2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

DVT: Decoupled Dual-Branch View Transformation for Monocular Bird's Eye View Semantic Segmentation

Jiayuan Du, Xianghui Pan, Mengjiao Shen, Shuai Su, Jingwei Yang,
Chengju Liu* and Qijun Chen*, *Senior Member, IEEE*

***Abstract*— Monocular Bird's Eye View (BEV) semantic segmentation is critical for autonomous driving for its inherent advantages in spatial representation and downstream tasks. However, it is challenging to simultaneously learn view transformation and pixel-wise classification. Previous works suffer from non-flat region distortion, distant depth ambiguity, and visual occlusion. To address these aforementioned concerns, we propose dual-branch view transformation (DVT), a novel framework for monocular BEV semantic segmentation. Our method consists of: (i) A dual-branch view transformation to decouple features into flat region and non-flat region and process them independently. (ii) A depth-aware weighting method to make the model pay more attention to the distant depth. (iii) An auxiliary task to introduce more inductive biases to alleviate the inaccuracy caused by visual occlusion. Furthermore, we design a class-aware weighting method to address the class and size imbalance of datasets. Experimental results on nuScenes and KITTI-360 datasets demonstrate that DVT outperforms previous state-of-the-art (SOTA). Our codes are available at `https://github.com/MrPicklesGG/DVT`.**

I. INTRODUCTION

Monocular Bird's Eye View (BEV) semantic segmentation is a crucial task in autonomous driving, enabling vehicles to understand their surroundings from a top-down view. This generally involves transforming monocular images from perspective view to bird's eye view (PV-BEV) and conducting pixel-wise classification, yielding a 2D occupancy grid map with dense semantics. Such a compact representation is friendly for downstream tasks such as trajectory planning, motion estimation and collision avoidance.

The PV-BEV transformation implicitly includes the core task of recovering 3D information from 2D images, but the lack of depth information due to the perspective mapping makes it a mathematically ill-posed problem. The pioneer work IPM [1] introduces a constraint that the inversely mapped points should lie on the ground plane, such that the inverse perspective mapping can be described by a homography matrix. However, the assumption of IPM addresses only the inverse mapping of flat regions, leading to distortion in non-flat regions. As a result, objects above the ground plane such as buildings, vehicles, and pedestrians, are often inaccurately segmented. Some works [2]–[7] explore semantic information with the help of neural networks to compensate for the distortion, but due to the coupling between features of flat and non-flat regions, these methods cannot completely solve the distortion caused by non-flat regions. Other works [8]–[13] recover 3D spatial information by estimating the depth value or depth distribution of the images. However, due to the perspective effect, the distant depth features are sparse in the BEV space, resulting in poor segmentation results at a distance. There are also some works [14]–[24] that use fully connected bottleneck networks or attention-based networks for end-to-end learning. These methods perform better than previous methods in reasoning about invisible areas, but are not accurate enough in the geometric positions and boundaries of objects. In general, current methods for monocular BEV segmentation suffer from projection distortion, depth ambiguity, and visual occlusion. Meanwhile, due to the imbalance problem of the datasets, the above problems are difficult to be naturally optimized in the learning process.

To overcome the aforementioned problems, we propose DVT, a novel framework for monocular BEV semantic segmentation via decoupled **D**ual-branch **V**iew **T**ransformation. (i) We adopt a two-branch structure to decouple the view transformation into two independent processes of flat regions and non-flat regions, which solves the problem of projection distortion in non-flat regions and also introduces geometric priors for the whole network. (ii) To improve the poor distant depth estimation caused by perspective effect, we introduce a depth-aware weighting method. It leverages the Jacobian determinant to reflect the sensitivity of PV-BEV transformation and re-weights the loss function. (iii) Considering that additional inductive bias can improve the reasoning ability of the model for invisible areas. Therefore, instance segmentation is used as an auxiliary task to improve the accuracy of object position and boundary estimation. Furthermore, to alleviate the problem of size imbalance and class imbalance of the datasets, we introduce a class-aware weighting method, including online hard example mining factor, inverse frequency weight and boundary augmentation weight. In order to verify the effectiveness of our proposed method, we conduct extensive experiments on two large-scale datasets nuScenes [25] and KITTI-360 [26]. The experimental results show that DVT outperforms the previous state-of-the-art (SOTA). Our contributions can be summarized as follows:

- We propose DVT, a novel monocular BEV semantic segmentation framework via decoupled dual-branch

