Orientation of ambiguous image sequences with similar and repeated structures

Giulio Perda[1], Luca Morelli[1], Fabio Remondino[1]

[1] 3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy – (gperda, lmorelli, remondino)@fbk.eu

Keywords: orientation, matching, ambiguous scenes, repeated pattern, graph-based, doppelgangers

Abstract
Image orientation, nowadays called Structure from Motion (SfM), is still an open research topic in particular in case of scenes featuring visual aliasing, or doppelgangers. Indeed, visually similar but distinct elements of the scene can cause incorrect matches, not detected by geometric or learning-based outliers removal methods, leading to misplaced camera poses and wrong 3D reconstructions. The paper reviews various state-of-the-art approaches to orient ambiguous image sequences and determination correct camera orientation parameters. We also present an in-house graph-based approach to reliably and precisely orient sets of images with doppelgangers. Different experiments on common ambiguous datasets are reported and commented.

Figure 1: Examples of ambiguous image sequences, including symmetries, repeated structures, lack of features or low-texture surfaces.

1. Introduction

The problem of reconstructing static 3D scenes from images has been explored extensively in the past three decades (Agarwal et al., 2009; Ozyesil et al., 2017; Remondino et al. 2017) and still represents an important focus of the photogrammetric and computer vision communities. The entire pipeline consists of multiple steps, the first being camera calibration and image orientation, generally referred to Structure from Motion (SfM).
Originally coined in neuroscience in reference to the human visual system, it was Ullman who first used it within a computational context, albeit at the intersection of neuroscience. In his seminal work (Ullman, 1979), the interpretation of structure from motion is examined from a computational point of view. "The question addressed is how the 3-D structure and motion of objects can be inferred from the 2-D transformations of their projected images when no 3-D information is conveyed by the individual projections". He articulated the so-called SFM theorem: given three orthographic views of four non-coplanar points, the structure and motion compatible with these views are uniquely determined. Afterwards, Longuet-Higgins (1981) introduced the first correspondences-based automated methodology to solve SfM for a pair of images using the epipolar constraint. Later advances in reconstruction of multiple unordered image sequences served as the foundation for the reconstruction of large-scale datasets in urban scenarios or from internet photo collections (Pollefeys et al., 2004; Frahm et al., 2010; Heinly et al., 2015). The process of SfM generally consists of an initial feature extraction and matching from 2D images, carried out with either local, SIFT-like descriptors (Hartmann et al., 2015) or by more recently developed learned descriptors (Jin et al., 2021; Morelli et al., 2024), geometry verification of correspondences by outliers filtering and removal, bundle adjustment for camera pose and 3D points estimation (Triggs et al., 1999; Weber et al., 2023). Nowadays, existing SfM can handle hundreds of thousands of unordered images using conjugate gradient method (Byrod, and Astrom, 2010), visibility-based preconditioner (Kushal and Agarwal, 2012), dense factorization (Zhou et al., 2020), square root bundle adjustment (Demmel et al., 2021), or search-space (Weber et al., 2021), parallelization (Ren et al., 2022). SfM has followed three main research directions and implementations: Incremental SfM (Wu, 2013; Schönberger and Frahm, 2016; Wang et al., 2018), which involves a sequential chain of resection and intersection; Hierarchical SfM (Farenzena et al., 2009; Toldo et al., 2015), which clusters images into overlapping subsets afterwards oriented in a hierarchical manner; Global SfM (Jiang et al., 2013; Cui and Tan, 2015; Wang and Heipke, 2020), which simultaneously estimates all unknown parameters at the same time. Hyrbid SfM was also proposed (Cui et al., 2017), mentioning efficiency, accuracy and robustness in a unified framework which takes the advantages of both incremental and global methods. The vast majority of SfM methods work in an offline mode, but 3D reconstruction with collaborative (Nocerino et al., 2017) and on-the-fly (Gan et al., 2024; Zhan et al., 2025) approaches were also presented. Most of the methods retrieve correct camera poses in various scenarios, although robustness and scalability can still be improved, in particular when ambiguous scenes are present. Most popular methods rely on the correctness of the feature matching step and a subsequent outliers removal. Ambiguous datasets (Figure 1) consist of repeated structures such as similar building facades, repetitive patterns and symmetric objects. These characteristics can lead to unregistered or misregistered images, folded or incomplete reconstructed point clouds with heavy consequences on time and costs.

1.1 Paper's Aim

The paper aims to report state-of-the-art approaches to orient image sequences featuring symmetries, repeated structures and visually similar (but distinct) features (Figure 1) which are generally hampering the extraction of correct image correspondences and the determination of precise camera orientation parameters. Moreover, the work presents a developed method, based on graphs, to reliably and precisely orient sets of images of ambiguous scenarios.

