

Self-Supervised Video Object Segmentation for Autonomous Driving: Cross-Domain Evaluation on KITTI and Waymo

Yoseph Naoom
School of Engineering
Jönköping University
Jönköping, Sweden

Zakariya Mohamed Shafe
School of Engineering
Jönköping University
Jönköping, Sweden

Abstract—Self-supervised video object segmentation offers a promising alternative to annotation-dependent methods for autonomous driving perception, yet its behaviour under domain shift to real-world driving data remains poorly understood. This paper presents a systematic cross-domain evaluation of SSL-UVOS, an unsupervised video object segmentation framework built on a frozen DINO ViT-S/8 backbone, across twelve pipelines varying pretraining domain, evaluation dataset, and input resolution. Experiments are conducted on two autonomous driving benchmarks: KITTI MOTS and the Waymo Open Dataset. Domain-specific pretraining is performed using SimCLR on unlabelled driving frames. Results are reported using FG-ARI and J&F, with dedicated zone-stratified analysis targeting far-object segmentation, a safety-critical capability for detecting small, distant objects such as pedestrians and cyclists. SSL-UVOS generalizes considerably better to one dataset than the other, a disparity attributed to differences in scene complexity and distributional proximity to the backbone’s pretraining data. Higher input resolution emerges as the most effective intervention for far-zone segmentation, while domain-specific SimCLR pretraining provides only marginal and inconsistent gains across both datasets. These findings identify resolution as the most actionable factor for improving far-object segmentation and highlight the limitations of frozen-backbone contrastive adaptation under large domain shifts in driving scenarios.

Index Terms—video object segmentation, self-supervised learning, autonomous driving, DINO, SimCLR, KITTI MOTS, Waymo, FG-ARI, J&F metric

I. INTRODUCTION

Autonomous driving systems must reliably detect and segment dynamic objects such as pedestrians, cyclists, and vehicles across continuous video streams. Video object segmentation addresses this by assigning dense per-pixel labels to objects across video frames, enabling detailed spatiotemporal understanding beyond bounding-box based tracking [1]. Supervised approaches to this problem have shown strong performance, but depend on large volume of manually annotated data that is costly and time-consuming to collect [2]. This limitation is non compliant with the high demands of scaling to real-world driving scenarios.

Self-supervised methods offer a promising alternative by learning directly from unlabelled video. DINO [3] demonstrated that Vision Transformers pretrained through self-

distillation develop attention maps that naturally highlight semantically meaningful foreground regions without any supervision. Building on this, SSL-UVOS [4] extends a frozen DINO ViT-S/8 backbone with sparse cross-frame attention and hierarchical clustering to produce per-frame segmentation masks entirely without labels. While SSL-UVOS achieves competitive results on the DAVIS 2017 benchmark, its behaviour under domain shift to real-world driving data remains unexplored.

This paper evaluates SSL-UVOS on two autonomous driving datasets, KITTI MOTS and the Waymo Open Dataset, under twelve systematically designed pipelines that vary pretraining domain, evaluation domain, and input resolution. Domain-specific pretraining is performed using SimCLR [5], a contrastive self-supervised learning framework. Results show that SSL-UVOS generalises considerably better to KITTI than to Waymo, driven primarily by differences in scene complexity and distributional proximity to the backbone’s pretraining data. Higher input resolution emerges as the most effective intervention for segmenting small distant objects on KITTI, while domain-specific pretraining with a frozen backbone provides marginal and inconsistent gains across both datasets.

II. RELATED WORK

Ding et al. [4] propose a self-supervised video object segmentation method that capitalises on emergent properties of DINO-pretrained Vision Transformers. By extending a frozen ViT-S/8 backbone with sparse cross-frame attention and hierarchical KL-divergence clustering, the method produces per-frame segmentation masks without any labels or parameter updates at inference. Evaluated on DAVIS 2017 using FG-ARI and J&F, it achieves competitive performance against supervised baselines, demonstrating that self-supervised features alone are sufficient for general-domain object segmentation. However, the evaluation is confined to short, object-centric video clips, leaving the transferability of these features to complex driving scenes an open question.

Yao et al. [6] extend self-supervised video object segmentation to the amodal setting, where the objective is to recover the full geometric extent of an object including hidden regions.