This paper is supported by the National Natural Science Foundation of China under Grants (62073245, 62173248, 62333017). Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities. (Corresponding author: Chengju Liu, Qijun Chen)

The authors are with the Robotics & Artificial Intelligence Laboratory (RAIL), Tongji University, Shanghai, China (E-mails: {dujiayuan, panxianghui, 1910680, sushuai, jw_yang, liuchengju, qjchen}@tongji.edu.cn).

view transformation.
- We come up with a depth-aware weighting method on the cross entropy loss, alleviating the uncertainty of distant depth estimation.
- We introduce instance segmentation as an auxiliary task, utilizing its inductive bias to improve the prediction of position and boundary.
- We design a class-aware weighting method to address the class and size imbalance of monocular BEV semantic segmentation datasets.

II. RELATED WORK

Monocular BEV semantic segmentation works generally follow a similar pipeline including feature-extraction, PV-BEV transformation and semantic segmentation [27], [28]. Based on different view transformation strategies, the methods can be grouped into four categories as follows.

Homograph-Based Methods: The homograph-based methods use IPM [1] to perform PV-BEV transformation. An early work [2] conducts semantic segmentation in the PV space and transforms the results into the BEV space via IPM. Due to the plane assumption of IPM, the flat regions are mapped correctly while the non-flat regions suffer from distortion. To reduce the IPM distortion, a series of subsequent works [3]–[7] explore the semantic information.

Depth-Based Methods: The depth-based methods explicitly estimate depth map or distribution in PV images to recover 3D information, then perform view transformation from 3D space to BEV space. Pseudo-LiDAR series [8], [9] estimate pixel-wise depth maps from stereo or monocular images and convert depth maps into pseudo-LiDAR point clouds. In AM3D [29], the pseudo-LiDAR point clouds are enhanced by corresponding RGB features. To allow the pipeline to be trained in an end-to-end manner, E2E Pseudo-LiDAR [10] introduces a Change-of-Representation module. Rather than predicting a dense depth map, LSS [11] paradigm predicts the depth distribution instead. Subsequent works [12], [13] follow this LSS paradigm as well. To improve the accuracy of depth distribution, CaDDN [12] uses depth derived from LiDAR point cloud as supervision.

MLP-Based Methods: The MLP-based methods learn the PV-BEV transformation by multi-layer perceptron (MLP). Adopting MLP to perform 3D-2D feature projection is initially introduced by OFT [14], which projects 2D features to 3D voxel space. VED [15] introduces a variational encoder-decoder (VED) architecture with an MLP bottleneck to convert PV images to semantic BEV occupancy grid maps in an end-to-end manner. VPN [16] utilizes a two-layer MLP to perform the view transformation. STA-ST [17] and PON [18] make use of Feature Pyramid Network (FPN) [30] to extract multi-scale image features, and compress the features along the height axis and expand along the depth axis to leverage the vertical context. HFT [20] combines a camera model-based branch with a camera model-free branch via a hybrid feature transformation to utilize geometric prior and capture global context respectively. PoBEV [31] combines the bottleneck network with EfficientPS [32] to predict BEV panoptic segmentation. SkyEye [21] proposes a novel self-supervised framework for BEV semantic segmentation.

Transformer-Based Methods: The transformer-based methods learn the PV-BEV transformation using the transformer model. PYVA [19] puts forward a bidirectional self-supervision scheme with stronger attention-based BEV features. The geometric priors are used by GKT [22] to guide the transformer to concentrate on discriminative regions and unfold kernel features to produce BEV representation. TIIM [23] treats 1-1 correspondence between each image column and BEV polar ray as a sequence-to-sequence translation. Such geometric constraint avoids the dense cross attention between 2D image features and BEV queries, making it more suitable for the transformer-style architecture. GitNet [24] obtains coarse pre-aligned BEV features by a geometry-guided pre-alignment module and refines the features via a ray-based transformer like TIIM.

III. METHODOLOGY

The overview of our approach is shown in Fig. 1, including a shared backbone, a dual-branch PV-BEV transformer, a semantic segmentation head, and an auxiliary task head. Monocular frontal view images are first fed into the backbone network and FPN [30] to obtain multi-scale PV features. The PV features of each scale are fed into the dual-branch view transformation in parallel. In the dual-branch transformer, PV features are decoupled into flat and non-flat regions to be processed separately by different branches, and then re-coupled into the complete features in BEV space. Multi-scale BEV features are then fed into the semantic segmentation head to yield a BEV semantic map. Simultaneously, we conduct instance segmentation as an auxiliary task.

