100	67
Caterpillar	383	✗	✗	100	81	100	67	100	67
Church	507	✗	✗	✗	✗	100	78	100	78
Courthouse	1106	✗	✗	100	76	✗	✗	100	75
Ignatius	263	100	60	100	82	100	72	100	72
Meetingroom	371	✗	✗	100	73	100	61	100	61
Truck	251	100	51	100	72	100	67	100	67

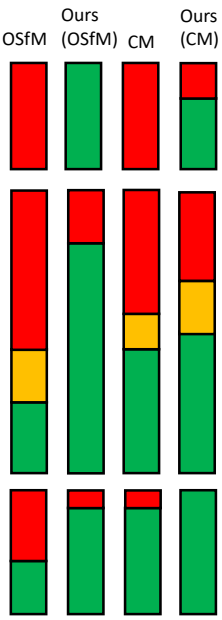

Table 1. We apply our approach to improve OpenSfM[2] and COLMAP[28], labeled OOS and OCM respectively. **%R** and **%O** indicate the percentage of images and observations reconstructed and "✗" indicates a failed reconstruction. Bars on the right show the number of success (green), partial (orange), and failure (red) cases for each method and dataset. Our method improves both SfM systems for all datasets without any parameter tuning.

counting matching) is 25% with the OpenSfM system. See the supplemental material for more details.

4.2. Overall Results

Fig. 1 shows per dataset results, comparing the base systems OpenSfM and COLMAP to the modified systems with our proposed resectioning approach. The three sections, from top to bottom, correspond to the Repeated Structures dataset, the UIUCTag dataset, and Tanks and Temples. The ideal result is to have 100% of images registered for each scene (%R) and a high percentage of observations reconstructed (%O). On the right, we summarize performance with bar charts showing the fraction of "success" (green), "partial" (orange), and "failure" (red) reconstructions in each dataset. Reconstructions with noticeable misregistrations are failures, and the %R and %O are not shown for failures, since some are incorrect.

Our approach improves both SfM systems for all datasets: For all three datasets, the inclusion of our approach in both OpenSfM and COLMAP provides a significant boost in successful reconstructions. Notably, for the Duplicate Structures dataset (extreme ambiguity), our approach using ambiguity-adjusted matches improves OpenSfM from $0/6$ to $6/6$ and COLMAP from $0/6$ to $4/6$ successful reconstructions. Performance for UIUCTags and Tanks and Temples also improves in both systems, with larger improvement in OpenSfM. Overall, our method leads to OpenSfM improving from 7 to 25 successful reconstructions in these 29 challenging scenes. Our method improves COLMAP from 13 to 19 successful reconstructions.

Fig. 5 shows examples of scenes with mismatched images (first column). While the base SfM systems were unable to achieve good reconstructions (second column), our improved resectioning order and local pose initialization lead to correct reconstructions (last column). Our supplemental material includes additional qualitative results.

4.3. Evaluation on Multi-View Stereo

To quantitatively assess the accuracy of the generated reconstructions, we employ the estimated camera poses in the downstream task of multi-view stereo (MVS). The Tanks and Temples [17] dataset provides ground-truth laser scans for their training scenes. We run the COLMAP MVS [29] pipeline using poses generated from the OpenSfM [2] and COLMAP [28] SfM pipelines with and without our proposed improvements. We use default MVS parameters. We use the benchmark's code to compute accuracy (precision) and completeness (recall) of the generated dense models after aligning the reconstructed and ground truth point clouds. Results are in Table 2.

For the OpenSfM [2] pipeline, the baseline system was unable to produce a dense model for $4/7$ scenes while our method produced a dense model for all scenes (though "Church" has misregistrations). The improvements obtained using our method matched our qualitative inspection of SfM output for all scenes except "Barn", where the models produced by baseline and our method look nearly identical, but our method has lower precision and recall, likely due to a slight misregistration in part of the model that is difficult to perceive (see supplemental material for views of the

	OpenSfM[2]			OOS			COLMAP [28]			OCM		
	P	R	F	P	R	F	P	R	F	P	R	F
Barn	0.49	0.61	0.55	0.33	0.46	0.38	0.41	0.55	0.47	0.43	0.56	0.49
Caterpillar	✗	✗	✗	0.40	0.68	0.51	0.39	0.65	0.49	0.40	0.65	0.50
Church	✗	✗	✗	0.25	0.16	0.19	0.53	0.43	0.48	0.53	0.44	0.48
Courthouse	✗	✗	✗	0.16	0.34	0.22	0.31	0.43	0.36	0.36	0.56	0.44
Ignatius	0.67	0.74	0.70	0.61	0.79	0.69	0.72	0.82	0.77	0.73	0.83	0.77
Meetingroom	✗	✗	✗	0.41	0.32	0.36	0.42	0.31	0.35	0.42	0.31	0.36
Truck	0.62	0.73	0.67	0.63	0.74	0.68	0.60	0.70	0.65	0.61	0.70	0.65
Mean	0.25	0.30	0.27	**0.40**	**0.5**	**0.43**	0.48	0.56	0.51	**0.5**	**0.58**	**0.53**

