LAP3R: LARGE-SCALE PANOPTIC 3D RECONSTRUCTION FROM AERIAL IMAGERY

Anonymous authors
Paper under double-blind review

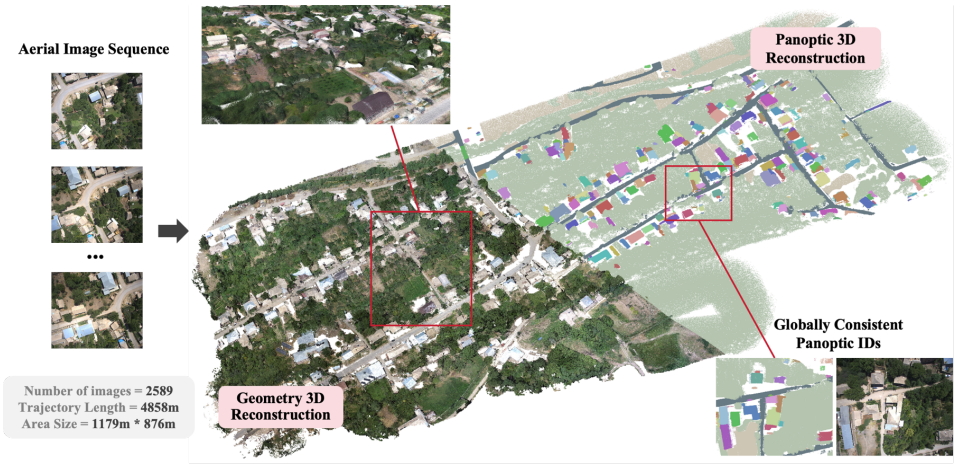


Figure 1: **Joint geometry and panoptic 3D reconstruction on a large-scale scene.** Our method reconstructs geometrically coherent 3D scenes while simultaneously producing panoptic predictions with consistent object identities across long spatial ranges.

ABSTRACT

Large-scale panoptic 3D reconstruction from aerial imagery requires globally consistent geometry and persistent object identities across distant views. Existing feed-forward methods struggle with both: depth inconsistencies distort local reconstructions, repetitive layouts confound loop closure, and cross-view instance supervision is scarce. We present Lap3R, a framework that combines semantic-prior geometry rectification, spatially and temporally verified loop mining, and cross-view co-visibility mask linking. Lap3R first rectifies local geometric distortions using semantic priors. It then verifies long-range loop candidates to retain genuine revisits while rejecting false matches caused by repetitive aerial layouts. To establish persistent instance identities, Lap3R aggregates mask agreement across co-visible views, without temporal tracking or manual cross-view annotations. On large scenes, Lap3R achieves the lowest trajectory error and Chamfer distance among the evaluated methods, while substantially improving semantic and instance reconstruction over feed-forward baselines. We further use Lap3R to generate pseudo-labels from aerial videos; a model trained solely on these labels surpasses the evaluated panoptic baselines on all reported metrics.

1 INTRODUCTION

Large-scale panoptic 3D reconstruction aims to recover a globally consistent scene from a long image sequence while assigning each reconstructed point a semantic category and an instance identity. For aerial imagery, this representation can support applications such as urban digital twins, asset inventory, and autonomous UAV inspection, where understanding *what* a structure is and *which* structure it is matters alongside its geometry. Feed-forward models have substantially accelerated multi-view geometry prediction (Cabon et al., 2025; Wang et al., 2025a; 2026b), and recent work has begun to integrate panoptic scene understanding into reconstruction (Žust et al., 2025; Li et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2026; Barberteguy et al., 2026; Koch et al., 2026). Yet extending these advances from bounded image sets to long aerial sequences requires consistency over spatial ranges far beyond those seen within a single model input.

The difficulty is not confined to one stage of reconstruction. At the local scale, chunk-based methods (Deng et al., 2025; Zhang et al., 2026) reconstruct overlapping views independently, but depth disagreements across views can distort a chunk before global alignment begins (Fig. 3). At the scene scale, aligning chunks requires long-range loop constraints, yet repetitive roads and building layouts make appearance-based retrieval prone to false matches, while conservative filtering can miss genuine revisits (Fig. 4). Even an aligned reconstruction does not complete the panoptic task: an object must retain its identity across observations that may be far apart in the sequence. Establishing these identities is particularly difficult in aerial imagery, where cross-view instance annotations are scarce and per-image masks may split one object, merge several objects, or fail to associate an object after a long gap. Large-scale panoptic reconstruction therefore calls for consistency at three connected levels: reliable geometry within chunks, trustworthy correspondences between chunks, and persistent instance identities across views.

We present **Lap3R**, a framework that addresses these requirements through geometric rectification, verified global alignment, and cross-view instance association. Its geometric component, *semantic-prior rectification and verified loop mining* (SRLM), first uses predicted ground semantics to identify structural support for correcting distorted local reconstructions. A semantic-aware as-rigid-as-possible deformation propagates the correction while preserving local structure. SRLM then retrieves potential long-range revisits and retains loop constraints only when they pass spatial feature verification and a temporal-neighborhood check. This separates the broad search for possible revisits from the more stringent decision of which constraints should influence global alignment.

For panoptic consistency, we introduce *cross-view co-visibility mask linking* (CCML). CCML measures whether two masks are repeatedly associated with the same target mask across views in which they are both visible. It groups associations supported by this multi-view evidence and filters masks that persistently show one-to-many overlaps, which could otherwise merge distinct objects. We use the resulting cross-view associations to construct instance supervision without manual cross-view instance annotations. At inference time, the trained model predicts local panoptic reconstructions, SRLM aligns their geometry, and CCML links their local instance identities across chunks. The same linking procedure can also use predicted geometry to generate pseudo-panoptic labels from RGB-only aerial videos.

Experiments on UAVScenes and UrbanBIS show that Lap3R achieves the lowest absolute trajectory error and Chamfer distance among the evaluated methods across all six scenes. It also improves semantic and instance reconstruction over the evaluated feed-forward panoptic baselines. Furthermore, a separately trained model using only pseudo-labels generated from unlabeled RGB aerial videos surpasses these baselines on all reported panoptic metrics, indicating that the proposed association procedure can provide useful supervision beyond geometry-annotated sequences.

Our contributions are threefold:

- We formulate large-scale aerial panoptic reconstruction around consistency within local geometry, across global alignments, and across instance observations, and present Lap3R as an integrated framework to address these requirements.
- We develop SRLM to rectify local geometry using semantic ground priors and verify long-range loop constraints, together with CCML to establish persistent instance identities from repeated agreement across co-visible views.
- We demonstrate improved geometric and panoptic reconstruction on large aerial scenes and show that CCML can produce useful pseudo-panoptic supervision from RGB-only videos.

