IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Uni-DVPS: Unified Model for Depth-Aware Video Panoptic Segmentation

Kim Ji-Yeon, Oh Hyun-Bin, Kwon Byung-Ki, Dahun Kim, Yongjin Kwon, and Tae-Hyun Oh

***Abstract*— We present Uni-DVPS, a unified model for Depth-aware Video Panoptic Segmentation (DVPS) that jointly tackles distinct vision tasks, *i.e.*, video panoptic segmentation, monocular depth estimation, and object tracking. In contrast to the prior works that adopt diverged decoder networks tailored for each task, we propose an architecture with a unified Transformer decoder network. We design a single Transformer decoder network for multi-task learning to increase shared operations to facilitate the synergies between tasks and exhibit high efficiency. We also observe that our unified query learns instance-aware representation guided by multi-task supervision, which encourages query-based tracking and obviates the need for training extra tracking module. We validate our architectural design choices with experiments on Cityscapes-DVPS and SemKITTI-DVPS datasets. The performances of all tasks are jointly improved, and we achieve state-of-the-art results on DVPQ metric for both datasets. Project page: `https://jiyeon-k1m.github.io/uni-dvps`**

I. INTRODUCTION

Depth-aware Video Panoptic Segmentation (DVPS) involves the collaborative solution of three core tasks in computer vision: panoptic segmentation, monocular depth estimation, and object tracking. The integration of diverse capabilities into a single model can equip machines with a nuanced understanding of 3D scenes, from coarse layouts to fine-grained details. Recent efforts [1], [2], [3], [4] have been dedicated to constructing a multi-task learning system, attempting to benefit from the synergies between sub-tasks.

Despite the advancements in the pursuit of integration, a closer look at the existing methods reveals that they actually opt for distinct decoder networks tailored to each task, with the encoder backbone being the sole element shared across tasks. This configuration results in non-negligible computational load within these task-specific decoders, somewhat obscuring the core drive behind the quest for integration. To illustrate, ViP-DeepLab [1] employs semantic, instance, and next-frame instance decoders, further branching the semantic decoder into segmentation and depth pathways. Similarly, PolyphonicFormer [2] adopts heterogeneous panoptic and depth decoders, and MonoDVPS [3] uses a semantic-and-depth decoder, an instance decoder, and an optical flow decoder, separately. In response to this dissonance, we propose a unified Transformer decoder, facilitating the maximization of shared computations across tasks. We demonstrate that this unified decoder implementation surpasses the performance of multiple task-specialized decoder networks while streamlining redundant task-specific computations.

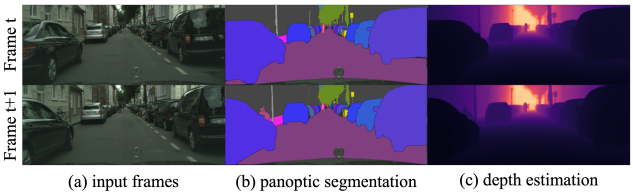


Fig. 1. **Depth-aware video panoptic segmentation**. Given an input frame, Uni-DVPS predicts both panoptic segmentation and depth estimation maps. We achieve temporally consistent object tracking by matching predicted queries across consecutive frames. Note that we eliminate the need for training extra tracking module.

In addition, the design of our unified query-based decoder is motivated by recent advances in query-based image detection [5]. By virtue of our unified query-based decoder design, query embeddings are learned to represent each instance in an independent frame. Combining our unified decoder with the sub-tasks of DVPS, the queries are enriched with both semantic and depth information which turns out helping to get more discriminative and temporally-consistent representations of the instances. This allows to obviate the need for a dedicated tracking head, and tracking is effortlessly enabled at inference through bipartite matching of the queries across consecutive frames. It also makes our model trainable in an image-based manner with no need for video-level training recipes, tracking losses, and tracking annotations. This hints that our query trained on the panoptic segmentation and depth sub-tasks is sufficient for tracking during the inference stage.

We demonstrate that our simple and unified design is efficient in memory and inference speed while providing enhanced performance, and our work is also distinguished from the existing DVPS approaches which require a separate

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government, MSIT, under Grant 2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network (60%), and in part by the Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government, MSIT, under Grant 2022-0-00290, Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense (40%). *(Corresponding author: Tae-Hyun Oh.)*

Kim Ji-Yeon is with the Department of Convergence IT Engineering, POSTECH, Pohang 37673, South Korea (e-mail: jiyeon.kim@postech.ac.kr).

Oh Hyun-Bin is with the Department of Electrical Engineering, POSTECH, Pohang 37673, South Korea (e-mail: hyunbinoh@postech.ac.kr).

Kwon Byung-Ki is with the Graduate School of AI, POSTECH, Pohang 37673, South Korea (e-mail: byungki.kwon@postech.ac.kr).

Dahun Kim is with Google DeepMind, Mountain view, CA 94043 USA (e-mail: mcahny@google.com).

Yongjin Kwon is with Superintelligence Creative Research Laboratory, ETRI, Daejeon, 34129 South Korea (e-mail: scocso@etri.re.kr).

Tae-Hyun Oh is with the Department of Electrical Engineering, Graduate School of AI, POSTECH, Pohang 37673, South Korea, and also with the Institute for Convergence Research and Education in Advanced Tech., Yonsei University, Seoul 03722, South Korea (e-mail: taehyun@postech.ac.kr).