Crucially for this work, they evaluate on the KINS-Video-Car dataset, situating their method within the autonomous driving domain and explicitly addressing generalisation under distribution shift. While this establishes that self-supervised segmentation can be applied to driving data, the evaluation focuses on amodal mask recovery rather than the detection of small, distant objects, which represents a distinct and safety-critical challenge in the domain.

Goyal et al. [7] take a semi-supervised approach, proposing a clip-based DETR-style encoder-decoder architecture for simultaneous segmentation and tracking. The method explicitly targets three challenges relevant to driving: maintaining object identity across long sequences, segmenting small objects, and handling large appearance changes over time. Evaluated on DAVIS 2017, it demonstrates that multi-scale features are particularly important for small and distant objects making it a finding that motivates the resolution experiments conducted in this work.

Taken together, these studies establish that self-supervised video object segmentation is effective in general domains [4], shows a promise in driving-adjacent settings [6], and that small distant objects remain a core open challenge [7]. What remains unaddressed is a systematic evaluation of self-supervised segmentation under controlled domain shift between driving datasets, with explicit analysis of far-object performance. This work directly targets that gap by evaluating SSL-UVOS on KITTI MOTS and Waymo under twelve pipelines varying pretraining domain and input resolution, with dedicated zone-stratified analysis of far-zone segmentation quality.

III. BACKGROUND

A. SimCLR

SimCLR [5] is a contrastive self-supervised learning framework that learns visual representations without labels. Given an input image, two augmented views are generated by randomly applying transformations such as cropping, color jittering and Gaussian blur. Both views are encoded by a shared backbone and projected into a contrastive embedding space via a small MLP projection head. The projection head is discarded after pretraining meaning only the backbone representations are retained for downstream tasks.

Representation learning is driven by the NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss, which maximizes agreement between the two views of the same image (positive pair) while pushing apart representations from different images (negative pairs). For a positive pair (i, j) , the loss is:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \quad (1)$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$ is cosine similarity, τ is a temperature parameter controlling the sharpness of the distribution over negatives, and the sum runs over all $2N - 1$ other examples in the batch. A lower τ enforces stricter separation of negative pairs. Because every other sample in

the batch serves as a negative, SimCLR requires large batch sizes to expose the model to sufficiently many negatives during training [5].

B. DINO and SSL-UVOS

DINO (Self-Distillation with **No** labels) [3] is a self-supervised learning method for Vision Transformers (ViTs) that eliminates the need for annotated training data. Rather than supervising the model with ground-truth labels, DINO employs a student-teacher architecture in which both networks share the same architecture but are updated differently: the student is trained via gradient descent while the teacher's weights are updated as an exponential moving average of the student. Both networks process differently augmented views of the same image, and the student is trained to match the teacher's output distribution. A key emergent property of this training scheme is that the self-attention maps of a DINO-pretrained ViT naturally segment semantically meaningful foreground regions without any supervision, making it particularly well-suited as a backbone for unsupervised object segmentation.

SSL-UVOS [4], introduced in *Betrayed by Attention*, extends a frozen DINO ViT-S/8 backbone for unsupervised video object segmentation. Patch-level query and key attention features are extracted for each frame and used to compute sparse cross-frame attention maps by correlating each frame's query features against key features sampled from every r -th keyframe. These spatio-temporal attention maps are then segmented via hierarchical clustering using KL-divergence as a merging criterion, iteratively grouping patches into coherent regions until convergence. The result is a per-frame segmentation mask produced entirely without labels or parameter updates at inference time. SSL-UVOS achieves competitive performance against supervised and semi-supervised methods on the DAVIS 2017 benchmark, establishing it as the base framework used throughout this work.

IV. METHODOLOGY

A. Data source

KITTI MOTS. [8] Provides pixel-accurate instance segmentation masks for cars and pedestrians across 21 training sequences of urban driving video.

Waymo Open Dataset. [9] Provides high-resolution video from a fleet of autonomous vehicles covering diverse urban and suburban environments. We use the camera-based segmentation annotations for evaluation.

B. Data Analysis

Table I summarises key properties of the two evaluation datasets. The most notable difference is instance density: Waymo averages 21 annotated instances per frame compared to 6 for KITTI, reflecting substantially more complex multi-object scenes. This difference in scene complexity provides a structural explanation for the performance gap observed across all pipelines, as denser scenes introduce more ambiguity for the clustering-based segmentation approach and increase noise in metric computation.

