LVG-SfM: Learning-based View-Graph generation for robust on-the-fly SfM

Wentian Gan[1†], Yifei Yu[1†], Giulio Perda[2], Luca Morelli[2], Rui Xia[1], Zongqian Zhan[1], Xin Wang[1✉], and Fabio Remondino[2]

[1] Wuhan University, Wuhan, China
[2] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy

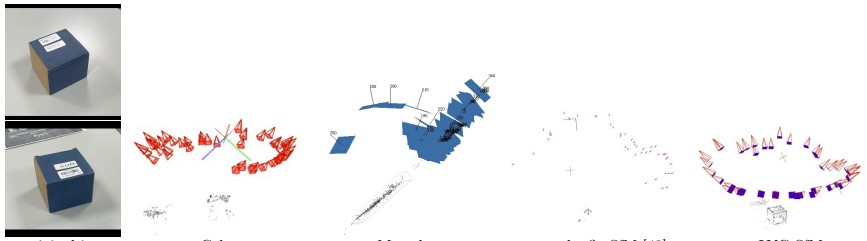
Abstract. Structure from Motion (SfM) has been widely studied in many fields, such as computer vision, photogrammetry, robotics, *etc*. Recent advancements focus on improving the real-time performance of SfM, which is crucial for applications in augmented reality, mixed reality, robotics, *etc*. However, the robustness of real-time processing is still limited by outliers in the feature extraction and matching process, stemming from challenging scenes depicting objects with poor texture, repetitive structures, and symmetric objects, which can cause blunders in the view-graph. Focusing on these scenes, a **L**earning-based **V**iew-**G**raph generation method (LVG-SfM) is investigated and integrated into the on-the-fly SfM pipeline [43]. First, to provide a higher number of reliable matches and generate a more robust view-graph, a set of SoTA learning-based feature extraction and matching methods [19] are tested. Then, the spuriously incorrect two-view geometries generated from repetitive structures are removed from the view-graph with the help of SoTA learning-based disambiguation network – Doppelgangers [3]. Experimental results demonstrate that our LVG-SfM can successfully work on-the-fly on challenging ambiguous scenes with poor textures and repetitive structures, achieving correct scene reconstructions and robustifying SfM. Project website at: https://sygant.github.io/lvgsfm

Keywords: Structure from Motion · View-Graph · Disambiguation

1 Introduction

In the past half-century, Structure-from-Motion (SfM) has been extensively studied in the fields of computer vision, photogrammetry, robotics, *etc*. It estimates camera poses and sparse point cloud using three different approaches: (i) Incremental SfM [1, 25, 38] solves images sequentially with recursive bundle adjustment (BA); (ii) Hierarchical SfM [9,18,33] divides images into small subsets and process them in parallel to improve time efficiency; (iii) Global SfM [6, 14, 36] takes all potential two-view geometries as input and outputs all camera poses

† These two authors contributed equally to this work.
✉Corresponding author. Email: xwang@sgg.whu.edu.cn



(a) original image (b) Colmap (c) Metashape (d) on-the-fly SfM [43] (e) LVG-SfM

Fig. 1: Reconstruction results of a scene with a symmetric object. LVG-SfM successfully generates a correct reconstruction while popular SfM software output incorrect camera poses.

simultaneously. However, most of these SfM methods operate offline, whereas the demand for real-time applications (such as quick disaster response, online measurements, collaborative 3D mapping, *etc.*) is increasing. Therefore, many researchers investigated online (or real-time) SfM solutions that aim to solve camera poses and sparse point cloud at speeds comparable to the image capturing rate, such as, online-feedback SfM [12], RTSfM [46], On-the-Fly SfM [42, 43].

The input to both offline and online SfM approaches are feature correspondences and eligible multi-view geometries. Relationships between images can be modeled with a view-graph: images represent the nodes while edges relate images through a metric derived from the correspondences or the two-view geometry. Compared to offline SfM, online SfM is more sensitive to incorrect view-graphs. Indeed, offline SfM first builds a complete view-graph which is then solved by various robust estimation strategies, whereas online SfM incrementally adds the newly captured image to already registered images, potentially causing two problems: first, given only a partial view-graph, a good initial stereo reconstruction is much harder to obtain; second, the registration of newly captured image degenerates if the partial view-graph contains a high rate of outliers. Therefore, an accurate view-graph is crucial for online SfM. Moreover, lack of texture, presence of repetitive structures, or highly symmetric scenes pose a challenge to the construction of a correct view-graph [11, 41].

Supported by recent advancements in learning-based feature extraction, matching and outlier detection methods, a more robust view-graph can be constructed, significantly enhancing the performances of online SfM. The paper presents a new real-time SfM solution, named **LVG-SfM**, which integrates and offers three operative processes:

- **Learning-based correspondence generation:** we leverage on [19] to extract and match sufficient and robust correspondences even in case of poor textures using learning-based image matching methods, such as SuperPoint [7], DISK [30], ALIKE [45], ALIKED [44], SuperGlue [24] and LightGlue [15].
- **Learning-based view-graph robustification for ambiguous edges elimination:** we leverage on Doppelgangers [3] to further prune, after the two-

view geometric verification, a view-graph by eliminating ambiguous edges due to repetitive structures.
- **LVG-SfM:** the proposed method builds upon [42] to offer an advanced and robust real-time multi-agent SfM pipeline able to tackle ambiguous image sequences with repetitive structures and poor texture scenarios.

As illustrated in Fig. 1, the newly presented method is successful in real-time reconstruction, even in scenarios with poor textures and repetitive structures. In contrast, existing online and offline methods, including commercial software, often yield incorrect camera poses and visual artifacts.

2 Related work

In this section, three relevant topics are reviewed, including SfM, local feature extraction and matching, and disambiguation of ambiguous scenes.

