Changes in Real Time: Online Scene Change Detection with Multi-View Fusion

Supplementary Material

6. Additional Experimental Results

6.1. Analysis on Inference Frame Registration

Existing offline approaches such as MV3DCD [13] assume access to all pre- and post-change images jointly and estimate their poses in a common coordinate frame using COLMAP [47]. This SfM-based pipeline is computationally expensive and inherently offline, as it requires processing the entire dataset at once.

In contrast, our lightweight PnP-based pose estimation module enables online operation by processing each incoming frame individually and on-the-fly. By leveraging the pre-change camera poses used to build the reference scene representation, we register each incoming inference frame directly to the same reference coordinate frame without requiring joint optimization over all images. This design allows real-time pose estimation with respect to the reference frame, enabling online pose-agnostic SCD.

In Tab. 6, we provide a comparison with other online pose estimation approaches against the poses estimated from COLMAP offline. We consider the poses from COLMAP as the pseudo-ground truth, as PASLCD [13] does not have ground truth poses. We report the difference [28] between the estimated and COLMAP poses (rotation and translation) averaged across all frames.

ChangeSim [36] uses a frame-retrieval heuristic that selects the nearest pre-change frame based on the L_1 distance between camera poses estimated via an off-the-shelf RGB-D SLAM system (RTAB-Map [24]). Although lightweight, this approach critically assumes that the pre- and post-change trajectories are closely aligned in both camera position and orientation. This assumption rarely holds in practical scenarios such as autonomous inspections, where the scene is revisited from unconstrained viewpoints and independent trajectories. Our method avoids these limitations by explicitly estimating each frame's pose relative to the reference scene, enabling robust registration under unconstrained viewpoints. As mentioned in Sec. 4, we use pseudo-ground truth poses from COLMAP for ChangeSim's frame retrieval for a fair comparison. Note that we cannot use ChangeSim's RGB-D SLAM system as PASLCD does not have depth information. Therefore, we do not add the timing for ChangeSim in Tab. 6.

SplatPose [23] estimates inference-frame poses using a coarse-to-fine pipeline. First, it retrieves the most similar reference image using a pre-trained LoFTR [48] model and adopts its pose as an initial coarse estimate. This pose is then refined via analysis-by-synthesis [23], where a photometric loss is minimized to solve for a single rigid transformation applied to all Gaussian primitives in the reference scene. However, the refinement heavily depends on the quality of the coarse estimate, often fails to converge in complex real-world scenes containing multiple changes, and does not run at real-time rates.

Table 6. Pose estimation results on PASLCD [13] under similar and different lighting conditions. Timing is averaged per-frame. Our lightweight PnP-based pose estimation achieves comparable or better accuracy while running in real-time.

Method	**Similar Lighting**		**Different Lighting**		**Time**
	R (°) ↓	**t** ↓	**R** (°) ↓	**t** ↓	
ChangeSim [36]	28.66	1.27	43.65	1.81	✗
SplatPose [23]	6.06	0.59	7.07	0.73	10.2 s
SplatPose+ [31]	**0.06**	**0.01**	*0.32*	**0.02**	*6.2 s*
Ours	*0.09*	**0.01**	**0.13**	**0.02**	**30 ms**

SplatPose+ [31] replaces the analysis-by-synthesis refinement with HLoc [44] for more reliable pose estimation. While this greatly improves accuracy, it is not sufficiently fast to achieve real-time online performance. For instance, SplatPose+ requires around 6 s per frame for pose estimation on average. In contrast, our lightweight PnP-based approach estimates poses in approximately 30 ms, offering over two orders of magnitude speed-up while maintaining competitive or better accuracy.

6.2. Properties of the Self Supervised Loss

Intuitively, our self-supervised loss, L_{SSF}, integrates complementary but potentially noisy change cues, present in independent views across modalities, into a persistent and explicit 3D representation. We explicitly design L_{SSF} to yield strictly bounded gradients. This contrasts with the self-supervised loss presented by Furukawa *et al*. [12], whose gradients can explode when predicting a confident change mask, thereby requiring carefully tuned hyperparameters and supervised semantic priors.