Table 2. The metrics used to compare the MVS model are Precision(P), Recall(R), and F1-score(F). The baseline OpenSfM pipeline fails to produce a model for 4/7 scenes (marked with "✗" and considered to be 0 precision and recall), while our method (OOS) produces a dense model for all scenes. For COLMAP[28], our method (OCM) produces a better model for *Courthouse* as indicated by the higher Precision, Recall, and F1-score while producing comparable models for the remaining scenes.

MVS results). For the COLMAP [28] pipeline, the baseline system and our method produced comparable results for 6 of 7 scenes. For "Courthouse", our method significantly outperformed the baseline.

4.4. Comparison to Disambiguation Methods

We compare our modified OpenSfM against two recent disambiguation approaches: Wilson et al. [36] and Yan et al. [39]. For each method, we use code provided by the authors to modify the tracks graph and otherwise use the OpenSfM system for reconstruction. For Yan et al. [39], we set the *Score* and *Coverage* parameters to 0.3 and 0.8, respectively, for small scenes ($\leq$ 60 images) and to 0.07 and 0.6, respectively, for large scenes ($>$ 60 images), following their recommendations. For [36], we leave the parameters unchanged from their default values.

Fig. 4 shows that we greatly outperform other disambiguation methods when applied to the same OpenSfM base system. Our method successfully reconstructs 25 scenes (of 29), compared to 9 succcesses for Yan et al. [39] and 4 successes for Wilson et al. [37]. Yan et al. sometimes produce a higher percentage of reconstructed observations than our method because they are able to salvage subsets of erroneous tracks. Wilson et al. achieves its best performance on larger scenes because it requires long tracks to analyze the tracks graph. The main failing of these methods is that their strategy to prune the tracks graph can either yield partial reconstructions, if too aggressive, or misregistrations, if not aggressive enough. While it might be possible to find parameters for each specific scene that yield better results, scene-specific tuning is not practical. Pruning the tracks graph based on our AAM similarity metric provides similar results to these methods (see supplemental material). Our improvements to initial pose estimation take advantage of structural knowledge without making decisions too early that risk failure and without requiring data-specific tuning.

4.5. Ablation Study

In Fig. 4, we display results of evaluating the impact of individual components of our method. "w/o Ambiguity-adjusted matches" means that unweighted matches are used as the similarity measure, instead of AAM. "w/o Local resectioning order" means that the order is based on the number of reconstructed tracks in common with an image, instead of the most similar image in reconstruction according to AAM. "w/o Local pose estimation" means that the initial pose estimate is based on all observed reconstructed points, rather than only the points that are also observed by the most reliable (according to AAM) images.

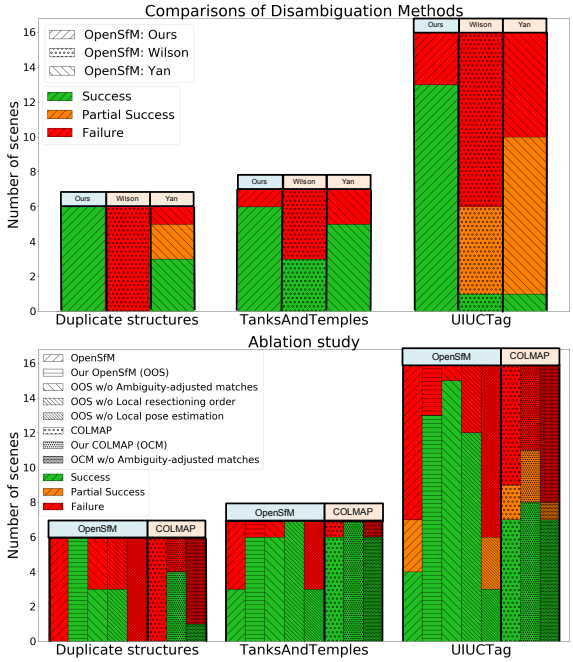

Figure 4. **Top**: Summary of performance for our method (first column) and two disambiguation approaches Wilson et al. [36] (second column), and Yan et al. [39] (third column) on the three datasets. Our method outperforms the others on each dataset. **Bottom**: Summary of reconstruction results with ablations. The bars for each dataset represent(from left to right): OpenSfM, Our OpenSfM (OOS), OOS w/o ambiguity adjust matches, OOS w/o local resectioning order, OOS w/o local pose estimation, COLMAP, Our COLMAP (OCM), and OCM w/o AAM. *Local Pose Estimation* has a large positive impact on results across all three datasets. *Local Resectioning Order* and *AAM* have a positive impact on results when repeated structures are present.

