ONE 4RROW, EVERY TARGET: UNIFIED 4D RECONSTRUCTION AND RECOGNITION

Anonymous authors
Paper under double-blind review

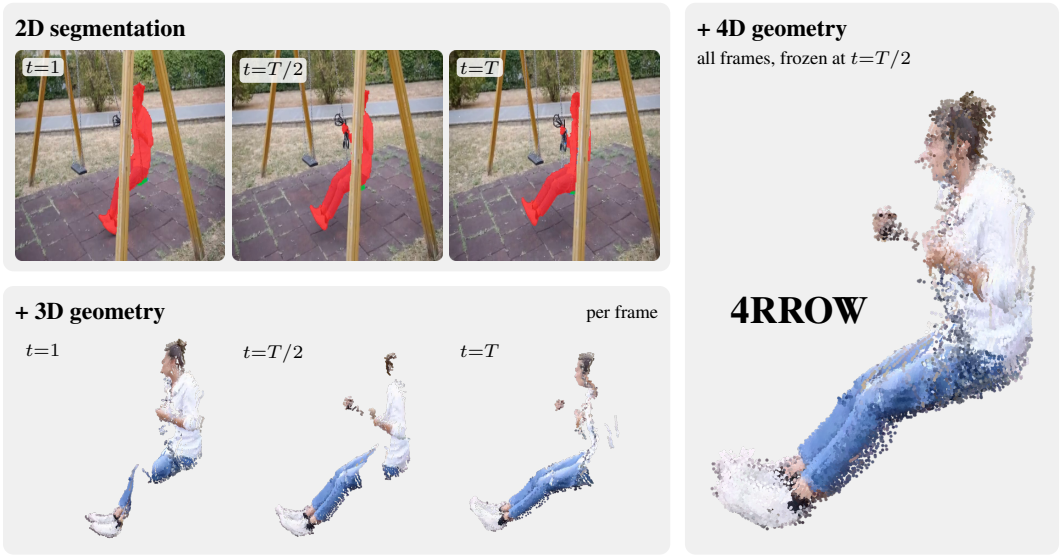


Figure 1: **4RROW** isolates and reconstructs a full dynamic subject in 4D from monocular video, despite the person not being fully visible in any given frame.

ABSTRACT

Feedforward models have made rapid progress on reconstructing 4D geometry and 4D tracking from dynamic video, and separately on incorporating semantics into static 3D reconstructions. However, no single model combines all three tasks. This paper introduces 4RROW, the first feedforward model to jointly perform 4D reconstruction, point tracking, and recognition from a single input video of arbitrary length, enabling extraction of semantic 4D instances of dynamic objects (Fig. 1). Following D4RT, we use a powerful encoder coupled with an efficient point-wise decoder. We introduce object queries for instance recognition in such a way that they seamlessly operate alongside the point queries, and generalize the point query mechanism to videos of arbitrary length. To enable strong semantics we adapt DINOv3 for video and make it temporally aware, turning it into TimeDINO. On combined geometry and semantic prediction, 4RROW outperforms prior work on the static Scannet++ dataset while establishing a baseline on the dynamic Waymo dataset. Finally, we show our unified model can even outperform specialized video instance segmentation models in an open-vocabulary setting while also producing strong 4D reconstruction results.

1 INTRODUCTION

Feedforward models for reconstructing 3D geometry and camera parameters in static scenes have been shown to be effective (e.g. Cabon et al. (2025); Wang et al. (2024; 2025a)). More recently, several methods successfully demonstrated that feedforward models can also handle 4D geometry in *dynamic* scenes (e.g. Chen et al. (2025); Wang et al. (2026a); Zhang et al. (2026)). Orthogonally, other works demonstrated that 3D geometry models, targeting *static* scenes, can be imbued with *semantics*—spanning both semantic and instance segmentation (e.g. Barberteguy et al. (2026); 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.

(2025a); Koch et al. (2026)). However, most real-world videos are inherently dynamic, with moving objects, occlusions, and evolving scene structures that static formulations cannot accommodate.

To address this challenge, this paper proposes an architecture which combines 4D geometry with semantics. To tackle dynamics effectively and efficiently, we adopt and extend the point-based query mechanism of D4RT (Zhang et al., 2026); by explicitly capturing continuous point trajectories over time, it provides a scalable formulation that is naturally well-suited for isolating and tracking *objects* through time. Simultaneously, we require rich semantic capabilities, for which we leverage a DinoV3 backbone (Siméoni et al., 2025) known for its strong visual representations and segmentation performance. However, because DinoV3 was originally designed for static images, we introduce an adaptation of the architecture to video that enables spatio-temporal reasoning while still preserving its powerful pre-trained semantic features.

We propose **4RROW**: **4**D **R**econstruction and **R**ecognition of **O**bjects in **V**ariable-length **V**ideos. To the best of our knowledge, 4RROW is the first method that jointly tracks points, segments object instances, and reconstructs dynamic 4D scenes within a single unified feedforward network. This tight coupling enables capabilities that neither pure reconstruction nor recognition models can achieve alone (Fig. 1). Consider the dynamic sequence of a woman on a swing: while she is partially occluded in every individual frame, no single view offers a complete observation of her shape. Nevertheless, because 4RROW simultaneously segments the object instance, tracks points over time, and predicts 3D geometry, it can aggregate visible points from *dynamic objects* across *all* video frames and map them into the coordinate system of any chosen reference frame at any chosen reference time. Thus, our model recovers a complete, unoccluded 3D point cloud of the woman "frozen in time" (Fig. 1, right) directly from an in-the-wild monocular video.

We evaluate 4RROW across three complementary settings: dynamic 4D scenes, static 3D benchmarks, and dynamic videos from popular 2D Video Instance Segmentation (VIS) benchmarks. For 4D datasets, where joint evaluation protocols are still unestablished, we propose a novel depth-aware AP metric and establish a baseline for simultaneous geometry prediction and instance segmentation on the Waymo Open dataset (Sun et al., 2020). On static scenes, our model compares favorably against specialized state-of-the-art baselines on ScanNet++ (Yeshwanth et al., 2023). Crucially, to our knowledge, we are the first to evaluate a joint geometry–recognition framework on standard VIS benchmarks. Although VIS is traditionally dominated by task-specific 2D architectures, our unified model establishes new SOTA results for open-vocabulary setting across all benchmarks and importantly, in the most challenging OVIS benchmark, even surpasses closed-vocabulary techniques.