The strictly bounded gradients of our L_{SSF} ensure optimization stability. Specifically, the gradients of the consistency and regularization terms are bounded by $C \in [0, 2]$ and $[0, 1]$, respectively. Under this formulation, strong visual evidence ($C \approx 2$) dominates the regularization term, pushing towards a higher recall. Conversely, partial evidence introduces competing forces, naturally resorting to multi-view consistency for validation. This stability allows us to avoid extensive hyperparameter tuning, which is often infeasible in online, self-supervised settings where ground-

Table 7. Quantitative SCD results on PASLCD [13] averaged over all 10 scenes. We evaluate performance under both similar and different lighting conditions. Our method achieves the best performance in both offline and online settings.

	Similar Lighting		Different Lighting	
Method	**mIoU ↑**	**F1 ↑**	**mIoU ↑**	**F1 ↑**
GeSCD [21]	0.469	0.602	*0.485*	*0.619*
MV3DCD [13]	*0.498*	*0.645*	0.457	0.611
Ours (Offline)	**0.569**	**0.706**	**0.535**	**0.682**
CS+CYWS2D [40]	*0.258*	*0.378*	0.229	0.342
SplatPose+ [31]	0.236	0.357	*0.237*	*0.359*
Ours	**0.503**	**0.652**	**0.469**	**0.624**

Table 8. Quantitative SCD results on PASLCD [13] averaged over all 10 scenes. We evaluate performance under indoor and outdoor environments. Our method achieves the best performance in both offline and online settings.

	Indoor		Outdoor	
Method	**mIoU ↑**	**F1 ↑**	**mIoU ↑**	**F1 ↑**
GeSCD [21]	*0.516*	*0.651*	0.428	0.569
MV3DCD [13]	0.479	0.634	*0.476*	*0.622*
Ours (Offline)	**0.548**	**0.697**	**0.556**	**0.690**
CS+CYWS2D [40]	0.167	0.247	*0.319*	*0.448*
SplatPose+ [31]	*0.213*	*0.326*	0.260	0.390
Ours	**0.483**	**0.638**	**0.489**	**0.638**

truth annotations are unavailable.

6.3. Instance-level Results for PASLCD

We provide an instance-level breakdown of the PASLCD results originally summarized in Table 1. Specifically, we detail performance under similar and varying lighting conditions (Table 7), as well as across indoor and outdoor environments (Table 8). Note that the varying lighting conditions encompass scenarios where the illumination is either better or worse than that of the reference scene. Consistent with our primary findings, both the online and offline formulations of our method significantly outperform the evaluated baselines across these diverse settings.

7. Discussion on Limitations and Future Work

Our performance depends on the reliability of the underlying vision foundation model [39], and future advances in these models will further strengthen our approach by producing more reliable and accurate change cues. Future improvements in lightweight keypoint extractors would similarly benefit our pose estimation module, enhancing both accuracy and speed. Beyond our current pixel- and feature-level cues, exploring more robust and complementary change cues offers a promising direction for boosting performance in both online and offline SCD settings.

Although XFeat [37] facilitates rapid, semi-dense feature extraction for lightweight PnP pose estimation, we observed that it struggles with unconstrained in-plane rotations. Furthermore, accurate image registration using XFeat is highly dependent on input image dimensions due to the architecture's sensitivity to raw pixel density. To ensure optimal feature extraction, input images must generally maintain a minimum of VGA resolution (640x480).

One of the constraints for run-time is the rendering speed (dependent on primitive count) of the base 3DGS, which may suffer in extremely large-scale environments or much greater image resolutions (with our reported results at 1008×560 resolution). However, our framework is agnostic to the specific 3DGS variant, and in such cases, scalable variants (e.g., Hierarchical-3DGS [20]) could be used to maintain real-time performance.

Our efficient representation update strategy is designed to bring the scene representation up to date with minimal computational overhead while preserving reconstruction fidelity, which is an essential property for scalable, repeatable autonomous inspections. This goal is complementary to continual-learning–based update methods that focus on long-term history recovery. Our current formulation does not support explicit history recovery; future work could explore unifying these objectives to enable both efficient updates and continual learning.