A. Backbone

We adopt a variant backbone of EfficientDet [33] to generate multi-scale PV features $\{\boldsymbol{F}_i^{pv} \mid i = 4, 8, 16, 32\}$, which comprises an ImageNet-pretrained EfficientNet-B3 [34] and a modified bi-directional feature pyramid network (BiFPN) [32]. The level 2-6 features $\{\boldsymbol{P}_n \mid n = 2, 3, 4, 5, 6\}$ of EfficientNet-B3 are at the $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$ and $\frac{1}{64}$ of the input image resolution, respectively. These features are taken by the first BiFPN layer as inputs, repeatedly conducting top-down and bottom-up bidirectional feature fusion. To yield desired PV features at scale 4, 8, 16, 32, we drop off the bottom-up feature fusion between features at scale 32 and 64 in each BiFPN layer. For the cascade between BiFPN layers, we downsample output feature map at scale 32 to get feature map at scale 64 and complement the input for the rest of the layers.

B. Dual-Branch View Transformation

Due to the lack of depth information caused by perspective mapping, projecting the pixels on image back to 3D space is ill-posed. IPM [1] solves this ill-posed problem by adding constraint that the back-projected points fall on the ground plane, which means that the mapping can be described via a homography matrix. IPM-based methods

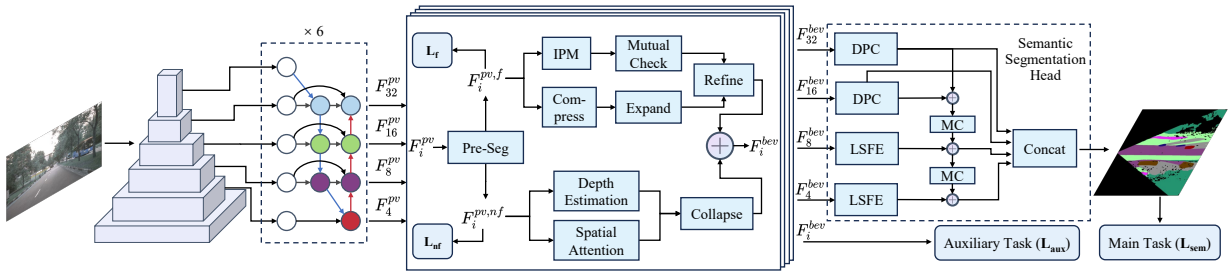

Fig. 1. **Overall architecture of our DVT**. Our approach consists of a shared backbone, a dual-branch view transformation module, a semantic segmentation head and an auxiliary task head. $\mathcal{L}_f$ and $\mathcal{L}_{nf}$ are the losses of pre-segmentation module to decouple the feature $\boldsymbol{F}_i^{pv}$ into flat region feature $\boldsymbol{F}_i^{pv,f}$ and non-flat region feature $\boldsymbol{F}_i^{pv,nf}$ in PV space. $\mathcal{L}_{sem}$ and $\mathcal{L}_{aux}$ denote losses for semantic segmentation and instance segmentation respectively.

are effective for view transformation in flat regions, but will produce large distortion in non-flat regions. We employ dual-branch architecture to transform flat region features and non-flat region features from PV to BEV, respectively. For the flat regions, we adopt IPM-based approach. As for the non-flat regions, we conduct implicit depth estimation using bottleneck network with attention mechanism to reduce projection distortion. For each scale PV feature $\boldsymbol{F}_i^{pv}$, we first conduct a pre-segmentation in the PV space to decouple it into flat region feature $\boldsymbol{F}_i^{pv,f}$ and non-flat region feature $\boldsymbol{F}_i^{pv,nf}$. $\boldsymbol{F}_i^{pv,f}$ and $\boldsymbol{F}_i^{pv,nf}$ are independently transformed to BEV features $\boldsymbol{F}_i^{bev,f}$ and $\boldsymbol{F}_i^{bev,nf}$ and then re-coupled into complete BEV feature $\boldsymbol{F}_i^{bev}$.

