A Survey on Learning-based Structure from Motion

David Yao
dyyao2@illinois.edu

Jing Wen
jw116@illinois.edu

Abstract

Structure-from-motion (SfM) is one of the fundamental tasks in 3D vision. It jointly estimates camera poses and structures in the form of dense depths or 3D points from RGB images. This task has largely advanced in recent years by introducing neural networks and large-scale training. In this survey, we study the recent progress in learning-based structure-from-motion. We summarize the core problems and their solutions, current results, and discuss the milestones, challenges, and potential research directions.

1. Introduction

Structure-from-motion (SfM) is one of the core problems in 3D vision, which has been studied for decades. It is a fundamental building block for many downstream tasks, e.g. AR/VR applications and robotics. As depicted in Figure 1, SfM aims to estimate the structure of the scene and the camera poses jointly given a set of unordered RGB or RGB-D images in an offline manner . Note that SfM is highly relevant to visual simultaneous localization and mapping (Visual SLAM) which also outputs camera poses and the scene structure. Differently, visual SLAM assumes the inputs are sequential and often operate in a real-time scenario. In this survey, we mainly focus on the strict SfM setting but do not strictly distinguish the two tasks. We slightly 'abuse' the term SfM to refer to the joint estimation of structures and camera poses from any visual inputs in this survey.

The conventional SfM pipeline begins with feature matching, which finds correspondences between images. Then, given the 3D correspondences, relative camera poses and 3D points are estimated in a two-view setting using RANSAC, triangulation, etc. At last, bundle adjustment is adopted to refine both the structure and the camera poses of *all* views. All the components in this pipeline have been well-studied and improved over the years. However, this Conventional SfM still faces the following issues: 1) It requires densely sampled views that have significant overlaps between each other. 2) It fails in textureless or non-Lambertian surfaces due to missing correspondences. 3) It requires large camera translations for proper triangulation.

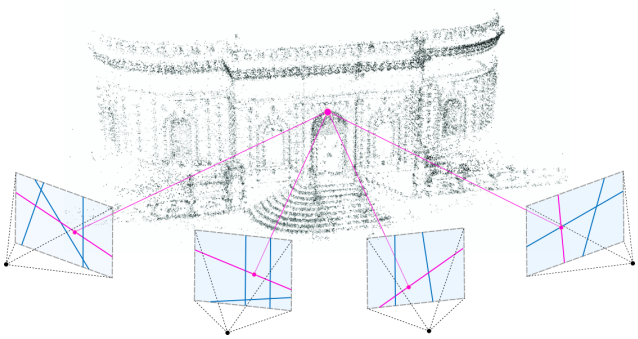

Figure 1. Structure-from-motion outputs camera poses and 3D structure using correspondences found in multi-view images of overlapping scene regions

To address the above issues, researchers have turned to learning-based approaches that benefit from large-scale training to replace part of or the entire pipeline. In feature matching, neural deep features are explored to replace the hand-crafted features [17–19, 24, 29]. Predictions from other computer vision tasks such as tracking and optical flow, are also explored to directly produce the correspondences [4, 36]. Given the 2D correspondences, learning-based SfM [2, 11, 22, 25, 28, 32, 35] is proposed as an alternative to bundle adjustment, the optimization problem that minimizes the projection error.

Some works aim to solve half of structure-from-motion problem with learning, and optimize for the other component. RelPose [33] and its followups [13, 34] as well as PoseDiffusion [27] and BARF [14] predict only the motions, i.e. camera poses, while DUSt3R [31] and MASt3R [12] focuses on the structure (the scene geometry). We summarize the methods in Sec. 2.

We further discuss the evaluation results in Sec. 3. We list the frequently used datasets and evaluation metrics to measure the performance. We then perform summarize evaluation results from representative works. We additionally answer two questions: To what extent the learning-based SfM has improved from conventional SfM? What are the scenarios whereun learning-based SfM outperforms/underperforms conventional SfM?

In Sec. 4, we conclude this survey by discussing the cur-

rent limitations and highlighting potential future directions.

2. Methods

Recent works introduce learning-based methods to replace the components such as correspondences and bundle adjustment in the Conventional SfM pipeline. Other approaches solve either the 'motion' or the 'structure' half of the structure-from-motion problem. In this section, we categorize the methods into neural feature matching, learning-based bundle adjustment, learning-based camera pose estimation, and structure estimation.

