

Reliable 3D reconstruction from long-sequence in-the-wild videos

Shuo Sun,¹ Torsten Sattler,² Malcolm Mielle,³ Achim J. Lilienthal,^{1,4} Martin Magnusson¹

Abstract—Casually recorded videos on the Internet can be a great source for many downstream tasks, such as view synthesis, controllable video generation, and photo-realistic scene reconstruction, if we can successfully achieve 3D reconstruction. However, reliable camera pose estimation and scene reconstruction from such unconstrained videos remains an open challenge. Existing visual-only SLAM/SfM methods perform well on benchmark datasets but struggle with real-world footage, which often exhibits uncontrolled motion including rapid rotations and pure forward movements, textureless regions, and dynamic objects. We analyse the limitations of current methods and introduce a robust pipeline designed to improve 3D reconstruction from casual videos. We build upon recent deep visual odometry methods but increase robustness in several ways. Camera intrinsics are automatically recovered from the first few frames using structure-from-motion. Dynamic objects and less-constrained areas are masked with a predictive model. Additionally, we leverage monocular depth estimates to regularize bundle adjustment, mitigating errors in low-parallax situations. Finally, we integrate place recognition and loop closure to reduce long-term drift and refine both intrinsics and pose estimates through global bundle adjustment. We demonstrate large-scale contiguous 3D models from several online videos in various environments. In contrast, baseline methods typically produce locally inconsistent results at several points, producing separate segments or distorted maps. In lieu of ground-truth pose data, we evaluate map consistency, execution time and visual accuracy of re-rendered NeRF models. Our proposed system establishes a new baseline for visual reconstruction from casual uncontrolled videos found online, demonstrating more consistent reconstructions over longer sequences of in-the-wild videos than previously achieved.

I. INTRODUCTION

It is widely recognized that synthetic data plays a crucial role in advancing embodied AI. Photo-realistic scene construction/generation and controllable video generation can be used to train robots in planning and navigation. However, this raises an important question: before we can generate synthetic data using these models, how do we train the models themselves? Similar to how large language models are trained on text data from the Internet, casually recorded online videos—such as those on YouTube—could serve as a valuable source, provided that we can perform reliable 3D reconstruction on them.

However, reliable 3D scene reconstruction and camera pose estimation from such in-the-wild videos is an open research challenge. Existing methods for visual-only SLAM

(simultaneous localization and mapping) and SfM (structure-from-motion) [1, 2, 3, 4, 5] work well on benchmark datasets (typically mostly stationary scenes with large camera baselines) but are computationally expensive and struggle in uncontrolled real-world settings, particularly in the presence of large camera rotations, textureless environments, and dynamic objects. More specifically—as we will show in our experiments and analysis (Section IV)—current visual SLAM and SfM methods often fail and generate multiple separate trajectories due to small-parallax motion during the recording and insufficient reliable correspondence estimations across frames.

In this work, we analyze the limitations of current methods and introduce a robust pipeline designed to handle these challenges, and push the boundaries of robust 3D scene reconstruction from unconstrained video data. We primarily focus on robustness, i.e., we want to generate one consistent trajectory. In order to work with in-the-wild videos shot with unknown cameras, we first initialize the reconstruction process by estimating camera intrinsics from the first few frames using structure-from-motion. In Section IV-D we present an evaluation of several recent methods for recovering focal length and show our method can get an accurate focal length for reconstruction. At the core of our pipeline is the deep visual odometry method DPVO [6], chosen because of its reliable correspondence estimation and efficiency. Instead of keypoint detection and feature matching, DPVO estimates frame-to-frame correspondence by evaluating deep optical flow, which can handle texture-poor regions well. To handle dynamic objects and unconstrained regions like the sky, we mask them out with a predictive model [7]. Unlike traditional methods that rely solely on 2D correspondences—which struggle with large rotations—we use mono-depth estimates [8] to regularize bundle adjustment (BA), making the pipeline much less sensitive to small-parallax situations, e.g., when the camera is far from the scene and rotation-only motion. Finally, we incorporate place recognition and loop closure to reduce long-term drift. We run a final refinement to refine camera intrinsics, to ensure high-quality results.

Our proposed system establishes a new baseline for visual reconstruction in challenging real-world settings, outperforming existing approaches in scenarios with extreme motion, dynamic elements, and sparse loop closures. In particular, we demonstrate consistent reconstruction over longer sequences from in-the-wild videos than existing methods can produce. These reconstruction results may be used for several downstream tasks, such as visual localization, novel view synthesis, and 3D scene understanding.

In summary, the contributions of this paper are:

¹ AASS research center, Örebro University, Sweden; ² Czech Technical University in Prague; ³ Independent researcher; ⁴ Technical University of Munich, Chair: Perception for Intelligent Systems. This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101017274 (DARKO) and the Czech Science Foundation under EXPRO grant UNI-3D (grant no. 23-07973X).