2.1 SfM

So far, SfM has obtained ample achievements, which is proved by many well-established open-sourced SfM frameworks. Popular incremental pipelines include Visual SfM [37], Colmap [25] and Micmac [23], while OpenMVG [20] and Theia [29] also support global SfM. Despite their popularity, all these SfM frameworks work offline, meaning that image acquisition and image processing happen separately. Online methods, on the other hand, concurrently solve SfM while images are being captured by one or multiple agents. To achieve the goal of online processing, [46] proposed real-time SfM (RTSfM), employing hierarchical feature matching and Bag-of-Words to boost image matching speed, however, spatiotemporal continuity between images is required. Recently, [42] proposed on-the-fly SfM with a novel online image retrieval method to alleviate the requirement of spatiotemporal continuity. [43] enhance this method to support multiple agents and achieve online SfM by efficiently merging multiple submaps. Our approach further improves the robustness of [43] by generating a more reliable view-graph for subsequent processing.

2.2 Local Feature Extraction and Matching

Local feature extraction and matching are vital for image-based localization and 3D reconstruction. Over the last several decades, many popular hand-crafted detectors have been developed, for example, SIFT [16], one of the most widely-used features in SfM, is invariant to rotation and scaling, SURF [2] is an improved version with higher detection speed and improved descriptor, ORB [22] is frequently applied in VSLAM (visual simultaneous location and mapping) methods due to its highly time-efficient detection and matching. However, all these methods show a certain degree of degeneration under very large viewpoint changes and poor texture scenarios. In recent years, learning-based image-matching approaches have been investigated to overcome these defects. LIFT [40] introduced

a Deep Network architecture encompassing several modules for detection, orientation estimation, and feature description. SuperPoint [7] presented a self-supervised network framework with two parallel heads to detect keypoints and output descriptors, respectively. ALIKED [44] introduces the Sparse Deformable Descriptor Head (SDDH) to efficiently extract strong expressiveness descriptors in challenging images. DISK [30] employs a four-layer U-Net architecture to ensure precise extraction of feature points, combined with advanced reinforcement learning algorithms for strategy optimization, allowing it to adaptively learn which pixels are most representative. DeDoDe [8] learns keypoints directly from 3D consistency, using large-scale SfM reconstructions as supervision. Instead of the conventional matching method based on the nearest neighbor search, recently, motivated by the attention mechanism, SuperGlue [24] provides a new learning-based approach by integrating the expressive capabilities of transformers with optimal transport to handle extremely large view changes. However, its training becomes more complex as the number of keypoints increases. Two SOTA approaches, Loftr [28] and LightGlue [15] take inspiration from SuperGlue. Loftr treats every pixel on the 1/8th resolution activation as a candidate keypoint which is then matched similarly to SuperGlue, namely, detector-free matching (and bundle adjustment [10]). LightGlue achieves faster inference speeds on easily matchable image pairs while maintaining higher accuracy on more challenging image pairs.

2.3 Disambiguation of ambiguous scenes

All SfM methods rely on the correctness of feature matching and outliers removal. Ambiguous datasets (Sec. 4.1, Tab. 2) pose a challenge for conventional SfM methods. Generally, ambiguity arises from repetitive structures ranging from duplicate instances of the same object caused by 3D rotational symmetries, separate but identical surfaces or mirrored structures – often found on building facades. Other sources of ambiguity are poor or repetitive textures, or the absence of distinctive background features [11, 27, 35, 41]. Ambiguity can result in large sets of wrong matches, translating to self-consistent incorrect epipolar geometries (EG) that pass standard geometric verification [34], and ultimately lead to folded or duplicated structures. Existing disambiguation methods are diverse. [41] analyze missing correspondences among various overlapping images to check the correctness of the corresponding image pairs, [5] exploit ambiguous matches to first recover symmetry and then impose it as an additional constraint in bundle adjustment, [11] proposes a post-processing solution, which detects conflicting observations in the reconstructed model stemming from images observing 3D points that should be occluded by other structures, [4] detect symmetry and repetitive structures in building facades and use a graph-based global analysis to recover correct 3D geometry, [26,34] both utilize the view-graph to explore the distribution of matched and non-matched features among all images and estimate a criterion for identifying repetitive structures, [31] detect ambiguity by leveraging on the consistency of image background by looking for visual contradictions

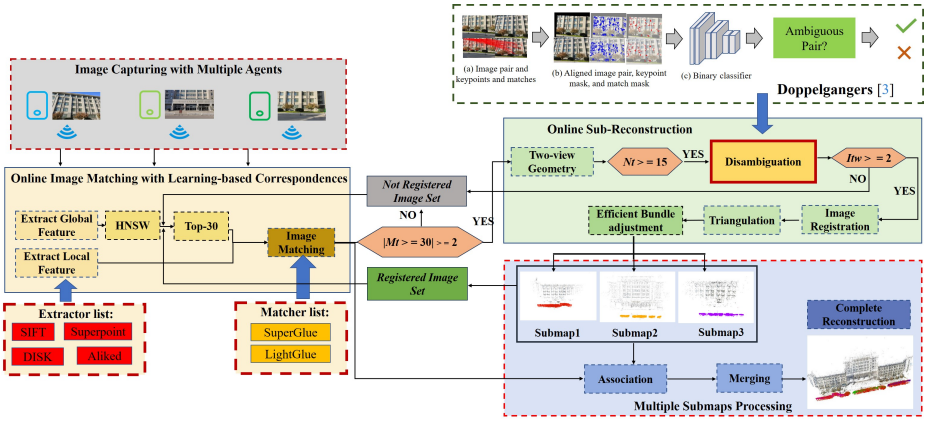


Fig. 2: Overall workflow of the proposed LVG-SfM. *Mt* denotes the number of matched features, Nt is the number of inlier correspondences after two-view geometry verification, *Itw* indicates the number of edges connecting new image and registered images

in scene segments, which may be few or difficult to detect. Recently, disambiguation methods exploited deep-learning techniques. In [39] a graph convolutional network is trained to perform image retrieval whereas [3] treat disambiguation as a binary classification problem by training a model with pairs of images. This model outputs the probability of ambiguity for the pair. Although this method has limitations in non-landmark scenes, it outperforms other approaches in most publicly available datasets with ambiguous image sequences [11, 21, 35] and for this reason we integrate into the on-the-fly framework.