2.1. Neural correspondences

As the very first step, robust 2D correspondences in the image space are essential for the success of structure-from-motion. Conventional SfM performs matching by naively comparing hand-designed herusitics based descriptors such as SIFT. SuperGlue [19] is one of the earliest papers that learns neural descriptors. It encodes CNN features with an attentional graph neural network to acquire the matching descriptors, and uses the Optimal Matching layer together with the Sinkhorn algorithm to find partial correspondences. Later works [17, 18, 24, 29] extend to predicting dense correspondence directly, i.e., identifying matching in image space directly without producing intermediate descriptors. Among them, LoFTR [24] relies on the coarse-to-fine module and Transformer to enlarge the receptive field for matching.

Correspondences can also come from other existing deep learning tasks. For example, optical flow [26] and point tracking [4, 6–8, 30] that track points between two consecutive frames or through video sequences respectively, inherently provide correspondences. ParticleSfM [36] demonstrates that dense point trajectories can be used in SfM. It predicts the optical flow with RAFT [26] and chains the flow into longer sequences. VGGSfM [28] changes the two-frame flow prediction to long-term point trajectories, considering the rich temporal context. To tackle SfM in dynamic environments, LEAP-VO [4] jointly improves matching and static/dynamic classification into a single framework, estimating correspondences and filtering for dynamics.

2.2. Learning-based structure-from-motion

Given the 2D correspondences, the next steps are predicting camera poses and structures. Conventional SfM first initializes the 3D points and camera poses from two-view geometry. Then it performs bundle adjustment to refine structures and motions jointly. Bundle adjustment is optimized using Levenberg-Marquardt (LM) algorithm. As one of the earliest works that integrates bundle adjustment into the deep learning pipeline, BA-Net [25] modifies the LM algorithm so that it is differentiable, achieving an end-to-end trainable network. DeepSfM [32] no longer requires the LM algorithm. Instead, it updates iteratively between camera poses and depths, utilizing cost volumes that introduce inductive biases of photo consistency and geometric consistency. Later works [2, 22, 28] dismiss the need for good initialization of camera poses and 3D points. For example, GASfM [2] directly predicts the camera poses and depths from 2D correspondences with the Graph Attention Network. VGGSfM [28] adds a second stage that refines the predictions with differentiable bundle adjustment but the two stages are end-to-end trainable. FlowMap [22] is supervised only with optical flow and point trajectories, choosing SGD as optimization algorithm. It is per-scene optimized and thus can work on more input frames than feed-forward approaches.

We usually assume that the scene is static when performing structure-from-motion. However, the assumption does not always hold. Conventional SfM fails when there are object motions in the scene. Kopf et al. [11] and Zhang et al. [35] develop in monocular videos that have dynamic objects, yet another scenario in which learning-based SfM exceeds the conventional SfM.

2.3. Camera pose estimation

Now we discuss the works that only focus on half of the structure-from-motion problem, either 'motion' or 'structure'. The two parts are highly coupled. If one of them is solved, the other one can be easily tackled as well. We start from the motion half, in other words, data-driven camera pose estimation from a set of images. Structure can be easily obtained using multi-view stereo methods like [21] An example is PoseDiffusion [27] which proposes the diffusion-based bundle adjustment given images and 2D correspondences. Differently, BARF [14] refines the camera poses by jointly optimizing a NeRF [15]. Most of the works in structure-from-motion predict camera motions from 2D correspondences. Therefore, it requires densely sampled images where there are large overlaps between adjacent images. RelPose [33] and RelPose++ [13] address the task that directly predicts the camera poses from sparsely sampled images. To capture the ambiguity in this ill-posed task, it models the poses with an energy-based model. Recently, this line of work has been further improved by RayDiffusion [34] which represents the camera pose as a bundle of rays and uses a diffusion model to generate the rays conditioned on input images.

2.4. Structure estimation

In this section, we review the approaches that estimate structure in the form of depths or 3D points. With rapid developments in depth estimation models, many recent SfM works try to leverage these foundation models to obtain

more robust camera pose results even in ill-posed camera movements (eg. pure rotations). Kopf et al. [10] finetunes a depth estimation model along with camera poses to obtain accurate poses in settings where traditional SfM [20] fails to converge. Instead of depth estimation, some methods have chosen to directly estimate 3D locations of points in a *shared* coordinate system by multiple cameras. Note that this setting is different from traditional depth estimation in which depths are predicted in each camera's coordinates. Suppose that the predicted 3D points are in shared coordinates, and the 3D to 2D correspondences are known, the camera poses can be trivially acquired by simply minimizing the projection error. So we categorize this line of work we call structure estimation as part of structure-from-motion. The recently proposed method DUSt3R [31] belongs to this category. DUSt3R takes a pair of images as inputs and estimates the dense depths of both images in the same coordinate system. The follow-up work MASt3R [12] extends DUSt3R for better precision and feature matching. ACE0 [1] tackles SfM through a scene coordinate regression model, alternating between training the scene coordinate regression model with existing predicted poses and registering additional image poses using the scene coordinate regression model.