2 Related Work

2.1 Scalable Feed-Forward 3D Reconstruction

Feed-forward models replace correspondence search, calibration, and iterative optimization with direct neural prediction, from pairwise pointmap regression (Wang et al., 2024; Leroy et al., 2024) to joint multi-view prediction of cameras, depth, and point maps (Cabon et al., 2025; Yang et al., 2025;

Wang et al., 2025a; 2026a), but attending over all views at once confines them to short sequences. Streaming methods bound the cost by compressing past observations into a recurrent state or spatial memory (Wang & Agapito, 2025; Wang et al., 2025b; Chen et al., 2025; Wu et al., 2025), a causal attention cache (Chen et al., 2026b; Zhuo et al., 2026; Lan et al., 2026; Chen et al., 2026a), or a test-time-trained representation (Elflein et al., 2026), at the price of lost geometric detail and drift over long trajectories. Chunk-based methods instead reconstruct overlapping chunks and register them into a global map with $\mathrm{Sim}(3)$ or non-rigid alignment and loop closure (Deng et al., 2025; Cheng et al., 2026; Liu et al., 2025; Maggio et al., 2025; Maggio & Carlone, 2026; Xie et al., 2026; Zhang et al., 2026). They differ in the alignment model, but all take the chunk reconstructions as given and propose loops by appearance retrieval (Izquierdo & Civera, 2024); both assumptions break in aerial scenes, where depth inconsistency distorts each chunk before alignment and repetitive layouts induce false loops. We rectify each chunk with semantic priors before alignment and verify loop candidates spatially and temporally. All of these methods also reconstruct geometry only, without semantic categories or persistent instance identities.

2.2 Panoptic 3D Reconstruction

Recent work integrates scene understanding directly into feed-forward reconstruction. PE3R (Hu et al., 2026) lifts SAM and CLIP embeddings onto DUSt3R-style geometry without training, which yields open-vocabulary semantic point clouds but neither a learned panoptic decoder nor persistent instance identities. PanSt3R (Žust et al., 2025) and Pano3D (Barberteguy et al., 2026) predict geometry and panoptic segmentation jointly, and IGGT (Li et al., 2026) and UNITE (Koch et al., 2026) learn instance-aware reconstruction features that support semantic, instance, and open-vocabulary queries. These models are developed for bounded scenes observed over short ranges, and none of them maintains instance identities across the distant views of a large-scale aerial sequence.

Extending these models to large-scale scenes is further constrained by the availability of panoptic supervision. PanSt3R and Pano3D rely on densely annotated semantic and instance datasets, which are scarce for outdoor scenes and rarely carry identities that persist across distant views. PE3R, IGGT and UNITE derive labels from foundation-model predictions through propagation, geometric lifting, and clustering, an offline pipeline whose cost and fragility grow with the number of images. We instead construct panoptic supervision by linking per-view foundation-model masks across views through co-visibility, which yields semantic and instance labels that stay consistent over long sequences without manual instance annotation.

3 Methods

3.1 Overview

We formulate local panoptic 3D reconstruction as a unified multi-view prediction problem, where scene geometry, semantic categories, and cross-view consistent instance identities are jointly recovered from the input views. We instantiate this formulation with a VGGT-based architecture and a multi-view panoptic prediction head adapted from PanSt3R (Žust et al., 2025).

Given a long aerial image sequence $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{|\mathcal{I}|}$, we partition it into a set of overlapping chunks $\{\mathcal{C}_k\}_{k=1}^{K}$, where adjacent chunks share a subset of input views. The model predicts depth maps $\mathbf{D} \in \mathbb{R}^{N\times H\times W}$, camera intrinsics $\mathbf{K}$, camera poses $\mathbf{T}$, semantic segmentation masks $\mathbf{M}^{\mathrm{cls}} \in \{1,\dots,C\}^{N\times H\times W}$, and instance segmentation masks $\mathbf{M}^{\mathrm{inst}} \in \{1,\dots,m\}^{N\times H\times W}$ for an input chunk. Let $\pi(\cdot;\mathbf{K},\mathbf{T})$ denote the camera projection. We then obtain the point cloud $\mathcal{P}_k = \{\mathbf{p}_i\}$ by unprojecting the depth maps $\mathbf{D}$ using $\pi^{-1}(\cdot;\mathbf{K},\mathbf{T})$.

Our framework follows a training-to-inference pipeline for large-scale panoptic 3D reconstruction (Fig. 2). We first construct cross-view consistent panoptic supervision using cross-view co-visibility mask linking (CCML), which associates two masks as the same instance when they are consistently matched across their co-visible views (Sec. 3.2). The resulting labels are then used to train the model to predict local geometry and panoptic outputs from image chunks. At inference, we divide the input image sequence into overlapping chunks and reconstruct each chunk independently with the trained model. We then apply semantic-prior rectification and verified loop mining (SRLM) to rectify the geometry within each chunk and establish reliable long-range constraints through spatial

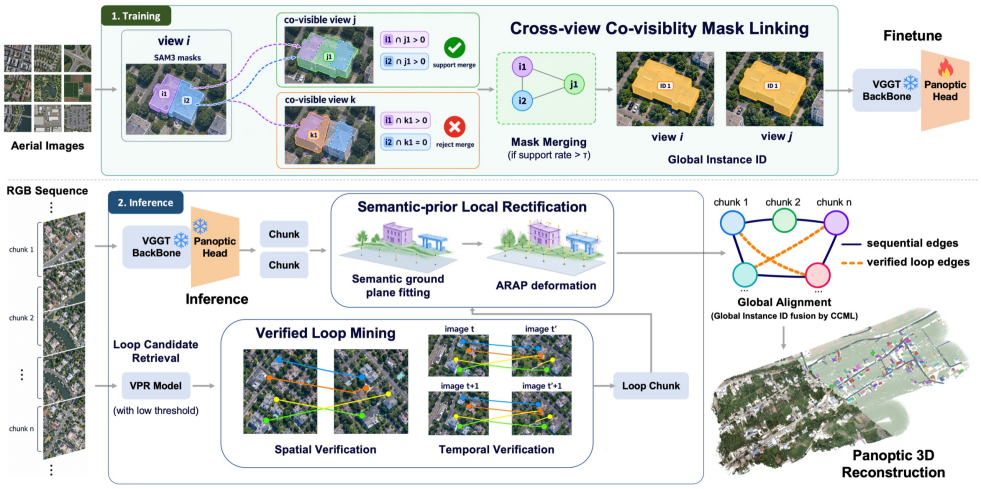


Figure 2: **Method overview.** During training, CCML evaluates mask consistency across co-visible views, merges masks with sufficient cross-view support, and assigns globally consistent instance IDs for panoptic-head finetuning. During inference, semantic-prior local rectification corrects local chunk distortion, while verified loop mining combines VPR-based retrieval with spatial and temporal verification to obtain reliable long-range constraints. Global alignment of the rectified chunks yields the final large-scale panoptic 3D reconstruction.

and temporal verification, yielding a globally aligned reconstruction (Sec. 3.3). Finally, CCML is applied across overlapping chunks to associate instance identities, producing the final large-scale panoptic 3D reconstruction. Details of training and inference are provided in Appendix A.

3.2 Cross-View Co-Visibility Mask Linking

To map local instance IDs to globally consistent ones, we propose cross-view co-visibility mask linking (CCML), which turns a set of instance masks into panoptic labels with cross-view consistent instance IDs. SAM3 (Carion et al., 2026) segments each image $\mathbf{I}_i$ into a set of instance masks $\mathcal{M}_i = \{\mathbf{M}_{i,a}\}_{a=1}^{A_i}$, each with a semantic label. Each mask $\mathbf{M}_{i,a}$ is back-projected once into a 3D point set $\mathcal{P}_{i,a}^{\text{mask}}$ using the ground-truth depth $\mathbf{D}_i^{\text{gt}}$, intrinsics $\mathbf{K}_i^{\text{gt}}$, and pose $\mathbf{T}_i^{\text{gt}}$, and warped to any view j through $\pi_j(\mathcal{P}_{i,a}^{\text{mask}}) = \pi(\mathcal{P}_{i,a}^{\text{mask}}; \mathbf{K}_j^{\text{gt}}, \mathbf{T}_j^{\text{gt}})$.

