DROID-SLAM in the Wild

Supplementary Material

Sequence	Number of Frames	Length of Trajectory [m]
Downtown 1	1427	90.83
Downtown 2	2200	122.25
Downtown 3	1438	62.33
Downtown 4	1794	85.19
Downtown 5	2157	129.93
Downtown 6	1900	104.99
Downtown 7	1900	109.35

Table 7. **Overview of Our DROID-W Dataset.**

In the supplementary material, we provide additional details about the following:

- More information about the DROID-W dataset and downloaded YouTube videos (Sec. 7).
- Details about the Jacobian derivation of our uncertainty optimization (Sec. 8).
- Additional qualitative comparisons for uncertainty, point clouds, and ablation study (Sec. 9).

7. Dataset

Prior benchmarks on dynamic SLAM are limited to indoor environments, exhibiting simple object motions and lacking truly challenging real-world conditions. To enable evaluation in more complex and unconstrained settings, we introduce an outdoor dataset, DROID-W, and additionally download 6 challenging videos from YouTube.

DROID-W Dataset. The DROID-W dataset is captured using a Livox Mid-360 LiDAR rigidly mounted with an RGB camera. It comprises 7 outdoor sequences (`Downtown 1-7`) with RGB frames at a resolution of 1200×1600, ground-truth camera poses, and synchronized IMU and LiDAR measurements. The RGB stream is recorded at 20 FPS, while RTK provides ground-truth poses at 10 Hz. We use the estimated trajectories from FAST-LIVO2 [65] as ground truth for `Downtown 1-2` due to the absence of RTK measurements. As shown in Table 8, FAST-LIVO2 provides sufficiently accurate estimates to serve as reliable ground truth for these sequences. Detailed information, including trajectory lengths and frame numbers, is reported in Table 7. The DROID-W dataset features long camera trajectories, high scene dynamics, and partial over-exposure – characteristics commonly encountered in real-world scenarios. We believe this dataset will provide significant value to the community and support future research on robust in-the-wild SLAM.

YouTube Videos. The framerates of the YouTube videos vary substantially. Therefore, we report the FPS and sequence duration for each sequence in Table 9, which provides a more meaningful characterization of camera motion and scene dynamics. Camera intrinsics are estimated using MonST3R [62] from the first 20 frames of each video. The sequences contain a large number of dynamic objects of diverse categories, with many of them moving simultaneously, leading to highly dynamic scenes. They also exhibit challenging and cluttered conditions, including motion blur, strong view-dependent effects, and low dynamic range.

8. Uncertainty Optimization and Jacobians

Given the definition in Sec. 3.3 of the main paper, we obtain the following uncertainty energy function:

$$\begin{aligned}\mathbf{u}'_i &= \log(\exp(\boldsymbol{\theta}\cdot\mathbf{F}_i)+1),\\ \mathbf{E}_{\text{uncer}}(\mathbf{u}') &= \sum_{(i,j)\in\mathcal{E}} \mathbf{e}_{ij} + \gamma_{\text{prior}}\sum_i \log(\mathbf{u}'_i+1.0),\\ &= \sum_{(i,j)\in\mathcal{E}} \frac{1-\frac{\mathbf{F}_i\cdot\mathbf{F}_{ij}}{\|\mathbf{F}_i\|_2\|\mathbf{F}_{ij}\|_2}}{\mathbf{u}'_i\cdot\mathbf{u}'_{ij}} + \gamma_{\text{prior}}\sum_i \log(\mathbf{u}'_i+1.0).\end{aligned} \tag{11}$$

Thus, we can derive the following Jacobians:

$$\begin{aligned}\frac{\partial \mathbf{e}_{ij}}{\partial \mathbf{u}'_i} &= -\frac{1-\frac{\mathbf{F}_i\cdot\mathbf{F}_{ij}}{\|\mathbf{F}_i\|_2\|\mathbf{F}_{ij}\|_2}}{(\mathbf{u}'_i)^2\cdot\mathbf{u}'_{ij}} = -\frac{\mathbf{e}_{ij}}{\mathbf{u}'_i},\\ \frac{\partial \mathbf{e}_{ij}}{\partial \mathbf{u}'_j} &= -\frac{1-\frac{\mathbf{F}_i\cdot\mathbf{F}_{ij}}{\|\mathbf{F}_i\|_2\|\mathbf{F}_{ij}\|_2}}{\mathbf{u}'_i\cdot(\mathbf{u}'_{ij})^2}\cdot\frac{\partial \mathbf{u}'_{ij}}{\mathbf{u}'_j} = -\frac{\mathbf{e}_{ij}}{\mathbf{u}'_j}\cdot\boldsymbol{\alpha}_{ij}.\end{aligned} \tag{12}$$