To summarize, our contributions are as follows: *(i)* We introduce 4RROW which, to the best of our knowledge, is the first method which unifies 4D reconstruction, point tracking and recognition into a single framework. This unlocks the ability to isolate and reconstruct dynamic objects frozen in time from multiple video frames (Fig. 1). *(ii)* We redesign point queries for variable-frame videos which makes our model practically applicable. *(iii)* We upgrade the image-pretrained DinoV3 encoder for time-aware 4D video encoding, resulting in TimeDINO. *(iv)* We establish a baseline on the dynamic Waymo dataset, outperform prior work on combined geometry and semantic prediction on the static ScanNet++ dataset, and also outperform highly specialized open-vocabulary 2D models in VIS while achieving strong reconstruction results.

2 Related Work

Feedforward Reconstruction Models (FRMs). DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) successfully pioneered 3D reconstruction as a dense feedforward regression problem for pairs of unposed views of a static scene. MUSt3R (Cabon et al., 2025), VGGT (Wang et al., 2025a), and π^3 (Wang et al., 2026b) generalized this to multi-view reconstruction. To handle longer videos, MUST3R (Cabon et al., 2025), CUT3R (Wang et al., 2025b), and Stream3R (Lan et al., 2026) used causal streaming transformer architectures. To adapt these methods to dynamic videos, Easi3R (Chen et al., 2025) and VGGT4D Hu et al. (2025) demonstrate that DUSt3R and VGGT can be adapted to dynamic scenes without retraining by inferring motion from their attention maps. MonST3R (Zhang et al., 2025a) trains on dynamic scenes but explicitly masks out dynamic objects to handle camera pose reconstruction. VGGT-Ω (Wang et al., 2026a) and Depth Anything 3 (DA3) (Lin et al., 2026) sidestep this problem by directly predicting camera pose. In MonST3R, VGGT-Ω, and DA3, 3D point or depth prediction is done at the time of each input frame hence

explicitly modeling dynamics is unnecessary. Instead, DynamicVGGT (He et al., 2026) explicitly handles moving objects by predicting point motion and by introducing a dynamic reconstruction task based on Gaussian Splatting (Kerbl et al., 2023). All these models operate under a dense prediction paradigm using DPT heads (Ranftl et al., 2021).

Recently, D4RT (Zhang et al., 2026) reconceptualizes 4D dynamic reconstruction as on-demand pointwise querying (Sajjadi et al., 2022; Jaegle et al., 2021). This results in computationally efficient training compared to dense prediction and is more efficient for sparse inference tasks such as point tracking and camera pose estimation from points. Moreover, their queries explicitly handles dynamic objects by enabling querying the 3D coordinate of a pixel at *any specified* time in the coordinate system of any given frame in the video, which by design enables tracking. Our work builds upon D4RT's pointwise query mechanisms since it naturally deals with dynamic scenes with moving objects. We demonstrate how pointwise querying can be extended to recognition and show which changes enable their pointwise query mechanism to work for variable-length videos.

Recognition. Video Instance Segmentation (VIS) requires joint detection, segmentation, and temporal tracking of object instances. Instead of building video-specific decoders from scratch, many modern VIS methods apply Mask2Former (Cheng et al., 2022) per-frame and using transformer modules and temporal memory mechanisms to merge and track object queries across time, as exemplified by VITA (Heo et al., 2022), DVIS++ (Zhang et al., 2025b), and LOMM (Lee et al., 2025). Most of these methods employ ViT-Adapters (Chen et al., 2023), pixel decoder and transformer decoders that, while effective, introduce significant architectural complexity. To simplify this, EoMT (Kerssies et al., 2025) successfully folded learnable object queries directly into the ViT image backbone. Similar to VidEoMT (Norouzi et al., 2026), we apply this approach to dynamic videos, but unlike their approach, we do not need the additional query propagation and fusion modules they introduced.

Unified reconstruction and recognition. Several prior methods to unify reconstruction and recognition were built upon implicit representations such as NeRF (Mildenhall et al., 2021) or Gaussian Splats (Kerbl et al., 2023). Examples are Panoptic Lifting (Siddiqui et al., 2023), Panoptic NeRF (Fu et al., 2022), DFF-NeRF (Kobayashi et al., 2022), and Feature 3DGS (Zhou et al., 2024). More recent work adds semantics to FRMs: UNITE (Koch et al., 2026) and IGGT (Li et al., 2025a) add instance embeddings trained with a contrastive loss which requires clustering to get the instances. UNITE additionally trains a semantic prediction head while IGGT relies on OpenSeg (Ghiasi et al., 2022). Panst3R Zust et al. (2025) and Pano3D Barberteguy et al. (2026) instead attach a Mask2Former-like (Cheng et al., 2022) prediction head. All these works build upon models trained for static scenes based on the dense prediction paradigm. Instead, we build on sparse pointwise queries which were designed for dynamic scenes and create a compatible object query mechanism.

3 METHOD

We propose 4RROW, a unified model for feed-forward 4D reconstruction, tracking, and semantic and instance segmentation. It builds on the D4RT framework (Zhang et al., 2026) which uses a powerful encoder coupled with an efficient point-wise decoder. As shown in Fig. 2, 4RROW consumes input videos of arbitrary length which are first processed globally via our modified *TimeDINO* encoder, resulting in a dense scene representation. A number of instance queries then interact with the TimeDINO features in an *Encoder-only Mask Transformer* module (EoMT, Kerssies et al. (2025)). The resulting scene representation can finally be queried at arbitrary points for their 3D position, motion, and instance assignments across the scene via an efficient cross-attention decoder.