Beyond instance density, both datasets represent a significant domain shift from the data distributions on which the base model was developed. The DINO ViT-S/8 backbone was pretrained on ImageNet, and SSL-UVOS was originally evaluated on DAVIS 2017, a dataset of short, object-centric clips with controlled camera motion and limited scene clutter. Both KITTI and Waymo differ substantially from this distribution in terms of camera motion, scene density, and visual diversity, motivating the domain-specific pretraining pipelines explored in this work.

TABLE I
DATASET STATISTICS FOR KITTI MOTS AND WAYMO.

	KITTI MOTS	Waymo
Sequences	21	20
Total frames	8,008	5,010
Avg. instances/frame	6	21
Classes	Car, Pedestrian	Truck, Traffic light, Bus
Annotation type	Pixel-accurate	Pixel bounding box

C. Base Model: SSL-UVOS

We use SSL-UVOS [4] as the base framework throughout all experiments. The model consists of three components:

- 1) **Feature extraction.** A frozen DINO ViT-S/8 backbone extracts patch-level query (Q) and key (K) attention features for each frame.
- 2) **Sparse cross-frame attention.** Key frames are sampled every r -th frame (ratio $r=8$). Query features from each frame are correlated with key features of the selected frames via scaled dot-product attention, producing a spatio-temporal attention map.
- 3) **Hierarchical clustering.** Patch-level attention distributions are iteratively merged using KL-divergence until convergence, yielding clusters corresponding to coherent image regions. Each patch is assigned to its nearest cluster, producing a per-frame segmentation map.

No labels are used at any stage and model weights are not updated during inference.

D. Domain-Specific Pretraining with SimCLR

For pipelines that include pretraining, we fine-tune the DINO ViT-S/8 temporal transformed head using SimCLR [5] on unlabeled frames from the target driving domain. SimCLR learns representations by maximizing the agreement between two augmented views of the same image using a contrastive loss in a projected embedding space, without requiring any labels.

All pretraining runs use the following configuration: input resolution 160×512 pixels, patch temperature $\tau=0.5$ (standard for SimCLR), 25 epochs, batch size 32, learning rate 5×10^{-5} , and `max_frames` set to `None`. The lower resolution was chosen due to computational resource constraints; to ensure fair comparison across all pipelines, the same resolution is used consistently at both pretraining and inference.

E. Ground Truth Preprocessing

To better reflect real-world driving conditions, we apply a preprocessing step to the ground truth annotation masks before evaluation. Specifically, a script estimates the approximate distance of each annotated object from the camera using its pixel area: objects occupying more pixels are assumed to be closer, while those occupying fewer pixels are assumed to be farther away.

We acknowledge this is a pragmatic approximation with known limitations. Most notably, a nearby pedestrian may occupy fewer pixels than a large vehicle at a greater distance, leading to incorrect distance estimates in such cases. Nevertheless, given the absence of depth annotations in the evaluation sets, this heuristic serves as a simple and reproducible preprocessing step, and its limitations are taken into account when interpreting results.

F. Evaluation Metrics

We adopt FG-ARI and J&F as our primary evaluation metrics, following the evaluation protocol of SSL-UVOS [4], which uses both metrics on the DAVIS 2017 benchmark. Beyond methodological consistency, both metrics are widely established in the video object segmentation literature, making our results directly comparable to prior work.

FG-ARI. The Foreground Adjusted Rand Index measures cluster assignment agreement with ground-truth instance labels on foreground pixels in a permutation-invariant manner. It requires at least two ground-truth instances per frame.

J (Jaccard / mean IoU.) Predicted clusters are matched to ground-truth instances using the Hungarian algorithm on the IoU cost matrix. J is the mean IoU across matched pairs, with unmatched instances scoring 0.

F (Boundary F-measure). Using the same matching, F measures precision and recall of predicted segment boundaries against ground-truth boundaries dilated by a tolerance proportional to the image diagonal.

J&F. The arithmetic mean of J and F, following the DAVIS evaluation protocol. This is our primary reported metric alongside FG-ARI.

1) *Metric Benefits and Limitations:* Both metrics share a common weakness: neither captures **temporal consistency**. Scores are computed per frame independently and averaged, so a model that jumps between segmentations frame-to-frame can score identically to one that tracks objects smoothly, this is a critical gap for autonomous driving applications. Both metrics also **weight objects by pixel area**, meaning large nearby vehicles dominate scores while small but safety-critical objects such as distant pedestrians contribute minimally.