3 Methodology

The presented **LVG-SfM** supports multi-agent online processing of image sequences (acquired with/without photogrammetric acquisition recommendations). The overall workflow (shown in Fig. 2) builds upon [42, 43] which contains four main modules: image capturing with multiple agents, online image matching, online sub-reconstruction and multiple submaps merging (more details can be found at https://yifeiyu225.github.io/on-the-flySfMv2.github.io) and adds new and revised functionalities: feature extraction and matching with various learning-based methods [19], disambiguation of incorrect two-view geometries resulting from repetitive structures using the Doppelgangers [3] approach.

3.1 Image Capturing with Multiple Agents

The image collector end supports a variety of image-capturing devices (e.g. mobile phones or tablets), working simultaneously. The captured images are transmitted to the processing end via local networks, 4G, or 5G. When simulating the acquisition of an already existing dataset, the images are incrementally sent to

Table 1: Characteristics of the employed local features descriptors.

Methods	Descriptor Dimension	Invariances	Complexity of training	Real-time Performance (with GPU)
SIFT	128	Scale, rotation	none	Fast
DISK	128	Scale and illumination changes	hard	Slow
ALIKED	256	Scale and illumination changes	relatively easy	Fast
SuperPoint	256	Scale, rotation and illumination changes	hard	Middle

the processing end in the same order as they are stored. Agents could focus on separate sections of the scene without mutual overlap. For this reason, it should be possible to reconstruct the scene into separate, incremental submaps which may eventually be joined.

3.2 Online image matching with learning-based methods

In general, image matching is the first step and also one of the most time-consuming processes. To achieve real-time performance for each new image, the original retrieval module of on-the-fly SfM [43] uses a pre-trained global feature extractor [13] and HNSW [17], selects up to 30 of the most similar pairs between the new image and images in the "*registered images*" set - containing successfully registered images and the "*not registered images*" set (see explanation below), whereby image matching is performed.

In addition to the handcrafted SIFT feature and nearest neighbour matcher employed by the original on-the-fly SfM, to cope with images of poor texture, as Fig. 2 shows, three learning-based local feature extractors including SuperPoint [7], DISK [30] and ALIKED [44] (Tab. 1) are tested, and two learning-based feature matching methods are integrated by SuperGlue [24] and LightGlue [15]. Note that all the chosen approaches are considered references between learning-based local features and matchers, but the current analysis could be extended also to others features.

For these similar pairs, local features are extracted and matched using the above-mentioned learning-based extractors. If the new image meets a requirement on the minimum number of matches ($\geq$ 30) for at least two pairs, the successful pairs proceed to the two-view geometry verification step. If not, the new image is temporarily rejected and added to the "*not registered images*" set, making it available for matching with future images.

3.3 Online sub-reconstruction with Disambiguation of two-view geometries

Ambiguity typically results in incorrect matches due to object symmetry, similarity or repetitive textures. These matches can sometimes pass the canonical two-view geometry verification [32,34], making disambiguation hardly achievable by focusing on a single pair. Thus, it is necessary to consider larger groups of images and provide a global understanding of the scene's ambiguity. Therefore,

the Doppelgangers approach [3] is utilized, which considers all image pairs passing geometric verification. The method trains a classifier network using image pairs of similar structures carefully selected from internet collections of world landmarks. Additionally, the spatial distribution of the matches for the pair is provided as input to the network. The classifier outputs the probability that the given pair is a true match. If the number of remaining pairs after disambiguation is above two, the newly captured image is oriented with the pipeline of [43] and added into the "*registered image*" sets, otherwise, it is inserted into the "*not registered image*" set. The result of the disambiguation process for each new image is the updated partial view-graph pruned by removing connections between ambiguous pairs.

For the online sub-reconstruction of our LVG-SfM, the canonical two-view epipolar geometry is firstly verified based on RANSAC, only the two-view geometries with more than 15 inliner correspondences are sent to doppelgangers for disambiguation (more details are described as above), then image with at least two edges connecting to a submap is solved by image registration, triangulation and efficient bundle adjustment.

3.4 Multiple submaps processing

Different agents can simultaneously work on separate parts of the scene, leading to non-overlapping image subsets. For each of these subsets, a distinct submap is created and updated in parallel. A new image could be added to one of the existing submaps based on the online sub-reconstruction or initiate a new submap if a good initial stereo reconstruction is found between it and the images of the"*not registered image*" set. As the number of common images between different submaps reaches a threshold, an attempt is made to merge the submaps using the solution described in [43].

4 Experiments

Extensive experiments are conducted to demonstrate the efficacy of the proposed approach. We differentiate the evaluation based on the type of datasets: poor texture datasets and repetitive structure datasets. For the former, we assess the performance of various learning-based feature extraction and matching methods without disambiguation, including both quantitative and qualitative evaluations. For the latter, we compare our LVG-SfM integrating both the SuperPoint+LightGlue combination and Doppelgangers [3] with three methods, SIFT+NearestNeighbouring on-the-fly SfM (vanilla) [43] and on-the-fly SfM with SuperPoint+LightGlue and a SIFT-based disambiguation pipeline [32] (see Sec. 4.3). We set the probability threshold of [3] to 0.8 and the repetitive structures threshold of [32] to 0.25. The difference in the evaluation is motivated by the fact that solutions for repetitive structure datasets, which exhibit rich textures, are less dependent on the descriptor choice, as all extractor-matcher combinations provide sufficient matches. All experiments are run on a machine with i9-12900K CPU and RTX3080 GPU.

