Towards Deeply Unified Depth-aware Panoptic Segmentation with Bi-directional Guidance Learning

Junwen He[1], Yifan Wang[1], Lijun Wang[*1], Huchuan Lu[1], Bin Luo[2],
Jun-Yan He[2], Jin-Peng Lan[2], Yifeng Geng[2], and Xuansong Xie[2]

[1]Dalian University of Technology
[2]DAMO Academy, Alibaba Group

Abstract

Depth-aware panoptic segmentation is an emerging topic in computer vision which combines semantic and geometric understanding for more robust scene interpretation. Recent works pursue unified frameworks to tackle this challenge but mostly still treat it as two individual learning tasks, which limits their potential for exploring cross-domain information. We propose a deeply unified framework for depth-aware panoptic segmentation, which performs joint segmentation and depth estimation both in a per-segment manner with identical object queries. To narrow the gap between the two tasks, we further design a geometric query enhancement method, which is able to integrate scene geometry into object queries using latent representations. In addition, we propose a bi-directional guidance learning approach to facilitate cross-task feature learning by taking advantage of their mutual relations. Our method sets the new state of the art for depth-aware panoptic segmentation on both Cityscapes-DVPS and SemKITTI-DVPS datasets. Moreover, our guidance learning approach is shown to deliver performance improvement even under incomplete supervision labels. Code and models are available at https://github.com/jwh97nn/DeepDPS.

1. Introduction

Scene understanding plays a crucial role in autonomous driving perception systems, but relying solely on 2D representations falls short for advanced systems. To address this limitation, Depth-aware Panoptic Segmentation (DPS) [44] has been proposed as a novel approach for geometric scene understanding, which enables the creation of 3D instance-
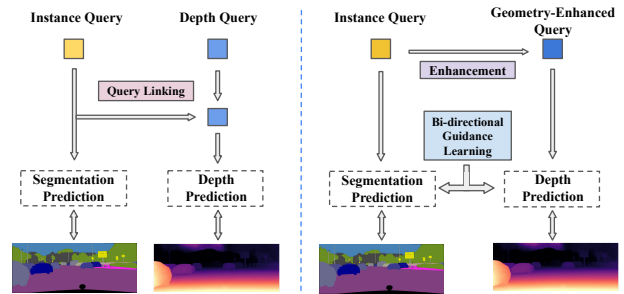
---
[*]Corresponding author

Figure 1: Pipeline comparison of prior work [55] (left) and ours (right). We integrate unified queries with geometric enhancement and mutual learning from cross-modality supervision, towards a deeper unified manner.

level semantic labels from a single image by means of inverse projection. More precisely, the simplified problem can be decomposed into two sub-tasks: panoptic segmentation and monocular depth estimation.

Early methods [44, 46] tackle this task by simply attaching a dense depth prediction head on top of the off-the-shelf panoptic segmentation model [7]. However, these methods are intuitively sub-optimal, because the separate task-oriented head design treats these two sub-tasks independently and ignores their mutual relation. Recent methods [40, 55] propose unified architectures that output both predictions in the same instance-wise manner, and utilize corresponding task-specific kernels (or queries) to jointly produce masks and depth maps for individual instances, which leverages the mutual benefits between semantic and depth information (shown in Fig. 1 left).

Despite recent efforts to unify the two sub-tasks, their learning processes remain largely separate. Specifically, they employ task-specific loss functions to guide individual predictions, which overlook the potential benefits of cross-

domain knowledge learning. While some attempts [16, 23] have been done to learn depth representations from semantic segmentation implicitly, the reciprocal relationship between the two tasks remains largely unexplored.

In this study, we introduce a new deeply unified framework for depth-aware panoptic segmentation, which leverages cross-modality knowledge not only at the architectural level but also during the learning phase. Rather than using separate queries for each task, we employ unified queries followed by geometry enhancement with latent representations. Furthermore, we design a bi-directional guidance learning approach to optimize multi-task feature learning, which can better leverage their interdependence by using the supervision of one to guide the other.

We propose a deeply unified encoder-decoder architecture, which performs joint panoptic segmentation and depth estimation in a per-segment manner with identical queries. We first generate instance-specific masks using unified per-segment queries, and enhance the queries with intermediate depth features as well as learned latent representations to integrate scene geometry, via the proposed Geometric Query Enhancement (Fig. 3). Subsequently, we predict depth maps from each enhanced query via a dot product with the depth embedding, and apply corresponding mask predictions to produce instance-wise depth predictions. To account for low-confidence filtering, which causes imperfect masks or blank segments, we introduce an extra backup query to cover up these regions.

Moreover, we present a novel approach to leverage cross-modality knowledge by refining intermediate feature representations through Bi-directional Guidance Learning. Our approach is based on the intuition that pixels crossing the semantic boundary are more likely to have a significant difference in depth and vice versa. To this end, we propose Semantic-to-Depth guidance to optimize relative depth feature distances using contrastive learning, and Depth-to-Semantic guidance to synchronize semantic feature continuity with depth annotations. The combination of both guidance mechanisms enables us to exploit their deeply-coupled relations and promote a more mutually-beneficial learning process.

Our method makes the following contributions:

1. We propose a new deeply unified architecture for depth-aware panoptic segmentation, which tackles both sub-tasks in a per-segment manner, by integrating scene geometry into unified queries with geometry enhancement.

2. We propose a new training method that refines both intermediate features simultaneously through bi-directional guidance learning, leveraging their mutual relations and boosting performance under incomplete supervision.