1) **Flat Region Branch:** The flat region feature map $\boldsymbol{F}_i^{pv,f}$ is converted to $\boldsymbol{F}_i^{bev,ipm}$ by IPM. However, merely using the IPM algorithm introduces errors into the view-transformation. Thus, we design a mutual-check module. We first conduct a simple self-attention to obtain the ambiguity of the IPM, denoted as $\boldsymbol{A}_i^{bev,ipm}$, and project it from the BEV space back into the PV space as $\boldsymbol{A}_i^{pv,ipm}$. The ambiguity attention $\boldsymbol{A}_i^{pv,ipm}$ is then applied to the flat region feature map $\boldsymbol{F}_i^{pv,f}$ to get the ambiguity feature map $\boldsymbol{F}_i^{pv,amb}$, retaining only the ambiguous regions. The $\boldsymbol{F}_i^{pv,amb}$ is compressed in the height dimension and expand in the depth dimension to generate $\boldsymbol{F}_i^{bev,amb}$. The $\boldsymbol{F}_i^{bev,ipm}$ and $\boldsymbol{F}_i^{bev,amb}$ are fused and refined together to generate the final BEV feature map $\boldsymbol{F}_i^{bev,f}$ of the flat region.

2) **Non-Flat Region Branch:** For the non-flat region feature map $\boldsymbol{F}_i^{pv,nf}$, we lift it into voxelized representation $\boldsymbol{F}_i^{3D,nf}$ using a 3D convolution layer. At the same time, we generate a spatial attention mask $\boldsymbol{A}_i^{bev,nf}$ by estimating the probability of a pixel being occupied by a non-flat element in the BEV space. We then mask out the voxels with low occupancy probability in the representation $\boldsymbol{F}_i^{3D,nf}$ to obtain $\boldsymbol{F}_i^{3D,masked}$. Subsequently, we collapse $\boldsymbol{F}_i^{3D,masked}$ in the height dimension and resample via intrinsic matrix to get the final BEV feature map $\boldsymbol{F}_i^{bev,nf}$ of the non-flat region.

C. Semantic Segmentation Head

BEV features $\left\{\boldsymbol{F}_i^{bev} \mid i = 4, 8, 16, 32\right\}$ are processed with a semantic segmentation head, including large scale feature extractor (LSFE) [32], dense prediction cells (DPC) [35] and mismatch correction (MC) [32] modules.

In order to enable the network efficiently capture fine features at large-scale, we employ LSFE module which consists of two 3 × 3 depth-wise separable convolutions with 128 output filters, each followed by a synchronized Inplace Activated Batch Normalization (iABN sync) and a Leaky ReLU activation function. Meanwhile, we adopt DPC to capture long-term context at small-scale. We replace the batch normalization layers and ReLUs on the original basis with iABN sync and Leaky ReLUs. To aggregate large-scale and small-scale features, we employ MC module to conduct mismatch mitigation. The MC module is composed by cascaded 3 × 3 depth-wise separable convolutions with 128 output channels and dilation rate of (1, 1). Each convolution is followed by iABN sync with Leaky ReLU and a bilinear upsampling layer. The bilinear upsampling layer upsamples the feature maps by a factor of 2.

D. Auxiliary Task

We conduct instance segmentation as an auxiliary task to improve the performance of semantic segmentation. The instance segmentation module follows the pattern of Mask R-CNN [36], which comprises Region Proposal Network (RPN) and ROI align [36] modules to generate bounding box, class, and mask predictions.

In the first stage of Mask-RCNN, the RPN module generalizes a set of rectangular object proposals and objectness scores for the given input FPN level via a fully convolutional network. Subsequently, ROI align extracts features from FPN encodings by directly pooling features from the n^{th} channel with a 14 × 14 spatial resolution bounded within a bounding box regression. In the second stage, the features that are extracted then serve as input to the bounding box regression, object classification and mask segmentation networks. To guide main task with better implicit depth estimation, we adopt Efficient-IoU (EIoU) [37] during the bounding box prediction instead of vanilla IoU:

$$\begin{aligned}\mathcal{L}_{EIoU} &= \mathcal{L}_{IoU} + \mathcal{L}_{dis} + \mathcal{L}_{asp} \\ &= 1 - IoU + \frac{\rho^2\left(\boldsymbol{b}, \boldsymbol{b}^{gt}\right)}{c^2} + \frac{\rho^2\left(w, w^{gt}\right)}{c_w^2} + \frac{\rho^2\left(h, h^{gt}\right)}{c_h^2}.\end{aligned} \tag{1}$$

$\rho(\cdot)$ indicates the Euclidean distance, $\boldsymbol{b}$ and $\boldsymbol{b}^{gt}$ denote the central points of predicted and groundtruth bounding boxes, w, h and w^gt, h^gt are their width and height. c, c_w and c_h are the diagonal length, width and height of the smallest enclosing box covering the two boxes, respectively. EIoU loss simultaneously considers the three geometric factors of bounding box regression: overlap area, center point distance, and aspect ratio, corresponding to the three parts of the loss function.