- Targeting the characteristics of uncalibrated settings, the presence of dynamic objects, small baselines(such as pure rotation), and long distances in the wild videos, we propose a robust visual SLAM system including quick calibration, dynamic object removal, depth-guided BA and pose graph optimization.
- Compared to current SOTA SfM methods [2, 3], our method achieves more robust results, generating smooth and continuous trajectories from 15-minute videos.
- We propose new metrics to evaluate the robustness and accuracy for in-the-wild video reconstruction, which can work as a baseline for future work.

II. RELATED WORK

There are many works on 3D reconstruction from images, but the most popular ones share similar pipelines: 1) estimating image correspondence, and 2) optimizing camera poses and scene geometry. According to the relevance to the paper, we divide the existing work based on whether prior knowledge (specifically, scene geometry estimation) is involved. We also discuss potential downstream applications of our methods in Section II-C.

A. Without explicit geometry estimation

Traditional vSLAM [1] and SfM [2] methods mainly rely on multiple view tracks to optimize camera poses and geometry. COLMAP [2] is one prominent example: COLMAP begins with image correspondence estimation by detecting and matching keypoints between images. After the scene initialization, COLMAP incrementally registers and triangulates new images until no image is successfully registered. Visual SLAM can be regarded as an online version with sequential inputs. However, traditional visual reconstruction methods are known to suffer from degeneracy in the presence of textureless regions or low-parallax motion.

B. With explicit geometry estimation

Due to the inherent ambiguity in the BA optimization, some prior methods seek to incorporate additional geometry estimates in the BA process. For example, *StudioSfM* [9] gets the depth image from an extra depth estimator and proposes to add depth regularization terms in the triangulation of TV show scenes with camera movement. StudioSfm builds upon the incremental SfM COLMAP, but still requires known intrinsics. The depth is used when registering new images; in our method, we fuse the prior depth in the BA stage to avoid instability in optimization. MegaSAM [10], building on DROID-SLAM [11], also uses prior depth during the BA optimization. However, its dense optical flow prediction is not scalable to large-scale scenes.

The recent work Dust3R [12] / Mast3R [13] and follow-up works [14, 5, 15] can also be regarded as using additional geometry estimation. The Dust3R model predicts the 3D point map directly. When aligning multiple frames together, 3D-3D matching is conducted first and followed by 2D-3D refinement. MAST3R-SLAM [15] utilizes Mast3R model prediction directly, tracking new camera frames by aligning

point-maps in 3D space. The dense prediction and multiple frame alignment consume a lot of GPU memory, and cannot be applied to large-scale scenes. In this work, we focus on long-sequence videos which often consist of thousands of images. To our knowledge, none of the methods before tried to directly reconstruct from long in-the-wild online videos.

C. Downstream applications

Successfully recovering 3D structure and camera poses can be applied to multiple downstream applications, such as feed-forward view synthesis [16, 17], controllable video generation [18, 19], photo-realistic scene construction [20, 21], and embodied AI [22, 23, 24]. Currently, most datasets [16] only support short sequences (less than 10 seconds) of video clips; our method can work as a tool to process long-sequence casual videos.

III. METHOD

As discussed above, current state-of-the-art SLAM/SfM methods often break trajectories into multiple segments when faced with challenging conditions, such as fast rotations commonly found in in-the-wild video sequences. This paper addresses these limitations by proposing a novel method capable of computing a single cohesive trajectory even under demanding conditions present in casual uncontrolled videos.

Our method is overviewed in Fig. 1. We first present our initialization strategy in Section III-A. In Section III-B, we describe the main reconstruction pipeline, adopting a deep learning-based optical flow estimation to determine image correspondences within a temporal sliding window, as well as introducing prior depth estimation as an extra regularization term in order to effectively reduce drift errors when the baseline between frames is small. Furthermore (Section III-C), to reduce the drift accumulated during reconstruction, we find loop closures by evaluating NetVLAD [25] descriptors on the image and optimizing on SIM(3) to fix both scale and pose drift. Optionally (Section III-D), a final bundle adjustment step can be performed on the entire sequence to optimize both camera parameters and the scene structure.

A. Initialization

Since we work with in-the-wild videos, camera parameters like focal length are usually unknown. Thus, we first process a short sequence of frames to set up the scene and to obtain an initial estimate for the camera intrinsics.

Assuming a pinhole camera model without distortion, the camera intrinsics can be estimated by selecting N_{init} frames that have sufficient difference in optical flow (evaluated by RAFT [26] between consecutive frames) in order to get enough parallax for reliable estimation. Then GLOMAP is run on the collected images, yielding an initial coarse estimate \mathbf{K}_{init} of the camera intrinsics (but note that GLOMAP is only used in the initialization stage). We evaluate the current focal length estimation methods in Section IV-D, our method achieved the most accurate reconstruction though at the cost of some extra seconds.