CCML reasons about one pair of masks at a time. The co-visible view set $\mathcal{O}(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'})$ of two masks $\mathbf{M}_{i,a}$ and $\mathbf{M}_{i',a'}$ contains the views in which both masks have sufficient valid projected support. For each co-visible view $j \in \mathcal{O}(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'})$, we project both masks into view j and compare them with the masks in $\mathcal{M}_j$. For a target mask $\mathbf{M}_{j,b} \in \mathcal{M}_j$, we define

$$r_{(i,a)\to(j,b)} = \frac{\left|\pi_j(\mathcal{P}_{i,a}^{\text{mask}}) \cap \mathbf{M}_{j,b}\right|}{\left|\pi_j(\mathcal{P}_{i,a}^{\text{mask}})\right|}, \tag{1}$$

which measures the fraction of the projected mask covered by $\mathbf{M}_{j,b}$. We regard $\mathbf{M}_{i,a}$ and $\mathbf{M}_{j,b}$ as associated when $r_{(i,a)\to(j,b)} \geq \tau_{\text{assoc}}$. View j then supports $\mathbf{M}_{i,a}$ and $\mathbf{M}_{i',a'}$ as the same instance when one target mask $\mathbf{M}_{j,b}$ is associated with both of them. Let $\mathcal{S}(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'}) \subseteq \mathcal{O}(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'})$ collect these supporting views. We only consider mask pairs with $|\mathcal{O}(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'})| \geq \tau_{\text{vis}}$, ensuring sufficient co-visible evidence for reliable consistency estimation. The cross-view co-association rate is then defined as

$$c(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'}) = \frac{|\mathcal{S}(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'})|}{|\mathcal{O}(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'})|}. \tag{2}$$

It measures the fraction of co-visible views that agree, so a high c indicates consistent support across multiple viewpoints rather than an accidental match in a single view.

An under-segmented mask covers several adjacent objects, so it can serve as a common target mask for the masks of all these objects and wrongly link them into one instance. Such a mask overlaps

several distinct masks after being warped into other visible views. We therefore classify a mask as under-segmented when this one-to-many overlap occurs in more than a fraction τ_{split} of its visible views, and exclude it from linking; we refer to this step as under-segmentation filtering.

The remaining masks are grouped into instances on a graph whose nodes are masks and whose edges connect $\mathbf{M}_{i,a}$ and $\mathbf{M}_{i',a'}$ if $c(\mathbf{M}_{i,a}, \mathbf{M}_{i',a'}) \geq \tau_{\text{con}}$. We perform iterative graph updating and merging to obtain globally consistent instances. We first extract connected components by depth-first search and merge the masks within each component into a new mask node. We then update the graph over these new nodes by reapplying the same cross-view association and consistency criterion. Newly connected components are merged again, followed by another graph update. This procedure is repeated until no further merges occur. All original masks represented by each final node are assigned the same instance ID while retaining their semantic labels.

Extension to RGB-only sequences. We further explore extending CCML to RGB-only aerial sequences through teacher–student pseudo-labeling. We first train a teacher model on sequences with ground-truth geometry using the CCML supervision described above. Given an unlabeled RGB sequence, the teacher predicts its geometry together with panoptic segmentation. We then apply CCML using the predicted geometry in place of ground-truth depth and camera poses to associate the predicted masks across views, yielding sequence-level pseudo-panoptic labels. Finally, a fresh student model is trained solely on these pseudo-labels, enabling the framework to leverage additional unlabeled aerial data.

3.3 Large-Scale Geometric Reconstruction

At inference, semantic-prior rectification and verified loop mining (SRLM) align the independently reconstructed local chunks into a globally consistent scene. This alignment faces two complementary sources of error: depth inconsistency distorts individual chunks (Fig. 3), while repetitive layouts make appearance-based VPR retrieval prone to false loop proposals (Fig. 4). SRLM addresses these issues in sequence: semantic-prior local rectification corrects the geometry within each chunk (Sec. 3.3.1), and verified loop mining filters long-range candidates using spatial and temporal consistency to obtain reliable alignment constraints (Sec. 3.3.2).

3.3.1 Local Geometry Rectification with Semantic Priors

Aerial scenes provide a strong structural prior through their dominant ground surfaces, which are expected to remain approximately planar even in visually weakly textured regions. For each local point cloud $\mathcal{P}_k$, we first extract reliable ground points $\mathcal{G}_k \subset \mathcal{P}_k$ using $\mathbf{M}^{\text{cls}}$ and estimate a dominant plane $\mathcal{H}_k = (\mathbf{n}_k, d_k)$ through robust weighted fitting,

$$\min_{\mathbf{n}_k, d_k} \sum_{\mathbf{p}_i \in \mathcal{G}_k} w_i \, \rho_\delta\big(\mathbf{n}_k^\top \mathbf{p}_i + d_k\big), \qquad \|\mathbf{n}_k\|_2 = 1, \tag{3}$$

where w_i denotes the prediction confidence and ρ_δ is the Huber loss used to suppress unreliable ground observations. To propagate these corrections to entire point cloud $\mathcal{P}_k$ while preserving object-level geometry, we construct a semantic-aware ARAP deformation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over $\mathcal{P}_k$. We partition $\mathcal{P}_k$ into semantic regions according to $\mathbf{M}^{\text{cls}}$ and sparsely sample control nodes $\mathcal{V} = \{\mathbf{g}_i\}$ within each region, with $\mathcal{E}$ connecting neighboring control nodes. Each node is associated with a local rigid transformation $(\mathbf{R}_i, \mathbf{t}_i)$, where $\mathbf{R}_i \in SO(3)$, and its deformed position is $\tilde{\mathbf{g}}_i = \mathbf{g}_i + \mathbf{t}_i$. Nodes within the same semantic region are coupled more strongly, while weaker connections are retained across semantic boundaries to preserve scene-level continuity without over-constraining different objects. We jointly optimize the semantic-aware ARAP deformation as

$$\begin{aligned} \min_{\{\mathbf{R}_i, \mathbf{t}_i\}} \quad & \sum_{(i,j)\in\mathcal{E}} \lambda_{s_i,s_j} w_{ij} \, \|(\tilde{\mathbf{g}}_j - \tilde{\mathbf{g}}_i) - \mathbf{R}_i(\mathbf{g}_j - \mathbf{g}_i)\|_2^2 \\ \text{s.t.} \quad & \tilde{\mathbf{g}}_i = \hat{\mathbf{g}}_i, \qquad i \in \mathcal{V}_{\text{ground}}, \\ & \tilde{\mathbf{g}}_i = \mathbf{g}_i + \mathbf{t}_i, \qquad \mathbf{R}_i \in SO(3), \end{aligned} \tag{4}$$

where λ_{s_i,s_j} assigns stronger rigidity to nodes within the same semantic region and weaker coupling across semantic boundaries. For each ground control node, $\hat{\mathbf{g}}_i$ denotes its orthogonal projection onto $\mathcal{H}_k$, given by $\hat{\mathbf{g}}_i = \mathbf{g}_i - (\mathbf{n}_k^\top \mathbf{g}_i + d_k)\mathbf{n}_k$.