2. Related works

In case of ambiguous image sequences, handcrafted or learning-based tie points extraction methods generally provide many outliers which hamper the recovery of correct camera poses. These outliers are eliminated with a conventional iterative sampling strategy based on RANSAC (Fischler and Bolles, 1981),

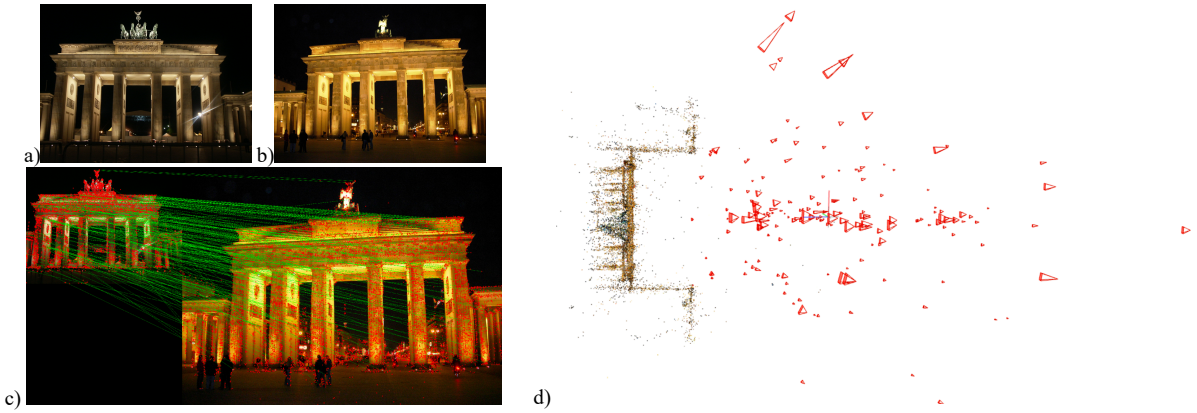

Figure 2: Front (a) and back (b) views of the Brandenburg Ture (Berlin, Germany) with distinct (at a certain scale) but quite similar features and shapes. Despite not being the same side of the monument, geometrically verified matches are established (c). Incorrect 3D reconstruction of the entire dataset without disambiguation (d).

relying on some geometric model, e.g. essential/fundamental matrix or homography. There were several attempts to improve RANSAC performances, such as spatial relations between image correspondences (Fotouhi et al., 2019), randomized model verification strategy (Chum and Matas, 2008), termination criterion to avoid the noise-free data assumption (Imre and Hilton, 2015), graph-cut approach (Barath and Matas, 2018), identification of independent inliers (Ivashechkin et al., 2021), etc. Alternatives to RANSAC include hand-crafted approaches such as LMedS (Rousseeuw and Leroy, 1987), MLESAC (Torr and Zisserman, 2000), PROSAC (Chum and Matas, 2005), DEGENSAC (Chum et al., 2005), MAGSAC (Barath et al., 2019), MAGSAC++ (Barath et al., 2020), AdaLAM(Cavalli et al., 2020) or learning-based approaches such as DSAC (Brachmann et al., 2017), CNe (Moo Yi et al., 2018) or OEAM (Ding et al., 2022). Despite all these possible approaches, in case of ambiguous image sequences, some problematic match pairs still remain as local feature matching methods easily confuse the scene (Figure 2).

Category	Papers	Highlights	Strengths	Limitations	Use Cases	Scalability
View-Graph and Cycle Consistency Enforcement	Chen et al., 2020 Cui et al., 2021 Gong et al., 2024 Havlena et al., 2010 Lee et al., 2020 Manam and Govindu, 2024 Michelini and Mayer, 2020 Shen et al., 2016 Shah et al., 2018 Sweeney et al. 2015 Wilson and Snavely, 2013 Xiao et al., 2021 Yan et al., 2017 Zach et al., 2010	Uses graph connectivity or loop constraints to ensure global pose consistency and remove ambiguous links. Graph theory techniques ensuring transitive consistency (3-cycle, loop closures), MST, clustering, graph pruning, and outlier rejection	- Improves global consistency - Handles ambiguous matchings effectively	- May require dense overlap - Computationally intensive for large scenes	- Urban environments - SfM from unordered or partial views	High
Pose averaging & motion robustification	Cui and Tan, 2015 Cui et al., 2019 Enqvist et al., 2011 Kataria et al., 2020 Wang and Heipke, 2020 Wang et al., 2019 Wilson and Snavely, 2014 Zhu et al., 2018	Focused on robustifying pairwise geometries and averaging rotations / translations in globally consistent ways	- Scalable to massive datasets - High tolerance to matching noise	- Assumes sufficient pairwise connectivity - Needs good initial geometry	- Aerial imagery - SLAM/SfM hybrid pipelines	Very High; often distributed or GPU-friendly
Feature quality & matching refinement	Lin et al., 2016	Improves matching quality through local semantic, statistical or geometric refinements	- Improves local precision - Generalizes across datasets	- Dependent on image quality and context - Sensitive to feature density	- Indoor mapping - Cultural heritage - Image-based modeling	High; match filtering scales linearly
Repetition & symmetry disambiguation	Heinly et al., 2014 Jiang et al., 2012 Roberts et al., 2011 Zach et al, 2008	Focuses on resolving structural aliasing via geometric priors, duplicate-aware processing, or missing data heuristics	- Resolves aliasing in man-made environments - Integrates structure priors	- Scene-specific (urban/facade) - Less useful in natural scenes	- Architecture-Urban scenes - Multi-object modeling	Moderate; limited by structure regularity
Learning-based disambiguation	Cai et al., 2023 Leroy et al., 2024 Peng et al., 2022 Wang et al., 2024 Xiangli et al., 2025	Deep learning for contextual disambiguation of structurally similar features	- Learn scene-level context - Adaptable to new domains	- Training data required - Potential dataset bias	- Autonomous driving - Real-world city reconstruction	High; compute resources dependence

Table 1: Proposed taxonomy and salient works covering the subject of image disambiguation.