3. Extensive experiments on Cityscapes-DVPS [44] and

SemKITTI-DVPS [44] demonstrate the effectiveness of our proposed method, leading to state-of-the-art performance on depth-aware panoptic segmentation and individual sub-tasks.

2. Related Works

2.1. Panoptic Segmentation

Early methods for panoptic segmentation [27] usually employ a two-stage (or proposal-based) approach [18, 29, 32, 35, 51, 59], which generates instance-level segmentation based on region proposals, followed by post-processing fusion [26, 27] to obtain the final panoptic segmentation results. Instead, bottom-up (or box-free) approaches attempt to group pixels to generate instance masks on top of semantic segmentation results. For example, DeeperLab [53] proposes to reformulate panoptic segmentation task into keypoint and multi-range offset heatmap predictions, with a fusion process followed by [42]. Panoptic-DeepLab [7] and its variants [5, 47] predict class-agnostic instance centers, together with pixel-level offsets to the corresponding center. More recently, unified architectures have been proposed and adopted by many studies, which surpass previous single-stage methods while avoiding complex post-processing. Panoptic FCN [33] encodes each instance or stuff into a specific kernel, achieving instance-aware and semantically consistent representations. K-Net [57] proposes a kernel update strategy to iteratively refine the kernels and mask predictions. MaskFormer [9] and Mask2Former [8] reformulate the segmentation task into mask classification, and use the DETR-like [3] architecture to predict a set of binary masks from learned object queries. When extending the unified framework to the video level, [25] and [21] incorporated a memory token to enable communication across multi-frame features. Inspired by these works, we extend the unified architecture to perform instance-level depth estimation together with panoptic segmentation, by incorporating both features through the learned latent representation and leveraging their mutual relations.

2.2. Monocular Depth Estimation

The estimation of depth from a single image is a challenging problem in 3D computer vision. Eigen *et al.*[12] proposes the first learning-based method, which uses multi-scale CNNs to predict depth maps directly. Follow-up methods exploit more powerful network architectures [2, 28, 31, 45] or reformulate the task as a classification problem and predict a fixed [14] or adaptive [2, 34] range for each pixel. Other methods perform multi-task learning (*e.g.* surface normal [30, 54] and semantic labels [11, 20]) to enhance depth predictions. Among them, SDC-Depth [49] decomposes the global depth prediction task into a series of category-specific ones, with the help of off-the-shelf seg-
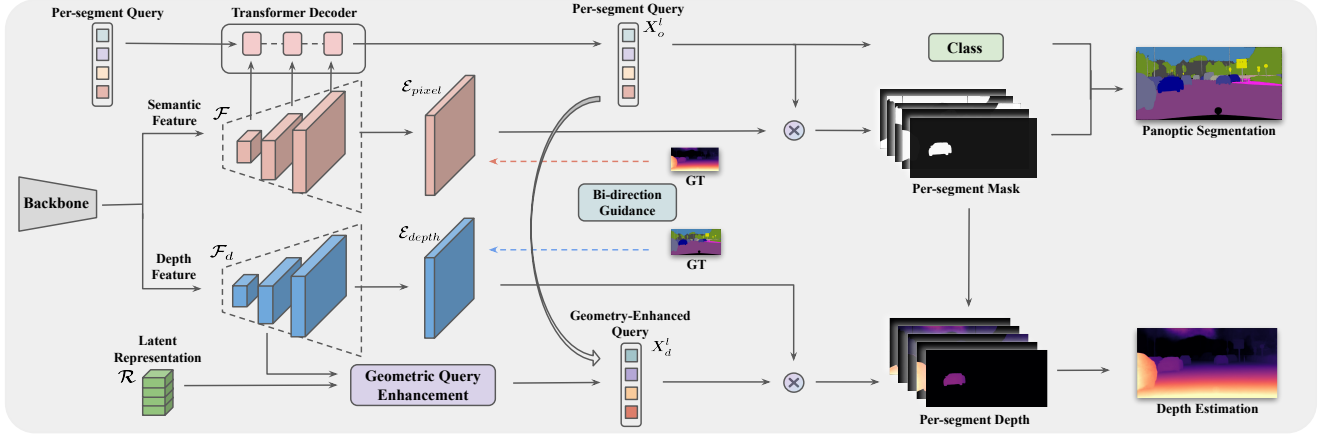
Figure 2: **Architecture Overview.** We learn unified per-segment queries X_o^l and obtain geometry enhanced queries X_d^l, by incorporating multi-scaled depth features $\mathcal{F}_d$ and learned latent representations $\mathcal{R}$ through **Geometry Query Enhancement**. We introduce a **Bi-directional Guidance Learning** to refine both features with cross modality supervisions, which includes Semantic-to-Depth (blue dashed arrow) and Depth-to-Semantic (orange dashed arrow) guidance. $\otimes$ denotes dot product.

mentation module. Our approach to depth estimation also involves predicting instance-wise depth maps. In contrast, this is achieved through a unified model that generates both instance masks and depth maps, and is further boosted from mutual learning between modalities.

2.3. Depth-aware Panoptic Segmentation

The task of combining panoptic segmentation and monocular depth estimation was initially proposed in [44], but numerous studies have examined the relationship between these two tasks prior. Earlier approaches have viewed this as a multi-task learning problem, typically utilizing a multi-branch model for simultaneous prediction and leveraging both supervisions for training [38, 52, 58]. Subsequent studies [16, 23, 60] have focused on improving the performance of one single task by exploiting information from the other. For instance, Guizilini *et al.* [16] propose to use pre-trained segmentation networks to guide depth representation learning. Jung *et al.* [23] present a novel training method that exploits semantics-guided local geometry to optimize intermediate depth representations.