3.1 VARIABLE-LENGTH VIDEO ENCODER

Given an arbitrary-length video $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$ of T ordered frames accompanied by timestamps $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_T) \in \mathbb{R}^T$, our objective is to jointly predict dense metric 3D point trajectories, point visibility states, camera poses, and instance-level semantic masks within a single forward pass. We upgrade the image-pretrained DinoV3 encoder (Siméoni et al., 2025) to time-aware 4D video using temporal RoPE (Su et al., 2021) encodings, dubbed *TimeDINO*. Our choice is based on

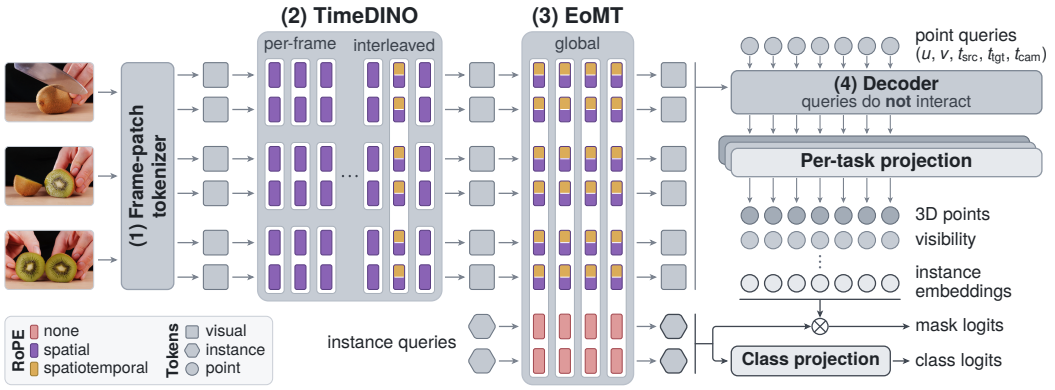


Figure 2: **Overview of 4RROW**. Given a dynamic video of any length, we first tokenize spatial frame patches. We process them with TimeDINO which transforms DinoV3 into a video model with temporal RoPE and per-frame and global attention at different stages (Sec. 3.1). We integrate object queries into the backbone (EoMT, Sec. 3.2) and use their output tokens to predict open-vocabulary class embeddings (Sec. 3.3). Our input queries specify a source location in the video, a target timestamp, and target camera viewpoint. We anchor them in time through projected mean frame embeddings, process them by a decoder, and use task-specific heads to predict geometric features, as well as instance embeddings which we use for inner-product mask segmentation (Sec. 3.3).

the fact that DinoV3 (i) has been shown to be well-suited for segmentation, (ii) natively uses 2D spatial RoPE, suggesting a natural extension to the time axis, (iii) has been pre-trained with register tokens $\boldsymbol{Z}_t^{\text{reg}} \in \mathbb{R}^{4\times D}$ to absorb background artifacts. Intuitively, networks trained with non-visual tokens facilitate the addition of extra task-specific (non-visual) tokens into the sequence without destabilizing self-attention or degrading visual features.

After extracting patch embeddings, we prepend structured prefix tokens to the N spatial patch tokens of each frame $t \in \{1, \dots, T\}$:

$$\boldsymbol{Z}_t = \left[\boldsymbol{Z}_t^{\text{meta}},\ \boldsymbol{Z}_t^{\text{extra}},\ \boldsymbol{Z}_t^{\text{reg}},\ \boldsymbol{Z}_t^{\text{patch}}\right] \in \mathbb{R}^{(N_{\text{prefix}}+N)\times D}. \tag{1}$$

Specifically, $N_{\text{prefix}} = 7$, comprising $\boldsymbol{Z}_t^{\text{meta}}$ which projects the image aspect ratio and known camera metadata, two learnable extra tokens $\boldsymbol{Z}_t^{\text{extra}}$, and the four register tokens $\boldsymbol{Z}_t^{\text{reg}}$. The extra tokens are temporally mean-pooled at the output of the final backbone layer and used to predict scene scale and focal length. All prefix and register tokens remain unrotated by spatial RoPE to preserve coordinate invariance, but are affected by the temporal RoPE introduced below.

To reduce computational requirements of directly applying the model on the entire video, we use interleaved global and frame-based attention (Arnab et al., 2021; Bertasius et al., 2021). Specifically, the first $1/3$ of all layers perform per-frame attention, the regime in which the model has been pre-trained in. Most remaining layers alternate between local per-frame and global attention, with the exception of the final $L_{\text{tail}} = 4$ *tail layers* which perform purely global attention.

To extrapolate seamlessly across variable frame counts and irregular sampling intervals in the global attention layers, we utilize RoPE to encode time and space using a factorized partial-channel formulation. Let $d_h = D/N_h$ denote the dimension of each of the N_h attention heads for D dimensional tokens. For normalized patch coordinates $(u, v) \in [-1, 1]^2$, we apply 2D spatial RoPE across *all* d_h channels by dedicating $d_h/2$ channels to each spatial axis. Because each frequency rotates a 2D channel pair, this yields $d_h/4$ rotation angles per axis with base period $b = 100$:

$$\theta_k^{(u)} = \frac{2\pi u}{p_k}, \quad \theta_k^{(v)} = \frac{2\pi v}{p_k}, \quad p_k = b^{\frac{4k}{d_h}}, \quad k \in \{0, \dots, d_h/4 - 1\}. \tag{2}$$

To encode time without corrupting image-pretrained spatial representations, we apply 1D temporal RoPE with frequencies

$$\omega_m = \Theta^{-2m/d_{\text{time}}}, \qquad m \in \{0, \dots, d_{\text{time}}/2 - 1\}, \tag{3}$$

where $\Theta = 10\,000$ and $d_{\text{time}} = d_h/2$ channels. For a frame at timestamp τ_t, these temporal rotations $\omega_m \tau_t$ compose with the spatial rotations on half of the channel pairs, encoding the *relative* temporal displacement $\Delta\tau = \tau_t - \tau_{t'}$ between frames t and t' in the attention inner products, while the remaining $d_h/2$ channels undergo spatial rotation only, preserving purely static 2D geometry.