Table 2: Datasets used in the experiments.

Dataset	#images	Image Size	Source	Scene	Poor texture	Repetitive structures
1002	37	840*840	[10]	Indoor	✓	✗
1003	36	840*840		Indoor	✓	✗
1008	53	840*840		Indoor	✓	✗
1021	47	840*840		Indoor	✓	✗
1025	56	840*840		Indoor	✓	✗
1026	48	840*840		Indoor	✓	✓
B3	199/342	1968*1312	[32][a]	Outdoor	✗	✓
Indoor	51/152	1200*800		Indoor	✗	✓
ToH	86/339	1310*873		Outdoor	✗	✓
Cup	64	1067*800	[21]	Indoor	✗	✓
ANC	448	various sizes with various image agents	[11]	Outdoor	✗	✓
AdT	434	various sizes with various image agents		Outdoor	✗	✓
BG	175	various sizes with various image agents		Outdoor	✗	✓
CSB	277	various sizes with various image agents		Outdoor	✗	✓
RC	282	various sizes with various image agents		Outdoor	✗	✓
CSWU	354	various sizes with various image agents	Self-captured	Outdoor	✗	✓

[a]For B3, Indoor and ToH, a subset of images was selected to reduce image overlap and make matching more difficult. ANC = Alexander Nevsky Cathedral, AdT = Arc de Triomphe, BG = Brandenburg Gate, CSB = Church on Spilled Blood, RC = Radcliffe camera.

4.1 Datasets and Evaluation Metrics

Datasets. As reported in Tab. 2, 16 public datasets are employed in our experiments. Six datasets (1002 to 1026) [10] are single-object indoor images featuring poor textures. B3, Indoor, ToH, Cup and CSWU are sequentially acquired images while the remaining 5 datasets from [11] are unordered Internet photo collections. These last 10 datasets plus 1026 all depict ambiguous indoor and outdoor scenes characterized by repetitive structures and object symmetry. Sample images of each dataset are shown in Fig. 4 and Fig. 6.

Evaluation Metrics. For the evaluation of extractor-matcher combination in poor texture datasets, five metrics are selected: number of reconstructed 3D points, Mean Track Length (**MTL**), number of reconstructed images (**RIs**), Average Mean Reprojection Error (**AMRE**) after each local bundle adjustment (BA) and Final Mean Reprojection Error (**FMRE**) of the final BA. The first three metrics correspond to the generated correspondences and correctness of the view-graph, i.e., more correspondences could result in more reconstructed 3D points, higher **MTL** and probably more accurate image pose estimated with more 2D-3D matches. The last two are common criteria for assessing the per-

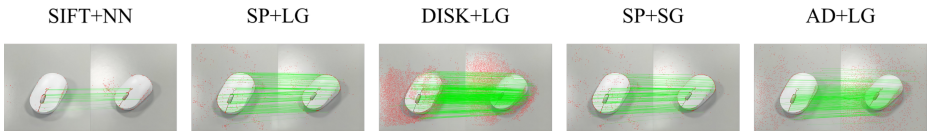


Fig. 3: Visual results of feature matching for a pair with poor textures in dataset 1008.

formance of bundle adjustment. The Disambiguation is evaluated qualitatively by looking at two results: (i) the final view-graph in matrix form, which shows the general distribution of the remaining two-view geometries; (ii) the camera poses and 3D points in the reconstructions, which may highlight blunders in the camera positions or artifacts in the resulting pointcloud.

4.2 Performance of learning-based feature extraction and matching on poor texture

Several learning-based feature extraction and matching methods are compared: SuperPoint + SuperGlue (SP+SG), SuperPoint + LightGlue (SP+LG), ALIKED + LightGlue (AD+LG), and DISK + LightGlue (DK+LG). Additionally, the default setting of on-the-fly SfM using SIFT and Nearest Neighbouring (SIFT+NN) is tested as well. For each image, a maximum of 8000 features are extracted.

According to Tab. 3, the handcrafted SIFT+NN method generally achieves the lowest AMRE and FMRE but with the lowest number of reconstructed points and images, even failing to provide a reconstruction on two datasets with poor textures (1002, 1008). In Fig. 3, a challenging example of matches for an image pair of object 1008 is shown. This can be explained by: first, SIFT is hard to detect keypoint that is salient on DoG (Difference of Gaussian) space for poor texture, resulting in insufficient features and correspondences; second, SIFT detects keypoints with higher 2D-position precision than the learning-based methods that start the detection from low-resolution activations. Thus, the limited number of reconstructed 3D points and higher-precision 2D observations involved in the BA might lead to over-fitting to the assumptions embedded in ray projection mode (internal precision of collinearity equation).

Learning-based feature extraction and matching methods achieve comparable metrics in Tab. 3. DISK+LightGlue outperforms in terms of reconstructed points thanks to its high number of matches as shown in Fig. 3). All learning-based methods are capable of registering nearly all images, producing sparse reconstructions with far more 3D points than SIFT+NN. These methods are indeed superior in extracting a higher amount of features in poor texture scenarios and in providing more reliable matches between pairs.