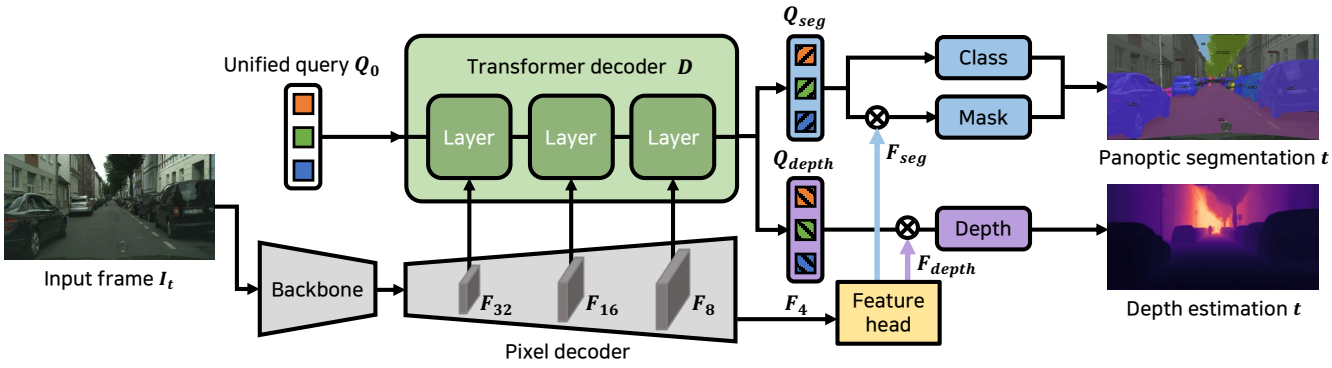


Fig. 2. **Architecture**. Given an input frame, the image encoders extract the multi-scale image features $\mathcal{F}$. The unified Transformer decoder takes the initialized unified query Q_0 and the extracted image features $\mathcal{F}$ to update the query with the Transformer attention mechanism. Then the unified query is diverged into each task and predicts per-segment segmentation mask and depth map by convolving the transformed features. The instance-wise results are merged to make the final prediction map. We use three iterative rounds of Transformer decoder blocks.

tracking module: a next-frame instance decoder [1], a tracking head [2], or an external pre-trained optical flow model [3].

We summarize our main contributions as follows:

- We propose a unified architecture, Uni-DVPS, for depth-aware video panoptic segmentation. Our architectural design choice of entangling decoders shows high efficiency with fewer parameters and faster inference speed.
- Our unified query learns instance-aware distinctive representation guided by multi-task supervision. Tracking by query-based matching eliminates the need for adopting and training extra tracking module.
- We disentangle the task-specific features and achieve state-of-the-art performances.

II. Related Work

Depth-aware video panoptic segmentation. Joint learning of video panoptic segmentation and monocular depth estimation was first introduced by ViP-DeepLab [1]. ViP-DeepLab partially integrates semantic and depth decoders, which was reported to cause sub-optimal results [4]. PolyphonicFormer [2] employs two strong and distinct task-specific Transformer decoder heads based upon DETR [5] with the instance-wise query learning. They also employ an instance tracking head to simultaneously predict temporally consistent panoptic segmentation and depth results, which requires annotations for tracking, a tracking loss, and a video-based training scheme. Recently, MonoDVPS [3] proposes a semi-supervised model consisting of separated decoder heads of semantic segmentation, instance segmentation, depth, optical flow, and camera pose estimation. While MonoDVPS emphasizes the self-supervised losses for their modules, it additionally leverages the extended dataset for segmentation and a separate optical flow module pre-trained with a supervised dataset. The existing DVPS approaches exploit multiple and separate task-specific decoders and manually-designed tracking heads to devise a system for the multi-task learning of DVPS, resulting in redundant computations and representations. Distinctively, we adopt a single unified Transformer decoder without any manually-designed tracking head.

Per-frame tracking in video segmentation. Recent lines of research, *e.g.*,[6], [7], employs additional tracking heads or proposes temporal loss functions, *e.g.*, a spatial extent aware loss of segments and pixels [8], and contrastive learning [9] to perform instance tracking. On the other hand, MinVIS [10] proposes a per-frame VIS model that shows comparable results without any tracking head or video-based training procedure. Our work applies per-frame mechanism to the DVPS task, where multiple sub-tasks compete with trade-offs [11]. In particular, we demonstrate that a similar instance query grouping phenomenon still holds even with the joint segmentation and depth multi-task setup and it is sufficient for tracking.

Unified architectures for multi-tasks. Recently, there have been advanced approaches [12], [13], [14], [15] to tackle multiple tasks with a single unified architecture. Rather than adopting task-independent experts that solely specialized within their domains, a unified model that handles multiple tasks at once benefits in performance and efficiency. Unified-IO [16] argues that the shared representation across different tasks enables to train a single Transformer-based architecture on almost ninety dense prediction datasets. TaskPrompter [15] suggests spatial-channel multi-task prompting Transformer framework that exploits the spatial-channel wise interaction between tasks, *i.e.*, semantic segmentation and boundary detection. We thus follow this research movement to build an effective unified model, particularly in DVPS domain.

III. Method

A. Overview

Our goal is to build a unified model that effectively and jointly tackles the multiple involved sub-tasks, *e.g.*, video panoptic segmentation, depth estimation, and object tracking (see Fig. 1). In contrast to existing DVPS methods [1], [2], [3], which rely on task-specific Transformer decoders tailored for segmentation, depth prediction and tracking, our Uni-DVPS employs a unified Transformer decoder architecture as shown in Fig. 2. We also leverage instance-aware representation, *i.e.*, query, to associate instances across the frames without

training extra tracking module. Since we eliminate the need for applying temporal losses to train the tracking module, our model is trained in a per-frame manner.