FG-ARI benefits from being permutation-invariant and label-free, making it well-suited to unsupervised evaluation where predicted cluster IDs carry no semantic meaning. Its main limitations are that it **ignores background errors** entirely, false positives on road surfaces go unpenalised, and it **excludes single-object frames**, which can discard a significant portion of light-traffic scenes and bias results.

J&F benefits from capturing both region overlap (J) and boundary accuracy (F), offering a more complete picture of segmentation quality. However, the F component is sensitive to **resolution**: at the reduced inference resolution of 160×512 used here, fine boundary details are lost and F scores are suppressed across all pipelines regardless of true segmentation quality. Additionally, its **binary one-to-one matching** penalises object splits disproportionately: if a single object is predicted as two segments, only the better half is matched and the other scores zero.

G. Evaluation Pipelines

We define twelve evaluation pipelines organised into three groups: four resolution-aware baselines using the original SSL-UVOS model without any pretraining, and two sets of four domain-specific pretraining pipelines run at different resolutions.

1) Baseline Pipelines (No Pretraining):

- 1) **No pretraining** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **KITTI** (160×512). The original SSL-UVOS model evaluated directly on KITTI MOTS at the lower resolution, establishing the out-of-the-box KITTI baseline.
- 2) **No pretraining** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **Waymo** (160×512). The same unmodified model evaluated on Waymo at the lower resolution, establishing the out-of-the-box Waymo baseline.
- 3) **No pretraining** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **KITTI** (240×768). The original SSL-UVOS model evaluated directly on KITTI MOTS at the higher resolution, isolating the effect of resolution relative to Baseline 1.
- 4) **No pretraining** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **Waymo** (240×768). The same unmodified model evaluated on Waymo at the higher resolution, isolating the effect of resolution relative to Baseline 2.

2) Pretraining Pipelines at 160×512 :

- 5) **Pretrain (KITTI)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **KITTI**. In-domain pretraining and evaluation on KITTI at the lower resolution, measuring the effect of domain adaptation against Baseline 1.
- 6) **Pretrain (KITTI)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **Waymo**. KITTI-pretrained model evaluated on Waymo, probing cross-domain generalisation at 160×512 .
- 7) **Pretrain (Waymo)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **Waymo**. In-domain pretraining and evaluation on Waymo at the lower resolution, measuring domain adaptation against Baseline 2.
- 8) **Pretrain (Waymo)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **KITTI**. Waymo-pretrained model evaluated on KITTI, completing the cross-domain pair symmetric to Pipeline 6.

3) Pretraining Pipelines at 240×768 :

- 9) **Pretrain (KITTI)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **KITTI**. In-domain pretraining and evaluation on KITTI at the higher resolution, isolating the effect of resolution relative to Pipeline 5.
- 10) **Pretrain (KITTI)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **Waymo**. KITTI-pretrained model evaluated on Waymo at $240 \times$

768 , testing whether higher resolution improves cross-domain transfer relative to Pipeline 6.

- 11) **Pretrain (Waymo)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **Waymo**. In-domain pretraining and evaluation on Waymo at the higher resolution, isolating the effect of resolution relative to Pipeline 7.

- 12) **Pretrain (Waymo)** $\rightarrow$ **DINO ViT-S/8** $\rightarrow$ **KITTI**. Waymo-pretrained model evaluated on KITTI at 240×768 , completing the higher-resolution cross-domain pair symmetric to Pipeline 10.

4) *Motivation*: The twelve pipelines address three orthogonal questions: (1) does domain-specific self-supervised pretraining improve on the unmodified baseline, (2) do features learned on one driving dataset transfer to another, and (3) does increasing inference resolution from 160×512 to 240×768 yield consistent gains. By holding all other factors fixed between paired pipelines (e.g. Pipelines 5 and 9 differ only in resolution), the design enables controlled, interpretable comparisons without any labelled data.

H. Implementation Details

All experiments use the publicly released SSL-UVOS checkpoint. Frames are resized to 160×512 for inference. Clustering uses $\tau=1.0$ and runs for up to 10,000 iterations or until convergence. Cross-frame attention is computed with key frame ratio $r=8$ and chunk size 1 to avoid memory overflow on long sequences. All experiments are run on a single NVIDIA L4 GPU with 23 GB VRAM.