More recently, there have been efforts to explore joint learning of both sub-tasks in a unified manner. MGNet [46] proposes a multi-task framework for monocular geometric scene understanding, which produces dense 3D point clouds with instance-aware semantic labels. ViP-DeepLab [44] extends Panoptic-DeepLab [7] with an additional depth prediction head and introduces two datasets, along with an evaluation metric for the new task. PanopticDepth [40] and PolyphonicFormer [55] propose unified architectures leveraging dynamic instance-specific kernels and query-based learning, respectively, to produce instance-level predictions. Our approach has a unique advantage compared to previous works as it implicitly guides the learning of intermediate features through cross-modality supervision, and utilizes their mutual relations.

3. Methods

We present a unified framework for depth-aware panoptic segmentation where the panoptic segmentation and depth prediction are deeply unified through the learned latent representation. To further improve the joint learning of these two sub-tasks, we propose a bi-directional guided contrast learning approach that leverages cross-modality knowledge for more effective feature learning.

3.1. Network Architecture

As shown in Fig. 2, our architecture is designed as an encoder-decoder structure with a shared encoder and two task-specific decoders. The shared backbone extracts low-resolution features, and then two separate decoders gradually upsample features to generate semantic and depth feature pyramid, denoted as $\mathcal{F}$ and $\mathcal{F}_d$, respectively, with resolutions of $\times 1/8$, $\times 1/16$ and $\times 1/32$. Additionally, it includes $\times 1/4$ pixel embedding $\mathcal{E}_{pixel}$ and depth embedding $\mathcal{E}_{depth}$. For panoptic segmentation, we adopt the mask classification idea [9] due to its efficacy. Specifically, we utilize a Transformer Decoder with $l = 9$ layers to process per-segment queries, where the input queries are sequentially interacted with the multi-scale semantic feature pyramid through masked attention [8], giving rise to the output unified per-segment queries X_o^l. A multi-layer perceptron (MLP) followed by a Softmax layer is applied to the processed queries to generate the classification probability of all the segments. Meanwhile, the binary mask prediction
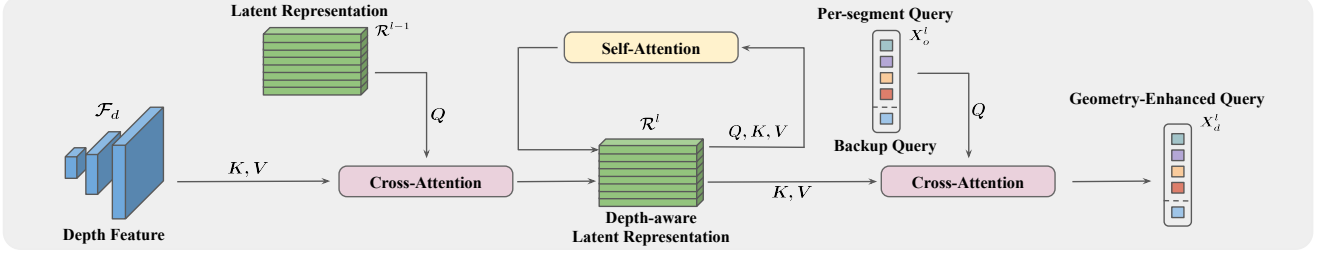
Figure 3: **Geometry Query Enhancement** with learned **Latent Representations**. We enhance per-segment queries with geometry information by operating masked cross-attentions and self-attentions alternatively, and set an extra **backup query** to cover up filtered-out regions in post-processing for better depth maps.

is conducted through a dot product between the processed queries and pixel embedding, i.e., $M = \sigma(\text{MLP}(X_o^l) \otimes \mathcal{E}_{pixel})$ with σ indicating the Sigmoid activation.

3.1.1 Per-Segment Depth Estimation

We perform per-segment depth prediction in a similar way as the above panoptic segmentation process, allowing for a unified pipeline of both tasks. To further bridge the gap between the two tasks, we propose a geometric query enhancement module to incorporate geometry information into the unified per-segment queries.

Geometric Query Enhancement. Instead of learning separate queries for different sub-tasks and linking them together as done in [55], we manage to use unified queries and enhance them with geometry information, towards a deeply unified architecture. Inspired by [17] and [22], we introduce a fixed-size latent representation (initialized as $\mathcal{R}^0$) to capture the global scene geometry, which serves as a middleware to communicate with per-segment queries and multi-scale depth features, as shown in Fig. 3. We first perform cross-attention between the masked depth features and the latent representation to project the geometry knowledge into the compressed latent space. To focus only on the regions of interest, we incorporate mask predictions from the segmentation branch, which results in faster convergence as shown in [8]. Next, we apply self-attention in the latent space to update the latent representation $\mathcal{R}^l$. Finally, cross-attention is performed between the original per-segment queries X_o^l and the updated latent representation $\mathcal{R}^l$ to generate the corresponding geometry-enhanced queries X_d^l. In this way, each geometry-enhanced query is refined by the depth features to produce consistent per-segment depth maps.

Depth Map Aggregation. Similar to mask predictions, we generate per-segment depth maps via dot product between the processed geometry-enhanced queries X_d^l and depth embedding $\mathcal{E}_{depth} \in \mathbb{R}^{C_d \times \frac{H}{4} \times \frac{W}{4}}$, as

$$d = D_{max} \times \sigma(\psi(X_d^l) \otimes \mathcal{E}_{depth}), \quad (1)$$

where σ and ψ denote the Sigmoid activation and feed-forward network (FFN) respectively, and D_{max} is the max distance which is set to 80 in all experiments. The final depth map can be obtained by aggregating the per-segment depth according to the segmentation masks

$$D(u, v) = \sum_{i \in \mathcal{H}} d_i(u, v) \cdot \mathbb{1}[M_i(u, v) > 0.5], \quad (2)$$

where M_i denotes i-th segmentation mask, $\mathbb{1}[\cdot]$ denotes the indicator function, (u, v) represents the spatial coordinate, and $\mathcal{H}$ only contains query ids with high-confidence segmentation masks[1].