When looking into the visual results in Fig. 4, SIFT+NN never achieves a correct reconstruction while no clear winner emerges between learning-based methods, showing that, for datasets with poor textures, the final reconstruction is strongly dependent on the descriptor choice. Dataset 1008 achieves the worst

Table 3: The quantitative results on several datasets with poor texture. "-" denotes failure, best is highlighted in bold. MTL = Mean Track Length, AMRE = Average Mean Reprojection Error, FMRE = Final Mean Reprojection Error.

Dataset	Method	#points	RIs	MTL	AMRE	FMRE
1002	SIFT+NN	-	-	-	-	-
	SP+LG	1684	**37/37**	3.43	2.33	0.77
	DK+LG	**10624**	**37/37**	6.94	1.26	0.68
	SP+SG	2437	36/37	3.98	1.36	0.70
	AD+LG	7138	**37/37**	**7.82**	**1.24**	**0.63**
1003	SIFT+NN	250	19/36	4.79	**1.18**	**0.43**
	SP+LG	3266	**36/36**	5.38	1.48	0.92
	DK+LG	**9527**	**36/36**	4.4	1.48	0.83
	SP+SG	3195	**36/36**	**5.62**	1.50	0.93
	AD+LG	4517	**36/36**	5.41	1.52	0.86
1008	SIFT+NN	-	-	-	-	-
	SP+LG	1887	**53/53**	4.42	**1.29**	0.79
	DK+LG	-	-	-	-	-
	SP+SG	1335	**53/53**	**4.88**	1.29	0.78
	AD+LG	-	-	-	-	-
1021	SIFT+NN	878	21/47	3.89	**0.70**	**0.17**
	SP+LG	1636	**47/47**	5.06	1.26	0.70
	DK+LG	**8137**	**47/47**	8.62	1.91	0.82
	SP+SG	1543	**47/47**	5.33	1.23	0.68
	AD+LG	3338	**47/47**	**12.4**	1.30	0.67
1025	SIFT+NN	567	15/56	3.90	**0.84**	**0.24**
	SP+LG	3976	**56/56**	5.43	1.38	0.82
	DK+LG	-	-	-	-	-
	SP+SG	3650	**56/56**	**5.78**	1.42	0.87
	AD+LG	2563	**56/56**	4.65	1.21	0.57

results because of the nature of its reflective and homogeneous texture, which is further complicated by the symmetry in the object.

Real-time performance. Tab. 4 provides the average processing time for each new fly-in image. According to these five datasets, it can be concluded that, although more matches and tie points are involved in the reconstruction processing using the learning-based method, the cost time basically stays the same with just very slight differences that could not affect the real-time performance. [1]

[1] In line with [43], we define "real-time" performance as solving each new image before the next one flies in. Capturing, storage, and transmission typically take 2-3 seconds for each image.

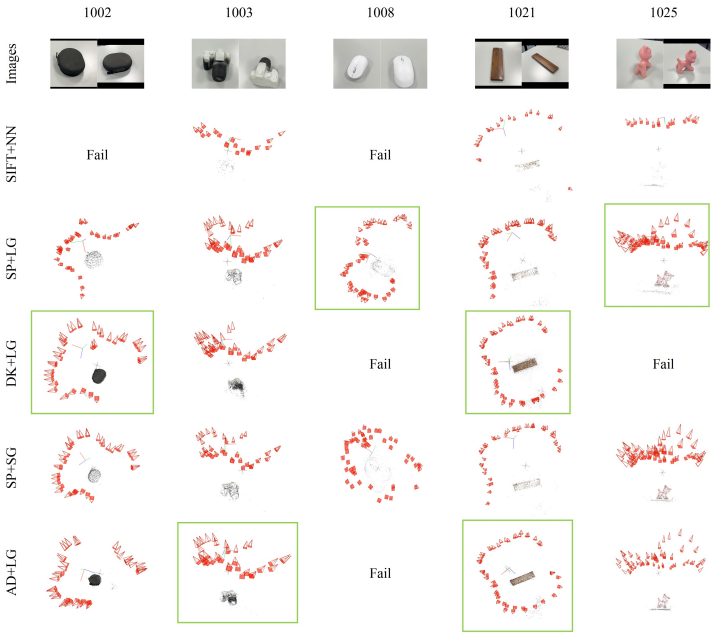


Fig. 4: Reconstruction results on poor texture datasets using various learning-based methods. Correct results are bounded by green boxes.

4.3 Performance of disambiguation on repetitive structures

The two-view geometry elimination method employed in this work is tested on 11 datasets (listed in Tab. 2) containing repetitive structures. Four methods are compared: the vanilla on-the-fly SfM [43], the vanilla on-the-fly SfM [43] with learning-based combination of SP+LG, the presented work (LVG-SfM), and the disambiguation method by Wang *et al.* [32] tested with SIFT feature and matching.

Fig. 5 illustrates the resulting view-graphs in matrix form using the four methods on the 11 datasets. In the vanilla on-the-fly SfM and its enhanced version with SP+LG where only the canonical two-view geometry verification is applied, it results in a dense view-graph that might contain many incorrect edges, in which the enhanced one generates more dense view-graph due to more generated matches shown as Fig. 3. The datasets from [11] yield irregular view-graph than the other six datasets, as these five datasets [11] are crowd-sourced images captured by different tourists in an arbitrary way. The other two methods exhibit much sparser view-graph with outliers removal of two-view geometry resulting from ambiguous texture, and their effect on 3D reconstruction is shown in Fig. 6. In this figure, it can be seen that, the baseline method of vanilla on-the-fly SfM without outlier elimination yields various degrees of artifacts illustrated by the red circles, even for the enhanced version, the reconstruction

Table 4: Average processing time (in seconds) for each newly added image.