Therefore, wrong tie points lead to wrong 3D reconstructions. This is normally called *visual disambiguation* or the *Big Ben problem* (Cai et al., 2023). Methods to remove disambiguation in scenes - also called illusory image matches or doppelgangers (Cai et al., 2023) - use heuristics-based analysis in the structure of the underlying scene graph (Zach et al., 2010; Wilson and Snavely, 2013), post-processing detection of incorrect reconstructions via minimal spanning tree (MST) and conflicting observations (Heinly et al., 2014), CNN-based feature consistency in the scene graph (Cai et al., 2023), training dataset that incorporates geo-tagged images (Xiangli et al., 2025) but also non-visual information, like timestamps or image ordering (Roberts et al., 2011). Methods based on geometric reasoning use the relative orientations (RO) of image pairs to either compute global rotation/translation averaging to estimate camera positions globally or check cycle consistency (CC) over image cycles (triplets or longer cycles) by chaining relative orientations and measuring the deviation from the identity or infer CC (CCI) over cycles in a Bayesian framework.

These methods then perform global, incremental or hierarchical SfM. Consistency is always checked on the view graph in which edges are represented by RO of image pairs. On the other hand, methods based on feature consistency do reasoning on the mutual absence or presence of features in image pairs or on their distribution in images. Features could be grouped together into clusters of correspondences, shared correspondences of an image pair are searched into a third image and if large portions of the correspondences are missing, that third image is likely matched incorrectly and its pose is an outlier.

As shown in Table 1, since almost decades new approaches are constantly proposed demonstrating how visual aliasing is challenging and still far from being fully solved (Havlena et al., 2010; Enqvist et al., 2011; Jiang et al., 2012; Heinly et al., 2014; Lin et al., 2016; Wang et al., 2019; Chen et al., 2020; Michelini and Mayer, 2020; Xiao et al., 2021; Morelli et al., 2022; Peng et al., 2022; Cai et al., 2023; Gong et al., 2024; Xiangli et al., 2025). All presented methods are based on assumptions which hamper generalization. These include the processing of ordered image sequences (Figure 2), the mandatory presence of rotations between image pairs, a relative orientation estimated from correct correspondences, geo-tagged metadata, presence of landmark images, etc. and none of the available solutions can correctly handle every possible example of ambiguous image sequence. Even newly presented SfM pipeline fully based on learning-based methods (transformer, differentiable reconstruction function, etc. (Leroy et al., 2024; Wang et al., 2024; Yang et al., 2025) suffer visual aliasing and generally cannot retrieve correct camera poses.

As there are no datasets with real ground truth data in object space to evaluate methods performances, beside visual inspections (Table 3), metrics used in the literature include: number of oriented images, weighted average of inlier ratio in the different components of a 3D reconstruction (Xiangli et al., 2025), orientation error with respect to given geolocations or, for learning-based approaches, the ROC AUC (Area Under the Receiver Operating Characteristic Curve) (Cai et al., 2023).

3. Methodology

The proposed method is a graph-based solution similar to other approaches (Sweeney et al., 2015; Shen et al., 2016; Chen et al., 2020). It begins with the construction of a weighted, undirected view graph, where each node represents an image and each edge indicates a pairwise correspondence between images. Edges are weighted solely by the number of verified feature matches but information on the pairwise geometry, such as the epipolar error, can be combined in the weight score. A community identification stage is then applied to the view graph using a greedy modularity-maximization algorithm. It begins with each node in its own community and repeatedly joins the pair of communities that lead to the largest modularity until no further increase in modularity is possible. The modularity function to be optimizes is defined as:

$$Q = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \gamma \left(\frac{k_c}{2m} \right)^2 \right]$$

where L_c denotes the total weight of edges within community c, k_c is the sum of the degrees (number of edges connected to the node) of all nodes in c, m is the sum of all edge weights in the graph, and γ is a resolution parameter that controls the granularity of the partition. A typical choice is $\gamma=1$, which balances the trade-off between the density of intra- and inter-community connections. The community identification step effectively groups images capturing visually repetitive or structurally similar regions into distinct clusters. Within each community, a maximum spanning tree (MST) is extracted using the edge weights as the selection criterion. The MST ensures that all nodes in a community remain connected with the minimum number of edges, while avoiding cycles that could propagate reconstruction errors in highly repetitive scenes. To increase local connectivity without reintroducing significant ambiguity, the MST is subsequently expanded in two stages. First, all original intra-community edges that were not part of the MST are reinstated as they are deemed robust and non-ambiguous. This enriches the local structure of each cluster, promoting stability in downstream geometric computations such as pose estimation and triangulation, which would not work properly for a too sparse view graph. Second, inter-community connectivity is restored by reintroducing all edges that originally connected neighbouring communities, that is communities which are connected by an edge in the MST. These edges provide the necessary global linkage between clusters, enabling coherent scene-wide reconstruction while maintaining robustness against the formation of spurious loops. Nodes of the graph which lie inside communities which depict different parts of the scene are hence not connected between them.