where $\boldsymbol{\alpha}_{ij}\in\mathcal{R}^{(\frac{H}{8}\times\frac{W}{8})\times(\frac{H}{8}\times\frac{W}{8})}$ is the bilinear interpolation weight matrix whose non-zero elements are of size $\frac{H}{8}\times\frac{W}{8}\times 4$. The final Jacobians are defined as follows:

$$\begin{aligned}\frac{\partial \mathbf{E}_{uncer}}{\mathbf{u}'_l} &= -\sum_{(l,m)\in\mathcal{E}}\frac{\mathbf{e}_{lm}}{\mathbf{u}'_l} - \sum_{(k,l)\in\mathcal{E}}\frac{\mathbf{e}_{kl}}{\mathbf{u}'_k}\cdot\boldsymbol{\alpha}_{kl} + \gamma_{\text{prior}}\cdot\frac{1}{\mathbf{u}'_l+1.0},\\ \frac{\partial \mathbf{u}'_l}{\partial\boldsymbol{\theta}} &= \frac{1}{1+\exp(-\boldsymbol{\theta}\cdot\mathbf{F}_l)}\cdot\mathbf{F}_l,\\ \frac{\partial \mathbf{E}_{\text{uncer}}}{\partial\boldsymbol{\theta}} &= \sum_{l=0}^{N}\frac{\partial \mathbf{E}_{\text{uncer}}}{\partial \mathbf{u}'_l}\cdot\frac{\partial \mathbf{u}'_l}{\partial\boldsymbol{\theta}}.\end{aligned} \tag{13}$$

Here, frame l serves as the reference frame in all edges (l,m) and as the target frame in all edges (k,l).

9. Additional Experiments

9.1. Uncertainty Estimation

Uncertainty Comparisons on YouTube Videos. We provide additional qualitative results on uncertainty estimation

Method	Inputs	Downtown 3	Downtown 4	Downtown 5	Downtown 6	Downtown 7	**Avg.**
FAST-LIVO2 [65]	RGB + IMU + LiDAR	0.06	0.06	0.09	0.09	0.06	0.071
DROID-W (Ours)	RGB	0.15	0.32	0.24	0.43	0.07	0.242

Table 8. **Tracking performance of FAST-LIVO2 [65] on the DROID-W dataset** (ATE RMSE ↓ [m]). FAST-LIVO2 is an efficient and accurate LiDAR-inertial-visual fusion system capable of delivering centimeter-level localization accuracy. Its performance on `Downtown 3-7` demonstrates sufficient accuracy to serve as ground truth for `Downtown 1-2`.

Sequence	Time	FPS	Resolution
`Elephant Herd`	00:08	24	1280 × 720
`Giraffe`	00:09	24	1280 × 720
`Taylor`	01:13	30	1280 × 720
`Tomyum 1`	01:40	30	1280 × 720
`Tomyum 2`	01:40	30	1280 × 720
`St. Moritz`	30:00	50	1280 × 720
`Tokyo Walking 1`	00:50	60	1920 × 1080
`Tokyo Walking 2`	00:22	60	1920 × 1080
`Tokyo Walking 3`	00:40	60	1920 × 1080

Table 9. **Overview of Downloaded YouTube Videos.**

in Fig. 5. As shown in Fig. 5, WildGS-SLAM [66] always fails to construct the reliable Gaussian map, leading to inaccurate and noisy uncertainty predictions. In contrast, our method leverages frame-to-frame feature alignment, demonstrating significantly greater robustness in visually complex and truly in-the-wild environments.

Moreover, our approach effectively handles strong view-dependent effects such as reflections and shadows (e.g. reflections in `Taylor 22` and `Tokyo Walking 2`). It is also highly sensitive to small dynamic objects, enabling precise uncertainty estimation even under challenging real-world conditions. Our method further exhibits strong robustness to severe motion blur and low dynamic range (e.g. `Tomyum 1` and `Tomyum 2`), which are extremely difficult for conventional segmentation or detection approaches.