Backup Query. In order to reduce the false positive rates in panoptic segmentation, low-confidence mask predictions are filtered out [9], but this may lead to blank segments in depth estimation. Therefore, we introduce a backup query that produces a global depth map to address this issue. Similar to the latent representations, the backup query uses cross-attention to update itself by querying multi-scale depth features, but without mask constraints, thereby enabling it to perceive global geometry knowledge instead of being limited to a specific region. The resulting depth values, computed from the dot product of the backup query and the depth embedding $\mathcal{E}_{depth}$, can be utilized to supplement blank regions. This simple idea not only enhances depth estimation performance but also mitigates the impact of inaccurate segmentation outcomes.

3.2. Bi-directional Guidance Learning

Based on the tight-coupled relationship between semantic and geometry information, adjacent pixels that cross semantic boundaries are more likely to have large differences in depth, and vice versa. The cross-domain knowledge learning has been employed in several methods [16, 23], mainly by leveraging scene semantics to produce semantically consistent intermediate depth representations. However, these one-way methods only guide the learning of

---

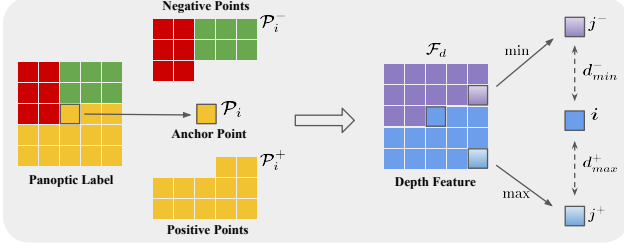[1]We follow the same post-processing as [8] to filter out mask predictions with low confidence.

Figure 4: Overview of Semantic-to-Depth Guidance. We optimize relative depth feature distances following a max-min strategy, inside each $K \times K$ local patch.

depth representations from semantic labels but not the other way around. Therefore, we propose a bi-directional guidance learning method to refine both semantic and depth representations simultaneously, which not only boosts performance on both sub-tasks but also achieves improvement under incomplete supervision, where the ground truth for only one task is available. The learning process occurs only during training, and thus no additional computation is required during inference.

3.2.1 Semantic-to-Depth Guidance

We use contrastive learning to refine depth representations, which aims to minimize the feature distance between pixels within the same instance while maximizing those across instance boundaries. This ensures that depth features are more discriminating on boundary regions, consistent with semantic information. Inspired by prior works [16, 23], we constrain the learning process to local patches since pixels far away from each other may not meet the geometry consistency assumption. In contrast to [23], we employ a max-min strategy to enforce the maximum feature distance within the same instance to be smaller than the minimum distance of different instance, with only two pixels being selected for contrastive learning (illustrated in Fig. 4).

First, we divide the panoptic labels into patches of size $K \times K$ with a stride of 1, and select the center point of each patch as the anchor point $\mathcal{P}_i$. Subsequently, we consider points with the same panoptic label as the anchor point to be positive points $\mathcal{P}_i^+$, and all other points as negative points $\mathcal{P}_i^-$. To exclude patches that do not contain instance boundaries, we only select patches with $|\mathcal{P}_i^-| > 0$. We define the maximum positive distance d_{max}^+ and minimum negative distance d_{min}^- as the L2 distance of the normalized depth feature, which are formulated as

$$d_{max}^+(i) = \max(\|\hat{\mathcal{F}}_d^l(i) - \hat{\mathcal{F}}_d^l(j^+)\|_2), \ j^+ \in \mathcal{P}_i^+ \quad (3)$$

$$d_{min}^-(i) = \min(\|\hat{\mathcal{F}}_d^l(i) - \hat{\mathcal{F}}_d^l(j^-)\|_2), \ j^- \in \mathcal{P}_i^-, \quad (4)$$

where l denotes the index of feature layer and $\hat{\mathcal{F}}_d$ denotes the normalized depth features, i.e., $\hat{\mathcal{F}}_d = \mathcal{F}_d/\|\mathcal{F}_d\|$.

We aim to enforce the learning process towards the objective of $d_{max}^+(i) < d_{min}^-(i)$, following the intuition that the maximum depth difference within an instance is likely to be smaller than the minimum depth difference across the boundary of the instance region. Our semantic guidance loss embodies the above intuition and adopts the following triplet-based form [48]:

$$\mathcal{L}_{sg} = \frac{1}{N} \sum_i \max(0, \alpha + d_{max}^+(i) - d_{min}^-(i)), \quad (5)$$

where α is a gap parameter that regularizes the margin between the two distances, i only includes patches that contain boundaries, and N is the number of such patches. The final semantic guidance loss is averaged across multiple layers.

3.2.2 Depth-to-Semantic Guidance

Following a similar idea, we in turn use depth supervision to guide the learning of semantic representations. Unlike panoptic labels, depth annotations cannot be simply grouped into different segments. Hence, we aim to enforce the continuity consistency between the depth supervision and intermediate semantic feature representations, based on the observation that pixels with continuously varying depth values are usually located within the same instance, while dramatic discontinuity usually occurs around instance boundaries. Considering the local geometry consistency, we restrict the learning process to local patches as well and choose the same patch size K as the semantic guidance learning.