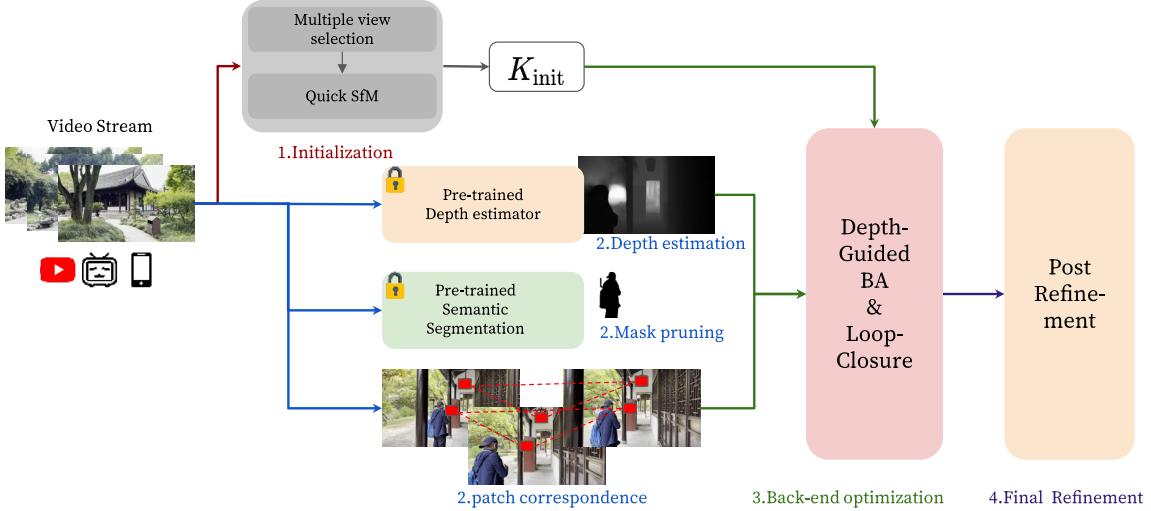


Fig. 1: Overview of our method. Given a video stream, we extract frames from the video sequentially. We first run an efficient global SfM process to estimate the camera intrinsic parameters K_{init} (Section III-A). Using an off-the-shelf semantic segmentation model, we prune the potential objects in the image (Section III-B.2) when estimating correspondence between frames. Correspondences are estimated across frames by DPVO; we use a monocular depth estimation model to get the prior depth, which can be used to guide the bundle adjustment optimization to improve robustness (Section III-B.3). SIM(3) pose graph optimization is conducted if a loop closure is detected (Section III-C). Finally, we run a re-triangulation and optionally global BA to refine the camera parameters and scene geometry (Section III-D).

B. Incremental Reconstruction

1) *Preliminaries:* At the heart of our visual SLAM pipeline, following the acquisition of the initial coarse camera intrinsic parameters K_{init} , lies the feature extraction and matching module from DPVO [6]. This module uses a recurrent neural network (RNN) to estimate correspondences without the need for explicit keypoint detection.

In DPVO, N_{patch} patches of $p \times p$ pixels are extracted from each image. Patch k in frame i is represented by $\mathbf{P}_{ik} = [\mathbf{x}, \mathbf{y}, 1, \mathbf{d}]^\top$, where $\mathbf{x}, \mathbf{y}, \mathbf{d} \in \mathbb{R}^{p^2 \times 1}$. Here, \mathbf{x}, \mathbf{y} denote the 2D coordinates of the extracted patches, and \mathbf{d} represents the associated inverse depths. Patches are randomly extracted from each image and each patch is connected to adjacent frames. When estimating the correspondence of the extracted patch in other frames, feature correlation is conducted between the patch feature and the image feature. For further details on the neural network architecture used to process the image and estimate correspondence, we point the reader to the original DVPO paper [6].

Once patches are extracted from the frames and frame-to-frame correspondences have been estimated, bundle adjustment is conducted by minimizing the re-projection error. The re-projection \mathbf{P}_{ik}^j of a given patch \mathbf{P}_{ik} to the frame j is expressed as:

$$\mathbf{P}_{ik}^j = \boldsymbol{\Pi}_{\text{init}}(\mathbf{G}_{ij} \boldsymbol{\Pi}_{\text{init}}^{-1}(\mathbf{P}_{ik})); \mathbf{G}_{ij} = \mathbf{G}_j \cdot \mathbf{G}_i^{-1}$$

where $\boldsymbol{\Pi}_{\text{init}}$ is the camera model constructed from K_{init} , which projects 3D points to 2D pixel coordinates, while $\boldsymbol{\Pi}_{\text{init}}^{-1}$ is the inverse projection which re-projects 2D pixels to 3D points in the local coordinate frame. \mathbf{G}_i is the camera pose of frame i , representing the *world-to-camera* order. Assuming that according to the DPVO network prediction, the image

correspondence of patch \mathbf{P}_{ik} on frame j is $\hat{\mathbf{P}}_{ik}^j$, which is the 2D re-projected coordinates and $\hat{\mathbf{P}}_{ik}^j = [\hat{\mathbf{x}}^j, \hat{\mathbf{y}}^j] \in \mathbb{R}^{p^2 \times 1}$. Then the bundle adjustment aims to minimize the re-projection error

$$\mathcal{E}_{(\mathbf{G}, \mathbf{d})} = \sum_{(i,j)} \sum_k \|\mathbf{P}_{ik}^j - \hat{\mathbf{P}}_{ik}^j\|. \quad (1)$$

2) *Semantic Masking:* In casual videos, objects, such as pedestrians or vehicles, frequently appear within the frame. To avoid potential wrong correspondences caused by dynamic objects, we avoid extracting patches on them. Using a semantic segmentation model, areas predicted to contain dynamic objects are masked. Although semantic masks are a simple strategy, they offer an efficient and direct method for excluding potentially dynamic elements. To further stabilize our optimization, we also prune regions lacking strong constraints, such as sky.