The result of this process is a sparsified yet sufficiently connected view graph that preserves meaningful geometric relationships within and across image clusters. The structure offers a strong foundation for incremental structure-from-motion, particularly in environments featuring strong repetition or limited texture.

The proposed method, like many graph-based approaches, perform well if the image dataset features a sequential acquisition whereas it can fail in case of unordered sets of images.

4. Datasets

The community has prepared and shared various datasets with ambiguous scenes to be disambiguated. Our tests used Cup, Cereal, Street, Temple of Heaven, Arc the Triomphe and A. Nevsky Cathedral available at:
https://snsinha.github.io/proj/DupSFM/index.html,
https://github.com/yanqingan/SfM_Disambiguation,
https://github.com/cvg/sfm-disambiguation-colmap.
Further datasets (Big Ben, Berliner Dome, Brandenburg Gate, Doppelgangers, MegaScenes, VisymScenes, etc.) are available at:
https://www.cs.unc.edu/~jheinly/duplicate_structure.html,
https://megascenes.github.io/,
https://github.com/doppelgangers25/doppelgangers-plusplus.
Many of these datasets are collections of Internet images. None of the ambiguous datasets features ground truth data for metric evaluation in object space.

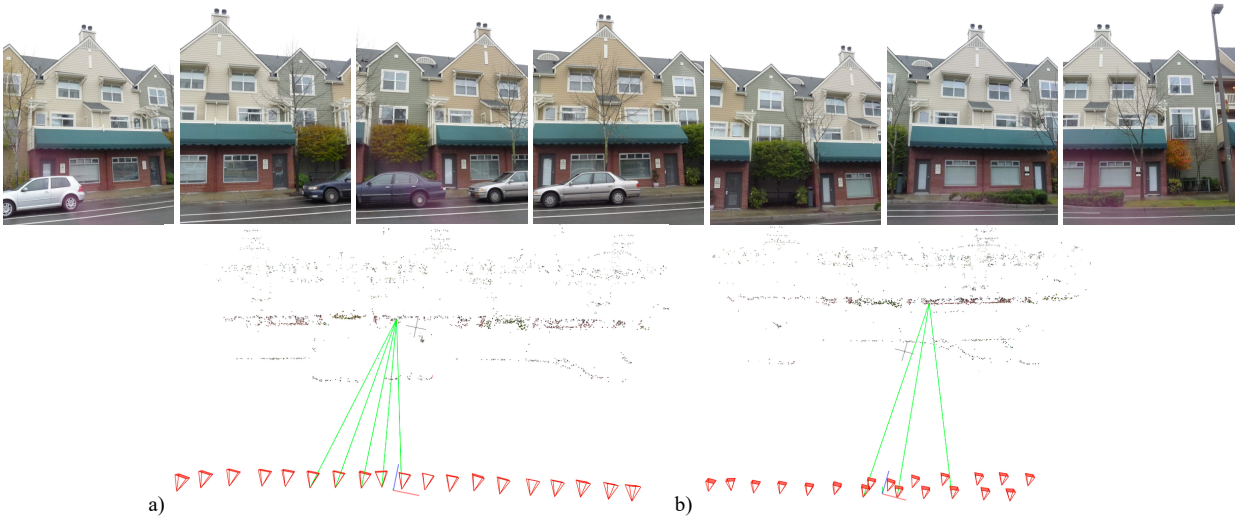

Figure 2: Images of the ambiguous Street sequence (top) and orientation results with "sequential" (a) and with no (b) assumptions.

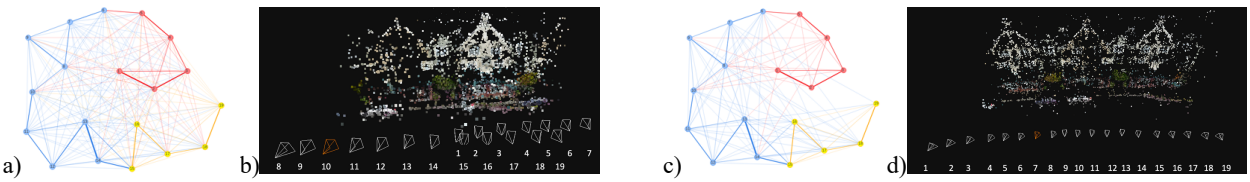

Figure 3: Results of the proposed graph-based approach. Initial graph (a) and derived incorrect camera poses (b). Refined graph with expanded MST (c) and correct camera poses (d).