B. Uni-DVPS Architecture

Uni-DVPS is composed of three modules: the feature extractor, unified Transformer decoder, and prediction heads. For the feature extractor, the backbone and pixel decoder extract the multi-scale pixel features $\mathcal{F} = \{\mathbf{F}_{32}, \mathbf{F}_{16}, \mathbf{F}_{8}, \mathbf{F}_{4}\}$ where $\mathbf{F}_s \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C}$ denotes different scale features.

Unified Transformer decoder. Distinguished from existing DVPS methods [1], [2], [3] that adopt individual decoders tailored for each segmentation, depth and tracking task, we propose a unified Transformer decoder to perform multi-task learning. Inspired by DETR [5], our unified Transformer decoder guides the object queries to learn segment-wise representations. Unlike early segmentation models [17], [18], our unified decoder embeds both segmentation and depth representations into the *unified queries* using multi-task supervision. The unified Transformer decoder takes the scaled pixel features and initial unified queries $\mathbf{Q}_0$ as input and then iteratively updates the queries to learn task-mutual representations at each layer. We set the initial queries $\mathbf{Q}_0 \in \mathbb{R}^{(N+1) \times C}$ as $(N+1)$ embeddings of dimension C and randomly initialize them following [10]. In each layer of decoder, the unified queries are updated by the Transformer attention mechanism. In particular, the queries are guided to attend to localized object regions in features, *i.e.*, masked-attention [18], and to share the context information through self-attention. Note that our unified queries represent distinctive features of each segment, *e.g.*, shape, texture, and depth. Such instance-aware queries learned by the unified decoder are utilized to associate objects across the frames, obviating the need for a video-based training regime. The distinctive and instance-aware queries are visualized in Fig. 3.

Integrating the separated Transformer decoders into a unified entity offers additional advantages. The well-known challenge of exponential growth in computational costs and complexity in Transformer-based architectures according to the size of an input image can significantly affect the DVPS task, especially when handling higher-resolution images. Streamlining the task-specific Transformer decoder heads into the unified structure increases shared components throughout the network, effectively reducing computational demands. Uni-DVPS, which embraces the unified Transformer decoder, exhibits a significant reduction in both parameter count and inference time compared to one of the DVPS methods with the task-specialized decoder networks, PolyphonicFormer [2]. More details about the model efficiency will be covered in Sec. IV-B.

Task predictions. In the task heads, the queries learned by the unified Transformer decoder are decoded into each task prediction with the features. The queries from the unified decoder are non-linearly transformed into segmentation queries $\mathbf{Q}_{seg} \in \mathbb{R}^{N \times C}$ and depth queries $\mathbf{Q}_{depth} \in \mathbb{R}^{(N+1) \times C}$ using three Multi-Layer Perceptron (MLP) layers. The feature

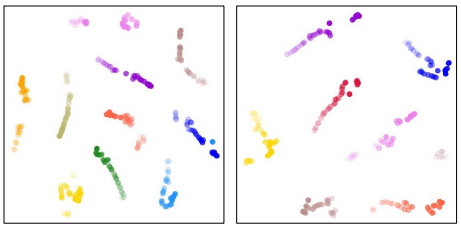


Fig. 3. **Visualization of our learned query embeddings.** We extract the learned queries from video clips and visualize them in each t-SNE [19] plot. Two plots show that the queries learn the instance-aware representations, which appear as distinctive colored clusters. The change of the color intensity implies the time change in video.

head, consisting of two MLP layers, transforms the highest resolution feature $\mathbf{F}_4$ into the segmentation feature $\mathbf{F}_{seg} \in \mathbb{R}^{N \times H \times W \times C}$ and depth feature $\mathbf{F}_{depth} \in \mathbb{R}^{(N+1) \times H \times W \times C}$. This division helps to disentangle the image features into each task and we empirically find that adopting the feature head helps to improve the performance.

For the segmentation task head, the masks $\hat{m} \in [0,1]^{N \times H \times W}$ of N instances are generated by convolving the segmentation queries with the segmentation feature as:

$$\hat{m} = \sigma(\mathbf{Q}_{seg} * \mathbf{F}_{seg}), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. Moreover, the classification scores $\hat{p} \in \mathbb{R}^{N \times (D+1)}$ of instances are predicted from the segmentation queries, where D denotes the number of ground-truth classes. Following [18], [10], we find the assignments between the predicted instances and the ground-truth instance using bipartite matching with the estimation of mask and class of each instance.

For the depth estimation, we first predict the normalized depth map $\hat{d} \in \mathbb{R}^{(N+1) \times H \times W}$ by convolving the depth queries with the depth feature as:

$$\hat{d} = \sigma(\mathbf{Q}_{depth} * \mathbf{F}_{depth}). \tag{2}$$

We leverage an additional backup query to handle non-segment regions that are difficult for segmentation queries to handle, *e.g.*, objects of unseen categories. Then we unnormalize the depth maps using fixed range as:

$$\hat{d} \cdot (d_{max} - d_{min}) + d_{min}, \tag{3}$$

where $[d_{min}, d_{max}]$ is given by the dataset. The estimated per-segment maps are merged into a final depth map using the binary segmentation maps.