3) *Depth-regularized BA:* Given the frame-to-frame 2D correspondence estimates, we jointly optimize the geometry (i.e., the depth of the patches) and camera poses using bundle adjustment (BA). As mentioned before, small-parallax views can introduce ambiguity in depth and pose estimation. This is why popular SfM implementations [2] exclude points lacking sufficient triangulation angles. Multiple view constraints—i.e. observing the same objects from different positions—can help solve the problem. However, in uncontrolled walking-tour videos, collecting enough views from varied positions is challenging because the movement (and video stream) is mostly going forwards. On the other hand, SLAM pipelines using range sensors—such as lidar or RGBD—are not affected by depth ambiguity, even in sequences with pure rotation.

Inspired by recent advances in monocular depth estimation models, we integrate prior depth information into our optimization process. Specifically, given an image $\mathcal{I} \in \mathbb{R}^{3 \times H \times W}$, we input \mathcal{I} and the estimated intrinsic parameters K_{init} into the monocular depth estimation model to obtain the corresponding depth map $\mathcal{D} \in \mathbb{R}^{H \times W}$. When registering the new image \mathcal{I}_i into the current reconstruction, we first rescale the depth \mathcal{D}_i to align with the existing reconstruction. The scale factor α_i is determined by evaluating \mathcal{D}_i with the median depth of the latest three keyframes' patches:

$$\alpha_i = \frac{\text{median}(\mathcal{D}_i)}{\text{median}(\mathbf{P}[d]_{(i-3:i)})}.$$

After computing the rescaled depth, we add a depth regularization term to the bundle adjustment to guide the optimization Eq. (1):

$$\mathcal{E}_{(\mathbf{G}, \mathbf{d})} = \sum_{(i,j)} \sum_k \|\mathbf{P}_{ik}^j - \hat{\mathbf{P}}_{ik}^j\| + \mu \|\mathbf{P}_{ik}[d] - \alpha_i \mathcal{D}_{ik}\|, \quad (2)$$

where μ is the regularization term weight.

C. Loop Detection and Pose-graph optimization

For large-scale scene reconstruction, accumulation of drift is unavoidable, so place recognition and loop closure are needed to correct the trajectory. We use NetVLAD [25] to extract feature descriptors $\mathbf{V}_i \in \mathbb{R}^D$ for each image and, to avoid false positive loop detection, we follow the common practice of requiring 3 consecutive matching frames to register a loop closure; sequences with less than N consecutive frames are disregarded.

When a loop is detected, we run a scale-aware pose graph optimization [27]. This process transform each camera pose from $\text{SE}(3) = [\begin{smallmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{smallmatrix}]$ to $\text{SIM}(3) = [\begin{smallmatrix} s\mathbf{R} & \mathbf{t} \\ 0 & 1 \end{smallmatrix}]$ by introducing the scale factor $s \in \mathbb{R}^+$. Given ΔS_{ij} the transformation between frame i and the detected loop frame j , the loop closure residual is defined as:

$$r = \log_{\text{SIM}(3)}(\Delta S_{ij}^{-1} S_i S_j),$$

where S_i and S_j are the absolute similarity poses. The pose graph optimization is run synchronously, with the current detected loop frame held fixed while optimizing all previous frames.

D. Post-Refinement

Finally, after running the SLAM pipeline on the video sequence, we perform a post-refinement to get better results. We first run the feature matching on all images to create a feature database, then we re-triangulate points with achieved camera poses from the SLAM above. We use the `point_triangulator` function in COLMAP to conduct the re-triangulation, enabling refinement of camera intrinsic parameters. Optionally, we can run `ba_adjuster` to refine both camera poses and scene geometry at the cost of time.

IV. EXPERIMENTS

In this section, we present both quantitative and qualitative experiments demonstrating the robustness of our method on in-the-wild videos. Our method outperforms current state-of-the-art SfM methods, producing more continuous trajectories with fewer breaks and reduced computation time. Additionally, we conduct ablation studies to validate the effectiveness of the individual components of our proposed pipeline.

A. Experiment Setup

1) *Baselines*: As the baselines, we choose the SOTA structure-from-motion methods COLMAP [2] and GLOMAP [3]. We did not compare with some visual SLAM methods [1, 11] because they are known to be fragile with in-the-wild videos [28]. For some recent modern SfM methods [4, 5], due to the heavy GPU requirement, they are not scalable to large-scale scenes.