Overall, Fig. 5 highlights the accuracy and robustness of our uncertainty optimization in unconstrained in-the-wild settings, effectively delineating uncertain regions while maintaining high confidence in static areas.

Uncertainty Visualization for Consecutive Keyframes. We visualize the optimized uncertainties across consecutive keyframes in Fig. 6 and Fig. 7. Our method optimizes frame-wise uncertainty by exploiting multi-view feature similarity, allowing it to fully leverage static-scene information whenever available. As shown in Fig. 6, the system effectively utilizes the door region for camera tracking prior to keyframe 280. In addition, our uncertainty estimation integrates multi-view cues from frames connected through the frame graph, i.e. it captures multi-view inconsistency within a local window. Thus, our approach assigns the reasonable higher uncertainty to the door region of keyframes 280/281 before the door begins to move.

Method	ATE RMSE [cm]
w/o prior term	5.18
Full	**2.30**

Table 10. **Ablation Studies on Bonn RGB-D Dataset [38].**

We observe strong mirror-reflection effects in Fig. 7, with keyframe 283 exhibiting the largest appearance change. Accordingly, our method assigns the highest uncertainty to keyframe 283, while maintaining relatively low uncertainty for the remaining keyframes that show only minor appearance differences. This behavior demonstrates the effectiveness of our approach and the precision of the resulting uncertainty estimates.

9.2. Point Cloud Reconstruction

Static Reconstruction Comparisons. We present 3D reconstruction comparisons between DROID-SLAM [48] and our method in Fig. 8. In the top row, DROID-SLAM fails to recover consistent geometry – erroneously reconstructing a single corridor as two separate structures – and exhibits noticeable tracking drift in sequences with the presence of strong dynamic distractors. In contrast, our approach produces coherent geometric reconstructions and accurate pose estimates. The second row further highlights the robustness of our method: it reconstructs cleanly visible and highly accurate white lane markings on the asphalt road, even under challenging real-world conditions.

Static / Dynamic Reconstruction. We visualize the static and dynamic point clouds in Fig. 9 and Fig. 10, providing complementary perspectives from both top-down and interior viewpoints. The high geometric consistency between our static reconstruction and the static regions within the dynamic point clouds illustrates the precision of our uncertainty estimation. These results demonstrate that our method effectively suppresses dynamic or uncertain regions while preserving the underlying static scene structure.

9.3. Additional Ablation Study

Table 10 shows that the prior regularization term effectively avoid the trivial solution $\mathbf{u} \to \infty$. Without this prior, the system assigns uniformly large uncertainties to all pixels, resulting in results similar to the *w/o Uncertainty-aware BA* configuration reported in Table 6 of the main paper.