Inside each patch, we choose the center pixel i as the reference pixel and all the rest as neighboring pixels. Then we calculate, for each neighboring pixel j, its relative depth distance and semantic feature distance from the reference pixel. The depth guidance loss is defined as

$$\mathcal{L}_{dg}(i, j) = -\frac{1}{N} \sum_i \sum_j e^{-\|\hat{d}_i - \hat{d}_j\|/\tau} \cdot e^{-\|\mathcal{F}_i - \mathcal{F}_j\|_2} \quad (6)$$

where $\hat{d}$ and $\mathcal{F}$ denote the groundtruth depth and semantic features, respectively, and τ is the scaling factor to balance the difference in magnitude between two distances, which is set to 10 throughout the experiment. i and j only contain pixels whose depth annotations are available, and N is the number of available patches. The final depth guidance loss is also averaged across multiple layers.

3.3. Losses

Following prior works [8, 9], we first find one-to-one matching between per-segment queries and ground-truth masks via bipartite matching, then adopt the same loss as [8]: a cross-entropy classification loss $\mathcal{L}_{cls}$, and a combination of binary cross-entropy loss and dice loss [39] for mask predictions: $\mathcal{L}_{mask} = \mathcal{L}_{ce} + \mathcal{L}_{dice}$.

Cityscapes-DVPS	Backbone	Extra Data	$\lambda = 0.5$			$\lambda = 0.25$			$\lambda = 0.1$			DPQ			FLOPs
PanopticDepth[40]	ResNet-50	-	65.6	59.2	70.2	62.3	57.0	66.1	43.2	40.7	45.1	57.0	52.3	60.5	619G
PolyphonicFormer[55]	ResNet-50∗	MV	64.3	56.0	70.3	59.7	53.3	64.4	39.3	31.8	44.7	54.4	47.0	59.8	-
Ours	ResNet-50	-	**69.3**	**61.4**	**75.0**	**66.8**	**59.1**	**72.4**	**52.8**	**46.9**	**57.1**	**63.0**	**55.8**	**68.2**	510G
ViP-DeepLab†[44]	WR-41∗	MV, CSV	68.7	61.4	74.0	66.5	60.4	71.0	50.5	45.8	53.9	61.9	55.9	66.3	4,725G
PolyphonicFormer[55]	Swin-B∗	MV	**70.6**	**63.0**	**76.0**	67.8	61.0	72.8	50.2	43.4	55.2	62.9	55.8	68.0	837G
Ours	Swin-B	-	69.8	62.3	75.3	**68.1**	**61.4**	**73.0**	**55.0**	**48.7**	**59.5**	**64.3**	**57.5**	**69.3**	1,037G
SemKITTI-DVPS	Backbone	Extra Data	$\lambda = 0.5$			$\lambda = 0.25$			$\lambda = 0.1$			DPQ			FLOPs
PanopticDepth[40]	ResNet-50	-	-	-	-	-	-	-	-	-	-	46.9	**46.0**	47.6	144G
PolyphonicFormer[55]	ResNet-50∗	MV	50.5	44.0	55.3	47.9	42.2	52.1	35.9	33.6	37.6	44.8	39.9	48.3	-
Ours	ResNet-50	-	**54.7**	**48.8**	**59.0**	**51.4**	**46.5**	**54.9**	**37.7**	**34.0**	**40.4**	**47.9**	43.1	**51.5**	164G
ViP-DeepLab†[44]	WR-41∗	MV, CSV	54.7	46.4	60.6	52.0	44.8	57.3	40.0	34.7	**43.8**	48.9	42.0	**53.9**	1,133G
PolyphonicFormer[55]	Swin-B∗	MV	58.5	55.1	61.0	**56.3**	54.0	**57.9**	41.8	41.1	42.4	52.2	50.1	53.8	201G
Ours	Swin-B	-	**59.7**	**57.1**	**61.5**	**56.3**	**54.8**	57.4	**42.4**	**42.0**	42.8	**52.8**	**51.3**	**53.9**	281G

Table 1: Depth-aware panoptic segmentation results on Cityscapes-DVPS and SemKITTI-DVPS. 'MV': Mapillary Vistas [41]. 'CSV': Cityscapes videos with pseudo labels [4]. †: test-time augmentation. ∗: Recursive Feature Pyramid (RFP) [43]. Each cell shows DPQ$^\lambda$ |DPQ$^\lambda$-Thing |DPQ$^\lambda$-Stuff, where λ is the threshold of relative depth error.

As our geometry-enhanced queries correspond to per-segment queries as well, the depth predictions are assigned with the same matching results. We use the scale-invariant loss [13], which is simple yet efficient, and formulated as

$$\mathcal{L}_{depth} = \frac{1}{n}\sum_i g_i^2 - \frac{\lambda}{n^2}\left(\sum_i g_i\right)^2 \qquad (7)$$

where $g_i = \log d_i - \log \hat{d}_i$ with the predicted depth d and ground-truth depth $\hat{d}$, and λ is set to 0.85.

The total loss is the weighted sum of all loss terms, as

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{depth}\mathcal{L}_{depth} \\ + \lambda_{sg}\mathcal{L}_{sg} + \lambda_{dg}\mathcal{L}_{dg}, \qquad (8)$$

where we use $\lambda_{cls} = 2$, $\lambda_{mask} = 5$, $\lambda_{depth} = 2.5$ and $\lambda_{sg} = \lambda_{dg} = 0.1$ in our experiments.