This mechanism allows us to directly inject continuous time-delta information into the encoder without needing to anchor the time axis to any particular frame. This is a critical design choice to enable efficient random frame sub-sampling during training, see Sec. 3.4.

3.2 4D Object Instance Tokens

Standard video instance segmentation pipelines rely on heavy auxiliary decoders such as multi-scale deformable attention pixel decoders (Cheng et al., 2022) or dedicated mask transformers. In 4RROW, we eliminate separate segmentation decoders by extending the Encoder-only Mask Transformer (EoMT) (Kerssies et al., 2025) paradigm into dynamic 4D video.

Specifically, we instantiate K learnable object queries $\boldsymbol{Q}_0^{\text{obj}} \in \mathbb{R}^{K\times D}$ and inject them as prefix tokens for all frames just before the last $L_{\text{tail}} = 4$ layers. This yields:

$$\boldsymbol{Z}_t^{\text{tail}} = \left[\boldsymbol{Q}_t^{\text{obj}}, \boldsymbol{Z}_t^{\text{meta}}, \boldsymbol{Z}_t^{\text{extra}}, \boldsymbol{Z}_t^{\text{reg}}, \boldsymbol{Z}_t^{\text{patch}}\right] \in \mathbb{R}^{(K+N_{\text{prefix}}+N)\times D} \tag{4}$$

where initially object queries are the same for each frame and are treated as prefix tokens receiving temporal but not spatial RoPE. Across the final L_{tail} global layers, standard self-attention yields object-query-to-patch, patch-to-patch, and object-query-to-object-query attention. At the output of the final encoder layer, the sequence is split into the refined spatial scene features $\mathbf{F} \in \mathbb{R}^{T\times N\times D}$ while the refined object queries are mean-pooled across frames resulting in $\boldsymbol{Q}_{\text{obj}} \in \mathbb{R}^{K\times D}$.

3.3 Content-Aware Pointwise Decoding

In pointwise decoding, each query must specify an *exact* spatio-temporal location: a continuous 2D coordinate (u, v) on source frame t_{src}, tracked to its 3D state at target timestamp t_{tgt}, and evaluated from the camera viewpoint of t_{cam}. While continuous 2D Fourier embeddings generalize spatial coordinates (u, v) seamlessly across all frames, anchoring the three temporal locations $(t_{\text{src}}, t_{\text{tgt}}, t_{\text{cam}})$ is fundamentally challenging under variable-length sequences. For this reason, D4RT retrieves temporal embeddings from learned lookup tables $\boldsymbol{E}[t]$, permanently locking inference to the exact frame count observed during training.

To overcome this limitation, we propose *Content-Aware Frame Pooling*. Instead of static indices that fail to generalize, we anchor temporal locations directly in the scene content of each frame. Specifically, we compute a content summary vector for each frame t by spatially mean-pooling its refined output patch tokens from the backbone features $\mathbf{F}$ into $\bar{\boldsymbol{f}}$. We then dynamically generate continuous time embeddings through learned linear projections of these content-aware vectors: $\boldsymbol{e}_{t_{\text{src}}} = \boldsymbol{W}_{\text{src}}\bar{\boldsymbol{f}}_{t_{\text{src}}}$, $\boldsymbol{e}_{t_{\text{tgt}}} = \boldsymbol{W}_{\text{tgt}}\bar{\boldsymbol{f}}_{t_{\text{tgt}}}$, $\boldsymbol{e}_{t_{\text{cam}}} = \boldsymbol{W}_{\text{cam}}\bar{\boldsymbol{f}}_{t_{\text{cam}}}$, for learned projection matrices $\boldsymbol{W}_{\text{src}}, \boldsymbol{W}_{\text{tgt}}, \boldsymbol{W}_{\text{cam}} \in \mathbb{R}^{D\times D}$.

For any point query $\boldsymbol{q} = (u, v, t_{\text{src}}, t_{\text{tgt}}, t_{\text{cam}})$, the input query token $\boldsymbol{z}_{\boldsymbol{q}} \in \mathbb{R}^D$ is assembled as

$$\boldsymbol{z}_{\boldsymbol{q}} = \text{FourierEmbed}(u, v) + \boldsymbol{e}_{t_{\text{src}}} + \boldsymbol{e}_{t_{\text{tgt}}} + \boldsymbol{e}_{t_{\text{cam}}} + \text{MLP}\left(\text{BilinearPatch}_{9\times 9}(u, v, t_{\text{src}})\right), \tag{5}$$

in which D4RT's static lookup tables are replaced by our content-aware embeddings $\boldsymbol{e}_{t_{\text{src}}}, \boldsymbol{e}_{t_{\text{tgt}}}, \boldsymbol{e}_{t_{\text{cam}}}$, $\text{FourierEmbed}(u, v)$ encodes continuous 2D position, and the 9×9 bilinear patch provides high-frequency local appearance conditioning. The resulting query representation is fully content-aware across both space and time, generalizing gracefully to arbitrary sequence lengths T at inference.

The query tokens $\boldsymbol{z}_{\boldsymbol{q}}$ pass through an 8-layer cross-attention decoder where queries attend strictly to the backbone scene features $\mathbf{F}$ with zero inter-query self-attention. From the decoded query representation $\boldsymbol{z}_{\boldsymbol{q}}^{\text{dec}}$, lightweight linear heads produce the model predictions. To enable semantics, for every point query i we also predict a compact instance embedding $\boldsymbol{e}_i$. To predict the mask logit m_i we take the inner product with projected object queries. For semantic category classification, we use an open-vocabulary approach and apply normalized CLIP (Radford et al., 2021) embeddings $\boldsymbol{T}_{\text{cls}} \in \mathbb{R}^{C\times D_{\text{text}}}$ and predict class logits $\boldsymbol{c}_i$ for all C classes. More formally this results in:

$$m_i = \left\langle \boldsymbol{W}_{\text{mask}}\boldsymbol{Q}_{\text{obj}}, \boldsymbol{e}_i \right\rangle \tag{6}$$

$$\boldsymbol{c}_i = \left\langle \boldsymbol{W}_{\text{cls}}\boldsymbol{Q}_{\text{obj}}, \boldsymbol{T}_{\text{cls}} \right\rangle \tag{7}$$

Following D4RT we also predict a normalized 3D point position, 2D image coordinate, 3D surface normal, target point visibility, point motion, and a confidence weight for each query.