	# images	COLMAP		Metashape		Reality Capture	Heinly et al., 2014	Wilson & Snavely., 2013	Cui & Tan, 2015	Yan et al., 2017	Cai et al., 2023	Wang et al., 2024	Manam & Govindu, 2024	Ours	
		E	S	E	S									MST	Exp. MST
Cup	64	X	X	X	X	X	X	X	X	✓	✓	✓	X	✓	✓
Cereal	25	X	✓	✓	✓	✓	✓	X	X	✓	X	✓	X*	✓	✓
Street	19	X	✓	X	X	✓	X	X	✓	✓	X*	X	✓	X	✓
Temple of Heaven	338	X	✓	X	X	✓	X	-	-	X*	✓	**OoM**	✓	✓	✓
Arc de Triomphe	435	X	X	X	X	X	✓	X	X	✓	✓	**OoM**	✓	X	X
A. Nevsky Cathedral	448	X	X	X	X	X*	✓	X	✓	✓	✓	**OoM**	✓	X	X

Table 2: Disambiguation results on the tested datasets across different methods. E = exhaustive. S = sequential. ✓/X/X*/- = correct disambiguation/incorrect disambiguation/reconstruction split or incomplete/not tested. OoM = Out of Memory (NVIDIA GeForce RTX 4080 16GB VRAM).

5. Experiments

The proposed disambiguation method (feature extraction and matching, view graph extraction, view graph refinement and bundle adjustment) and some state-of-the-art open or commercial methods are tested on various ambiguous sequences. Table 2 and Table 3 summarize the outcomes, showing image orientation results when the scene present repeated structures creating many illusory image matches. Assumption like "sequential" image acquisition or "exhaustive" image matching cannot generally help in case of complex scenarios. Graph-based approaches seem to be the reliable even if unordered datasets can still cause incorrect disambiguation and wrong or split 3D reconstruction. For sure the importance of using proper image acquisition protocols (Pierrot-Deseilligny et al., 2011; Nony et al., 2012; Remondino et al., 2013) in not neglectable as proper camera networks could reduce the ambiguities and support image matching.

6. Conclusions

When two (or more) images observe a diverse but visually similar scene, illusory image matches are generally created. Such scenarios are already challenging for human eyes to be differentiate therefore reliable and robust processes are needed to derive correct camera poses and sparse 3D point clouds.
The paper presented a review of visual disambiguation methods presented in the literature in the last two decades. Experiments clearly show that image orientation is still an open research task, in particular when acquired images contains ambiguous scenes and repeated structures. No method consistently works well over various ambiguous datasets and assumptions are always crucial.