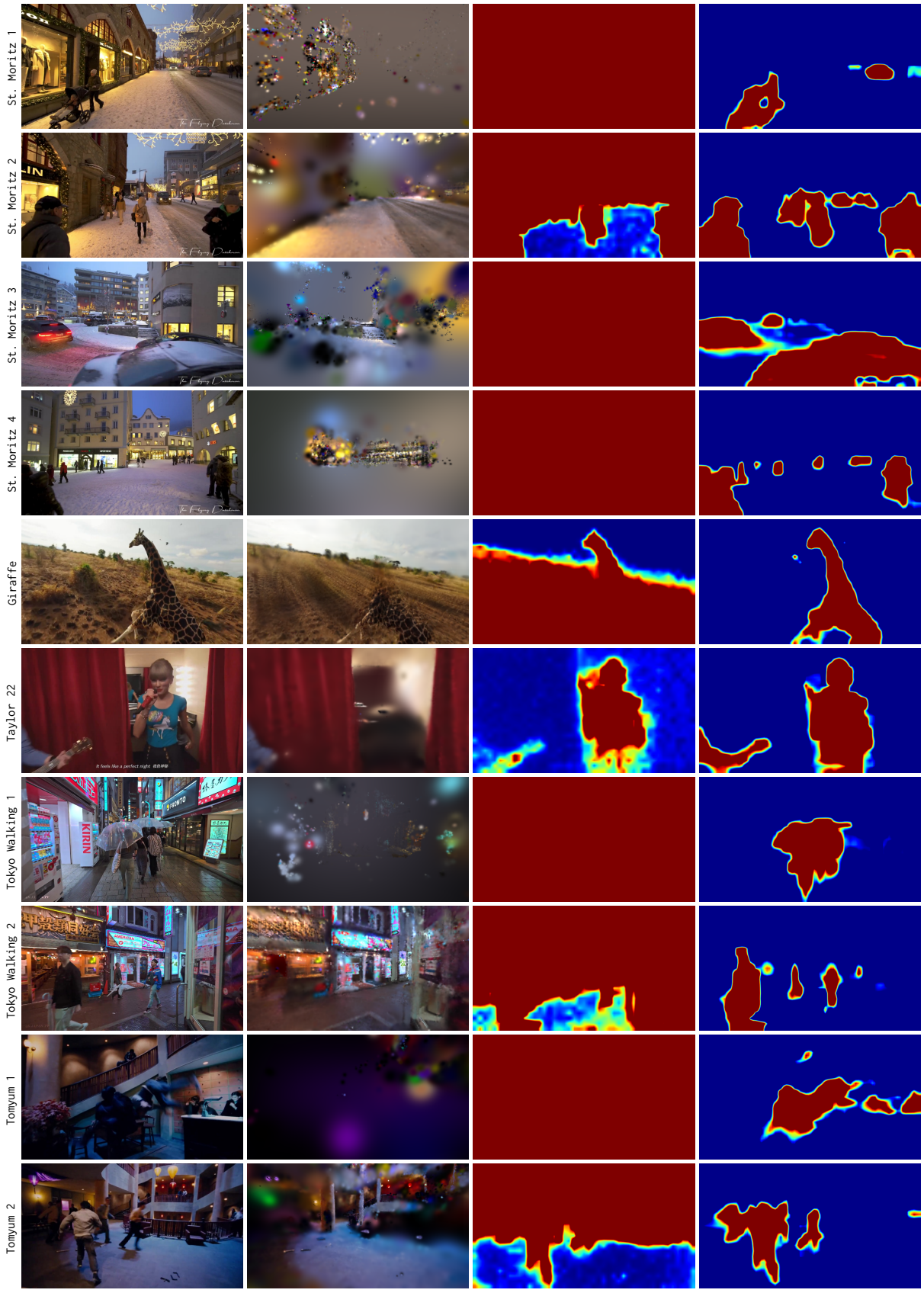

Figure 5. **Uncertainty Estimation.**

Figure 6. **Uncertainty Visualization for Consecutive Frames of YouTube `Tomyum 1`.** We observe that our method robustly handles scenarios in which an object transitions from a static state to dynamic motion, such as a door being pushed open by a person. Prior to the onset of motion, our approach leverages stable visual correspondences on the door to help tracking, since our uncertainty optimization is based on frame-to-frame feature alignment.

Figure 7. **Uncertainty Visualization for Consecutive Frames of YouTube `Tomyum 2`.** Our approach assigns high uncertainty to regions exhibiting strong view-dependent effects (e.g., the mirror).