3.4 TRAINING

4RROW is trained end-to-end using a joint multi-task objective:

$$\mathcal{L} = \lambda_{\text{inst}}\mathcal{L}_{\text{inst}} + \lambda_{\text{inst_embed}}\mathcal{L}_{\text{inst_embed}} + \frac{1}{Q}\sum_{i=1}^{Q}\Big[\gamma_i\,\lambda_{\text{3D}}\mathcal{L}_{\text{3D}}^{(i)} - \lambda_{\text{conf}}\log\gamma_i + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}}^{(i)} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}}^{(i)} + \lambda_{\text{displ}}\mathcal{L}_{\text{displ}}^{(i)} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}}^{(i)}\Big], \tag{8}$$

where $\gamma_i > 0$ is the predicted confidence for query point i (Kendall & Gal, 2017), and the $\mathcal{L}_*^{(i)}$ represent the per-query geometric losses from D4RT.

For segmentation and classification we follow Mask2Former (Cheng et al., 2022) and compute an optimal bipartite assignment between predicted object queries and ground-truth instances using the Hungarian algorithm based on mask and category similarity. The instance segmentation loss (weighted by $\lambda_{\text{inst}} = 0.3$) is defined as:

$$\mathcal{L}_{\text{inst}} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}(\boldsymbol{M}, \boldsymbol{M}^*) + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}(\boldsymbol{M}, \boldsymbol{M}^*) + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}(\boldsymbol{S}, \boldsymbol{y}^*), \tag{9}$$

where $\boldsymbol{M}^* \in \{0,1\}^{K\times Q}$ denotes the Hungarian-assigned ground-truth point-instance masks, $\boldsymbol{y}^* \in \{1,\dots,C\}^K$ denotes the ground-truth category labels, $\mathcal{L}_{\text{focal}}$ is the point-level sigmoid focal loss (Lin et al., 2017) ($\lambda_{\text{focal}} = 1.0$), $\mathcal{L}_{\text{dice}}$ is the scale-invariant Dice loss ($\lambda_{\text{dice}} = 1.0$), and $\mathcal{L}_{\text{cls}}$ is category cross-entropy ($\lambda_{\text{cls}} = 1.0$). In addition, we found it useful to apply a contrastive loss on the learned embedding space $\boldsymbol{e_q}$ using binary NCE (Gutmann & Hyvärinen, 2010).

4 RESULTS

As there are no established benchmarks for 4D reconstruction and recognition on dynamic scenes, we first target video instance segmentation and reconstruction separately with a single model in Sec. 4.1. On combined geometry and semantic prediction we compare to related work on the static ScanNet++ dataset in Sec. 4.2 and finally establish a baseline on the dynamic Waymo dataset in Sec. 4.3.

Implementation Details. All experiments use a ViT-H+/16 TimeDINO encoder and 8-layer decoder. We train on video clips of 24 frames at a resolution of 384×384. To be robust to time variations we sample these clips from $\approx$ 16s intervals, keeping the original timestamps. For synthetic static datasets there are no true timestamps, but we assume a virtual 24 fps. Additionally, we log-uniformly scale fps by a factor 0.5–$2\times$ to equally likely to slow or speed up the clip.

4.1 VIDEO INSTANCE SEGMENTATION AND 3D/4D RECONSTRUCTION

Video instance segmentation (VIS) requires jointly detecting, segmenting, and tracking instances through time in dynamic videos. At the same time, we also want strong 4D tracking and reconstruction performance. Therefore we co-train 4RROW on the 3D/4D dataset mixture of D4RT alongside the VIS datasets YTVIS'19 (Yang et al., 2019), YTVIS'21 (Yang et al., 2021), OVIS (Qi et al., 2022). Additionally we train on COCO Lin et al. (2014) images which we turn into 'videos' by cropping and panning images following a smooth trajectory. By jointly training on this specific combination of different vocabularies, represented by their respective text embeddings for labels, we replicate the open-vocabulary setting as defined by DVIS++. We use $K = 64$ object queries. During inference we process each video in a single forward pass at its native frame count (max 292 frames) and at 448×448 resolution. On VIS we report standard Average Precision (AP) while for geometry we measure depth prediction in Absolute Relative Error (AbsRel) (Eigen et al., 2014) and point cloud prediction as mean L1 distance. See results in Tab. 1.

Remarkably, on open-vocabulary VIS, 4RROW sets a new state of the art across all three benchmarks (Tab. 1, left), outperforming the highly specialized and complex DVIS++ model by $+7.9$ AP on YTVIS'19 and $+5.4$ AP on YTVIS'21. Moreover, on the challenging occlusion-heavy long video benchmark OVIS, our method offers the largest gap, $+11.7$ AP, and both YTVIS'19 and OVIS seem to be on par with the highly specialized closed-vocabulary results.

Simultaneously, we demonstrate the same 4RROW model still yields strong 3D/4D reconstruction accuracy on dynamic Sintel (Butler et al., 2012) and static ScanNet (Dai et al., 2017) videos (Tab. 1,

Table 1: **Video Depth, Point Cloud, and Video Instance Segmentation.** 4RROW achieves strong results on video depth estimation and point cloud reconstruction, while outperforming open-vocabulary methods on VIS and remaining competitive with specialized 2D closed-vocabulary systems. †Uses MinVis (Huang et al., 2022) for association.