Local Pose Estimation has the biggest impact. Comparing *OOS* (our method in OpenSfM) to *OOS w/o Local Pose Estimation* across all three datasets shows that performance suffers when *Local Pose Estimation* is not used (from $25/29$ successes to $6/29$ successes). This indicates that *Lo-*

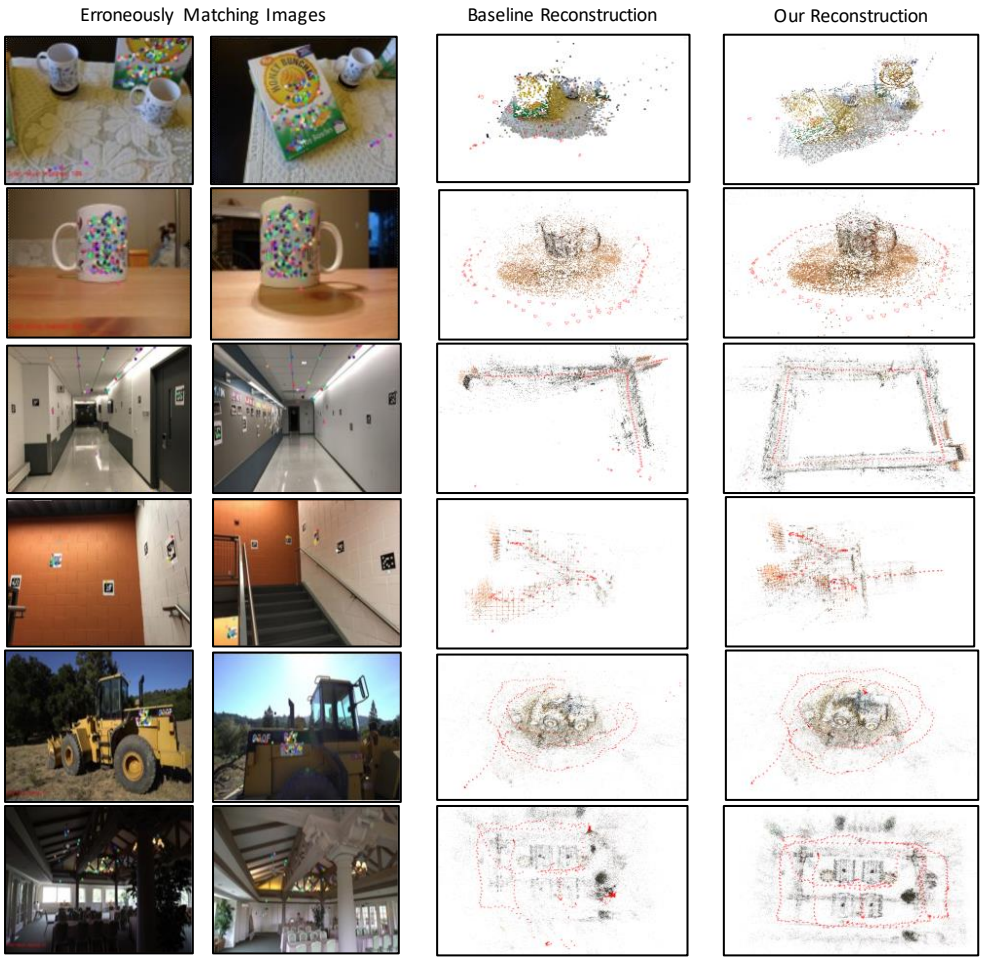

Figure 5. Qualitative results for: *Cereal*, *Cup*, *ece_floor3_loop_ccw*, *ece_stairs*, *Caterpillar*, and *Meetingroom*. The first column shows a pair of erroneously matching images. The second and third columns show the reconstructions (red frustra for camera locations) from the base OpenSfM [2] system and after modification with our approach. Bad matches like these cause misregistrations in the baseline, while our system's local pose estimation ignores the displayed incorrect matches in local pose estimation, leading to correct reconstructions.

cal Pose Estimation is useful across the range of extreme duplicate structure to minimal duplicate structure.

Local Resectioning Order helps with duplicate structure. *OOS w/o Local Resectioning Order* performs worse than *OOS* for the Duplicate Structures and the UIUCTag datasets (from total $19/22$ successes to $15/22$ successes), but performs better for Tanks and Temples (one more success). This shows that *Local Resectioning order* is beneficial for scenes with moderate to high prevalence of duplicate structures but may not be advantageous for scenes that have different challenges.