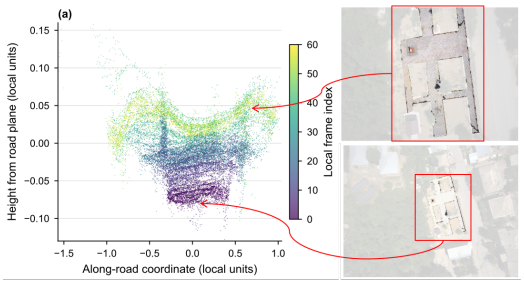


Figure 3: **Depth inconsistency in a local chunk.** Points are colored by local frame index, revealing depth inconsistency on the same road plane across frames.



Figure 4: **False loop retrieved by VPR.** Similar appearances across different locations cause false VPR retrievals in repetitive aerial scenes.

Finally, the optimized control-node transformations are propagated to the dense point cloud $\mathcal{P}_k$. For each dense point $\mathbf{p}$, we define $\mathcal{N}_s(\mathbf{p}) \subseteq \{1, \ldots, M\}$ as the indices of its three nearest control nodes belonging to the same semantic region. The final deformation is obtained by blending their local rigid motions,

$$\mathcal{T}(\mathbf{p}) = \sum_{i \in \mathcal{N}_s(\mathbf{p})} \alpha_i(\mathbf{p}) \left[\mathbf{R}_i(\mathbf{p} - \mathbf{g}_i) + \tilde{\mathbf{g}}_i\right], \tag{5}$$

where $\alpha_i(\mathbf{p})$ denotes the normalized distance-based weight associated with control node $\mathbf{g}_i$. Applying this deformation to $\mathcal{P}_k$ yields the rectified local reconstruction $\tilde{\mathcal{P}}_k$.

3.3.2 Verified Loop Mining

Given the rectified local point clouds $\{\tilde{\mathcal{P}}_k\}_{k=1}^{K}$, we first estimate relative $\mathrm{Sim}(3)$ transformations between adjacent chunks from their overlapping views. To reduce long-range drift, we retrieve loop candidates with a permissive VPR threshold and retain only those that pass spatial and temporal verification. For a candidate frame pair $(\mathbf{I}_{t_1}, \mathbf{I}_{t_2})$, we match RootSIFT features using bidirectional ratio and mutual-consistency tests, estimate a robust fundamental matrix, and require sufficient geometrically consistent inliers. We further require at least one synchronized neighboring pair, $(\mathbf{I}_{t_1-1}, \mathbf{I}_{t_2-1})$ or $(\mathbf{I}_{t_1+1}, \mathbf{I}_{t_2+1})$, to satisfy the same spatial test. Thus, retrieval provides recall while geometric and temporal consistency determines which loops are accepted.

We group the verified frame pairs by their source chunks. For each verified chunk pair $(\mathcal{C}_A, \mathcal{C}_B)$, we form an auxiliary loop chunk from the matched views, adding randomly sampled views when necessary to reach the predefined chunk size. We estimate $\mathrm{Sim}(3)$ transformations from the loop chunk to both source chunks and use them to derive a long-range relative constraint between $\mathcal{C}_A$ and $\mathcal{C}_B$. We then jointly optimize the sequential and loop constraints with pose graph optimization, following VGGT-Long (Deng et al., 2025) for global alignment. Finally, we apply CCML across the globally aligned chunks to associate local instance masks with globally consistent instance IDs, yielding the final large-scale 3D panoptic reconstruction.

4 Experiments

4.1 Experimental Setup

Datasets. For training, we use UAVFF3D (Yang et al., 2026) to construct panoptic supervision. For evaluation, we use four large-scale aerial scenes from UAVScenes (Wang et al., 2025c) (AMtown01, AMvalley01, HKairport01, and HKisland01), captured primarily from near-orthographic views, and two scenes from UrbanBIS (Yang et al., 2023) (Longhua and Yingrenshi), which contain more oblique aerial views. Together, they cover diverse urban environments and viewing configurations. We further generate pseudo-panoptic labels on RGB-only videos from A2Seek (Mo et al., 2025) with the teacher–student procedure of Sec. 3.2, using *Ours* as the teacher, and denote the student trained solely on these labels as *Ours-pseudo*.

Baselines. For large-scale geometric reconstruction, we compare with two families of methods in Tab. 1. Streaming methods include CUT3R (Wang et al., 2025b), TTT3R (Chen et al.,

Table 1: Quantitative comparison of large-scale geometric reconstruction. We report ATE (m) and CD (m) for streaming and chunk-based methods, with lower values indicating better performance. **Bold** and underlined numbers denote the best and second-best results, respectively.

	UAVScenes								UrbanBIS					
	AMtown		AMvalley		HKairport		HKisland		Longhua		Yingrenshi		Avg.	
Method	ATE ↓	CD ↓	ATE ↓	CD ↓	ATE ↓	CD ↓	ATE ↓	CD ↓	ATE ↓	CD ↓	ATE ↓	CD ↓	ATE ↓	CD ↓
Streaming Methods														
CUT3R	270.84	30.11	248.94	46.30	113.81	30.32	145.78	26.86	274.75	28.29	181.65	33.10	205.96	32.50
TTT3R	258.79	8.74	234.03	49.89	61.59	25.87	128.69	32.42	245.79	47.93	67.37	14.95	166.04	29.97
LingBot-Map	97.97	11.45	85.83	22.64	14.53	10.08	41.96	11.24	119.77	25.91	63.47	14.44	70.59	15.96
Chunk-based Methods														
VGGT-Long	45.20	8.13	83.91	25.70	8.39	6.18	27.06	21.83	91.80	30.41	79.59	23.47	55.99	19.29
MERG3R	269.79	50.00	116.90	11.09	9.28	3.36	34.53	10.90	288.06	35.02	141.02	26.19	143.26	22.76
Scal3R	53.31	6.15	158.06	39.42	5.76	5.94	33.77	17.14	70.65	30.84	33.24	15.51	59.13	19.17
VGGT-SLAM 1.0	287.06	44.40	258.00	46.57	121.10	44.27	146.52	42.58	334.10	50.00	194.32	49.95	223.52	46.30
VGGT-SLAM 2.0	260.20	10.00	230.82	50.00	110.42	49.84	91.91	48.45	310.47	44.15	126.22	49.00	188.34	41.91
TALO	203.20	8.31	55.60	14.62	12.15	7.93	67.88	19.88	47.49	26.66	48.18	15.43	72.42	15.47
Ours	**8.64**	**5.27**	**25.65**	**10.58**	**5.73**	**3.33**	**9.61**	**10.24**	**37.18**	**5.86**	**29.80**	**13.96**	**19.44**	**8.21**

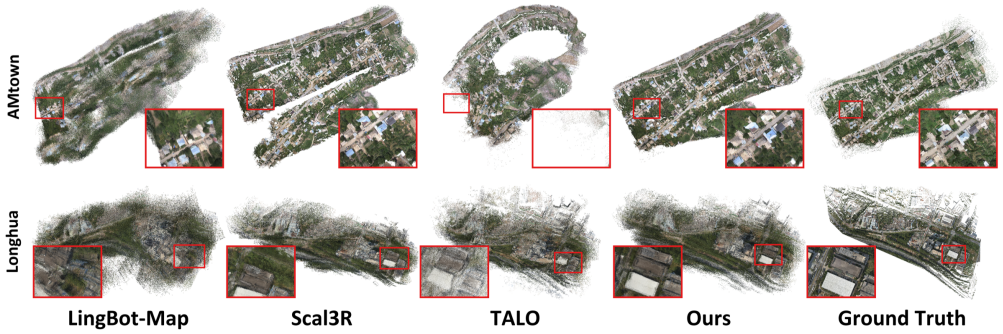


Figure 5: **Qualitative point cloud reconstruction.** Compared with other methods, our method produces more globally consistent scene geometry with reduced drift and local chunk distortion, yielding reconstructions that more closely match the ground-truth structure.