V. RESULTS

A. Overall Performance

Tables II and III report overall and zone-stratified J&F for all six runs on KITTI. Tables IV and V show the same for Waymo. **Green bold** marks the highest value in each highlighted column (FG-ARI and J&F for overall tables; best per zone for zone tables).

On KITTI (Table II), the 240×768 baseline achieves the strongest overall J&F (0.1402), and the KITTI 240×768 pretrain matches it closely (0.1388) while attaining the highest FG-ARI of any run (0.1690). Waymo-pretrained checkpoints underperform the baselines on every KITTI metric, suggesting that Waymo features do not transfer well to KITTI in this setting.

On Waymo (Table IV), the 160×512 baseline is the top-performing run, leading all others in both FG-ARI (0.3364) and J&F (0.0409). Domain-specific pretraining on Waymo data does not help—the Waymo 240×768 pretrain is the weakest Waymo run across all metrics—pointing to possible overfitting in the projection head or degradation of DINO’s general-purpose features.

B. Zone-Stratified Performance

The zone breakdowns (Tables III and V) reveal important differences between proximity ranges. On KITTI, the 160×512 baseline leads near-zone J&F (0.5775), but the 240×768 baseline takes the top far-zone score (0.0897),

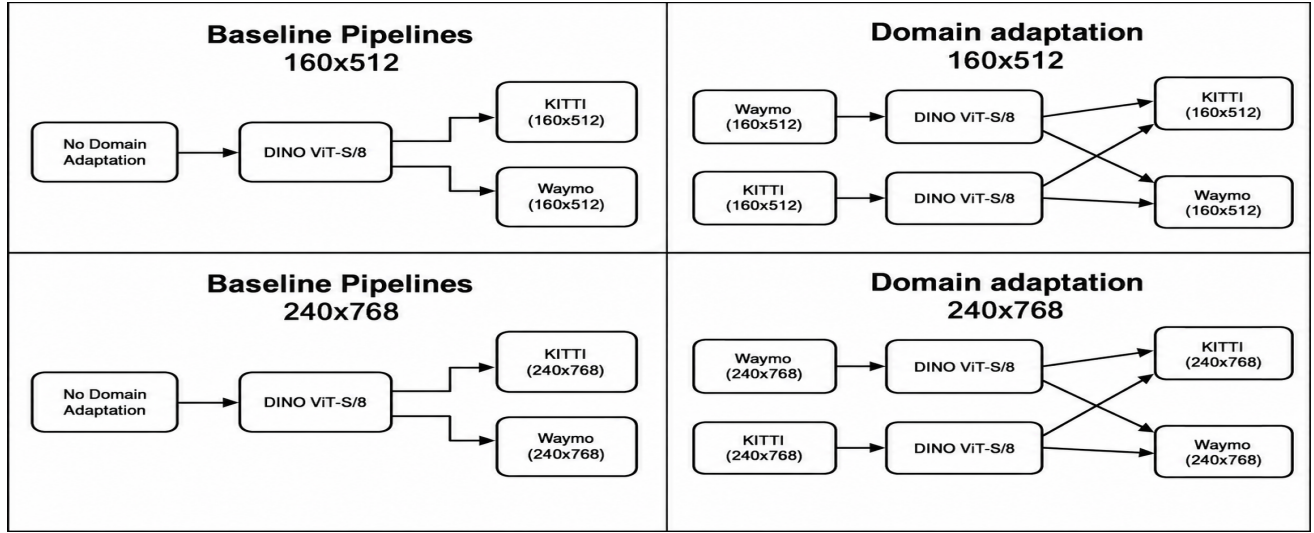


Fig. 1. Visual representation of evaluation pipelines

TABLE II
KITTI OVERALL METRICS.

(a) KITTI — overall

Run	FG-ARI	J	F	J&F
Baseline 160×512	0.1576	0.1079	0.1648	0.1363
Baseline 240×768	0.1665	0.1109	0.1694	0.1402
Pretrain KITTI 160×512	0.1646	0.1041	0.1556	0.1298
Pretrain KITTI 240×768	0.1690	0.1121	0.1656	0.1388
Pretrain Waymo 160×512	0.1605	0.1013	0.1520	0.1267
Pretrain Waymo 240×768	0.1575	0.1017	0.1527	0.1272

TABLE III
KITTI ZONE-STRATIFIED J&F.