2) *Datasets*: We select tour videos (see Table II) from YouTube channels with permissive licenses or with the approval of authors. We manually remove the “preview” section from each video, segment the videos into approximately 15-minute clips, and extract frames at 3 FPS to ensure all methods can run within a reasonable timeframe. For drone-recorded videos, we extract frames at 10 FPS due to their higher motion speed. We rescale images to 512×288 , which is compatible with our method. For the baselines COLMAP and GLOMAP, we rescale images to 2K resolution because they work better on high-resolution images.

3) *Evaluation Metrics*: Different from some existing datasets [29, 30], we do not have ground-truth camera poses as references for evaluation as we focus on uncontrolled videos. We propose the following metrics for evaluation:

- 1) The number of registered images (counting the images in the largest reconstructed model.)
- 2) The number of separate models generated (should ideally be one).
- 3) Breaks along the trajectory. In certain reconstructions, abrupt jumps (indicating significant drift) may occur along the trajectory, without the method recognizing faulty registration and initiating a new model. Given two consecutive camera positions t_i and t_{i+1} in the global frame with $\Delta t = \|t_i - t_{i+1}\|$, we compute the ratio $\widehat{\Delta t}_i = \|t_i - t_{i+1}\| / \text{mean}(\|\Delta t_{i-k:i+k}\|)$, which means normalizing the position difference based on local scale to remove the scale drift effect. We define a break if $\widehat{\Delta t}_i > 10 \text{ mean}(\Delta t)$
- 4) Rendering quality measured by PSNR (peak signal-to-noise ratio). Following the quantitative evaluation done in ACE0 [31], we build a NeRF model [32] along the trajectory and take every eighth view as a test view. For unregistered images, we set their poses as an identity matrix to penalize incomplete reconstructions.
- 4) *Implementation Details*: We use *Metric3D* [8] for monocular depth estimation and *Mask2Former* [7] for semantic segmentation. For the post-refinement, we enable intrinsic refinement when retriangulation. For

COLMAP/GLOMAP, we apply SIFT feature matching, each image is matched with 20 frames before and after, and we enable vocabulary tree matching to allow loop detection. COLMAP requires a lot of time when running on thousands of images. In our experiments, we adopt the fast version parameters¹ when running COLMAP.

B. Quantitative Results

We demonstrate the robustness of our method in Table II. In all video sequences, our method registers the most images in each sequence without breaking the scene into multiple models. In contrast, COLMAP can generate multiple models due to failed image registration. Though GLOMAP also always generates one single model, there are often multiple breaks in the generated results, i.e., failed registrations that are not detected by the method. Note that our method misses two images in the sequence of “Uppsala”. This is due to our post-refinement stage, where two images fail to generate enough SIFT feature correspondences. It is also worth mentioning that, though we use a fast configuration of COLMAP, its running time is much longer than ours. Our method is both more robust and more efficient.

If the 3D map has been accurately reconstructed from the input video, it should be possible to create a NeRF model from the registered frames and compute the rendering quality of a held-out image to the corresponding NeRF rendering. Since plain NeRF does not have enough capacity for outdoor unbounded large-scale scenes, we cut the long sequences into short sequences with 500 frames each. Then we build a NeRF model (specifically, *Nerfacto*) on each short sequence, with every 8th frame held out and used as a test frame. The novel view synthesis performance for the test frames is reported in Table I. Our method achieves better rendering results due to fewer breaks in the trajectory. When COLMAP or GLOMAP aligns well, their estimated camera poses are more accurate. However, our method might be less precise, but more robust overall. As shown in Table III, for the sequence “Helsingborg Seq-1”, COLMAP and GLOMAP register images well for the first 500 frames, and achieve better rendering results; while in frame 500–1000, COLMAP and GLOMAP have breaks in the trajectory, resulting in low rendering results. As shown in Fig. 2, there is a break in the trajectory generated by COLMAP. In the NeRF evaluation (Fig. 3), we can see the resulting rendering artifacts at the break pose. It should be noted that all reported PNSR values are rather low, in part because the trajectories generated by the methods are not perfect but also because the NeRF model used to evaluate struggles with the large-scale outdoor scenes [33].

C. Qualitative Results

In addition to the quantitative evaluations above, we also compare the recreated paths with approximate GPS tracks from the data sets, where available. As shown in Fig. 4, our method can achieve smooth and consistent trajectories. Though GLOMAP can register more frames than COLMAP

TABLE I: Rendering results on selected sequences. The reported results are the average/min/max PSNR values computed for segments of 500 frames.