C. Query-based Tracking

Recall that Uni-DVPS is designed as an image-based network that takes individual frames as an input and outputs segmented mask and depth in a frame-by-frame manner. Instead of employing an additional tracking head, we adopt a simple query-based tracking method in Uni-DVPS. We are motivated by MinVIS [10], while we establish multi-task queries and apply them to an entirely different domain, depth-aware video panoptic segmentation. We assume that if each query embedding is well-trained to represent individual instances within a single frame, we can associate instances

TABLE I

DEPTH-AWARE VIDEO PANOPTIC SEGMENTATION ON CITYSCAPES-DVPS VALIDATION SET.

Method	$k=1$			$k=2$			$k=3$			$k=4$			DVPQ$^k_\lambda$ Avg.		
	All	Thing	Stuff	All	Thing	Stuff	All	Thing	Stuff	All	Thing	Stuff	All	Thing	Stuff
ViP-DeepLab[1] [1]	47.4	38.8	53.7	44.0	28.1	51.6	39.0	23.3	50.5	37.5	20.2	50.0	42.0	27.6	51.5
PolyphonicFormer [2]	54.4	47.0	59.8	48.1	35.9	57.0	45.5	30.9	56.2	44.1	28.6	55.4	48.1	35.6	57.1
MonoDVPS[2] [3]	57.2	48.4	63.6	51.0	37.0	61.0	47.9	31.0	60.0	45.7	27.0	59.3	50.4	35.9	61.0
Uni-DVPS $\lambda = 0.50$	65.7	55.6	73.1	59.5	44.7	70.3	56.2	38.7	69.0	54.0	34.4	68.2	58.9	43.4	70.1
Uni-DVPS $\lambda = 0.25$	62.5	51.9	70.2	56.8	42.1	67.4	53.5	35.9	66.3	51.3	31.7	65.5	56.0	40.4	67.4
Uni-DVPS $\lambda = 0.10$	45.9	36.8	52.4	40.9	28.0	50.2	38.6	24.0	49.3	37.0	21.4	48.4	40.6	27.5	50.1
Uni-DVPS Avg.	58.0	48.1	65.2	52.4	38.3	62.7	49.5	32.9	61.5	47.4	29.2	60.7	**51.8**	**37.1**	**62.5**

by matching query embeddings across frames. We validate the concept with the visualization of the learned query embeddings in Fig. 3. Given two sampled video clips, we extract the learned query embeddings from each clip and visualize them in each plot. The same instances across the frames are indicated with the same color. We observe that the query embeddings representing the same identity across the frames get clustered and temporally consistent even if they are not trained with temporal cues. We apply the Hungarian matching algorithm between $\mathbf{Q}_t$ and $\mathbf{Q}_{t+1}$ and use the cosine similarity as the matching cost. Tracking by query matching allows us to remove the tracking head in our Uni-DVPS model, which simplifies the pipeline and distinguishes our method from the existing DVPS approaches.

D. Training and Loss functions

Our Uni-DVPS takes individual frames during training, all losses are calculated at the frame level. For the segmentation, we use the cross-entropy loss for the classification loss $\mathcal{L}_{cls}$, and a combination of the binary cross-entropy loss and the dice loss [20] for the mask loss as follows:

$$\mathcal{L}_{seg} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{mask}\mathcal{L}_{mask}. \tag{4}$$

For the depth estimation, the depth loss $\mathcal{L}_{depth}$ incorporates scale-invariant logarithmic loss [21], absolute relative loss, and square relative loss [22] as follows:

$$\mathcal{L}_{depth} = \lambda_{si}\mathcal{L}_{si} + \lambda_{abs}\mathcal{L}_{abs} + \lambda_{sq}\mathcal{L}_{sq}. \tag{5}$$

The total loss $\mathcal{L}_{total}$ comprises both segmentation loss $\mathcal{L}_{seg}$ and depth loss $\mathcal{L}_{depth}$ as:

$$\mathcal{L}_{total} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{depth}\mathcal{L}_{depth}. \tag{6}$$

Here, $\{\lambda_*\}$ denotes the weight parameters for each loss term.

IV. EXPERIMENTS

A. Experiment Setup

Datasets. Cityscapes-DVPS dataset [1] is an extended version of Cityscapes-VPS dataset [23], which provides depth annotations derived from the stereo disparity maps. The Cityscapes-DVPS dataset has 19 semantic categories, which include 8 'thing' and 11 'stuff' classes. In total, the dataset contains 3,000 annotated frames: 2,400 for training, 300 for validation, and 300 for test sets. SemKITTI-DVPS [1] extends SemanticKITTI [24] by projecting its 3D point clouds with panoptic labels (*i.e.*, semantic class and instance ID) to the image plane. The SemKITTI-DVPS dataset also has 19 semantic categories, comprising 8 'thing' and 11 'stuff' classes. In total, the dataset includes 19k training images (11 sequences), 4k evaluation images (sequence 08 in training sequences), and 4k test images (11 sequences).

Evaluation metrics. Following the established evaluation protocol [1], we employ Depth-aware Video Panoptic Quality (DVPQ) method to evaluate our model. DVPQ focuses on evaluating the quality of video panoptic segmentation varying the number of frames k with absolute relative depth errors within a pre-defined threshold λ. Additionally, we utilize the Panoptic Quality (PQ) metric for panoptic segmentation [25], and Absolute Relative Error (Abs. rel) and Root Mean Squared Error (RMSE) for depth estimation.