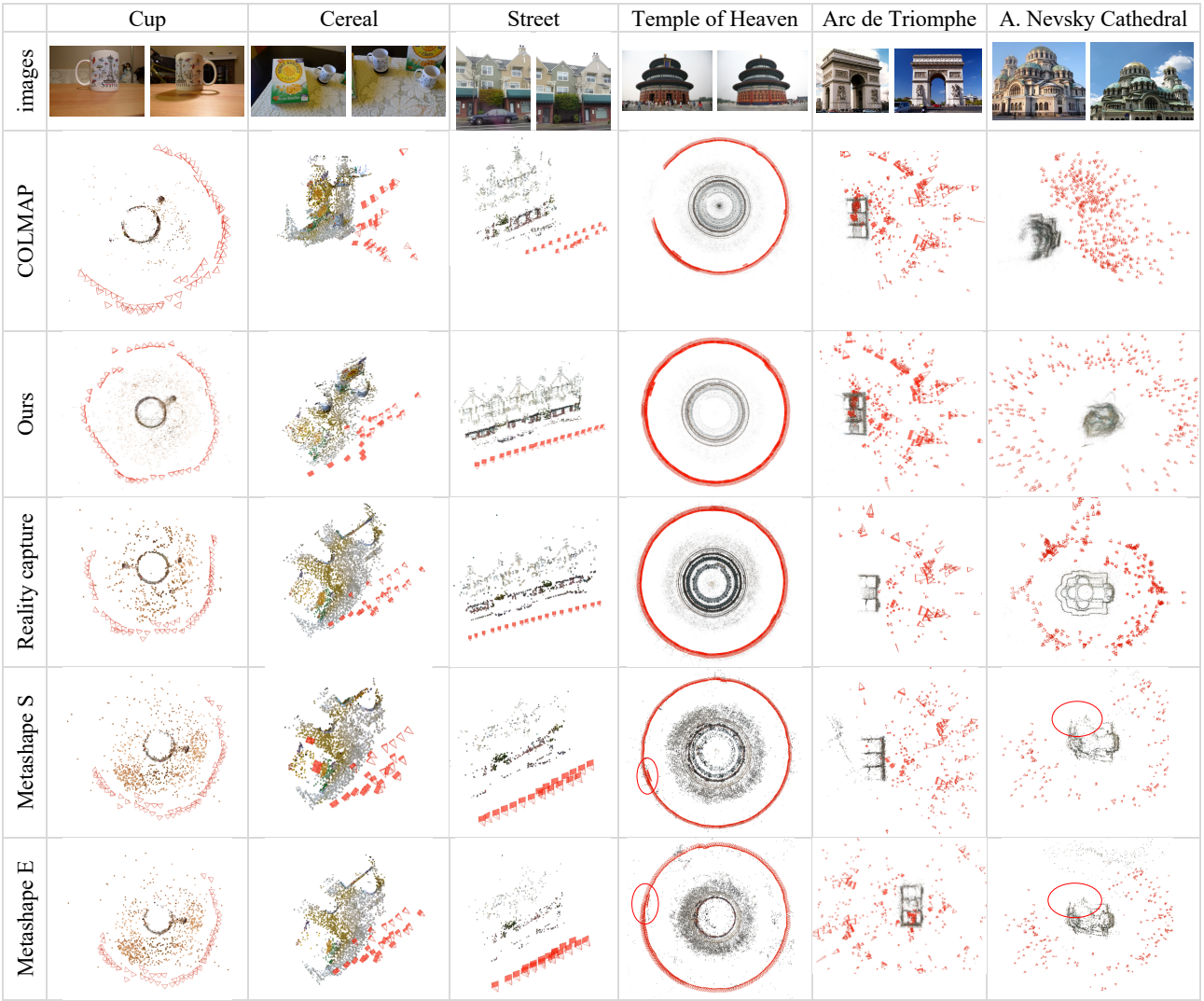

Table 3: Visual results of recovered camera poses and sparse point clouds for the tested datasets across different methods. S = sequential; E = Exhaustive.

Datasets generally considered unordered sets of images coming from the web, hampering ground truth data for metric evaluations. Newly developed learning-based methods are promising though customized on specific set of images and might generalize badly in case of context different from the training sets. Future research should not neglect conventional geometric approaches or an integration with learning ones to take the advantages of both sides.

References

Agarwal, S., Snavely, N., Seitz, S., Szeliski, R., 2009. Bundle adjustment in the large. Proc. *ECCV*, pp. 29-42

Barath, D., Matas, J., 2018. Graph-Cut RANSAC. Proc. *CVPR*, pp.4961-4974.

Barath, D., Matas, J., Noskova, J., 2019. MAGSAC: Marginalizing Sample Consensus. Proc. *CVPR*, pp. 10197-10205.

Barath, D., Noskova, J., Ivashechkin, M., Matas, J., 2020. MAGSAC++, a fast, reliable and accurate robust estimator. Proc. *CVPR*, pp. 1304-1312.

Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C., 2017. DSAC - differentiable RANSAC for camera localization. Proc. *CVPR*, pp. 6684-6692.

Byrod, M., Astrom, K., 2010. Conjugate gradient bundle adjustment. Proc. *ECCV*, pp. 114-127.

Cai., R., Tung, J., Wang., Q., Averbuch-Elor, H., Hariharan, B., Snavely, N., 2023. Doppelgangers: Learning to Disambiguate Images of Similar Structures. Proc. *ICCV*, pp. 34-44.

Cavalli, L., Larsson, V., Oswald, M.R., Sattler, T., Pollefeys, M., 2020. Handcrafted outlier detection revisited, Proc. *ECCV*, pp. 770–787

Chen, Y., Shen. S., Chen, Y., Wang., G., 2020. Graph-based parallel large scale structure from motion. *Pattern Recognition,* Vol. 107, 107537.

Chum, O., Matas, J., 2005. Matching with PROSAC – Progressive Sample Consensus. Proc. *CVPR*, pp. 220-226.

Chum, O. Werner, T., Matas, J., 2005. Two-View geometry estimation unaffected by a dominant plane. Proc. *CVPR*, Vol. 1, pp. 772-779.

Chum, O., Matas, J., 2008. Optimal randomized RANSAC. IEEE *Trans. Pattern Anal. Mach. Intell*. Vol. 30(8), pp. 1472-1482.

Cui, Z., Tan, P., 2015. Global structure-from-motion by similarity averaging. Proc. *ICCV*, pp. 864-872.

Cui, H., Gao, X., Shen, S., Hu, Z., 2017. HSfM: Hybrid Structure-from-Motion. Proc. *CVPR*.

Cui, H., Shen, S., Gao, W., Liu, H., Wang, Z., 2019. Efficient and robust large-scale structure-from-motion via track selection and camera prioritization. *ISPRS Journal of Photogrammetry and Remote Sensing,* 156, pp. 202-214.

Cui, H., Shi, T., Zhang, J., Xu, P., Meng, Y., Shen, S., 2021. View-graph construction framework for robust and efficient structure-from-motion. *Pattern Recognition*, 114, p.107712.

Demmel, N., Sommer, C., Cremers, D., Usenko, V., 2021. Square Root Bundle Adjustment for Large-Scale Reconstruction. Proc. *CVPR*, pp. 11723-11732.

Ding, X., Luo, Y., Jie, B., Li, Q., Cheng, Y., 2022. Using outlier elimination to assess learning-based correspondence matching methods. *Information Sciences*, Vol. 659, 120056

Enqvist, O., Kahl, F., Olsson, C., 2011. Non-sequential structure from motion. Proc. *ICCV*, pp. 264-271.

Farenzena, M., Fusiello, A., Gherardi, R., 2009. Structure-and-motion pipeline on a hierarchical cluster tree. Proc. *ICCV* Workshops, pp. 1489-1496

Fischler, M., Bolles, R., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, Vol. 24(6), pp. 381-395.

Fotouhi, M., Hekmatian, H., Kashani-Nezhad, M.A., Kasaei, S., 2019. SC-RANSAC: spatial consistency on RANSAC. *Multimed. Tools Appl.* 78 (7), pp. 9429-9461.