Sequence	Metrics	COLMAP [2]	GLOMAP [3]	Ours
Yanshan Park	Avg	13.64	13.35	13.95
	min	12.12	10.32	12.51
	max	14.48	15.08	15.55
Taicang Park	Avg	16.68	16.77	17.14
	min	13.36	12.79	15.64
	max	18.67	18.97	18.73
Uppsala	Avg	12.71	14.77	15.25
	min	7.7	12.13	14.43
	max	14.83	19.75	17.62
Helsingborg-1	Avg	15.18	12.61	14.72
	min	13.01	9.26	13.66
	max	16.47	16.22	15.77
Helsingborg-2	Avg	14.21	14.56	14.60
	min	10.86	12.18	13.17
	max	17.74	17.18	17.68
Lund	Avg	12.74	13.04	13.75
	min	11.7	12.06	13.16
	max	13.99	14.22	14.28

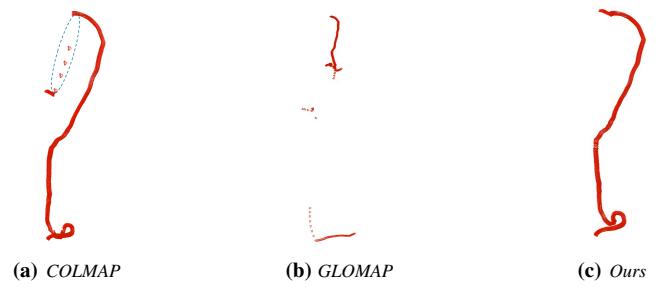


Fig. 2: The camera poses (red in the map) on the sequence “Helsingborg Seq-1” across frames 500–1000. COLMAP has a break in the trajectory (circled in blue); GLOMAP tacitly fails to register images; our method produces smooth and continuous trajectories.

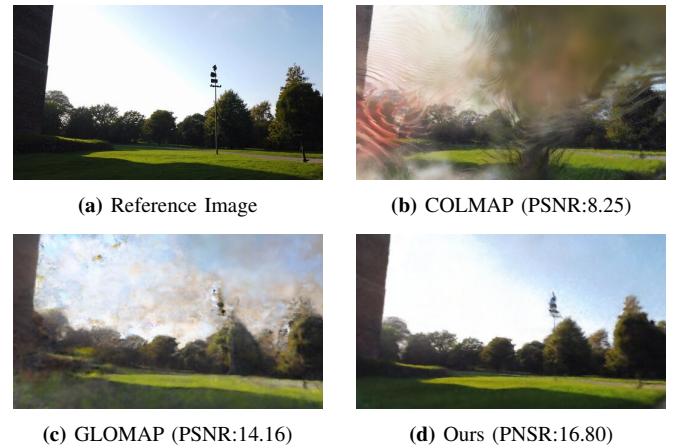


Fig. 3: The rendering results at the camera pose where COLMAP breaks.

¹<https://github.com/colmap/colmap/issues/116>

TABLE II: Reconstruction results on in-the-wild videos. Our method achieves most robust performance while using the least time for long sequence videos.

Sequence	Screenshot	#frames	Metrics	COLMAP [2]	GLOMAP [3]	Ours
Yanshan Park, China https://youtu.be/D8B30GIX-8s		3327	# Registered # Models # Breaks Time(min)	2989 2 3 665	3327 1 5 149	3327 1 0 18
Taicang Park, China https://youtu.be/LJf7LKLvmUc		2597	# Registered # Models # Breaks Time(min)	2534 2 10 385	2597 1 1 183	2597 1 0 8
Uppsala, Sweden, https://youtu.be/aVh.jTIP2cE?t=1262		2533	# Registered # Models # Breaks Time(min)	2206 3 2 200	2528 1 3 120	2531 1 1 12
Nanxun Ancient Town, China https://youtu.be/Owukwe_8OGw		1026	# Registered # Models # Breaks Time(min)	1026 1 0 57	1026 1 0 30	1026 1 0 6
Helsingborg, Sweden https://youtu.be/wUZ_zslH3vY?t=300 https://youtu.be/wUZ_zslH3vY?t=1200		Seq-1 2700 Seq-2 2700	# Registered # Models # Breaks Time(min)	2381 2 3 303	2382 1 29 154	2700 1 0 16
Lund, Sweden https://youtu.be/Nhc5BNlfDms?t=1800		2700	# Registered # Models # Breaks Time(min)	2689 2 1 258	2279 1 13 140	2700 1 0 18
The Backyard, USA https://youtu.be/OtkZJbW_sO0		578	# Registered # Models # Breaks Time(min)	1437 3 1 300	2697 1 17 180	2700 1 0 16
Time Average (min)				577	577	578
				14.05	9.26	15.74

TABLE III: NeRF rendering results on smaller clipped parts on “Helsingborg Seq-1”.

Sequence	COLMAP [2]	GLOMAP [3]	Ours
frame 0–500	16.41	16.22	14.05
frame 500–1000	14.99	9.26	15.74

(as seen in Table II), the resulting path and 3D model is often inconsistent for these kinds of uncontrolled input videos. We do not show “Taicang Park” and “The Backyard” sequence because all methods perform well on these two ones.

D. Ablation Studies

We conducted experiments to demonstrate the effectiveness of the different parts of our proposed reconstruction pipeline. Fig. 5 shows the reconstruction results for the “Yanshan Park” sequence with and without depth regularization, masking dynamic objects, and loop closure. Referring to the reconstruction of COLMAP and GLOMAP on the same sequence in Fig. 4, as we said, when COLMAP successfully registers images, the overall result is reliable. Based on this comparison and checking the video, we can say our method achieves good and accurate reconstruction on this sequence.