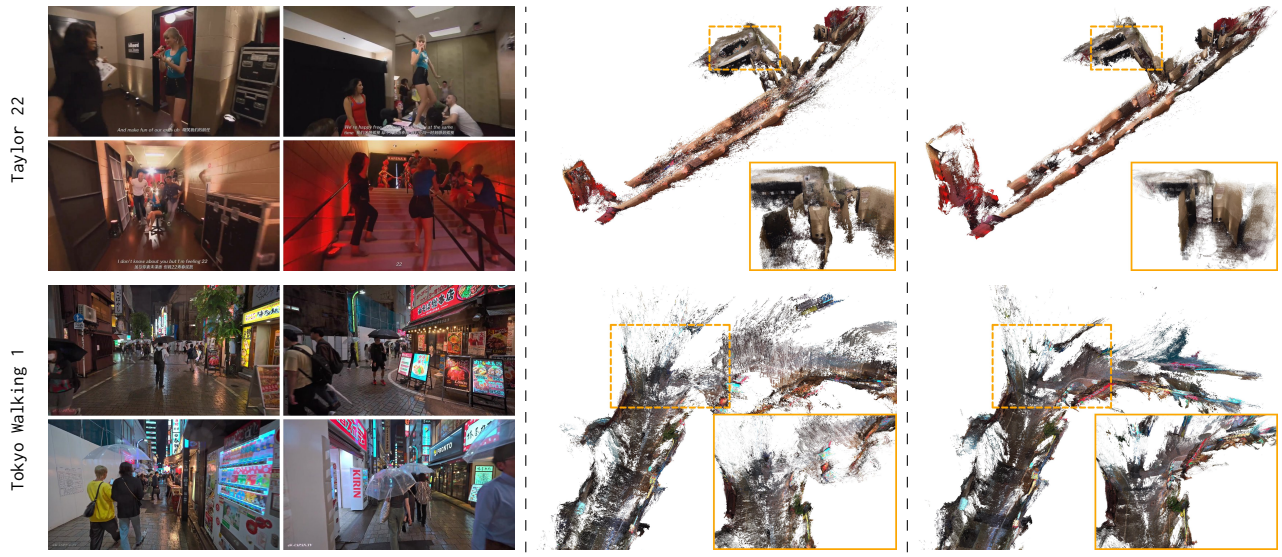

Figure 8. **3D Reconstruction Comparisons on YouTube Sequences.** We compare reconstructed static point clouds between DROID-SLAM [48] and our method. Our approach produces more accurate and consistent reconstructions across highly dynamic and visually challenging real-world sequences. For the `Taylor 22` scene, DROID-SLAM reconstructs two separate corridor structures due to inaccurate pose tracking, as highlighted in the boxed region, whereas our method produces a consistent geometric reconstruction. Moreover, the white lane marking on the asphalt road of `Tokyo Walking 1` is faint and fragmented in the point cloud reconstructed by DROID-SLAM, while in our reconstruction it remains cleanly visible, continuous, and highly accurate.

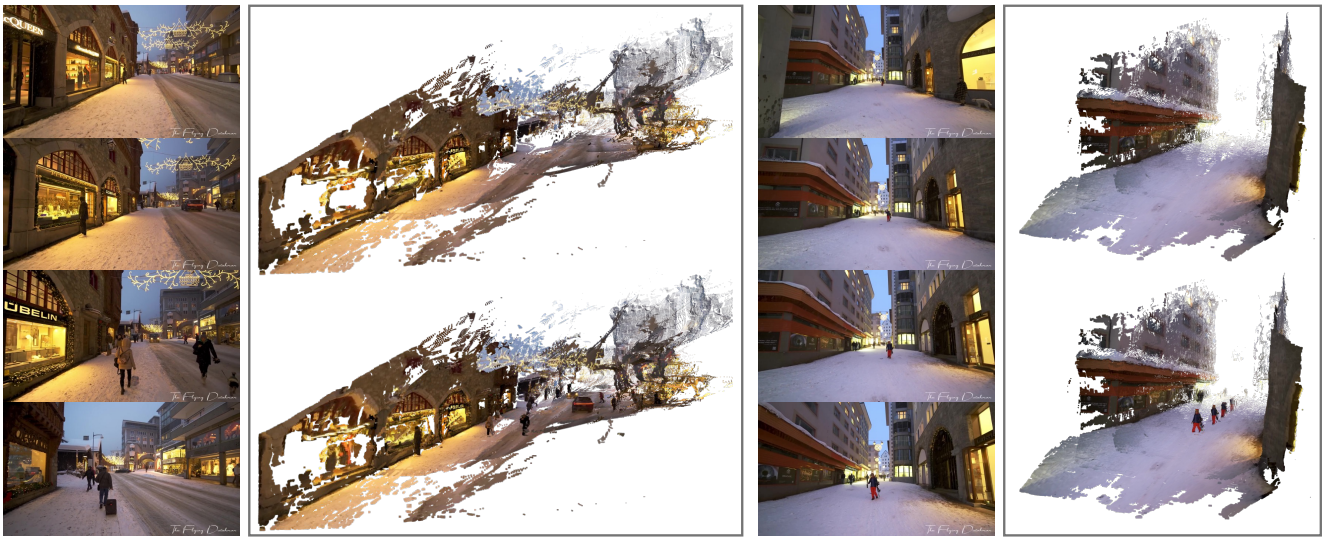

Figure 9. **Point Clouds Visualization from 4 Views.** *Top view*: reconstructed static scene from 4 input views shown in the figure. *Bottom view*: reconstructed dynamic point clouds. The comparisons between the dynamic and static point clouds further demonstrate the effectiveness of our uncertainty estimation. In both visualizations, the static point clouds remain highly consistent, indicating that our method reliably preserves static geometry while filtering dynamic regions.