Frahm, J.M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.H., Dunn, E., Clipp, B., Lazebnik, S., 2010. Building Rome on a cloudless day. Proc. *ECCV*, pp. 368-381

Gan, W., Yu, Y., Perda, G., Morelli, L., Xia, R., Zhan, Z., Wang, X., Remondino, F., 2024. LVG-SfM: Learning-Based View-Graph Generation for Robust on-the-Fly SfM. Proc. *ECCV* Workshops, LNCS, Vol 15623, pp. 158–174.

Gong, Y., Zhou, P., Liu, C., Yu, Y., Yao, J., Yuan, W. and Li, L., 2024. A cluster-based disambiguation method using pose consistency verification for structure from motion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209, pp. 398-414.

Hartmann, W., Havlena, M., Schindler, K., 2015. Recent developments in large-scale tie-point matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 155, pp. 47-62.

Havlena, M., Torii, A., Pajdla, T., 2010. Efficient Structure from Motion by Graph Optimization. Proc. *ECCV*, pp. 100-113.

Heinly, J., Dunn, E., Frahm, JM., 2014. Correcting for Duplicate Scene Structure in Sparse 3D Reconstruction. Proc. *ECCV*, pp. 780-795.

Heinly, J., Schonberger, J.L., Dunn, E., Frahm, J.-M., 2015. Reconstructing the World* in six days *(as captured by the Yahoo 100 million image dataset). Proc. *CVPR*, pp. 3287-3295.

Imre, E., Hilton, A., 2015. Order statistics of RANSAC and their practical application. *Int. J. Comput. Vis*., Vol. 111(3), pp. 276-297.

Ivashechkin, M., Barath, D., Matas, J., 2021. VSAC: efficient and accurate estimator for H and F. Proc. *ICCV*, pp. 15243-15252.

Jiang, N., Tan., P., Cheong, L., 2012. Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. Proc. *CVPR,* pp. 1458-1465

Jiang, N., Cui, Z., Tan, P., 2013. A global linear method for camera pose registration. Proc. *ICCV*, pp. 481-488.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E., 2021. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, *129*(2), pp.517-547.

Kataria, R., DeGol, J., Hoiem, D., 2020. Improving structure from motion with reliable resectioning. Proc. *3DV*, pp. 41-50.

Kushal, A., Agarwal, S., 2012. Visibility based preconditioning for bundle adjustment. Proc. *CVPR*, pp. 1442-1449

Lee, S., Lim, J., Suh, I.H., 2020. Progressive feature matching: Incremental graph construction and optimization. *IEEE transactions on image processing*, 29, pp. 6992-7005.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3D with MASt3R. Proc. *ECCV*.

Lin, W.Y., Liu, S., Jiang, N., Do, M.N., Tan, P., Lu, J., 2016. Repmatch: Robust feature matching and pose for reconstructing modern cities. Proc *ECCV*, pp. 562-579

Longuet-Higgins, H., 1981. A computer algorithm for reconstructing a scene from two projections. *Nature,* 293, pp. 133-135.

Manam, L., Govindu, V.M., 2024. Leveraging camera triplets for efficient and accurate Structure-from-Motion. Proc. *CVPR*, pp. 4959-4968.

Michelini, M., Mayer, H., 2020. Structure from motion for complex image sets. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 166, pp. 140-162.

Moo Yi, K., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P., 2018. Learning to find good correspondences. Proc. *CVPR*, pp. 2666-2674.

Morelli, L., Karami, A., Menna, F., Remondino, F., 2022. Orientation of images with low contrast textures and transparent objects. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLVIII-2/W2-2022, 77-84.

Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., Remondino, F., 2024. Deep-image-matching: a toolbox for multiview image matching of complex scenarios. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-2/W4-2024, 309-316.

Nocerino, E., Poiesi, F., Locher, al, Y.T. Tefera, Remondino, F., Chippendale, P., Van Gool, L., 2017. 3D reconstruction with a collaborative approach based on smarthphones and a cloud-based server. ISPRS *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XLII-2/W8, pp. 187-194.

Nony, N., De Luca, L., Godet, A., Pierrot-Deseilligny, M., Remondino, F., van Dongen, A., Vincitore, M., 2012. Protocols and assisted tools for effective image-based modeling of architectural elements. *LNCS*, Vol. 7616, pp. 432-439, Springer.

Ozyesil, O., Voroninski, V., Basri, R., Singer, A., 2017. A survey of structure from motion. *Acta Numerica*, Vol. 26, pp. 305-364.

Peng., Y., Yan, S., Liu, Y., Liu, Y., Zhang, M., 2022. View graph construction for scenes with duplicate structures via graph convolutional network. *IET Computer Vision*, Vol. 16(5), pp. 389-402.

Pierrot-Deseilligny, M., De Luca, L., Remondino, F., 2011. Automated image-based procedures for accurate artifacts 3D modeling and orthoimage generation. *Geoinformatics FCE CTU Journal*, vol. 6, pp. 291-299.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R., 2004. Visual modeling with a hand-held camera. *Int. Journal of Computer Vision*, Vol. 59, pp. 207-232.