We also tested current camera intrinsic parameter estimation methods. We tried *Mast3R* [13], *MoGE* [34] and

GeoCalib [35], where *Mast3R* and *MoGE* recover focal length from the predicted point cloud while *GeoCalib* directly predicts from the image. We select the first image (first two for *Mast3R*) in the video and feed it into the above-mentioned methods. We assume a pinhole camera model where the principal point is in the center. For the sequence “Yanshan Park”, the estimated focal lengths and corresponding reconstruction results are presented in Fig. 6. Locally, imprecise intrinsic parameter estimation will not result in a severe failure, but with longer trajectories, it will incur more drifts. Compared to the aforementioned methods that rely on only one or two images to estimate the focal length, our method is more computationally expensive but achieves higher accuracy, making it particularly beneficial for long sequences.

V. CONCLUSIONS

Robust and accurate 3D reconstruction from in-the-wild videos is a very challenging problem. Compared to standard SLAM datasets, where the camera is typically carefully moved through a scene to ensure being able to track its movement, there is no control over the camera motion and we often observe pure rotations or pure forward motion, which are challenging. At the same time, there often are moving objects in the scene, complicating the process. Addressing these challenges represents the next frontier in SLAM, and

Sequence	COLMAP [2]	GLOMAP [3]	Ours	Reference
Yanshan Park				None
Helsingborg-1				
Helsingborg-2				
Lund				
Uppsala				

Fig. 4: Reconstruction result visualization. Our method generally achieves smooth and continuous trajectories without breaks. COLMAP often produces a model only for part of the path. GLOMAP struggles to produce consistent results in these large-scale environments. GPS tracks are provided for Lund and the two Helsingborg sequences, but note that the provided GPS data is rather inaccurate. For Uppsala, the path has been drawn manually while referencing the video. No reference data is available for Yanshan Park.

progress in this direction will lead to more robust and adaptable systems, crucial for real-world robotics applications.

With the present work we take strides towards consistent 3D mapping of large-scale environments from uncontrolled videos: over 1 km in length, thousands of video frames, over 10 min duration. Specifically, we have investigated robust methods for recovering the focal length from in-the-wild videos, leverage semantic masks to improve data association in scenes with moving objects, and use monocular depth cues to regularize bundle adjustment in order to be more robust to difficult camera motion and features near the horizon.

Comparing our pipeline to GLOMAP [3] and COLMAP [2], our proposed method robustly produces longer sequences without breaks, and does so in a fraction of the time.

Future research directions include methods for further reducing drift while maintaining consistent reconstructions and strategies for cases where the view is severely covered by moving objects.

REFERENCES

- [1] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. “ORB-SLAM: a versatile and accurate monocular SLAM system”. In: *IEEE transactions on robotics* 31.5 (2015), pp. 1147–1163.

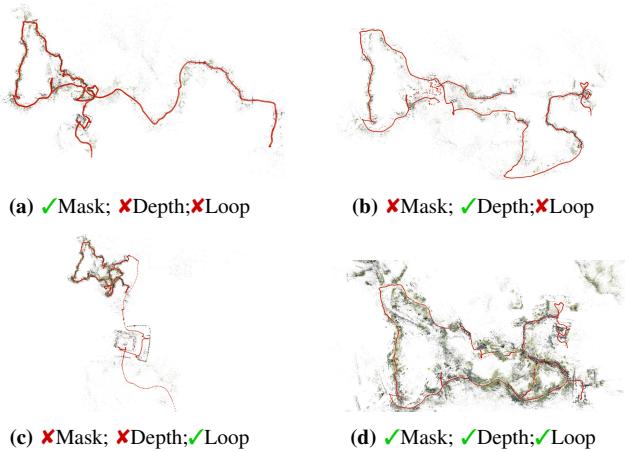


Fig. 5: Ablation study: on the “Yanshan Park” sequence, we show the effectiveness of the proposed modules. To handle in-the-wild videos, prior depth and pruning dynamics in the view greatly help to improve the robustness.

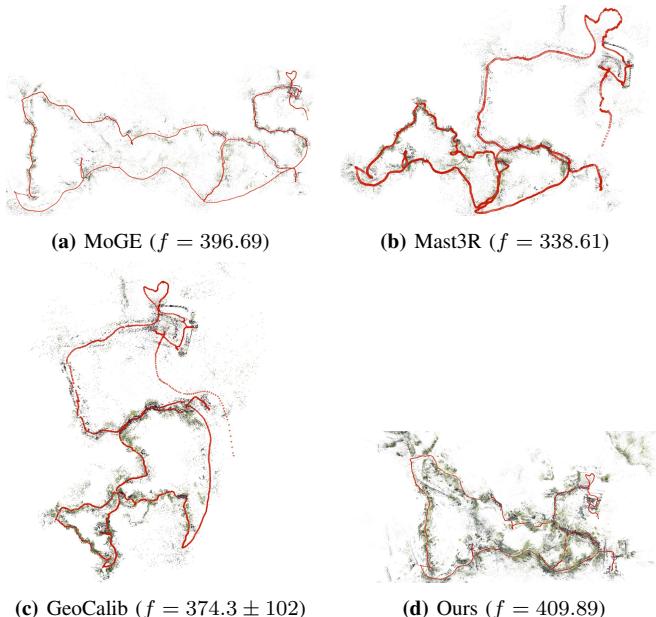


Fig. 6: Reconstruction results by different ways of estimating focals.