	Video Instance Seg. (AP ↑)			Video Depth (AbsRel ↓)		Point Cloud (L_1 ↓)	
Method	**YTVIS'19**	**YTVIS'21**	**OVIS**	**Sintel**	**ScanNet**	**Sintel**	**ScanNet**
3D Reconstruction							
MegaSaM	–	–	–	0.342	0.050	1.531	0.072
VGGT	–	–	–	0.318	0.044	1.582	0.063
MapAnything	–	–	–	0.397	0.043	1.718	0.064
SpatialTrackerV2	–	–	–	0.209	0.027	1.375	0.036
π^3	–	–	–	0.241	0.021	1.139	0.030
D4RT	–	–	–	0.171	0.020	0.768	0.028
2D Video Instance Segmentation (Closed-Vocab)							
DVIS++	68.3	63.9	53.4	–	–	–	–
LOMM	70.1	66.2	51.7	–	–	–	–
2D Video Instance Segmentation (Open-Vocab)							
FC-CLIP†	57.3	52.8	31.5	–	–	–	–
DVIS++	61.1	56.7	40.6	–	–	–	–
Unified Video Instance Segmentation (Open-Vocab) + 3D Reconstruction							
4RROW (Ours)	69.0	62.1	52.3	0.251	0.025	0.897	0.032

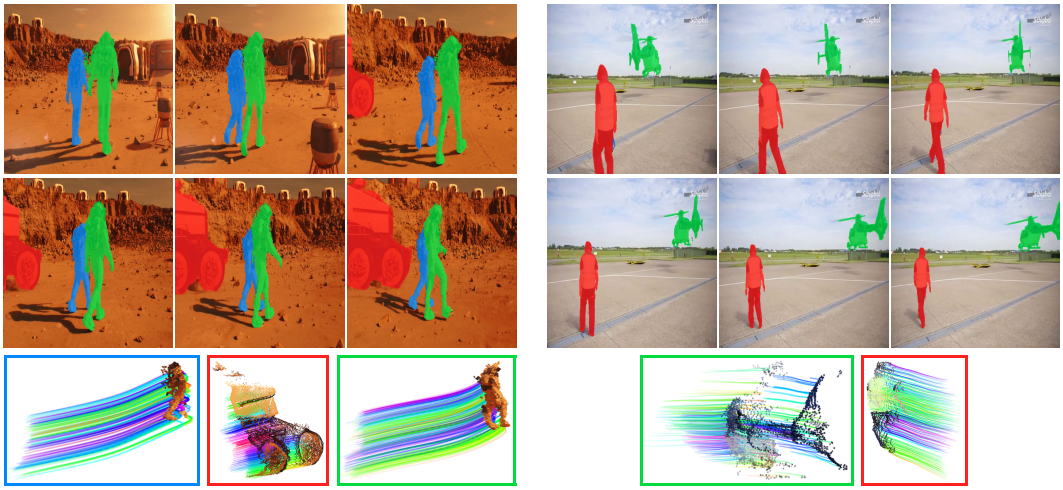


Figure 3: **Extracting how each object moves in 3D**, shown for two videos (left and right). Top two rows: six uniformly sampled frames with the instance masks predicted by 4RROW. Bottom row: 3D point map of each object at the last frame, with its 3D point tracks. Outline colors match the instance masks. See Sec. 4.1 for details.

right) using the evaluation protocol and baseline numbers of D4RT. 4RROW outperforms every dedicated reconstruction baseline except D4RT on Sintel point cloud reconstruction yielding 0.897 L_1. It also closely tracks pure-geometry D4RT on static ScanNet (0.025 vs. 0.020 depth AbsRel; 0.032 vs. 0.028 point cloud L_1) while outperforming MegaSaM (Li et al., 2025b), VGGT, and MapAnything (Keetha et al., 2025) on Sintel depth (0.251 AbsRel), demonstrating that a single feedforward model can unify state-of-the-art open-vocabulary VIS with competitive 3D/4D reconstruction.

Qualitative results. To demonstrate the joint tracking and instance segmentation capabilities of our model, we apply 4RROW to in-the-wild videos and show qualitative results in Fig. 3. From a single forward pass, 4RROW segments and tracks each object instance through the video and lifts it to 3D, which makes it possible to extract how each object moves in 3D. For each object, we take its 3D point map at the last frame and query these points across all frames. This gives 3D point tracks in world coordinates. In Fig. 4 we compare to D4RT on multiframe reconstruction of dynamic objects: points from all frames are tracked into a frozen reference time (the middle frame). D4RT predicts geometry and tracks, but has no notion of objects. The foreground can therefore only be obtained from motion, by thresholding point motion relative to the static background. As a result,

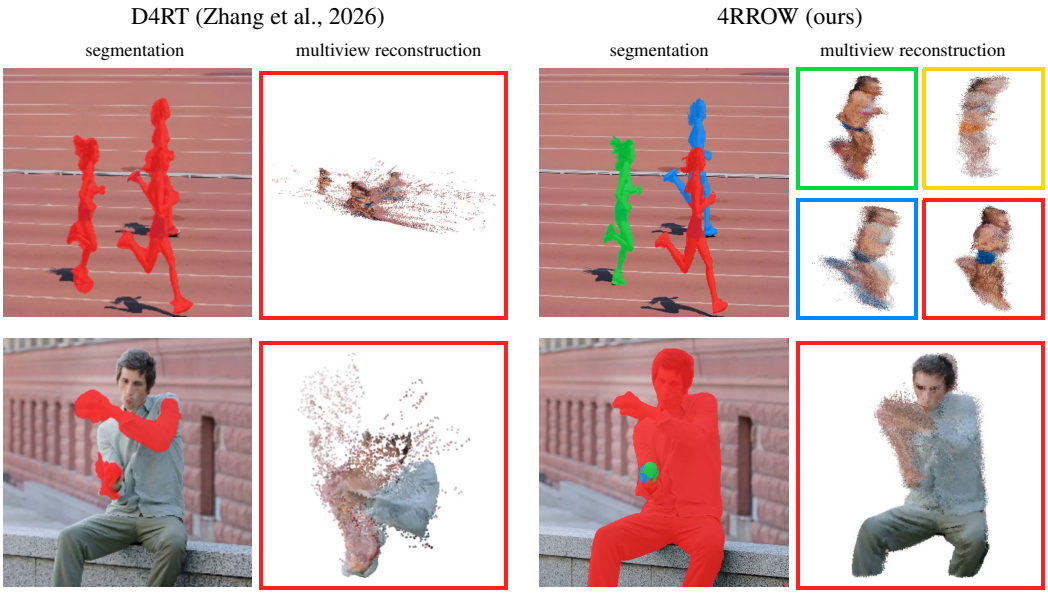


Figure 4: **Multiframe reconstruction of dynamic objects: comparison with D4RT.** For each sequence: object segmentation at the middle frame, and the reconstruction frozen at that time. D4RT segments by motion, while 4RROW segments object instances. See Sec. 4.1 for details.