Remondino, F., Menna, F., Koutsoudis, A., Chamzas, C., El-Hakim, S., 2013: Design and implement a reality-based 3D digitisation and modelling project. Proc. *IEEE Conference "Digital Heritage 2013"*, Vol. 1, pp. 137-144.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* Vol. XLII-2/W5, pp. 591-599

Ren, J., Liang, W., Yan, R., Mai, L., Liu, X., 2022. MegBA: A high-performance and distributed library for largescale bundle adjustment. Proc. *ECCV*, Vol. 2.

Roberts, R., Sinha, S.N., Szeliski, R., Steedly, D., 2011. Structure from motion for scenes with large duplicate structures. Proc. *CVPR*, pp. 3137-3144.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. *Wiley Interscience*, New York, 329 pages.

Schönberger, J.L., Frahm, J.M., 2016. Structure-from-motion revisited. Proc. CVPR, pp. 4104-4113.

Shah, R., Chari, V., Narayanan, P.J., 2018. View-graph selection framework for sfm. Proc. *ECCV*, pp. 535-550.

Shen, T., Zhu, S., Fang, T., Zhang, R., Quan, L., 2016. Graph-based consistent matching for Structure-from-Motion. Proc. *ECCV*, pp. 139-155.

Sweeney, C., Sattler, T., Höllerer, T., Turk, T., Pollefeys, M., 2015. Optimizing the Viewing Graph for Structure-from-Motion. Proc. *ICCV*, pp. 801-809.

Toldo, R., Gherardi, R., Farenzena, M., Fusiello, A., 2015. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*. Vol. 140, pp. 127-143

Torr, P., Zisserman, A., 2000. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Computer Vision and Image Understanding*, Vol. 78(1), pp. 138-156.

Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W., 1999. Bundle adjustment - a modern synthesis. Proc. *International workshop on Vision Algorithms*, pp. 298-372.

Ullman, S., 1979. *The Interpretation of Visual Motion*. MIT Press.

Xiangli, Y., Cai, R., Chen, H., Byrne, J., Snavely, N., 2025. Doppelgangers++: Improved visual disambiguation with geometric 3D features. Proc. *CVPR*.

Wang, X., Rottensteiner, F., Heipke, C., 2018. Robust image orientation based on relative rotations and tie points. *ISPRS Ann. Photogram. Remote Sens. Spatial Inf. Sci.*, IV-2 4(2), 295-302.

Wang, X., Xiao, T., Gruber, M., Heipke, C., 2019. Robustifying relative orientations with respect to repetitive structures and very short baselines for global SfM. Proc. *CVPR workshop PCV*.

Wang, X., Heipke, C., 2020. An improved method of refining relative orientation in global structure from motion with a focus on repetitive structure and very short baselines. *Photogram. Eng. Remote Sens*., Vol. 86(5), 299-315.

Wang, J., Karaev, N., Rupprecht, C., Novotny, D., 2024. VGGSfM: Visual Geometry Grounded deep Structure From Motion. Proc. *CVPR*, pp. 21686-2169.

Weber, S., Demmel, N., Cremers, D., 2021. Multidirectional conjugate gradients for scalable bundle adjustment. Proc. *German Conference on Pattern Recognition* (GCPR), pp. 712-724.

Weber, S., Demmel, N., Chan, T.C., Cramer, D., 2024. Power bundle adjustment for large-scale 3D reconstruction. Proc. *CVPR*, pp. 281-289.

Wilson, K., Snavely, N., 2013. Network principles for SfM: disambiguating repeated structures with local context. Proc. *ICCV*, pp. 513-520.

Wilson, K., Snavely, N., 2014. Robust global translations with 1dsfm. Proc. *ECCV*, pp. 61-75.

Wu, C., 2013. Towards linear-time incremental structure from motion. Proc. *International Conference on 3D Vision*, pp. 127-134.

Xiao, T., Wang, X., Deng, F., Heipke, C., 2021. Sequential cycle consistency inference for eliminating incorrect relative orientations in Structure from Motion. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 89(3), pp. 233-249.

Yan, Q., Yang, L., Zhang, L., Xiao, C., 2017. Distinguishing the indistinguishable: Exploring structural ambiguities via geodesic context. Proc. *CVPR*, pp. 3836-3844.

Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M., 2025. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. Proc. *CVPR*.

Zach, C., Irschara, A., Bischof, H., 2008, June. What can missing correspondences tell us about 3d structure and motion?. Proc. *CVPR*, pp. 1-8.

Zach, C., Klopschitz, M., Pollefeys, M., 2010. Disambiguating visual relations using loop constraints. Proc. *CVPR*, pp. 1426-1433.

Zhan, Z., Yu, Y., Xia, R., Gan, W., Xie, H., Perda, G., Morelli, L., Remondino, F., Wang, X., 2025. SfM on-the-fly: A robust near real-time SfM for spatiotemporally disordered high-resolution imagery from multiple agents. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224, pp. 202-221.

Zhou, L., Luo, Z., Zhen, M., Shen, T., Li, S., Huang, Z., Fang, T., Quan, L., 2020. Stochastic bundle adjustment for efficient and scalable 3D reconstruction. Proc. *ECCV*, pp. 364-379.

Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L., 2018. Very large-scale global sfm by distributed motion averaging. Proc. *CVPR,* pp. 4568-4577.