- [2] Johannes L Schonberger and Jan-Michael Frahm. “Structure-from-motion revisited”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4104–4113.
- [3] Linfei Pan et al. “Global structure-from-motion revisited”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 58–77.
- [4] Jianyuan Wang et al. “VGGSM: Visual geometry grounded deep structure from motion”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 21686–21697.
- [5] Bardienus Duisterhof et al. “MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion”. In: *arXiv preprint arXiv:2409.19152* (2024).
- [6] Zachary Teed, Lahav Lipson, and Jia Deng. “Deep patch visual odometry”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [7] Bowen Cheng et al. “Masked-attention mask transformer for universal image segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1290–1299.
- [8] Wei Yin et al. “Metric3d: Towards zero-shot metric 3d prediction from a single image”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 9043–9053.
- [9] Sheng Liu, Xiaohan Nie, and Raffay Hamid. “Depth-guided sparse structure-from-motion for movies and TV shows”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 15980–15989.

- [10] Zhengqi Li et al. “MegaSaM: Accurate, fast, and robust structure and motion from casual dynamic videos”. In: *arXiv preprint arXiv:2412.04463* (2024).
- [11] Zachary Teed and Jia Deng. “DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras”. In: *Advances in neural information processing systems* 34 (2021), pp. 16558–16569.
- [12] Shuzhe Wang et al. “Dust3r: Geometric 3d vision made easy”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 20697–20709.
- [13] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. “Grounding image matching in 3d with MAST3R”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 71–91.
- [14] Sven Elflein et al. “Light3R-SfM: Towards Feed-forward Structure-from-Motion”. In: *arXiv preprint arXiv:2501.14914* (2025).
- [15] Riku Murai, Eric Dexheimer, and Andrew J Davison. “MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors”. In: *arXiv preprint arXiv:2412.12392* (2024).
- [16] Tinghui Zhou et al. “Stereo magnification: Learning view synthesis using multiplane images”. In: *arXiv preprint arXiv:1805.09817* (2018).
- [17] Kai Zhang et al. “Gs-lrm: Large reconstruction model for 3d gaussian splatting”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 1–19.
- [18] Hao He et al. “Cameractrl: Enabling camera control for text-to-video generation”. In: *arXiv preprint arXiv:2404.02101* (2024).
- [19] Hao He et al. “Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models”. In: *arXiv preprint arXiv:2503.10592* (2025).
- [20] Mustafa Khan et al. “Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction”. In: *arXiv preprint arXiv:2407.02598* (2024).
- [21] WAYVE. *Ghost Gym: A Neural Simulator for Autonomous Driving*. 2023. URL: <https://wayve.ai/thinking/ghost-gym-neural-simulator/> (visited on 08/12/2025).
- [22] Kunyang Lin et al. “Learning vision-and-language navigation from youtube videos”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 8317–8326.
- [23] Amir Bar et al. “Egopet: Egomotion and interaction data from an animal’s perspective”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 377–394.
- [24] Amir Hossain Raj et al. “Rethinking social robot navigation: Leveraging the best of two worlds”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 16330–16337.
- [25] Relja Arandjelovic et al. “NetVLAD: CNN architecture for weakly supervised place recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5297–5307.
- [26] Zachary Teed and Jia Deng. “RAFT: Recurrent all-pairs field transforms for optical flow”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer. 2020, pp. 402–419.
- [27] Hauke Strasdat, JMM Montiel, and Andrew J Davison. “Scale Drift-Aware Large Scale Monocular SLAM”. In: *Robotics: Science and Systems*. Vol. 2. 3. 2010, p. 5.
- [28] Suvam Patra et al. “EGO-SLAM: A robust monocular SLAM for egocentric videos”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 31–40.
- [29] Jürgen Sturm et al. “A benchmark for the evaluation of RGB-D SLAM systems”. In: *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2012, pp. 573–580.
- [30] Julian Straub et al. “The Replica dataset: A digital replica of indoor spaces”. In: *arXiv preprint arXiv:1906.05797* (2019).
- [31] Eric Brachmann et al. “Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocator”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 421–440.
- [32] Matthew Tancik et al. “Nerfstudio: A modular framework for neural radiance field development”. In: *ACM SIGGRAPH 2023 Conference Proceedings*. 2023, pp. 1–12.
- [33] Matthew Tancik et al. “Block-NeRF: Scalable large scene neural view synthesis”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 8248–8258.
- [34] Ruicheng Wang et al. “MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision”. In: *arXiv preprint arXiv:2410.19115* (2024).
- [35] Alexander Veicht et al. “GeoCalib: Learning Single-image Calibration with Geometric Optimization”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 1–20.