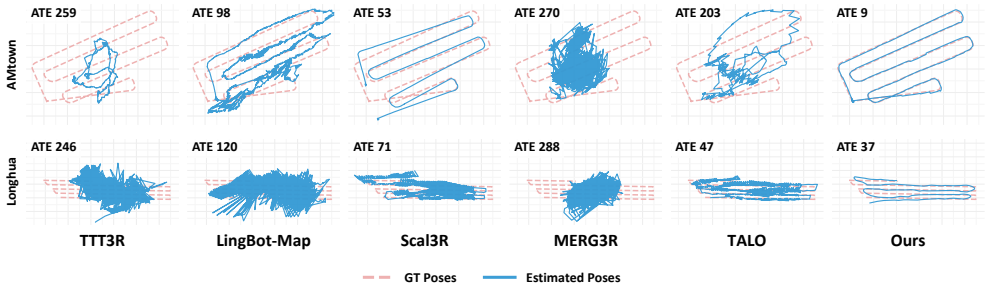


Figure 6: **Camera trajectory comparison.** Compared with prior methods, our approach more faithfully recovers the global trajectory structure and maintains better consistency with GT.

2026b), and LingBot-Map (Chen et al., 2026a), which process long sequences with a persistent sequential state. Chunk-based reconstruction methods include VGGT-Long (Deng et al., 2025), MERG3R (Cheng et al., 2026), Scal3R (Xie et al., 2026), VGGT-SLAM 1.0 (Maggio et al., 2025), VGGT-SLAM 2.0 (Maggio & Carlone, 2026), and TALO (Zhang et al., 2026), which recover overlapping local windows and then solve for a global alignment. All chunk-based methods use a chunk size of 60 frames with a 30-frame overlap. For panoptic reconstruction, we compare semantic feed-forward models including PE3R (Hu et al., 2026), PanSt3R (Žust et al., 2025), and IGGT (Li et al., 2026). To avoid an unfair advantage from adapting the reconstruction backbone to aerial data, we freeze all pretrained components and train only the newly added panoptic head.

Metrics and Protocol. Following prior works (Deng et al., 2025; Zhang et al., 2026), we evaluate geometric reconstruction using Absolute Trajectory Error (ATE) and Chamfer Distance (CD). CD

Table 2: **Quantitative comparison of panoptic reconstruction.** Our method consistently outperforms existing approaches on both datasets, while *Ours-pseudo* remains competitive and surpasses prior methods on all metrics. **Bold** and underlined numbers denote the best and second-best results.

	UAVScenes					UrbanBIS				
Method	mIoU_{3D}	mAcc_{3D}	AP	AP_{50}	AP_{25}	mIoU_{3D}	mAcc_{3D}	AP	AP_{50}	AP_{25}
PE3R	6.13	13.29	2.35	4.71	10.03	6.33	7.85	0.07	0.21	1.25
IGGT	4.63	17.39	0.35	0.35	0.35	10.62	22.28	0.16	0.56	1.72
PanSt3R	28.80	42.24	5.97	7.04	9.27	12.19	24.49	0.11	0.86	1.88
Ours	**41.29**	**58.94**	**25.33**	**39.72**	**47.82**	**27.27**	**38.17**	**8.20**	**12.32**	**21.71**
Ours-pseudo	34.03	57.80	9.95	16.50	29.36	26.27	36.64	2.59	7.44	18.88

is computed from bidirectional nearest-neighbor errors clipped at a truncation distance of 50 m. For panoptic reconstruction, we follow UNITE (Koch et al., 2026) to report mIoU_{3D} and mAcc_{3D} for semantic segmentation, and AP, AP_{50}, and AP_{25} for instance segmentation. More details are provided in the appendix.

4.2 Large-Scale Geometric Reconstruction

Quantitative Comparison. As shown in Table 1, our method consistently achieves the best ATE and CD across all evaluated scenes on both datasets. Streaming methods generally exhibit large trajectory errors on these long sequences, reflecting the difficulty of preventing accumulated drift through purely local propagation. Chunk-based methods such as VGGT-Long, Scal3R, and TALO substantially improve over these approaches, but their performance remains scene-dependent, and low trajectory error does not always translate into accurate reconstructed geometry. In contrast, our method consistently improves both metrics, indicating that the proposed geometric stabilization not only reduces global drift but also preserves the consistency of the reconstructed scene structure.

Qualitative Results. Figure 5 shows two major failure modes. Streaming methods accumulate drift, causing severe misregistration: LingBot-Map loses the street structure in AMtown and merges buildings with the terrain in Longhua. Chunk-based methods better preserve the scene but still suffer from local distortion and weak loop constraints: TALO curls AMtown into a ring, while Scal3R fragments AMtown and thickens structures in Longhua. Our method avoids both failure modes, preserving the global layout and separable building structures.

The camera trajectories (Fig. 6) provide a complementary view of global alignment quality. Streaming methods (TTT3R and LingBot-Map) accumulate severe drift over long sequences, often collapsing the trajectory into a compact region. Chunk-based approaches (Scal3R and TALO) better preserve the overall flight layout but still suffer from inaccurate or insufficient long-range loop constraints, leading to substantial trajectory deformation in some scenes. In contrast, our method maintains consistent long-range alignment across both AMtown and Longhua, producing trajectories that closely follow the ground truth with substantially reduced drift.

4.3 Panoptic Reconstruction

Quantitative Comparison. Table 2 compares panoptic reconstruction performance. Our method consistently improves both semantic and instance reconstruction over existing feed-forward approaches. On UAVScenes, 3D mIoU increases from 28.80 to 41.29, while AP_{50} improves from 7.04 to 39.72 over the strongest baseline, showing particularly clear gains in instance-level reconstruction. Similar improvements are observed on UrbanBIS, indicating that the proposed supervision remains effective across both near-orthographic and oblique aerial views. Notably, *Ours-pseudo* also surpasses prior methods on all metrics, showing that pseudo-labels from unlabeled aerial video alone can train a panoptic model that outperforms prior methods.

Qualitative results. Figure 7 compares semantic reconstruction qualitatively. PE3R produces incomplete semantic layouts with substantial unlabeled regions, while PanSt3R exhibits noticeable semantic confusion and inconsistent object boundaries. In contrast, our method recovers a more complete semantic layout with cleaner building boundaries and predictions that more closely match the ground truth.