Implementation details. We use ResNet-50 [26] as a backbone. We pre-train our model on Mapillary [27] and Cityscapes [28] datasets as suggested by [1] and [2]. We use AdamW Optimizer [29] and the polynomial learning rate decay. We train our model for 9K iterations with a batch size of 16. We apply data augmentation for Cityscapes-DVPS with horizontal flipping and random color augmentations. We do not change image resolution. For segmentation, we set $\lambda_{cls} = 2.0$ and $\lambda_{mask} = 5.0$. For depth estimation, we set $\lambda_{si} = 3.0$, $\lambda_{abs} = 3.0$ and $\lambda_{sq} = 3.0$. Both tasks are balanced by setting $\lambda_{seg} = 1.0$ and $\lambda_{depth} = 1.0$. Frames per Second (FPS)[3] is measured on a single NVIDIA Tesla A100 GPU with a batch size of 1, including post-processing time.

B. Quantitative results

Depth-aware video panoptic segmentation. We evaluate our Uni-DVPS on Cityscapes-DVPS [1] dataset with Depth-aware Video Panoptic Quality (DVPQ) metric in Table I. As the number of frames k increases and the relative depth error threshold λ decreases, the evaluating conditions that the model should satisfy become harsh so that the performances drop

[1]ViP-DeepLab with the ResNet-50 backbone evaluated from the author's code and pre-trained model.

[2]Fully supervised version of MonoDVPS (S-MDE).

[3]For assessing the inference time, we employ the evaluation tools provided by Detectron2 for Uni-DVPS and MMdetection for [2], respectively.

TABLE II

DEPTH-AWARE VIDEO PANOPTIC SEGMENTATION ON SEMKITTI-DVPS VALIDATION SET.

Method	$k=1$	$k=5$	$k=10$	$k=20$	DVPQ^k_λ Avg.
PolyphonicFormer [2]	44.8	40.0	38.7	37.8	40.3
MonoDVPS [3]	43.3	38.1	36.9	36.0	38.6
Uni-DVPS $\lambda=0.50$	52.7	49.2	46.8	43.1	48.0
Uni-DVPS $\lambda=0.25$	50.4	46.9	44.5	40.8	45.7
Uni-DVPS $\lambda=0.10$	38.0	34.3	32.0	28.7	33.3
Uni-DVPS Avg.	47.0	43.5	41.1	37.5	**42.3**

compared to easier case, *i.e.*, $\lambda=0.5$ and $k=1$. Compared to the recent methods, our averaged performances outperform all other methods across all categories, *i.e.*, DVPQ^k_λ-All, Thing, and Stuff. We observe that Uni-DVPS surpasses the most recent approach with a large margin of +1.4 DVPQ-All, +1.2 DVPQ-Thing, and +1.5 DVPQ-Stuff.

We also evaluate our model on SemKITTI-DVPS [1] dataset with DVPQ metric in Table II. Uni-DVPS outperforms all the competing methods in DVPQ^k_λ Avg. We show comparable results in $k=20$ compared to PolyphonicFormer [2]. Note that our model is only trained with independent frames without any temporal annotations. Without any post-processing for video tracking, our per-frame learned model can surpass the models trained with temporal losses.

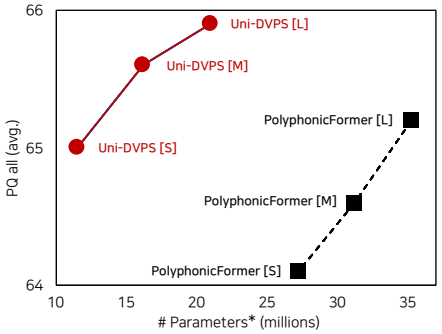
Model efficiency. The benefit of unifying the decoder is removing the redundant components, *i.e.*, parameter efficiency. We illustrate the trade-off between Panoptic Quality (PQ) and the number of parameters in Table IIIa. All the models in Table IIIa share a ResNet-50 backbone. Note that the number of parameters presented in the table is measured except for the backbone in the network. S, M, and L refer to 1, 2, and 3 iterative rounds of the transformer decoder block. Regardless of the number of rounds, our model consistently performs favorably against PolyphonicFormer [2] in terms of Panoptic Quality (PQ) while utilizing fewer parameters. Notably, in the case of large models (*i.e.*, 3 iterative rounds of Transformer decoder blocks), Uni-DVPS achieves superior results on DVPQ^k_λ Avg. while having approximately 40% fewer Transformer decoder parameters compared to PolyphonicFormer (See Table IIIb). Additionally, we measure the inference time and observe that Uni-DVPS operates at least $7.6\times$ faster than PolyphonicFormer when comparing models with the same iterative rounds. This shows, in contrast to PolyphonicFormer, Uni-DVPS's unified Transformer decoder effectively tackles multiple tasks without imposing additional computational burden.

C. Qualitative results

We visualize the results of panoptic segmentation and depth estimation on Cityscapes-DVPS validation set in Fig. 4. We observe that the same instances in the video represent the same color across the three consecutive frames. Training in a per-frame manner without employing temporal losses, Uni-DVPS combined with query-based tracking performs

TABLE III

MODEL EFFICIENCY.



(a) Trade-off between Panoptic Quality (PQ) and # Parameters.