Ambiguity Adjusted Matches helps overcome extreme duplicate structure. For both *OOS* and *OCM*, the removal of ambiguity adjusted matches (AAM) decreases the number of successful reconstructions on the Duplicate Structures dataset ($6/6$ to $3/6$ and $4/6$ to $1/6$ respectively). This matches our intuition that AAM is particularly important when visually dominant textures are repeated (e.g. the Oats container from Fig. 2).

5. Conclusion

Our new method addresses the matching problem in incremental SfM caused by duplicate structures. We determine a neighborhood of reliable images using an ambiguity-adjusted similarity measure and use these images to determine resectioning order and initial pose estimates. Our approach does not require dataset-dependent parameters, in contrast to existing disambiguation methods. Results show that our method improves two state-of-the-art SfM systems, producing more complete and accurate scene reconstructions, and outperforming recent disambiguation methods. Our method is easy to implement, reduces runtime, applies to any incremental SfM system, and improves reconstruction results for a wide variety of challenging scenes.

Acknowledgments

This work is supported in part by ONR MURI Award N00014-16-1-2007 and the AWS Machine Learning Research Awards program.

References

[1] Drone deploy. https://www.dronedeploy.com/. 1

[2] OpenSfM. https://github.com/mapillary/OpenSfM. 1, 2, 4, 5, 6, 7, 8

[3] Pix4d. https://pix4d.com/. 1

[4] Reconstruct. https://www.reconstructinc.com/. 1

[5] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Commun. ACM*, 2011. 2

[6] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski. Bundle adjustment in the large. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 2

[7] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart. Receding horizon "next-best-view" planner for 3d exploration. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1462–1468. IEEE, 2016. 3

[8] S. Chen, Y. F. Li, J. Zhang, and W. Wang. *Active Sensor Planning for Multiview Vision Tasks*. 2008. 3

[9] A. Cohen, C. Zach, S. N. Sinha, and M. Pollefeys. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[10] J. DeGol, T. Bretl, and D. Hoiem. Improved structure from motion using fiducial marker matching. In *ECCV*, 2018. 2, 5

[11] E. Dunn and J.-M. Frahm. Next best view planning for active model improvement. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009. 3

[12] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981. 1, 3

[13] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, 2010. 2

[14] S. Haner and A. Heyden. Covariance propagation and next best view planning for 3d reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 2

[15] J. Heinly, E. Dunn, and J.-M. Frahm. Correcting for Duplicate Scene Structure in Sparse 3D Reconstruction. In *European Conference on Computer Vision (ECCV)*, 2014. 2, 3, 5

[16] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, 06 2009. 3

[17] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 2, 5, 6

[18] S. Kriegel, C. Rink, T. Bodenmüller, and M. Suppa. Efficient next-best-scan planning for autonomous 3d surface reconstruction of unknown objects. *Journal of Real-Time Image Processing*, 10, 12 2013. 3

[19] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 2008. 2, 4

[20] A. Locher, M. Havlena, and L. Van Gool. Progressive structure from motion. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 2004. 3

[22] M. Mendoza, J. I. Vasquez-Gomez, H. Taud, L. E. Sucar, and C. Reta. Supervised learning of the next-best-view for 3d object reconstruction. 2019. 3

[23] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006. 5

[24] R. Raguram, J.-M. Frahm, and M. Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *European Conference on Computer Vision (ECCV)*, 2008. 2

[25] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, June 2011. 2, 3, 4, 5

[26] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[27] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "how do i organize my holiday snaps?". In *European Conference on Computer Vision (ECCV)*, 2002. 2

[28] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 4, 5, 6, 7

[29] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5, 6

[30] R. Shah, V. Chari, and P. J Narayanan. View-graph selection framework for sfm. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[31] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph-based consistent matching for structure-from-motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3

[32] K. N. Snavely. *Scene Reconstruction and Visualization from Internet Photo Collections*. PhD thesis, Seattle, WA, USA, 2009. 1

[33] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 2006. 1, 2

[34] E. Uriza, F. Gómez-Fernández, and M. Rais. Efficient large-scale image search with a vocabulary tree. *Image Processing On Line*, 2018. 3, 5

[35] J. Wetzel. Image based 6-dof camera pose estimation with weighted ransac 3d. In *GCPR*, 2013. 2

[36] K. Wilson and N. Snavely. Network principles for sfm: Disambiguating repeated structures with local context. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 3, 5, 7

[37] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 7

[38] C. Wu. Towards linear-time incremental structure from motion. In *Proceedings of the International Conference on 3D Vision*, pages 127–134, 2013. 2, 5

[39] Q. Yan, L. Yang, L. Zhang, and C. Xiao. Distinguishing the indistinguishable: Exploring structural ambiguities via geodesic context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3, 5, 7

[40] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. pages 1426–1433, 06 2010. 3