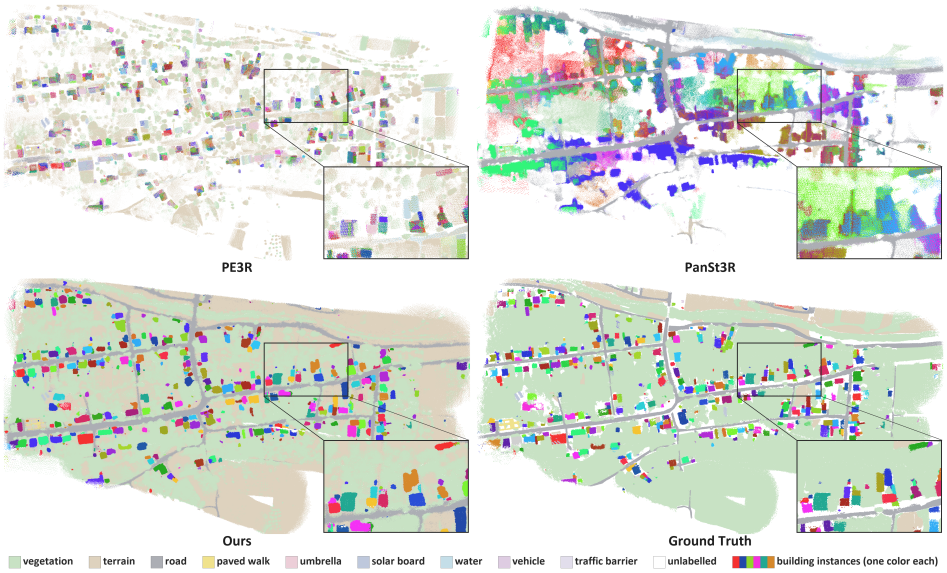


Figure 7: **Qualitative semantic reconstruction.** Following UNITE-style evaluation, predictions are remapped to the ground-truth point cloud to avoid confounding effects from reconstruction errors. Compared with existing methods, our method produces a more complete semantic layout with cleaner object boundaries and better agreement with the ground truth.

4.4 Ablation Studies and Analysis

Geometric components. Table 3 shows that both local rectification and verified loop mining are essential to reconstruction quality. Removing temporal verification degrades the reconstruction, while removing both spatial and temporal verification causes a much larger failure, confirming that appearance-based loop retrieval alone is unreliable in repetitive aerial scenes. The rectification ablations further show that semantic ground priors and ARAP play complementary roles: the former corrects local chunk distortion, while the latter preserves local geometric structure.

Semantic components. Table 4 directly evaluates CCML-generated labels on Longhua against ground-truth instances, isolating label quality from training. Both components contribute, with iterative updating having the larger effect. Under-segmentation filtering removes masks spanning multiple objects and prevents erroneous instance merging. Iterative updating further merges fragmented instances by pooling co-visible evidence across already linked masks, revealing associations that are too weak at the pairwise level.

Table 3: Ablation on SRLM.

Configuration	ATE↓	CD↓
Full SRLM	**37.18**	**5.86**
w/o Temporal	54.66	15.72
w/o Spatial & Temporal	160.10	35.90
w/o ARAP	47.93	8.12
w/o Semantic & ARAP	73.70	10.84

Table 4: Ablation on CCML.

Configuration	AP_{50} ↑	AP_{25} ↑
Full CCML	34.76	42.92
w/o Under-seg. Filtering	31.50	38.98
w/o Iterative Updating	27.81	35.63

5 Conclusion

We presented Lap3R to address the coupled geometric and panoptic-consistency challenges in large-scale aerial reconstruction. We propose SRLM to stabilize local geometry and global alignment, while CCML constructs persistent instance identities without temporal tracking. On large-scale aerial scenes, Lap3R achieves the lowest ATE and Chamfer distance and improves semantic and instance reconstruction over existing feed-forward methods. Moreover, we explore using Lap3R to generate pseudo-labels from unlabeled RGB videos for training a separate model, which also outperforms prior methods on all reported panoptic metrics. These results show that Lap3R can scale panoptic supervision beyond geometry-annotated sequences, reducing dependence on costly cross-view annotation and enabling broader use of aerial video. One limitation is that the current method assumes sufficiently visible ground structure and cross-view support for rectification and mask linking.

AI USE DISCLOSURE

We used generative AI tools to assist with the design and implementation of the proposed method, and to improve the readability and LaTeX formatting of this manuscript. All AI-assisted code, methodological suggestions, and manuscript text were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Victor Barberteguy, Ahmet Iscen, Mathilde Caron, Alireza Fathi, Gül Varol, and Cordelia Schmid. Pano3D: Unified 3D reconstruction and panoptic segmentation. In *European Conference on Computer Vision*, 2026.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *International Conference on Learning Representations*, 2026b.

Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. LONG3R: Long sequence streaming 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5273–5284, 2025.

Leo Kaixuan Cheng, Abdus Shaikh, Ruofan Liang, Zhijie Wu, Yushi Guan, and Nandita Vijaykumar. MERG3R: A divide-and-conquer approach to large-scale neural visual geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28969–28978, 2026.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. VGGT-Long: Chunk it, loop it, align it—pushing VGGT's limits on kilometer-scale long RGB sequences. *arXiv preprint arXiv:2507.16443*, 2025.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3D reconstruction at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jie Hu, Shizun Wang, and Xinchao Wang. PE3R: Perception-efficient 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26877–26887, 2026.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Sebastian Koch, Johanna Wald, Hidenobu Matsuki, Pedro Hermosilla, Timo Ropinski, and Federico Tombari. Unified semantic transformer for 3D scene understanding. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=eB7oHCJzud.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *International Conference on Learning Representations*, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024.

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, and Ziwei Liu. IGGT: Instance-grounded geometry transformer for semantic 3D reconstruction. In *International Conference on Learning Representations*, 2026.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Dominic Maggio and Luca Carlone. VGGT-SLAM 2.0: Real-time dense feed-forward scene reconstruction. In *Robotics: Science and Systems*, 2026.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the $\mathrm{SL}(4)$ manifold. In *Advances in Neural Information Processing Systems*, 2025.

Mengjingcheng Mo, Xinyang Tong, Mingpi Tan, Jiaxu Leng, Jiankang Zheng, Yiran Liu, Haosheng Chen, Ji Gan, Weisheng Li, and Xinbo Gao. A2Seek: Towards reasoning-centric benchmark for aerial anomaly understanding. In *Advances in Neural Information Processing Systems (NeurIPS 2025), Datasets and Benchmarks Track*, 2025. URL `https://openreview.net/forum?id=apHeLNdl9g`.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21486–21499, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Sijie Wang, Siqi Li, Yawei Zhang, Shangshu Yu, Shenghai Yuan, Rui She, Quanjiang Guo, JinXuan Zheng, Ong Kang Howe, Leonrich Chandra, Shrivarshann Srijeyan, Aditya Sivadas, Toshan Aggarwal, Heyuan Liu, Hongming Zhang, Chujie Chen, Junyu Jiang, Lihua Xie, and Wee Peng Tay. UAVScenes: A multi-modal dataset for UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026b.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Advances in Neural Information Processing Systems*, 2025.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, and Xiaowei Zhou. Scal3R: Scalable test-time training for large-scale 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21760–21771, 2026.

Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang. UrbanBIS: A large-scale benchmark for fine-grained urban building instance segmentation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xiang Yang, Yongli Wang, HaiFeng Li, and Yunsheng Zhang. UAVFF3D: A geometry-aware benchmark for feed-forward UAV 3D reconstruction. *arXiv preprint arXiv:2605.17942*, 2026.

Fengyi Zhang, Tianjun Zhang, Kasra Khosoussi, Zheng Zhang, Zi Huang, and Yadan Luo. TALO: Pushing 3D vision foundation models towards globally consistent online reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21870–21879, 2026.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. In *International Conference on Learning Representations*, 2026.