background points leak into the reconstruction. Objects that move together are merged into a single region (the sprinters), and static parts of an object are missed (only the moving arms and apple are segmented, and the rest of the person is missing). The resulting frozen-in-time reconstructions therefore mix distinct objects or miss parts of them. By explicitly segmenting object instances, 4RROW produces clean masks that separate instances whether or not they move, and obtains a multiframe reconstruction of each object frozen in time. We encourage readers to view the accompanying videos and interactive 3D reconstructions in the supplemental material. We also note that such dynamic tracking capabilities are absent from architectures based on DUSt3R or VGGT.

4.2 Static Dataset: ScanNet++

ScanNet++ (Yeshwanth et al., 2023) provides high-fidelity indoor scans of static scenes with aligned 3D geometry and dense semantic labels of 100 classes. We measure combined reconstruction and recognition using 3D mIoU as introduced by IGGT. This voxelizes the 3D space in 10cm cubes and considers a voxel to be a true positive if the model correctly predicts both the spatial occupancy and its class. The mean Intersection-over-Union (mIoU) is first calculated per frame and is then averaged over all frames. There is also a 2D mIoU averaged per frame. However, because the IGGT metric selected a very small set of 9 scenes spaced widely apart and drawn from the training set, we also evaluate on the full ScanNet++ validation set where we use 48 linearly spaced frames per scene.

For this scenario we train our model only on static 3D dataset mixture of D4RT plus COCO 'videos', and ScanNet++, removing the 9 IGGT ScanNet++ training scenes. Only for COCO and ScanNet++ we use the instances and class labels. Results are presented Tab. 2. On the ScanNet++ validation we outperform Pano3D by a significant margin: our method yields 25.8 3D mIoU while Pano3D yields 20.6. To better understand where the difference comes from we also run the metric with oracle semantic labels and with oracle geometry. With oracle semantics 4RROW significantly outperforms Pano3D, suggesting that 4RROW is better in geometry. In contrast, Pano3D is better on semantics.

On the 9 IGGT scenes, 4RROW significantly outperforms all methods except Pano3D in 3D mIoU. Upon closer inspection, 4RROW is actually better on 7 of the 9 scenes, but fails badly on the other two scenes, hurting the average. We hypothesize that this is because the IGGT scenes consist of only 6–9 frames, spaced widely apart in time. This makes this setting more of a multi-view reconstruction problem than a monocular video and tracking problem for which 4RROW is designed.

4.3 Dynamic Sequences: Waymo

We now provide a baseline for joint semantics and geometry prediction for the Waymo Open dataset (Sun et al., 2020). This is a dynamic dataset of autonomous driving sequences pairing continuous video with synchronized LiDAR measurements. We use front-camera sequences. We

Table 2: **Combined Reconstruction and Recognition on static ScanNet++ scenes.** 4RROW achieves especially strong performances for 3D mIoU on both the validation and IGGT (Li et al., 2025a) evaluation split which measures the joint prediction of geometry and semantics.

	validation - 48 frames				*9 train scenes*	
	3D mIoU			**2D mIoU**	**3D mIoU**	**2D mIoU**
Method	full	oracle semantics	oracle geometry			
LSM (multi-view) (Fan et al., 2024)	–	–	–	–	15.2	17.9
Feature-3DGS (Zhou et al., 2024)	–	–	–	–	10.6	22.5
IGGT (Li et al., 2025a)	–	–	–	–	20.1	31.3
Pano3D (Pi3) (Barberteguy et al., 2026)	20.6	26.9	49.2	54.9	29.5	52.5
4RROW (Ours)	25.8	44.9	46.9	51.0	25.9	50.6

Table 3: **Dynamic Reconstruction and Recognition on Waymo sequences.** 4RROW strongly outperforms our adapted static baseline Pano3D (Barberteguy et al., 2026) demonstrating the benefit of combining 4D tracking and instance understanding.

Method	DAP	DAP_{50}	DAP_{75}	DAP_S	DAP_M	DAP_L	DAR_1	DAR_{100}
Pano3D	3.7	9.3	2.6	1.2	5.8	29.3	1.6	6.5
4RROW	7.3	19.2	4.9	2.9	19.0	44.7	1.9	14.3

restrict our evaluation to Pedestrian and Car classes, as the other classes contain very few instances. We discard from the evaluation per-frame masks that occupy less than 0.125% of the image area.

Depth-aware AP (DAP). To measure segmentation and depth together we propose a novel, depth-aware generalization of the AP metric used for VIS (Yang et al., 2019). In standard AP for video, a predicted track is a true positive if its spatio-temporal mask IoU with a ground-truth track is above a threshold. In DAP, we additionally require the overlapping pixels to agree in depth. The IoU is computed only over pixels with a LiDAR depth g_i. A pixel inside both masks counts toward the intersection only if the predicted depth p_i is within a factor $1+\theta$ of g_i:

$$\text{IoU}_\theta(d,g) = \frac{|T_\theta|}{|M_d| + |M_g| - |T_\theta|}, \qquad T_\theta = \left\{ i \in M_d \cap M_g \;:\; \max\left(\tfrac{g_i}{s\,p_i}, \tfrac{s\,p_i}{g_i}\right) < 1+\theta \right\}, \quad (10)$$

where M_d and M_g are the predicted and ground-truth masks restricted to LiDAR pixels and pooled over all frames. s is one scale per video, the median of g_i/p_i, fitted before matching. A pixel with the right mask but the wrong depth therefore counts as both a false positive and a false negative, just like a 3D point in the wrong place. Matching, IoU thresholds and precision–recall averaging are the same as in standard AP. Every COCO-style variant therefore carries over directly, e.g. AP_{50} becomes DAP_{50}, AR@100 becomes DAR@100, etc. We set θ to 25%. Note that because DAP scores each LiDAR ray on its own with a relative depth tolerance, it avoids voxelization and its pitfalls: choosing a voxel size, sparse or single-point voxels, and voxels that mix objects.