(b) KITTI — zone J&F

Run	Near J&F	Mid J&F	Far J&F
Baseline 160×512	0.5775	0.3421	0.0828
Baseline 240×768	0.5397	0.3391	0.0897
Pretrain KITTI 160×512	0.5348	0.3396	0.0774
Pretrain KITTI 240×768	0.5130	0.3600	0.0861
Pretrain Waymo 160×512	0.5440	0.3170	0.0767
Pretrain Waymo 240×768	0.5466	0.3279	0.0755

TABLE IV
WAYMO OVERALL METRICS.

(c) Waymo — overall

Run	FG-ARI	J	F	J&F
Baseline 160×512	0.3364	0.0374	0.0444	0.0409
Baseline 240×768	0.3143	0.0365	0.0449	0.0407
Pretrain KITTI 160×512	0.3301	0.0360	0.0430	0.0395
Pretrain KITTI 240×768	0.3237	0.0353	0.0417	0.0385
Pretrain Waymo 160×512	0.3232	0.0352	0.0418	0.0385
Pretrain Waymo 240×768	0.3127	0.0321	0.0383	0.0352

TABLE V
WAYMO ZONE-STRATIFIED J&F.

(d) Waymo — zone J&F

Run	Near J&F	Mid J&F	Far J&F
Baseline 160×512	0.1430	0.0213	0.0036
Baseline 240×768	0.1445	0.0212	0.0027
Pretrain KITTI 160×512	0.1396	0.0196	0.0032
Pretrain KITTI 240×768	0.1397	0.0178	0.0020
Pretrain Waymo 160×512	0.1357	0.0199	0.0030
Pretrain Waymo 240×768	0.1268	0.0183	0.0015

TABLE VI
FAR-ZONE METRICS — KITTI ($n=2,143$ FRAMES).

(e) KITTI — far-zone

Run	Far J	Far F	Far J&F	Δ vs. B ₁₆₀
Baseline 160×512	0.0567	0.1089	0.0828	—
Baseline 240×768	0.0619	0.1175	0.0897	+0.0069
Pretrain KITTI 160×512	0.0537	0.1011	0.0774	−0.0054
Pretrain KITTI 240×768	0.0606	0.1116	0.0861	+0.0033
Pretrain Waymo 160×512	0.0536	0.0998	0.0767	−0.0061
Pretrain Waymo 240×768	0.0519	0.0991	0.0755	−0.0073

TABLE VII
FAR-ZONE METRICS — WAYMO ($n=3,830$ FRAMES).

(f) Waymo — far-zone

Run	Far J	Far F	Far J&F	Δ vs. B ₁₆₀
Baseline 160×512	0.0024	0.0048	0.0036	—
Baseline 240×768	0.0018	0.0035	0.0027	−0.0009
Pretrain KITTI 160×512	0.0021	0.0042	0.0032	−0.0004
Pretrain KITTI 240×768	0.0014	0.0026	0.0020	−0.0016
Pretrain Waymo 160×512	0.0021	0.0039	0.0030	−0.0006
Pretrain Waymo 240×768	0.0010	0.0020	0.0015	−0.0021

confirming that higher resolution specifically helps for small distant objects. The KITTI 240×768 pretrain leads the mid-zone (0.3600). On Waymo, the 240×768 baseline leads near-zone (0.1445), while the 160×512 baseline dominates mid and far zones.

C. Far-Zone Focus

Far-object segmentation is of particular safety relevance in autonomous driving, since small distant objects (pedestrians, cyclists) must be detected early. Tables VI and VII report far-zone J, F, and J&F in detail, together with the absolute change relative to the 160×512 baseline. **Green bold** marks the highest Far J and Far J&F within each table.

On KITTI (Table VI), higher resolution is the single most effective lever for far-zone performance: the 240×768 baseline leads with Far J&F of 0.0897, a gain of +0.0069 over its 160×512 counterpart. Among pretrained runs, only the KITTI 240×768 pretrain improves on the 160×512 baseline (+0.0033); all Waymo-pretrained checkpoints regress.

On Waymo (Table VII), far-zone scores are extremely low across all runs (below 0.004), and the 160×512 baseline remains the best performer. Higher resolution does not help here: the 240×768 baseline drops to 0.0027 (−0.0009). Every pretrained checkpoint degrades far-zone performance, with the Waymo 240×768 pretrain reaching the lowest score (0.0015), less than half the baseline.

VI. CONCLUSION

We evaluated SSL-UVOS under twelve pipelines varying pretraining domain, evaluation dataset, and input resolution, with a particular focus on far-object segmentation in autonomous driving scenarios. Our experiments yield five main findings.