Dataset	SIFT+NN	SP+LG	DISK+LG	SP+SG	AD+LG
1002	-	2.95	2.78	2.84	2.92
1003	3.22	2.78	2.81	2.83	2.83
1008	-	3.21	-	3.19	-
1021	3.17	3.09	3.04	3.11	3.06
1025	3.25	3.25	-	3.32	3.25

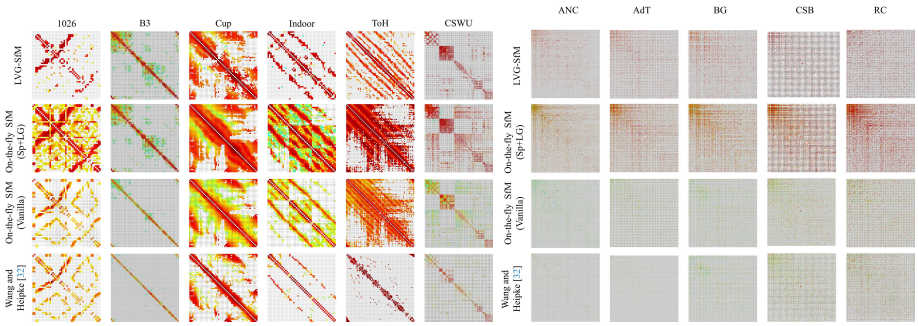


Fig. 5: View-graph results of different methods. The vertical and horizontal axis denote the image ID, whereby the darker the pixel is the more inlier correspondences the corresponding two-view geometry includes. White pixels indicate removed two-view geometries

is not improved by SP+LG as it is achieved in the tests on poor texture (see Tab. 3), this in turn demonstrates that learning-based correspondences fail to handle images with repetitive structures. Nevertheless, after integrating with doppelgangers and [32] that are tailored for classifying image pairs of repetitive structure, the reconstruction performance is obviously boosted. Particularly, for the 1026 dataset which contains both poor and ambiguous textures, LVG-SfM notably outperforms the other three methods. Compared to [32], the proposed method generates many superior results on ToH, Cup, ANC, AdT, BG, CSB, RC and CSWU, and nearly the same results on B3 and Indoor, this is due to that the Doppelgangers [3] used in our work was pre-trained on outdoor datasets, it is hard to completely eliminate the impact of ambiguous textures in the indoor dataset (Indoor) and B3 contains critical two-view configuration of very short baseline that might lead to negative influence for SfM.

In addition, analogous to Tab. 3, the numerical results including the averaging processing time per new image are provided in Tab. 5. In general, our LVG-SfM and the enhanced version reconstruct nearly the same level number of 3D points, both are higher than the SIFT-based methods. For BG, CSB and RC, our LVG-SfM has less 3D points than the enhanced version does, as two independent sub-reconstructions are generated due to the removed edges by dop-

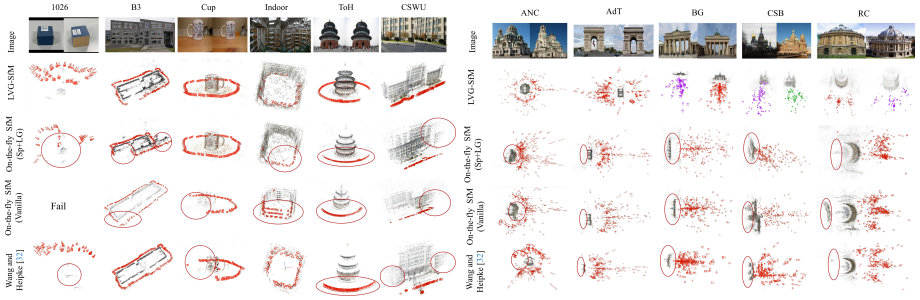


Fig. 6: Reconstruction results of datasets with repetitive structures. Blunders are highlighted in circles. The split reconstructions of BG, CSB, and RC are identified by different colors.

pelgangers and these kind of split reconstruction is consistent with the results shown in [3], which is in accordance with the dataset itself. Comparing the values of RIs, without disambiguation, the enhanced method and vanilla method can generally register more images, whereas the obtained 3D points and camera poses are incorrect as Fig. 6 visualizes. After disambiguation, the proposed LVG-SfM performs better than [32] with respect to reconstructing more images correctly. The MTL is indeed decreased a little by disambiguation of view-graph, but it basically stays in the second best place for our LVG-SfM. In the bundle adjustment, all the four methods achieves final mean reprojection error of subpixel. Although better reconstruction can be achieved in this work, as Tab. 5 shows, comparing LVG-SfM and on-the-fly SfM (SP+LG), the computation involved in the inference of Doppelgangers results in an additional average delay of 0.5 seconds per newly captured image to eliminate outliers in two-view geometry caused by ambiguous textures. This indicates that LVG-SfM is capable of handling images with ambiguous textures, although there is a slight decrease in real-time performance.

5 Conclusions and Future Works

This work focused on generating a robust view-graph via learning-based feature extraction and matching methods to improve the performance of the online SfM pipeline [43], especially in scenes with poor texture or repetitive structure, which generally result in incorrect camera poses and folded or duplicated point clouds. SOTA learning-based feature extraction and matching methods [19] provided a higher amount of matches between pairs in these challenging scenarios, hence generating a more reliable input view-graph which we further robustified by pruning incorrect matches stemming from ambiguous pairs with the learning-based Doppelgangers disambiguation method [3]. As the experimental results show, our LVG-SfM successfully recovers correct camera poses and outputs more reasonable 3D points with on poor texture areas.

Table 5: The quantitative results on several datasets with repetitive texture. "-" denotes failure, best is highlighted in bold.