4. Experiments

4.1. Datasets

Cityscapes-DVPS [44] is an extension of Cityscapes-VPS [24] that includes depth annotations converted from disparity maps using stereo images. It consists of 3,000 annotated frames, with 2,400, 300, and 300 frames in the training, validation, and test sets, respectively. The dataset maintains the same semantic classes as the original Cityscapes [10] dataset, which include 8 thing and 11 stuff classes.

SemKITTI-DVPS [44] is derived from the odometry split of the KITTI dataset [15]. The dataset includes 11 training sequences, 11 test sequences, and validation is performed on sequence 08. Sparse semantic annotations are obtained by projecting panoptic-labelled 3D point clouds from SemanticKITTI [1] onto the image plane. The dataset comprises 19,130 training images, 4,071 evaluation images, and 4,342 test images.

4.2. Evaluation Metrics

We evaluate results using standard evaluation metrics, with Panoptic Quality (PQ) [27] for panoptic segmentation and Depth-aware Panoptic Quality (DPQ) for depth-aware panoptic segmentation introduced by [44]. Specifically, let P and Q be the prediction and ground-truth, we use P and P^d to denote segmentation and depth estimation respectively, and the same notation also applies to Q. Then, DPQ^λ is defined as

$$DPQ^\lambda(P, Q) = PQ(P^\lambda, Q),$$

where $P^\lambda = P$ for pixels that have absolute relative depth errors under the threshold λ (i.e., $|P^d - Q^d| \leq \lambda Q^d$), and other pixels will be assigned to *void*. DPQ^λ is evaluated under three values of $\lambda = \{0.1, 0.25, 0.5\}$, and averaged to obtain the final DPQ.

4.3. Implementation

We use Detectron2 [50] to implement our model and follow [8] to choose multi-scale deformable attention Transformer [61] as multi-scaled decoders. We adopt ResNet-50 [19] and Swin-B [37] as the shared backbone, and do not use extra datasets for pre-training or Recursive Feature Pyramid (RFP) [43] to enhance the backbone, which differs from prior methods [44, 55].

Following prior works [40, 55], the full training process consists of two steps. First, we train the segmentation branch using panoptic labels only for 50 epochs. Then, we finetune the entire model with both supervisions for 10 epochs. During the first step, we resize images with a random scale from 0.5 to 2.0, followed by a fixed size crop to 512×1024 and 384×1280 on Cityscapes-DVPS and SemKITTI-DVPS, respectively. After the pre-training phase, we use full-resolution for finetuning. Large-scale jit-

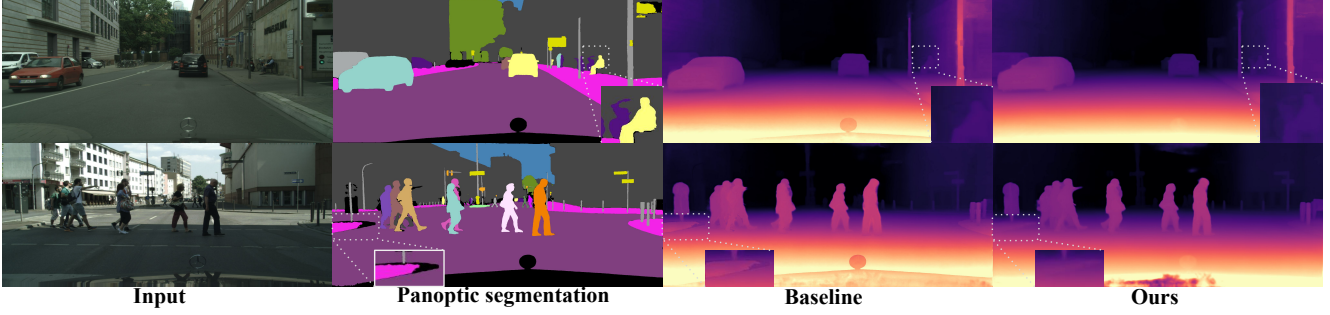**Input** **Panoptic segmentation** **Baseline** **Ours**

Figure 5: Visualization results on Cityscapes-DVPS. Top row: unified architecture distinguishes boundaries for better depth estimation. Bottom row: backup query alleviates the case of imperfect mask predictions.

Method	Backbone	Depth	PQ	PQTh	PQSt
VPSNet [24]	ResNet-50		65.0	-	-
ViP-DeepLab [44]	ResNet-50	✓	60.6	-	-
PanopticDepth*[40]	ResNet-50	✓	66.9	60.1	71.9
PolyphonicFormer [55]	ResNet-50	✓	65.4	-	-
Ours	ResNet-50	✓	**69.5**	**62.1**	**75.7**

Table 2: Panoptic Segmentation results on Cityscapes-DVPS validation set.

Method	abs rel	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
DPT-Hybrid [45]	0.0697	0.1106	0.9434	0.9914	0.9976
PanopticDepth [40]	0.0711	0.1125	0.9359	0.9919	0.9982
PolyphonicFormer [55]	0.0647	0.1013	0.9524	0.9950	0.9985
Ours	**0.0597**	**0.0940**	**0.9616**	**0.9953**	**0.9988**

Table 3: Monocular Depth Estimation results on Cityscapes-DVPS validation set.

tering (LSJ) [36] and horizontal flipping are also employed during training. Final predictions are obtained from a single inference, while no test-time augmentation is employed. More details are provided in supplementary material.

4.4. Main Results

Depth-aware Panoptic Segmentation. We evaluate our model on two datasets, Cityscapes-DVPS and SemKITTI-DVPS, and compare it with recent methods. Our method achieves state-of-the-art results on both datasets as presented in Tab. 1. It is worth mentioning that we did not employ any additional techniques, such as pre-training on larger datasets [41], training with pseudo-labels [4], backbone enhancement [43], or test-time augmentation [44].