3. Evaluations

3.1. Datasets

Given that SfM has been a long-standing problem, and obtaining pose and depth is often a prerequisite to many downstream 3D problems, there are many datasets available. To evaluate SfM outputs, i.e. pose and structure, the dataset needs to have ground truth pose and depth. For synthetic datasets [3, 23], these can be directly obtained from software. For real-world datasets [9], camera pose can be obtained using Different metrics and evaluation techniques have also been proposed to estimate them when they are not available. Some of the most popular datasets are listed here.

Tanks and Temples [9] contains Real-world video sequences that cover both outdoor and indoor scenes. Ground-truth camera poses are obtained through positioning the camera via a precisely controlled robotic arm. The dense ground-truth structures are obtained using a high-end laser scanner. However, the dataset only contains around 20 scenes.

CO3D [16] provides 360 object-oriented image sequences. Contains 6 million frames from 37k videos, covering 61 MS-COCO categories. Ideal for deep-learning purposes due to the dataset size. Groundtruth pose is provided in the form of COLMAP estimates which is only an estimate of ground truth camera pose.

MPI Sintel [3] is a synthetic dataset produced in the Blender. This dataset contains 23 outdoor scenes containing highly dynamic elements. This is ideal for evaluating pose estimation in highly dynamic environments.

Scannet [5] is a large-scale RGB-D video dataset that contains 2.5 million views in more than 1500 real-world scans. Depth is obtained using depth sensors and camera pose is obtained using on-device sensors. Even though depth and pose are not as accurate as synthetic datasets, the size of this dataset allows for large-scale training.

3.2. Metrics

3.2.1 Camere poses

Pose estimation is often evaluated directly when ground truth is available. Translation and rotation components are evaluated separately. Sfm generates estimated camera pose trajectories up to a scale difference, and in a different coordinate system. The evaluation pipeline often involves first aligning the camera poses using umeyama algorithm [4, 35, 36], which finds the best scale and rotation that minimizes the mean squared error of corresponding camera centers.

Average Translation Error (ATE). This computes the translation difference between predicted and ground truth camera poses.

Relative Pose Estimation (RPE-t). This computes the relative translation error. It measures the difference in translation between temporally adjacent frames. This can provide insights into the smoothness of the trajectory.

Relative Pose Estimation (RPE-r). Similar to RPE-t but computes the relative rotational error.

By thresholding translation/rotation errors, we can also quantify pose errors using accuracy [34] and area under the curve (AUC) [28]. This can be useful as pose estimation is often the input to further downstream tasks which require only a rough estimate of pose as small deviations can be finetuned later on.

For datasets without ground truth camera poses, camera pose evaluation can be done implicitly through measuring novel-view synthesis results [22]. A better pose will lead to better renderings.

3.2.2 Structures

Measuring structure is often performed by measuring depth. Similarly to pose, an ideal shift and scale is estimated that minimizes the least squared error between estimated and ground-truth depth maps to resolve scale ambiguities [31].

Absolute Relative Depth. Depth error is usually estimated by finding the relative depth error. This scales the error relative to the ground-truth depth, ensuring that error estimates are not biased towards distant regions.

Threshold Accuracy $\delta_{1.25}$. Measures the accuracy of predictions that are within 1.25 ratio of the ground-truth depth, i.e., $\max(d_{gt}/d_{pred}, d_{pred}/d_{gt}) < 1.25$.

	Conventional	Learning-based			
Method	COLMAP (SP + SG) [20]	PoseReg [27]	RelPose++ [13]	PoseDiffusion [27]	VGGSfM [28]
RRE@15°	31.6	53.2	82.3	80.5	**92.1**
RTE@15°	27.3	49.1	77.2	79.8	**88.3**
AUC@30°	25.3	45.0	65.1	66.5	**74.0**

Table 1. Camera Pose Estimation on Co3D results taken from VGGSfM [28]. VGGSfM performs much better than neural feature enhance COLMAP and other deep learning approaches. Co3D frames have wide baselines traditional methods like COLMAP perform poorly.

3.3. Results and Comparisons

Due to the slight differences in the experimental settings in each paper, fair comparisons among them are not available. We mainly focus on the works that are performed in the same setting as COLMAP [20], the most popular convention SfM baseline. We include evaluation results for representative work [28] in Tab. 1. Other quantitative results can be found in respective papers.