Lojze Žust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. PanSt3R: Multi-view consistent panoptic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

A Training and Inference Details

A.1 Training

Panoptic head and input sampling. We attach a shared-query panoptic head to the pretrained VGGT reconstruction backbone. The head uses 300 object queries, a hidden dimension of 256, a mask dimension of 256, a feed-forward dimension of 2048, eight attention heads, and nine decoder layers. Each training sample contains at most eight views, with the number of views randomly varied between two and eight. Images are resized to 518×378 pixels before entering the model. The same query set is shared across all views in a sample, so a matched query represents one instance consistently across the views available to that sample.

Panoptic-head optimization. The constructed labels are used to train the panoptic head with sequence-level Hungarian matching. The matching cost combines classification, point-sampled mask binary cross-entropy, and Dice terms. The final decoder output and the nine auxiliary decoder outputs are supervised with the same loss:

$$\mathcal{L}_{\text{pan}} = \sum_{\ell=0}^{9} \left(2\mathcal{L}_{\text{cls}}^{(\ell)} + 5\mathcal{L}_{\text{bce}}^{(\ell)} + 5\mathcal{L}_{\text{dice}}^{(\ell)} \right). \tag{6}$$

Here, $\ell = 0$ denotes the final decoder output and $\ell = 1, \ldots, 9$ denote the auxiliary outputs. The classification term uses a class-balanced foreground average and assigns a weight of 0.1 to unmatched no-object queries. For each matched mask, the mask losses are computed on 12,544 sampled points. Point sampling uses an oversampling ratio of 3.0 and an importance-sampling ratio of 0.75.

Initialization and trainable parameters. The geometry and reconstruction components are initialized from the pretrained VGGT checkpoint and kept frozen during panoptic training. The 300-query panoptic head is initialized from the compatible OMG-Seg weights. Thus, the training updates only the panoptic head parameters. The geometric loss is disabled by setting its coefficient to zero, so the optimization target in these experiments is the panoptic objective above.

Pseudo-label student training. To evaluate whether the constructed supervision can be transferred to RGB-only data, we use a teacher–student procedure on A2Seek. The teacher is first trained with the geometry-assisted CCML supervision described above. It then processes 1000 unlabeled A2Seek sequences.

For each RGB-only chunk, the teacher predicts depth, camera parameters, semantic masks, and local instance masks. These predictions replace the ground-truth geometry used during the initial CCML construction. We back-project the predicted masks using the predicted depth and camera parameters, evaluate their co-visibility agreement, remove masks with persistent one-to-many overlaps, and iteratively merge the remaining mask nodes into sequence-level instance components. The resulting semantic labels and instance IDs are used as pseudo-panoptic supervision. Pixels without valid geometric support or a retained instance assignment are ignored for the corresponding supervision terms.

We construct a fixed manifest of 50,000 distinct within-chunk training samples using random seed 0. Each sample contains between two and eight views, with the number of views varied during training. The student is initialized from the original pretrained VGGT checkpoint together with the compatible OMG-Seg panoptic-head initialization. It does not load the teacher weights, reuse the geometry-annotated seed sequences, or receive ground-truth depth and camera parameters during training. The student is trained only on the A2Seek pseudo-labels; the geometry loss remains disabled and only the panoptic head is updated.

Optimization schedule. We use AdamW with learning rate 10^{-4}, minimum learning rate 10^{-6}, weight decay 0.05, and gradient-norm clipping at 250. The learning rate is warmed up for the first 1,000 optimizer steps, remains at 10^{-4} until step 15,000, and then follows a geometric decay to 10^{-6} at step 47,500. Training stops after 50,000 optimizer steps with gradient accumulation set to one. Mixed-precision distributed training is used with a per-GPU limit of 32 input images.

A.2 INFERENCE

At inference time, the trained model takes only an RGB sequence as input. We partition the sequence into overlapping chunks $\{\mathcal{C}_k\}_{k=1}^{K}$, keeping the frame order and the overlap between neighboring chunks. The following steps are applied in order.

Local chunk prediction. For each chunk $\mathcal{C}_k$, the model predicts depth, camera parameters, semantic masks, and instance masks,

$$(\widehat{\mathbf{D}}_k, \widehat{\mathbf{K}}_k, \widehat{\mathbf{T}}_k, \widehat{\mathbf{M}}_k^{\mathrm{cls}}, \widehat{\mathbf{M}}_k^{\mathrm{inst}}).$$

The depth and camera predictions are unprojected into a local point cloud $\widehat{\mathcal{P}}_k$. The panoptic decoder uses the shared query predictions across all views in the chunk, so a retained query receives one instance label throughout that chunk whenever it is visible. These labels are local: an integer in $\widehat{\mathbf{M}}_k^{\mathrm{inst}}$ has meaning only within chunk k, and the same integer in another chunk is not assumed to denote the same object. No cross-chunk instance state is used at this stage.

Local geometric rectification. We apply the semantic-prior rectification from Sec. 3.3.1 to each $\widehat{\mathcal{P}}_k$. Ground points selected from $\widehat{\mathbf{M}}_k^{\mathrm{cls}}$ define a robust dominant plane. The orthogonal plane targets are propagated through the semantic-aware ARAP deformation, producing a rectified local cloud $\widetilde{\mathcal{P}}_k$. The semantic and local instance labels remain attached to their corresponding points during this deformation.

Global geometric alignment with SRLM. We next apply verified loop mining (Sec. 3.3.2). Similar views are first retrieved with a permissive VPR threshold. A candidate pair is kept only when RootSIFT matching and robust epipolar verification provide enough spatial inliers and at least one temporally neighboring pair passes the same test. Adjacent-chunk constraints and the verified long-range loop constraints are jointly optimized as a pose graph in $\mathrm{Sim}(3)$. Let $\mathbf{G}_k \in \mathrm{Sim}(3)$ denote the optimized transform for chunk k. We obtain the globally aligned geometry

$$\mathcal{P}_k^{\mathrm{global}} = \mathbf{G}_k\big(\widetilde{\mathcal{P}}_k\big).$$

This step aligns the coordinate systems of the chunks; it does not yet change their local instance IDs.

Cross-chunk instance-ID alignment with CCML. After geometric alignment, we run cross-view co-visibility mask linking (CCML) on the local panoptic outputs. A node is a pair $u = (k, a)$, where k is a chunk index and a is one of its local instance IDs. The node stores the pixels and 3D points carrying that label, along with their semantic category. For two nodes u and v with the same semantic category, we determine the set $\mathcal{O}(u, v)$ of views in which both supports have sufficient valid projected points. For a target mask b in a co-visible view j, the support ratio is

$$r_{u\to(j,b)} = \frac{|\pi_j(\mathcal{P}^u) \cap \mathbf{M}_{j,b}|}{|\pi_j(\mathcal{P}^u)|}.$$

The view supports u and v as the same object when one target mask is associated with both projections, using $r_{u\to(j,b)} \geq \tau_{\mathrm{assoc}}$. Let $\mathcal{S}(u, v)$ be the supporting views. We discard pairs with fewer than τ_{vis} co-visible views and compute

$$c(u, v) = \frac{|\mathcal{S}(u, v)|}{|\mathcal{O}(u, v)|}.$$

An edge is added between u and v when $c(u, v) \geq \tau_{\mathrm{con}}$. Before adding edges, masks exhibiting persistent one-to-many overlaps are removed by the under-segmentation filter with threshold τ_{split}. Thus a mask covering several adjacent objects cannot merge those objects solely because it is a common target mask.