On Cityscapes-DVPS, our model achieves superior performance compared to published methods that use larger backbones (WR-41 [56] and Swin-B [37] with RFP [43]) despite using the less powerful ResNet-50 [19] backbone. Although our model performs slightly worse than Poly-phonicFormer [55] with Swin-B on DPQ$^{0.5}$, it demonstrates significant improvement on DPQ$^{0.1}$, indicating the better

Figure 6: Latent representation attention maps.

depth quality, especially on challenging thresholds.

Learning on SemKITTI-DVPS is more challenging, because both ground-truths are sparse, leading to much fewer available training pixels. Tab. 1 reports our DPQ result on SemKITTI-DVPS validation set, and we outperform all state-of-the-art methods under the same backbone.

Individual Sub-tasks. We have evaluated our model on two individual sub-tasks and presented the results in tables Tab. 2 and Tab. 3. Our approach outperforms state-of-the-art methods on both sub-tasks, indicating that the unified architecture and cross-modality learning process are advantageous for both sub-tasks.

Visualization Results. The panoptic segmentation and depth estimation results on Cityscapes-DVPS are visualized in Fig. 5. The top row shows the benefit of the unified framework, which improves depth results with clearer boundaries, especially on small objects (*e.g.*, bicycle in the distance). The bottom row shows that the backup query could provide smoother depth values in blank mask regions, caused by imperfect mask predictions.

In addition, we choose 5 latent representation indices and visualize their attention maps in Fig. 6. We discover that certain latent representations specialize in various instances such as roads (bottom left), cars and buses (middle column), as well as different distance ranges (right cloumn). More visualization results are available in the supplementary materials.

Variants	Instance-wise	Backup Query	S.Guidance	D.Guidance	$\lambda=0.5$	$\lambda=0.25$	$\lambda=0.1$	DPQ	PQ↑	abs rel↓	#comp.
A					68.5	64.3	48.3	60.3	69.0	0.087	- 8.6%
B	✓				68.7	66.0	51.8	62.1	68.9	0.068	-
C	✓	✓			68.9	66.5	51.9	62.4	69.0	0.065	+ 4.2%
D	✓	✓	✓		**69.3**	**67.1**	51.9	62.8	**69.5**	0.064	+ 17.8%
E	✓	✓	✓	✓	**69.3**	66.8	**52.8**	**63.0**	**69.5**	**0.063**	+ 19.1%

Table 4: Ablation studies on Cityscapes-DVPS. 'Instance-wise': unified architecture instead of attaching a depth regression head to the baseline model. 'S/D.Guidance': Semantic-to-Depth/Depth-to-Semantic Guidance Learning.

	$\lambda=0.5$	$\lambda=0.25$	$\lambda=0.1$	DPQ	PQ↑	abs rel↓
Full-supervision	67.6	65.2	42.7	58.5	67.8	0.112
Semi-supervision	66.1	63.3	37.9	55.8	66.4	0.140
+ depth guidance	67.0	64.5	39.2	56.9	67.1	0.131
+ semantic guidance	66.6	64.7	**41.7**	57.7	66.7	**0.122**
+ both guidance	**67.1**	**65.0**	41.2	**57.8**	**67.4**	0.126

Table 5: Effect of Bi-directional Guidance Learning. 'Semi-supervision': incomplete ground-truth labels.

	$\lambda=0.5$	$\lambda=0.25$	$\lambda=0.1$	DPQ	PQ↑	abs rel↓
w/o guidance	67.5	63.9	41.2	57.5	67.7	0.132
SGT [23]	**67.5**	64.0	42.5	58.0	68.0	0.120
ours S.Guide	**67.5**	**64.3**	**42.9**	**58.2**	**68.1**	**0.117**
DGS	67.6	64.2	42.3	58.0	68.3	0.129
ours D.Guide	**67.7**	**64.3**	**42.6**	**58.2**	**68.5**	0.126

Table 6: Ablation studies on choices of Bi-directional Guidance Learning losses.

	$\lambda=0.5$	$\lambda=0.25$	$\lambda=0.1$	DPQ	PQ↑	abs rel↓
Single query	65.8	59.3	34.8	53.3	66.4	0.163
Double latent	67.4	64.3	40.4	57.4	67.8	0.119
Query Linking [55]	67.5	63.9	41.2	57.5	67.8	0.123
Latent representation	**67.6**	**65.2**	**42.7**	**58.5**	**67.8**	**0.112**

Table 7: Effect of Geometric Query Enhancement.

4.5. Ablation Studies

Depth-aware Panoptic Segmentation. To demonstrate the advantages of our proposed architecture and learning method, we conducted ablation studies on the Cityscapes-DVPS dataset in Tab. 4, with ResNet-50 backbone and full resolution images. Our baseline model, Variant-A, simply adds an additional CNN depth regression head to the segmentation branch. By replacing the basic depth head with the proposed instance-wise depth decoder, our model achieves a 1.8% improvement, and Backup Query further enhances it by 0.3%. Interestingly, we observed that the guidance learning process not only facilitates cross-domain feature learning but also benefits itself. We speculate that the self-improvement stems from the joint-learning framework, where both queries communicate with each other through intermediate latent representations. We also provide analysis of computation costs associated with different components (#comp. in Tab. 4), and found that less than 20% extra training cost brings boosting performance.

Another notable advantage we have observed over the previous method [44] is that, hyper-parameters have a limited effect on the final performance, which is consistent with the findings in [55], thanks to the unified framework. Therefore, we keep most hyper-parameters the same as in previous works (*e.g.*, segmentation loss weights [8], gap parameter α and patch size K [23]). More hyper-parameter ablations can be found in supplementary.