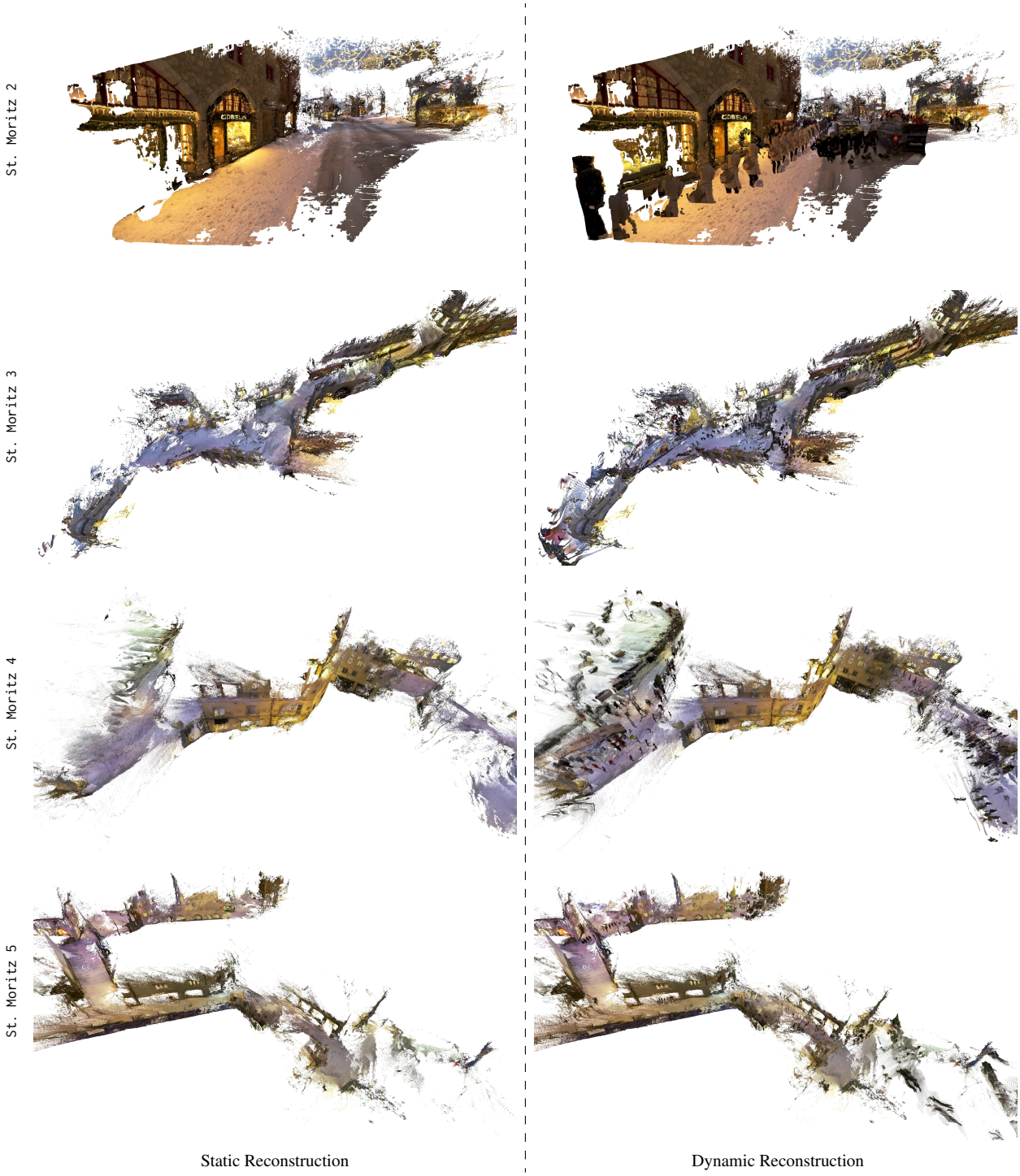

Figure 10. **Qualitative Results of Our Static and Dynamic Reconstruction.** We visualize globally aligned static reconstructions alongside dynamic point clouds across all keyframes. Notably, we apply the estimated per-frame dynamic uncertainty to filter out dynamic points. The left and right columns show the static and dynamic reconstructions, respectively. These comparisons highlight the accuracy of our uncertainty estimation, as our method effectively suppresses dynamic regions while preserving geometric fidelity in static areas.