We then repeatedly update the graph. Connected components are merged into super-nodes, the cross-view association scores are recomputed for the new nodes, and newly connected components are merged again. Iteration stops when one pass creates no new merge. Every retained final component receives a new scene-level instance ID, and all of its source pairs (k, a) are mapped to that ID. For a point or pixel x from chunk k, the final instance label is therefore

$$\widehat{m}^{\mathrm{global}}(x) = \Gamma\big(k, \widehat{m}_k^{\mathrm{local}}(x)\big)\,,$$

where Γ is the component-to-ID mapping. Stuff labels and pixels without a retained thing instance keep instance ID zero. Masks removed by the under-segmentation filter do not create cross-chunk links.

Final panoptic reconstruction. Finally, we merge the transformed point clouds $\{\mathcal{P}_k^{\text{global}}\}$, carry over their semantic labels, and replace each chunk-local instance label by its mapped scene-level ID. The result is a globally aligned panoptic point cloud whose geometry is determined by SRLM and whose persistent instance identities are determined by CCML. The two operations are deliberately separated: SRLM aligns chunk coordinates, whereas CCML aligns the labels attached to those coordinates.

B Implementation Details of CCML and SRLM

B.1 CCML Parameters

We use fixed CCML thresholds for geometry-assisted labels and teacher-generated pseudo-labels. A projected mask counts as having sufficient projected support in a view when it occupies more than 5% of that view's area. We set the association threshold $\tau_{\text{assoc}} = 0.8$, the under-segmentation filtering threshold $\tau_{\text{split}} = 0.3$, the cross-view consensus threshold $\tau_{\text{con}} = 0.9$, and the minimum number of co-visible views $\tau_{\text{vis}} = 5$. Components containing fewer than two source masks are not assigned a new instance ID.

B.2 SRLM Parameters

For semantic-ground plane fitting, we use the semantic labels `car_park`, `paved_walk`, and `road`. At most 256 ground points are sampled per frame. Points must have confidence at least 10^{-5}, and a frame must provide at least 64 valid ground pixels. The plane is fitted with confidence-weighted Huber IRLS using $\delta = 1.345$ in ρ_δ, at most 20 iterations, and tolerance 10^{-6}. A semantic-ground alignment is used only when at least $10{,}000$ mutual ground correspondences are available; otherwise, the implementation falls back to raw 3D correspondences. Robust weighted $\mathrm{Sim}(3)$ estimation uses Huber delta 0.1, at most 5 iterations, and tolerance 10^{-9}.

For spatial loop verification, RootSIFT uses at most $8{,}192$ SIFT features per image, with contrast threshold 0.04, edge threshold 10, and $\sigma = 1.6$. Descriptor matching applies a Lowe ratio of 0.7 together with mutual consistency. Fundamental matrices are estimated with the automatic OpenCV backend, using an inlier threshold of 0.6 pixels, confidence 0.999, and at most $10{,}000$ RANSAC iterations. A frame pair must contain at least 8 tentative matches and 8 geometric inliers. Temporal verification uses radius 1 and checks both synchronized directions.

C Evaluation Protocols

This section specifies the evaluation procedure used for the quantitative results. Geometric evaluation follows the Umeyama alignment and clipped bidirectional Chamfer convention of TALO (Zhang et al., 2026). Panoptic evaluation follows the UNITE point-transfer protocol (Koch et al., 2026). Geometry and panoptic quality are evaluated in separate stages. The former measures the accuracy of the recovered trajectory and 3D points, while the latter transfers predicted labels to a common ground-truth support before computing semantic and instance metrics. Unless stated otherwise, all distances are measured in metres and all class averages are macro averages.

C.1 Geometric Evaluation

Trajectory alignment and ATE. We estimate a single Umeyama similarity transform from the predicted camera centres to the ground-truth centres,

$$\mathbf{q}_i = s\mathbf{R}\mathbf{p}_i + \mathbf{t}, \qquad s > 0, \quad \mathbf{R} \in \mathrm{SO}(3), \tag{7}$$

and apply it to both the camera centres and the predicted camera rotations. The absolute trajectory error is the root mean squared aligned centre error,

$$\mathrm{ATE}_{\mathrm{RMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|s\mathbf{R}\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i\|_2^2}. \tag{8}$$

The implementation requires equal numbers of predicted and ground-truth frames and uses the default `sim3` alignment.

Point-cloud construction. For chunk-based methods, each local reconstruction is first mapped into the method-global frame with the optimized chunk-level $\mathrm{Sim}(3)$ graph. The trajectory-derived transform above is then applied to the resulting global points. When a local VGGT manifest is evaluated directly, points are kept only when their coordinates and confidence are finite, the confidence is larger than 10^{-5}, and, following VGGT-Long (Deng et al., 2025), it is at least 0.75 times the mean finite confidence of that chunk. The ground-truth cloud is restricted to the axis aligned bounding box of the ground-truth camera trajectory, expanded by $100\,\mathrm{m}$ on every side.

Both clouds are voxel-downsampled to $0.5\,\mathrm{m}$ cell size using voxel centroids. Let P and G denote the resulting prediction and ground-truth point sets. We query a nearest-neighbour distance in both directions,

$$d_{P\to G}(x) = \min_{y\in G} \|x - y\|_2, \qquad d_{G\to P}(y) = \min_{x\in P} \|y - x\|_2. \tag{9}$$

The directional errors clip each distance at $\tau = 50\,\mathrm{m}$,

$$\mathrm{Acc} = \frac{1}{|P|}\sum_{x\in P} \min(d_{P\to G}(x), \tau), \qquad \mathrm{Comp} = \frac{1}{|G|}\sum_{y\in G} \min(d_{G\to P}(y), \tau), \tag{10}$$

and the reported Chamfer distance is

$$\mathrm{CD} = \frac{1}{2}(\mathrm{Acc} + \mathrm{Comp}). \tag{11}$$

The evaluator also records a point-level F-score using a $2\,\mathrm{m}$ distance threshold, but the main geometric table reports only ATE and CD.

C.2 UNITE-style Semantic and Instance Evaluation

Semantic label transfer and metrics. Let (G, ℓ_G) be the GT points and their canonical semantic labels, and (P, ℓ_P) the aligned prediction points. For every valid GT point, the evaluator finds the nearest prediction point over the full prediction cloud, without conditioning the search on the GT class, and transfers its predicted label to the GT support. Unknown prediction labels remain in the search and therefore count as semantic errors. For class c, let TP_c be the number of transferred labels equal to c on GT points of class c, and let N_c^G and N_c^P be the corresponding GT and transferred-prediction counts. We compute

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{N_c^G + N_c^P - \mathrm{TP}_c}, \qquad \mathrm{Acc}_c = \frac{\mathrm{TP}_c}{N_c^G}, \tag{12}$$

and report

$$\mathrm{mIoU}_{3D} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathrm{IoU}_c, \qquad \mathrm{mAcc}_{3D} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathrm{Acc}_c, \tag{13}$$

where classes with an undefined denominator are omitted from the respective macro average.

Instance AP. Within each semantic class, the nearest prediction point transfers its predicted instance ID to every GT point. Prediction and GT segments are then matched one-to-one by descending IoU, and a match is accepted when its IoU exceeds the selected threshold. We report AP_{50} and AP_{25} at IoU thresholds 0.50 and 0.25, respectively. The aggregate AP is the mean of AP values at

$$\{0.50, 0.55, 0.60, \ldots, 0.95\}. \tag{14}$$