Bi-directional Guidance Learning. To further demonstrate the effectiveness of the proposed bi-directional learning method, we conduct extra experiments by simulating training with incomplete supervision labels. We separate the full training data into three subsets, where the first subset only contains panoptic labels, the second subset only

contains depth annotations, and the last one keeps both annotations unchanged. All experiments are trained on 512×1024 images using ResNet-50 backbone. As shown in Tab. 5, we find that training with incomplete supervision leads to dramatic performance drops (-3.3%) on DPQ and both sub-tasks as well. Introducing the proposed guidance loss terms during training results in performance boosts and achieves minimum performance gap with full supervision using bi-directional guidance.

We additionally show comparisons of different forms of guidance loss design in Tab. 6. "SGT" refers to semantic-guided triplet loss proposed in [23], and "DGS" refers to disparity-guided smoothness, which regularizes the smoothness of semantic features based on disparity gradients (vertically and horizontally), modified from [6]. We find that any implementation of such guidance learning is beneficial, and experimentally choose the best one. The result also indicates that the idea of bi-directional guidance learning is more important than the loss design choice.

Geometric Query Enhancement. We conduct several experiments to generate queries for depth estimation and show results in Tab. 7. "Single query" denotes using the same per-segment queries to incorporate both features. "Double latent" denotes adding an extra latent representation for semantic features. "Query Linking" denotes the method in [55], which adopts grouping to select salient fea-

tures followed by query update and reasoning. We found the geometric-enhanced queries with learned latent representations achieve the best performance among all alternatives.

4.6. Limitations

As shown in Fig. 5, inaccurate mask predictions affect the quality of the depth map, making decent pre-training for segmentation necessary for follow-up depth learning. This suggests that the model may not generalize well to other scenes with unseen objects. In the future, we hope to address the above limitations and improve the model towards the goal of a fully unified framework.

5. Conclusion

We present a deeply unified framework for depth-aware panoptic segmentation at both the architectural and learning levels, exploiting the natural correlation between these two sub-tasks. By leveraging cross-modality guidance learning, the intermediate feature representations not only benefit from each other but also from themselves in turn. With the above contributions, the proposed method achieves state-of-the-art results on two datasets and shows promise in improving performance even with incomplete supervision labels. We hope our approach will inspire future researches in geometric scene understanding.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 6

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[4] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision*, pages 695–714. Springer, 2020. 6, 7

[5] Liang-Chieh Chen, Huiyu Wang, and Siyuan Qiao. Scaling wide residual networks for panoptic segmentation. *arXiv preprint arXiv:2011.11675*, 2020. 2

[6] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2624–2632, 2019. 8

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 1, 2, 3

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2, 3, 4, 5, 6, 8

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 3, 4, 5

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[11] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 2

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2

[13] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 6

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 2

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[16] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. 2, 3, 4, 5

[17] Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rares Ambrus,

Greg Shakhnarovich, Matthew Walter, and Adrien Gaidon. Depth field networks for generalizable multi-view scene representation. In *European Conference on Computer Vision (ECCV)*, 2022. 4

[18] Jun-Yan He, Shi-Hua Liang, Xiao Wu, Bo Zhao, and Lei Zhang. Mgseg: Multiple granularity-based real-time semantic segmentation network. *IEEE Transactions on Image Processing*, 30:7200–7214, 2021. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[20] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 2

[21] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021. 2

[22] 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. 4

[23] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12642–12652, 2021. 2, 3, 4, 5, 8

[24] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020. 6, 7

[25] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13914–13924, 2022. 2

[26] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 2

[27] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2, 6

[28] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2

[29] Justin Lazarow, Kwonjoon Lee, Kunyu Shi, and Zhuowen Tu. Learning instance occlusion for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10720–10729, 2020. 2

[30] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 2

[31] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017. 2

[32] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2019. 2

[33] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2021. 2

[34] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 2

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 7

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6, 7

[38] Yue Meng, Yongxi Lu, Aman Raj, Samuel Sunarjo, Rui Guo, Tara Javidi, Gaurav Bansal, and Dinesh Bharadia. Signet: Semantic instance aided unsupervised 3d geometry perception. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 9810–9820, 2019. 3

[39] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[40] Gao Naiyu, He Fei, Jia Jian, Shan Yanhu, Zhang Haoyang, Zhao Xin, and Huang Kaiqi. Panopticdepth: A unified framework for depth-aware panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 6, 7

[41] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 6, 7

[42] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Person-

lab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–286, 2018. 2

[43] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10213–10224, 2021. 6, 7

[44] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021. 1, 2, 3, 6, 7, 8

[45] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2, 7

[46] Markus Schön, Michael Buchholz, and Klaus Dietmayer. Mgnet: Monocular geometric scene understanding for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15804–15815, 2021. 1, 3

[47] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Standalone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020. 2

[48] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014. 5

[49] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 541–550, 2020. 2

[50] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[51] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 2

[52] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 3

[53] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv preprint arXiv:1902.05093*, 2019. 2

[54] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2

[55] Haobo Yuan, Xiangtai Li, Yibo Yang, Guangliang Cheng, Jing Zhang, Yunhai Tong, Lefei Zhang, and Dacheng Tao. Polyphonicformer: Unified query learning for depth-aware video panoptic segmentation. *arXiv preprint arXiv:2112.02582*, 2021. 1, 3, 4, 6, 7, 8

[56] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 7

[57] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 2

[58] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 3

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[60] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023. 3

[61] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 6