Camera pose estimation. In few-view setting, VGGSfM [28] and RayDiffusion [34] shows superior performance results on CO3D [16]. COLMAP fails to run in extreme few-view situations as it requires a minimum set of correspondence points to initialize and register additional camera poses. Even with neural features, VGGSfM still shows superior results as seen from Tab. 1.

In challenging dynamic scenes, RCVD [10] and LEAP-VO [4] perform the best and fail gracefully on highly dynamic dataset MPI Sintel [3]. RCVD [10] relies on dense depth prediction and optical flow to obtain sufficient correspondence in static regions for optimization to proceed. LEAP-VO [4] learns static/dynamic classification and tracklets that provide correspondence across video. COLMAP was only able to run on 11 out of the 23 videos. However, LEAP-VO predicts tracklets in a sliding window fashion and is susceptible to drifting. Additionally, it is trained and evaluated only on synthetic datasets. Its performance does not generalize as well to real-world applications. RCVD finetunes a monocular depth model on each video, taking up to 20 minutes per video whereas COLMAP takes seconds.

Geometry/depth estimation. Deep learning techniques that leverage depth estimation models directly output a complete estimation result, whereas conventional SfM only provide partial/sparse reconstructions. DUSt3R [31] achieves more accurate results, especially in scenes with ambiguous textures or low-texture regions. FlowMap [22] follows a similar approach as RCVD [10] by finetuning a monocular depth estimation model into the SfM pipeline, outputting complete dense depth.

4. Conclusions

Learning-based structure-from-motion takes advantage of data priors from large-scale datasets and deep learning to address the remaining issues in the conventional SfM pipeline, such as COLMAP: To solve the failures in textureless regions and non-Lambertian surfaces, learning-based correspondences are adopted to replace the hand-craft features. To mitigate the requirements of densely sampling views, generative models are proposed which directly take as inputs the sparse views and output the camera poses. Other research directions such as SfM in dynamic scenes are also advanced by learning-based SfM.

4.1. Limitations

Bad generalization ability. Common to all deep learning techniques, a large and well distributed training dataset is required to obtain generalizable results. Many learning-based methods are proven to work well in specific conditions and on specific datasets, but their results may not translate as well to real-world applications.

Huge computational cost. Due to the use of deep learning, learning-based SfM consists of deep neural networks that require a large amount of GPU computation to train and inference, The heavy computation cost leads to difficulties when used in compute-constraint environments. Therefore, COLMAP remains the defacto SfM technique used to preprocess any downstream computer vision tasks.

4.2. Future directions

We propose two potential future directions addressing the limitations accordingly. First, a larger curated training dataset will help the model capture a better distribution and therefore generalize better to in-the-wild data. With the growing interest and increasing datasets available, many learning-based works such as DUSt3R [31] and ACE0 [1] have made the first attempts to combine available datasets which greatly reduced the performance gap with conventional SfM. Second, there is a great need to improve computational efficiency. Traditional SfM approaches like COLMAP can successfully do so on CPU in a reasonable time. With deep learning techniques, processing thousands of images simultaneously on the GPU is still impossible. Efficient architectures can be adopted as a solution.

References

[1] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 3, 4

[2] Lucas Brynte, José Pedro Iglesias, Carl Olsson, and Fredrik Kahl. Learning structure-from-motion with graph attention networks. In *CVPR*, 2024. 1, 2

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012. 3, 4

[4] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *CVPR*, 2024. 1, 2, 3, 4

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[6] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-vid: A benchmark for tracking any point in a video. *NeurIPS*, 2022. 2

[7] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *ICCV*, 2023.

[8] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *ECCV*, 2022. 2

[9] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 3

[10] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3, 4

[11] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 1, 2

[12] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv*, 2024. 1, 3

[13] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv*, 2023. 1, 2, 4

[14] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 1, 2

[15] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 2020. 2

[16] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 3, 4

[17] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *NeurIPS*, 2018. 1, 2

[18] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *ECCV*, 2020. 2

[19] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 2

[20] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3, 4

[21] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[22] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv*, 2024. 1, 2, 3, 4

[23] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[24] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 1, 2

[25] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *arXiv*, 2018. 1, 2

[26] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2

[27] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 1, 2, 4

[28] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *CVPR*, 2024. 1, 2, 3, 4

[29] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *ACCV*, 2022. 1, 2

[30] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *ICCV*, 2023. 2

[31] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 3, 4

[32] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *ECCV*, 2020. 1, 2

[33] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 1, 2

[34] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *ICLR*, 2024. 1, 2, 3, 4

[35] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *ECCV*, 2022. 1, 2, 3

[36] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *ECCV*, 2022. 1, 2, 3