Datasets	method	#points	RIs	MTL	AMRE	FMRE	Averaging time per Image (in second)
1026	LVG-SfM	1799	41/48	5.64	1.83	0.86	3.54
	On-the-fly SfM (SP+LG)	2162	48/48	5.24	1.23	0.74	3.03
	On-the-fly SfM (Vanilla)	–	–	–	–	–	–
	Wang and Heipke [32]	1244	46/48	4.54	0.57	0.11	3.81
B3	LVG-SfM	161313	199/199	4.94	0.97	0.45	3.64
	On-the-fly SfM (SP+LG)	164848	199/199	5.18	1.03	0.50	3.21
	On-the-fly SfM (Vanilla)	35083	198/199	4.34	0.43	0.06	2.61
	Wang and Heipke [32]	35694	184/199	4.00	0.43	0.17	3.94
Cup	LVG-SfM	23353	64/64	5.59	1.25	0.72	3.41
	On-the-fly SfM (SP+LG)	26126	64/64	5.58	1.27	0.74	2.91
	On-the-fly SfM (Vanilla)	5764	64/64	6.74	0.53	0.02	2.5
	Wang and Heipke [32]	6441	64/64	6.70	0.37	0.17	3.44
Indoor	LVG-SfM	34523	51/51	4.45	1.09	0.59	3.33
	On-the-fly SfM (SP+LG)	33335	51/51	4.89	1.07	0.68	2.94
	On-the-fly SfM (Vanilla)	7322	51/51	3.79	0.50	0.06	2.44
	Wang and Heipke [32]	8385	51/51	3.59	0.36	0.15	3.74
ToH	LVG-SfM	61280	86/86	7.63	1.73	0.74	3.55
	On-the-fly SfM (SP+LG)	40506	86/86	13.09	1.32	0.99	3.15
	On-the-fly SfM (Vanilla)	34997	86/86	5.06	0.87	0.04	3.12
	Wang and Heipke [32]	46542	71/86	2.10	0.64	0.11	4.02
CSWU	LVG-SfM	164110	354/354	14.38	2.10	0.99	3.74
	On-the-fly SfM (SP+LG)	155247	354/354	14.50	1.46	0.97	3.11
	On-the-fly SfM (Vanilla)	80413	345/354	8.40	1.05	0.04	2.73
	Wang and Heipke [32]	13473	76/354	7.10	0.81	0.21	4.22
ANC	LVG-SfM	171206	438/448	7.25	1.40	0.10	3.55
	On-the-fly SfM (SP+LG)	133529	448/448	8.08	1.43	0.37	2.84
	On-the-fly SfM (Vanilla)	77554	444/448	4.82	1.09	0.07	2.66
	Wang and Heipke [32]	38724	414/448	4.24	0.56	0.03	3.99
AdT	LVG-SfM	134234	347/389	7.31	1.41	0.74	3.44
	On-the-fly SfM (SP+LG)	121629	387/389	8.93	1.54	0.12	2.86
	On-the-fly SfM (Vanilla)	27776	213/389	7.19	1.08	0.05	2.23
	Wang and Heipke [32]	27777	327/389	5.81	0.65	0.16	3.77
BG	LVG-SfM	26978/60430	(67,185)/297	5.92/7.71	1.31/1.40	0.26/0.25	3.61
	On-the-fly SfM (SP+LG)	76680	296/297	8.54	1.47	0.13	3.01
	On-the-fly SfM (Vanilla)	23144	268/297	5.30	0.95	0.21	2.31
	Wang and Heipke [32]	11354	262/297	4.31	0.57	0.09	4.12
CSB	LVG-SfM	34732/60528	(94,154)/277	9.18/9.22	1.43/1.47	0.13/1.72	3.57
	On-the-fly SfM (SP+LG)	95961	276/277	9.80	1.54	1.14	3.03
	On-the-fly SfM (Vanilla)	44167	268/277	5.88	1.09	0.21	2.36
	Wang and Heipke [32]	33856	268/277	6.42	0.63	0.15	3.88
RC	LVG-SfM	42161/22713	(186,90)/282	11.95/8.63	1.44/1.39	0.92/0.07	3.42
	On-the-fly SfM (SP+LG)	54335	282/282	13.95	1.53	1.07	2.74
	On-the-fly SfM (Vanilla)	47695	277/282	7.10	1.09	0.02	2.35
	Wang and Heipke [32]	43866	275/282	6.62	0.67	0.04	4.23

In the future, we plan to further expand the LVG-SfM pipeline in two directions: first, exploring and integrating more learning-based matching and outlier removal methods; second, integrating learning-based methods for real-time dense point cloud and surface mesh generation.

Acknowledgement. This work was jointly supported Natural Science Foundation of Hubei Province, China (2022CFB727), National Natural Science Foundation of China (42301507) and ISPRS Scientific Initiatives 2023.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)
2. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. pp. 404–417. Springer (2006)
3. Cai, R., Tung, J., Wang, Q., Averbuch-Elor, H., Hariharan, B., Snavely, N.: Doppelgangers: Learning to disambiguate images of similar structures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 34–44 (2023)
4. Ceylan, D., Mitra, N.J., Zheng, Y., Pauly, M.: Coupled structure-from-motion and 3d symmetry detection for urban facades. ACM Transactions on Graphics (TOG) **33**(1), 1–15 (2014)
5. Cohen, A., Zach, C., Sinha, S.N., Pollefeys, M.: Discovering and exploiting 3d symmetries in structure from motion. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1514–1521. IEEE (2012)
6. Cui, Z., Tan, P.: Global structure-from-motion by similarity averaging. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 864–872 (2015)
7. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
8. Edstedt, J., Bökman, G., Wadenbäck, M., Felsberg, M.: Dedode: Detect, don't describe—describe, don't detect for local feature matching. In: 2024 International Conference on 3D Vision (3DV). pp. 148–157. IEEE (2024)
9. Farenzena, M., Fusiello, A., Gherardi, R.: Structure-and-motion pipeline on a hierarchical cluster tree. In: 2009 IEEE 12th international conference on computer vision workshops, ICCV workshops. pp. 1489–1496. IEEE (2009)
10. He, X., Sun, J., Wang, Y., Peng, S., Huang, Q., Bao, H., Zhou, X.: Detector-free structure from motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21594–21603 (June 2024)
11. Heinly, J., Dunn, E., Frahm, J.M.: Correcting for duplicate scene structure in sparse 3d reconstruction. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. pp. 780–795. Springer (2014)
12. Hoppe, C., Klopschitz, M., Rumpler, M., Wendel, A., Kluckner, S., Bischof, H., Reitmayr, G.: Online feedback for structure-from-motion image acquisition. In: BMVC. vol. 2, p. 6 (2012)
13. Hou, Q., Xia, R., Zhang, J., Feng, Y., Zhan, Z., Wang, X.: Learning visual overlapping image pairs for sfm via cnn fine-tuning with photogrammetric geometry information. International Journal of Applied Earth Observation and Geoinformation **116**, 103162 (2023)
14. Jiang, N., Cui, Z., Tan, P.: A global linear method for camera pose registration. In: Proceedings of the IEEE international conference on computer vision. pp. 481–488 (2013)
15. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17627–17638 (2023)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)

17. Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE transactions on pattern analysis and machine intelligence **42**(4), 824–836 (2018)
18. Mayer, H.: Efficient hierarchical triplet merging for camera pose estimation. In: German Conference on Pattern Recognition. pp. 399–409. Springer (2014)
19. Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., Remondino, F.: Deep-image-matching: a toolbox for multiview image matching of complex scenarios. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **48**, 309–316 (2024)
20. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: Openmvg: Open multiple view geometry. In: Reproducible Research in Pattern Recognition: First International Workshop, RRPR 2016, Cancún, Mexico, December 4, 2016, Revised Selected Papers 1. pp. 60–74. Springer (2017)
21. Roberts, R., Sinha, S.N., Szeliski, R., Steedly, D.: Structure from motion for scenes with large duplicate structures. In: CVPR 2011. pp. 3137–3144. IEEE (2011)
22. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011)
23. Rupnik, E., Daakir, M., Pierrot Deseilligny, M.: Micmac–a free, open-source solution for photogrammetry. Open geospatial data, software and standards **2**, 1–9 (2017)
24. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
25. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
26. Shah, R., Chari, V., Narayanan, P.: View-graph selection framework for sfm. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 535–550 (2018)
27. Snavely, N., Seitz, S.M., Szeliski, R.: Skeletal graphs for efficient structure from motion. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
28. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021)
29. Sweeney, C., Hollerer, T., Turk, M.: Theia: A fast and scalable structure-from-motion library. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 693–696 (2015)
30. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems **33**, 14254–14265 (2020)
31. Wang, L., Ge, L., Luo, S., Yan, Z., Cui, Z., Feng, J.: Tc-sfm: Robust track-community-based structure-from-motion. IEEE Transactions on Image Processing **33**, 1534–1548 (2024)
32. Wang, X., Heipke, C.: An improved method of refining relative orientation in global structure from motion with a focus on repetitive structure and very short baselines. Photogrammetric Engineering & Remote Sensing **86**(5), 299–315 (2020)
33. Wang, X., Rottensteiner, F., Heipke, C.: Robust image orientation based on relative rotations and tie points. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; IV-2 **4**(2), 295–302 (2018)

34. Wang, X., Xiao, T., Gruber, M., Heipke, C.: Robustifying relative orientations with respect to repetitive structures and very short baselines for global sfm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
35. Wilson, K., Snavely, N.: Network principles for sfm: Disambiguating repeated structures with local context. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 513–520 (2013)
36. Wilson, K., Snavely, N.: Robust global translations with 1dsfm. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13. pp. 61–75. Springer (2014)
37. Wu, C.: Visualsfm: A visual structure from motion system. http://www. cs. washington. edu/homes/ccwu/vsfm (2011)
38. Wu, C.: Towards linear-time incremental structure from motion. In: 2013 International Conference on 3D Vision-3DV 2013. pp. 127–134. IEEE (2013)
39. Yan, S., Zhang, M., Lai, S., Liu, Y., Peng, Y.: Image retrieval for structure-from-motion via graph convolutional network. Information Sciences **573**, 20–36 (2021)
40. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 467–483. Springer (2016)
41. Zach, C., Irschara, A., Bischof, H.: What can missing correspondences tell us about 3d structure and motion? In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
42. Zhan, Z., Xia, R., Yu, Y., Xu, Y., Wang, X.: On-the-fly sfm: What you capture is what you get. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **X-1-2024**, 297–304 (2024). https://doi.org/10.5194/isprs-annals-X-1-2024-297-2024, https://isprs-annals.copernicus.org/articles/X-1-2024/297/2024/
43. Zhan, Z., Yu, Y., Xia, R., Gan, W., Xie, H., Giulio, P., Luca, M., Fabio, R., Xin, W.: Sfm on-the-fly: Get better 3d from what you capture (2024), https://arxiv.org/abs/2407.03939
44. Zhao, X., Wu, X., Chen, W., Chen, P.C., Xu, Q., Li, Z.: Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. IEEE Transactions on Instrumentation and Measurement **72**, 1–16 (2023)
45. Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P.C., Li, Z.: Alike: Accurate and lightweight keypoint detection and descriptor extraction. IEEE Transactions on Multimedia **25**, 3101–3112 (2022)
46. Zhao, Y., Chen, L., Zhang, X., Xu, S., Bu, S., Jiang, H., Han, P., Li, K., Wan, G.: Rtsfm: Real-time structure from motion for mosaicing and dsm mapping of sequential aerial images with low overlap. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2021)