Method	# layers	PQ ↑	DVPQ ↑	# Params*	FPS
PolyphonicFormer [S]	3	64.1	-	27.2M	1.2
PolyphonicFormer [M]	6	64.6	-	31.2M	1.2
PolyphonicFormer [L]	9	65.2	48.1	35.2M	1.2
Uni-DVPS [S]	3	65.0	50.3	11.5M	12.8
Uni-DVPS [M]	6	65.6	50.5	16.2M	10.7
Uni-DVPS [L]	**9**	**65.9**	**51.8**	**21.0M**	**9.2**

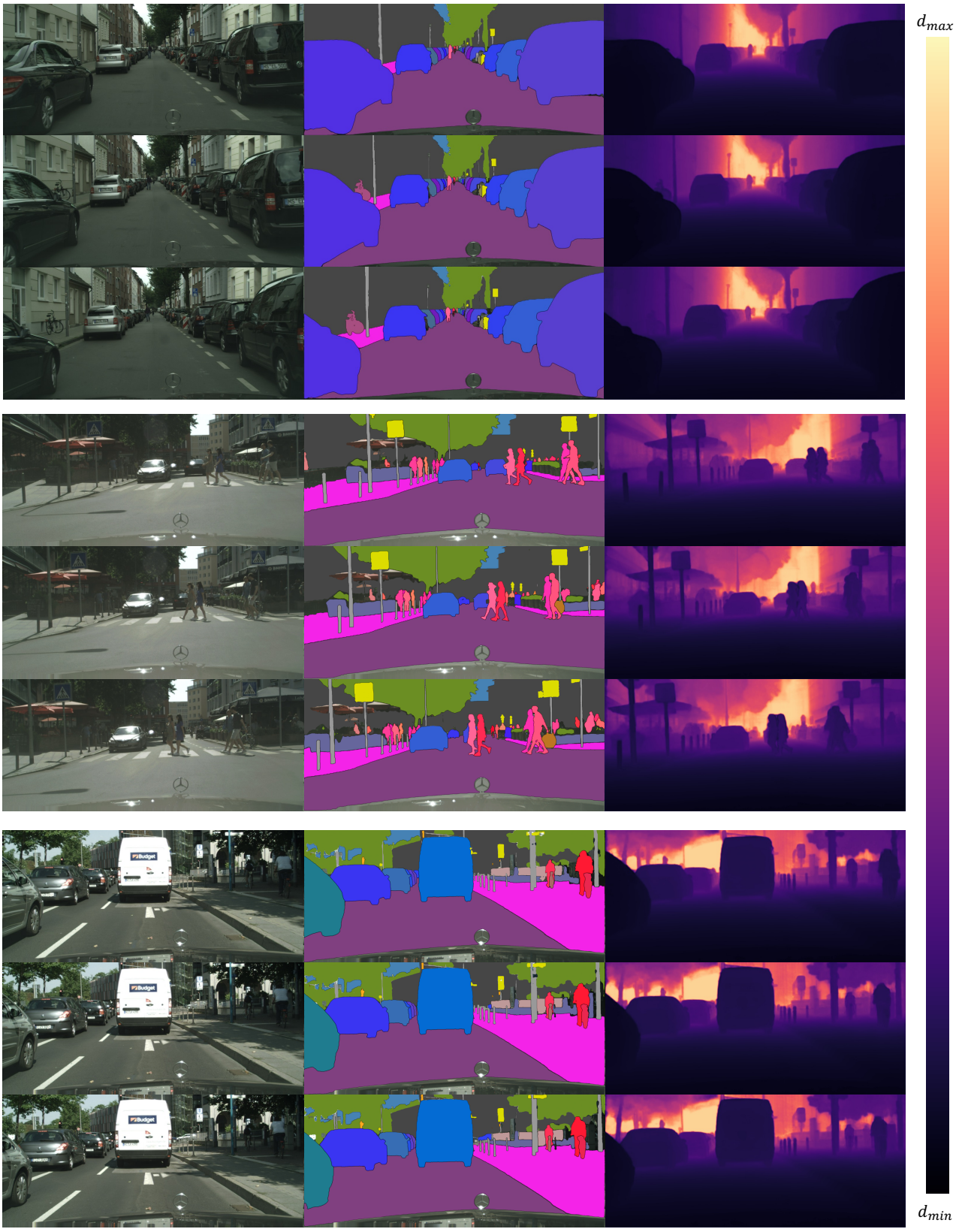
(b) PQ, DVPQ, and Performance varying model sizes.

effectively in DVPS tasks. In addition, we find that Uni-DVPS can handle dynamic objects in driving scenarios, *e.g.*, moving cars or walking people, with showing clear boundary in both segmentation and depth estimation results.

In Fig. 5, we visualize the results of panoptic segmentation and depth estimation on less-than-ideal conditions, *e.g.*, high dynamic range and occlusions. Compared to the competing method [2], our Uni-DVPS shows a clear boundary of distant and small objects in the scene. Also, our Uni-DVPS robustly performs instance segmentation even if the instance is far away and occluded by others. We also visualize the cross-attention maps of the learned unified queries in our unified Transformer decoder. As shown in Fig. 6, each query attends to the specific instance, which helps the query encode instance-aware representations. Such instance-aware queries are decoded with task-specific features, which yield instance-specific and detailed segmentation and depth maps and consistent object tracking.