First, SSL-UVOS generalises considerably better to KITTI than to Waymo. The model achieves FG-ARI of up to 0.169 and J&F of up to 0.140 on KITTI, compared to FG-ARI of 0.336 and J&F of 0.041 on Waymo. This asymmetry is not surprising given two structural factors. The Waymo Open Dataset is substantially more complex than KITTI MOTs: it contains a far larger number of annotated instances per scene, covers more diverse environments and exhibits richer label density, all of which increase evaluation difficulty and introduce more noise into metric computation. Additionally, the DINO ViT-S/8 backbone used by SSL-UVOS was pre-trained on ImageNet and evaluated extensively on DAVIS 2017, a dataset of short, object-centric video clips whose visual characteristics (controlled camera motion, prominent foreground objects, limited scene clutter) are considerably closer to KITTI than to Waymo. This prior distributional alignment gives the model a structural advantage on KITTI and makes generalisation to Waymo’s more complex, multi-instance driving scenes inherently harder.

Second, higher input resolution consistently improves far-object segmentation on KITTI. Increasing from 160×512 to 240×768 yields a far-zone J&F gain of +0.0069 on KITTI, the largest single improvement across all interventions tested.

This effect did not hold on Waymo, where the added resolution may amplify scene complexity and annotation noise rather than recovering fine object detail.

Third, domain-specific pretraining with SimCLR provides marginal and inconsistent benefits. On KITTI, the KITTI 240×768 pretrain modestly outperforms the unmodified baseline on FG-ARI and far-zone J&F, but the 160×512 counterpart regresses. On Waymo, no pretrained checkpoint surpasses the baseline on any metric; the Waymo 240×768 pretrain is the weakest run overall. We attribute this to the frozen backbone design: SimCLR fine-tunes only the temporal transformer head while leaving DINO’s spatial features fixed which is likely insufficient to adapt to Waymo’s substantially different data distribution.

Fourth, the model performs reasonably well at grouping foreground instances (FG-ARI), but struggles to achieve precise spatial matching (J&F), particularly at reduced resolution where boundary details are lost and the F component is systematically suppressed.

Fifth, despite the contributions made far-zone segmentation still remains an open challenge across both datasets. Scores on Waymo’s far zone are near zero across all pipelines, and even the best KITTI far-zone result (0.0897) reflects the inherent difficulty of segmenting small, distant objects at limited resolution and without any supervision.

Together, these findings suggest that resolution is currently the most actionable factor for improving far-zone segmentation, while the benefit of contrastive pretraining is limited by the frozen backbone design. The performance gap between KITTI and Waymo further highlights the importance of dataset complexity and domain proximity to the backbone’s pretraining distribution. Future work should explore end-to-end fine-tuning, larger and more diverse pretraining corpora, explicit handling of dense multi-instance scenes, and evaluation at native camera resolution to better characterise the potential of domain adaptation for safety-critical autonomous driving applications.

ACKNOWLEDGMENT

Both authors contributed equally to all aspects of this work, including experimental design, implementation, evaluation, and writing.

REFERENCES

- [1] M. Gao, F. Zheng, J. J. Yu, C. Shan, G. Ding, and J. Han, “Deep learning for video object segmentation: a review,” *Artificial Intelligence Review*, vol. 56, pp. 457–531, 2023.
- [2] T. Delatolas, V. Kalogeiton, and D. P. Papadopoulos, “Learning the what and how of annotation in video object segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 6951–6961.
- [3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
- [4] S. Ding, R. Qian, H. Xu, D. Lin, and H. Xiong, “Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation,” in *European Conference on Computer Vision*. Springer, 2024, pp. 215–233.

- [5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.
- [6] J. Yao, Y. Hong, C. Wang, T. Xiao, T. He, F. Locatello, D. P. Wipf, Y. Fu, and Z. Zhang, “Self-supervised amodal video object segmentation,” *Advances in neural information processing systems*, vol. 35, pp. 6278–6291, 2022.
- [7] R. Goyal, W.-C. Fan, M. Siam, and L. Sigal, “Tam-vt: Transformation-aware multi-scale video transformer for segmentation and tracking,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 8336–8345.
- [8] P. Voigtlaender, M. Krause, A. Ošep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, “MOTS: Multi-object tracking and segmentation,” in *CVPR*, 2019.
- [9] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.