Baselines and Results. Because joint 4D reconstruction and instance segmentation is an open challenge, there are currently no established baselines evaluated on DAP for dynamic driving datasets. To provide a comparison, we adapt Pano3D (Barberteguy et al., 2026) and retrain it directly on the Waymo Open dataset. For 4RROW, we initialized from an 80% checkpoint from the VIS setup (Sec. 4.1) and continue training adding Waymo data. As shown in Tab. 3, 4RROW significantly outperforms Pano3D across all metrics. While Pano3D provides a strong baseline for joint 3D reconstruction and semantics, it was designed and optimized specifically for static scenes. The performance gap highlights that dynamic sequences require explicitly handling moving objects and temporal persistence, which 4RROW achieves through its unified tracking and 4D instance queries.

5 Conclusions

In this work, we introduced 4RROW, the first unified feedforward model to jointly tackle 4D reconstruction, point tracking, and instance segmentation from monocular video. By combining D4RT's continuous pointwise decoding with a video-adapted DinoV3 backbone and sequence-level object queries, 4RROW unifies dynamic geometry with open-vocabulary semantics. This formulation unlocks unattempted capabilities, aggregating visible observations over time to recover complete, unoccluded 3D point clouds of dynamic instances frozen in time. Across diverse 4D, 3D, and 2D benchmarks, 4RROW establishes a foundational baseline on Waymo Open, outperforms prior baselines on ScanNet++, and rivals specialized 2D systems on competitive VIS benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools for

- propose or refine hypotheses,
- implement methods,
- clean and reformat dataset,
- support qualitative and thematic data analysis,
- interpret results,
- design or provide feedback on research methodology or experiments.

We have not used generative AI tools for

- Generate synthetic data sets,
- formulate mathematical claims,
- create or modify scientific figures or images,

and

- help develop theoretical models or conceptual frameworks,
- provide critical ingredients for proving mathematical claims,
- assist in the writing of proofs,
- assist with translation,

are not applicable to this work. Additionally, we used generative AI tools for

- suggest experimental parameters,
- create or edit software code,
- draft parts of a research paper,
- summarize or analyse existing literature,
- discover research topics or identify gaps,
- brainstorming,
- sourcing/searching for information,
- edit a research paper to improve readability,
- identify relevant literature,
- propose a title or keywords for a research paper.

We have reviewed all AI-assisted work, by using code reviews for automatically generated code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research does not involve active human participants or the collection of new private data, relying solely on publicly available research benchmarks in compliance with their applicable licenses. The primary applications of this work lie in foundational 3D vision, robotics, autonomous driving, and computer graphics. As this is core methodology research, we do not anticipate specific ethical concerns or adverse societal impacts.

Reproducibility Statement

To facilitate the independent reproduction and verification of our results, all components of our methodology and experimental setup are detailed throughout the paper. Specifically, the comprehensive network architecture, tokenization schemes, and layer configurations are described in Section 3 and illustrated in Fig. 2. Full details of the training objectives, loss formulations, hyper-parameter weightings (λ), and spatio-temporal frame sampling procedures are provided in Section 3.4. All experiments are conducted exclusively on established, publicly available research datasets (ScanNet++, Waymo Open Dataset, and standard VIS benchmarks) using the official splits, evaluation protocols, and preprocessing steps detailed in Sec. 4. Finally, to enable thorough qualitative verification of the predicted 4D geometry and dynamic point tracks, an interactive 3D visualization viewer and video sequences are provided in the supplementary material.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021.

Victor Barberteguy, Ahmet Iscen, Mathilde Caron, Alireza Fathi, Gül Varol, and Cordelia Schmid. Pano3D: Unified 3d reconstruction and panoptic segmentation. *ECCV*, 2026.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *CVPR*, 2025.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. *ICCV*, 2025.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *ICLR*, 2023.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3d. *NeurIPS*, 2024.

Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *3DV*, 2022.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010.

Zhuolin He, Jing Li, Guanghao Li, Xiaolei Chen, Jiacheng Tang, Siyang Zhang, Zhounan Jin, Feipeng Cai, Bin Li, Jian Pu, Jia Cai, and Xiangyang Xue. DynamicVGGT: Learning dynamic point maps for 4d scene reconstruction in autonomous driving. In *CVPR*, 2026.

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. VITA: Video instance segmentation via object token association. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23109–23120, 2022.

Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction. *ArXiv*, 2025.

De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint*, 2025.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is Secretly an Image Segmentation Model. In *CVPR*, 2025.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *ArXiv*, 2022.

Sebastian Koch, Johanna Wald, Hidenobu Matsuki, Pedro Hermosilla, Timo Ropinski, and Federico Tombari. Unified semantic transformer for 3d scene understanding. *TMLR*, 2026.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. In *ICLR*, 2026.

Seunghun Lee, Jiwan Seo, Minwoo Choi, Kiljoon Han, Jahoon Jeong, Zane Durante, Ehsan Adeli, Sang Hyun Park, and Sunghoon Im. LOMM: Latest object memory management for temporally consistent video instance segmentation. In *ICCV*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *ECCV*, 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 *ICLR*, 2025a.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *CVPR*, 2025b.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, 2017.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.

Narges Norouzi et al. VidEoMT: Your ViT is Secretly Also a Video Segmentation Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130:2022–2039, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *CVPR*, 2021.

Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*, 2022.

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *CVPR*, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, 2021.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo Open Dataset. In *CVPR*, 2020.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-Ω. *CVPR*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024.

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 *ICLR*, 2026b.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.

Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, June 2021.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *CVPR*, 2026.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *ICLR*, 2025a.

Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *TPAMI*, 2025b.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *CVPR*, 2024.

Lojze Zust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. Panst3r: Multi-view consistent panoptic segmentation. In *CVPR*, 2025.