D. Ablation Studies

Multi-task learning. Table IVa ablates the effect of the multi-task learning by comparing it with single task learning. We train Uni-DVPS model with only panoptic segmentation and only depth estimation datasets in Cityscpaes-DVPS for single-task learning. Along with the default panoptic segmentation learning, we compare the Panoptic Quality (PQ) and Video Panoptic Quality (VPQ) with and without the joint depth prediction learning. We also compare Absolute Relative Error (Abs. rel) and Root Mean Squared Error (RMSE) with and without the joint panoptic segmentation learning. We observe that Uni-DVPS enjoys a clear benefit of +3.5 PQ, +2.8 VPQ,



(a) input frames (b) video panoptic segmentation (c) depth estimation

Fig. 4. **Qualitative Results.** We visualize the the prediction results of panoptic segmentation and depth estimation for the consecutive ($T = 3$) input frames. Although Uni-DVPS is trained in a per-frame manner, the object tracking with query matching is well-performed in the video domain.

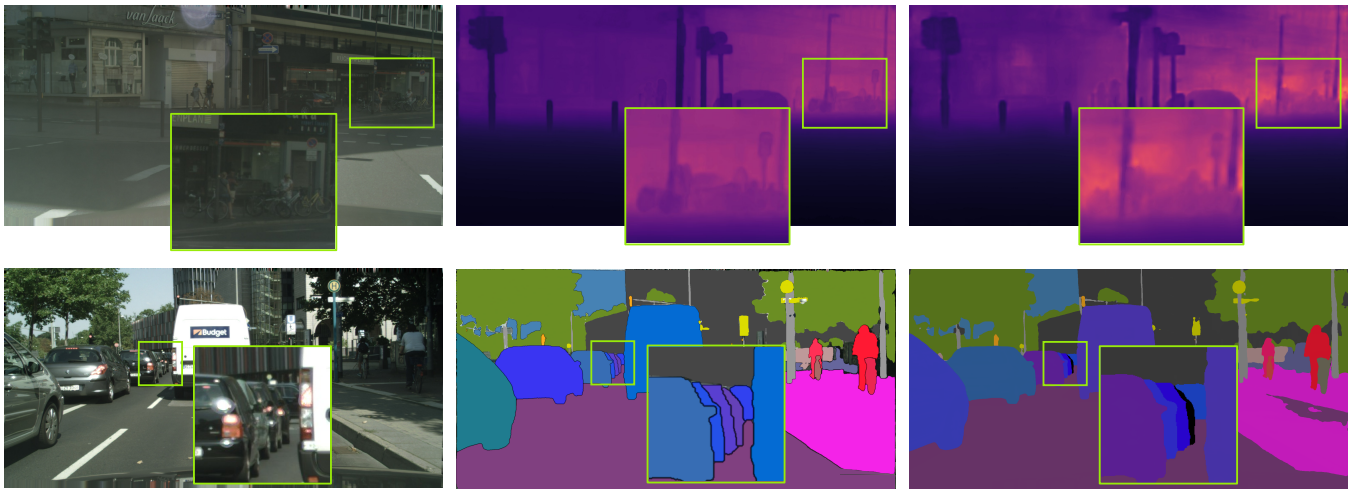


Fig. 5. **High dynamic range and occlusion scenarios.** We visualize the video panoptic segmentation and depth prediction results of Uni-DVPS and PolyphonicFormer on less-than-ideal conditions (*e.g.*, high dynamic range and occlusions) in the Cityscapes-DVPS dataset.

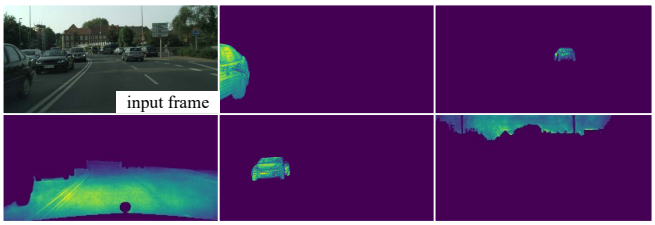


Fig. 6. **Visualization of cross-attention maps.** We visualize the cross-attention maps of learned queries in the unified Transformer decoder. Each query attends to the specific instance of the input frame, which helps to encode instance-aware representations.

+0.005 Abs. rel, and +0.3 RMSE from the joint multi-task learning, which validates the inter-connectedness between the segmentation and depth prediction tasks.

Unified Transformer decoder. In Table IVb, we evaluate the impact of our unified Transformer decoder on the depth-aware video panoptic quality (DVPQ) metric. To demonstrate, we establish a comparative model by substituting the unified Transformer decoder head in Uni-DVPS with two distinct Transformer decoder heads, leading to the increase of +14.0M parameters. We find that Uni-DVPS outperforms the two-decoder comparative model by a healthy margin of +0.9 DVPQ while saving significant number of parameters.

Architectural improvements. We verify our architectural design choices in two aspects: (1) the feature head module and (2) the scale of the unified Transformer decoder. In Table IVc, we compare Depth-aware Video Panoptic Quality (DVPQ) performances with and without task-specific features. We find that Uni-DVPS obtains at least +0.6 DVPQ gain from the feature head module, which well separates the shared features into task-specific features to focus on each task. On the other hand, we observe interesting results that stacking more layers in feature head module does not helpful for multi-task learning. We thus choose two layers of MLP as the feature head module for Uni-DVPS.

In Table IIIb, we evaluate the effect of scaling the unified transformer decoder varying the number of decoder layers. Note that a single round contains 3 Transformer decoder layers and we increase the number of rounds linearly. We find that both Panoptic Quality (PQ) and Video Panoptic Quality (VPQ) increases as the number of rounds increase. We set the round of Transformer decoder to three (*i.e.*, $L = 3$) considering the memory capacity in our computational environment.

We ablate the architectural design of the feature head module in Table IVd. We modulate the feature head module

TABLE IV
ABLATION STUDY OF UNI-DVPS.

Panoptic	Depth	PQ ↑	VPQ ↑	Abs. rel ↓	RMSE ↓
✓		62.4	56.2	-	-
	✓	-	-	0.0718	4.188
✓	✓	**65.9**	**59.0**	**0.0669**	**3.880**

(a) Impact of multi-task learning.

	# Params	DVPQ ↑
multiple task-specific decoders	35.0M	50.9
unified decoder	**21.0M**	**51.8**

(b) Impact of the unified Transformer decoder.

	# layers	DVPQ ↑
w/o feature head module	-	50.6
w/ feature head module	1	51.6
	2	**51.8**
	3	51.2

(c) The number of layers in feature head module.

	DVPQ ↑	Abs. rel ↓
feature head module	**51.8**	**0.0669**
+ ResidualExcite [30]	51.1	0.0690

(d) Design of feature head module.

by applying ResidualExcite [30] feature fusion module on top of our MLP-based feature head. In the feature fusion module, we reweight the segmentation features using the depth features and residual connection following [30]. We observe that such applying of the feature fusion module would not result in performance improvements. Instead, our simple MLP-based feature head is effective for multi-task learning and shows performance improvements in DVPQ and Abs. rel metrics.

V. Conclusion

We propose Uni-DVPS, a simple and effective unified model for depth-aware video panoptic segmentation that jointly performs video panoptic segmentation, depth estimation, and object tracking. Without utilizing task-specific decoders, we adopt a single Transformer decoder for multi-task learning, which increases the model efficiency and improves performance. Our architectural design choice facilitates the query to learn distinctive and instance-aware representations favorable to both segmentation and depth estimation. We disentangle task-specific features, which encourages the model to achieve state-of-the-art performances on DVPS benchmarks.

Considering that the current DVPS benchmarks tackle the driving scenarios, it is challenging to make the model perform in various scenes in real-world applications. Since the model may not segment and estimate depth for unseen objects, proposing open-vocabulary DVPS could be a promising further development. Moreover, even though we aim to build a unified architecture that streamlines the training regime of multi-task learning, leveraging the explicit temporal consistency could improve the performance of long-term object-tracking tasks. We hope Uni-DVPS can be a simple and effective baseline for further research developments, especially in visual scene understanding and various applications.

References

[1] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[2] H. Yuan, X. Li, Y. Yang, G. Cheng, J. Zhang, Y. Tong, L. Zhang, and D. Tao, "Polyphonicformer: Unified query learning for depth-aware video panoptic segmentation," in *Eur. Conf. Comput. Vis.*, 2022.

[3] A. Petrovai and S. Nedevschi, "Monodvps: A self-supervised monocular depth estimation approach to depth-aware video panoptic segmentation," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2023.

[4] N. Gao, F. He, J. Jia, Y. Shan, H. Zhang, X. Zhao, and K. Huang, "Panopticdepth: A unified framework for depth-aware panoptic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.*, 2020.

[6] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Int. Conf. Comput. Vis.*, 2019.

[7] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *Int. Conf. Comput. Vis.*, 2021, pp. 6910–6919.

[8] S. Woo, D. Kim, J.-Y. Lee, and I. S. Kweon, "Learning to associate every segment for video panoptic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[9] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai, "In defense of online models for video instance segmentation," in *Eur. Conf. Comput. Vis.*, 2022.

[10] D.-A. Huang, Z. Yu, and A. Anandkumar, "Minvis: A minimal video instance segmentation framework without video-based training," in *Adv. Neural Inform. Process. Syst.*, 2022.

[11] P. Ma, T. Du, and W. Matusik, "Efficient continuous pareto exploration in multi-task learning," in *Int. Conf. Machine Learning*, 2020.

[12] A. Kolesnikov, A. Susano Pinto, L. Beyer, X. Zhai, J. Harmsen, and N. Houlsby, "Uvim: A unified modeling approach for vision with learned guiding codes," in *Adv. Neural Inform. Process. Syst.*, 2022.

[13] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," 2023.

[14] D. Kim, J. Kim, S. Cho, C. Luo, and S. Hong, "Universal few-shot learning of dense prediction tasks with visual token matching," in *Int. Conf. Learn. Represent.*, 2023.

[15] H. Ye and D. Xu, "Taskprompter: Spatial-channel multi-task prompting for dense scene understanding," in *Int. Conf. Learn. Represent.*, 2022.

[16] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," in *Int. Conf. Learn. Represent.*, 2023.

[17] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Adv. Neural Inform. Process. Syst.*, 2021.

[18] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[19] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[20] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision (3DV)*, 2016.

[21] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Adv. Neural Inform. Process. Syst.*, 2014.

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012.

[23] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Video panoptic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[24] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Int. Conf. Comput. Vis.*, 2019.

[25] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[27] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Int. Conf. Comput. Vis.*, 2017.

[28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Int. Conf. Learn. Represent.*, 2018.

[30] M. Sodano, F. Magistri, T. Guadagnino, J. Behley, and C. Stachniss, "Robust double-encoder network for rgb-d panoptic segmentation," in *Int. Conf. Robotics and Automation*, 2023.