E. Loss Functions

For the pre-segmentation in the dual-branch view transformation module, we use binary cross entropy loss $\mathcal{L}_f$ and $\mathcal{L}_{nf}$ for training. For the main task semantic segmentation, we adopt the pixel-wise cross entropy loss for training:

$$\mathcal{L}_{ce} = -\sum_{ij} p_{ij}^{gt} \log p_{ij}, \tag{2}$$

where p_{ij} denotes the predicted probability for the pixel (i, j) being assigned class c, while p_{ij}^{gt} is the groundtruth.

Depth-Aware Weighting Method: To alleviate the uncertainty of distant objects and the distortion caused by perspective effect, we design a depth-aware weighting method on the cross entropy loss. Our derivation starts with the intrinsic parameters of the monocular camera. Given the focal length f_x, f_y and the nodal point (c_x, c_y), the image pixel (u, v) can be described by:

$$u = \frac{f_x x}{z} + c_x, \tag{3}$$

$$v = \frac{f_y y}{z} + c_y. \tag{4}$$

The Jacobian determinant gives the ratio of the area between PV image plane and the BEV plane, which simultaneously reflects the PV-BEV sensitivity:

$$\frac{\partial A_{pv}}{\partial A_{bev}} = |J| = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial z} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial z} \end{vmatrix} = \begin{vmatrix} \frac{\partial f_x}{\partial z} & \frac{\partial -f_x x}{\partial z^2} \\ 0 & \frac{\partial -f_y y}{\partial z^2} \end{vmatrix} = \frac{-f_x f_y y}{z^3}, \tag{5}$$

where y represents the height of the camera and can be obtained from the extrinsic matrix. And the depth-aware weight is given by:

$$w_{da} = 1 + \frac{\lambda_d}{\log\left(1 + \left|\frac{-f_x f_y y}{z^3}\right|\right)}. \tag{6}$$

Class-Aware Weighting Method: Semantic segmentation datasets often exhibit two types of imbalance: class imbalance, where some classes appear more frequently than others and size imbalance, where some objects occupy more pixels than others [38]. For the class imbalance, we adopt a general online hard example mining to focus on the worst 24 % prediction:

$$I_{ij} = \begin{cases} 1, \text{ if (i,j) belongs to worst 24 \% prediction,} \\ 0, \text{ otherwise .} \end{cases} \tag{7}$$

For the size imbalance, we first calculate pixel-wise weights of class c following [39]:

$$w_{ij}^c = 1/N_s^{1-n_c}, \tag{8}$$

where N_s represents the number of all samples, and n_c denotes the number of samples belong to class c. Subsequently, we adopt a linear interpolation to augment the fuzzy boundaries between infrequent class and frequent class. Then the whole class-aware weight is given by:

$$w_{ca} = \begin{cases} I_{ij}\left[(t - d_{ij})\, w_{ij}^{infreq} + d_{ij} w_{ij}^{freq}\right], \text{ if } d_{ij} \le t, \\ I_{ij} w_{ij}^c, \text{ otherwise ,} \end{cases} \tag{9}$$

where t denotes pixel threshold above the boundaries, d_{ij} is the L_1 distance between (i, j) and the boundaries, w_{ij}^{infreq} and w_{ij}^{freq} are weights of infrequent class and frequent class.

Take batch size n into consider, the semantic loss is given by :

$$\mathcal{L}_{sem} = \frac{1}{n} \sum w_{da} w_{ca} \mathcal{L}_{ce}. \tag{10}$$

For the auxiliary task instance segmentation, we adopt the standard Mask R-CNN loss. In particular, we take EIoU as the criterion for judging positive matches when calculating object proposal loss and bounding box loss.

The overall loss function is given by:

$$\mathcal{L} = \lambda_f \mathcal{L}_f + \lambda_{nf} \mathcal{L}_{nf} + \lambda_{sem} \mathcal{L}_{sem} + \lambda_{aux} \mathcal{L}_{aux}, \tag{11}$$

where $\lambda_f = \lambda_{nf} = 10$, and $\lambda_{sem} = \lambda_{aux} = 1$.

IV. EXPERIMENTS

A. Datasets

We benchmark DVT on two large-scale datasets nuScenes [25] and KITTI-360 [26]. For the nuScenes dataset, there are 702 training sequences and 148 validation sequences, which follows the same split specified in [18]. For the KITTI-360 dataset, we use sequences 0, 2-7 and 9 for training and use sequence 10 for validation, following the practice adopted in [31] and [21].

B. Training Details

Our DVT model is trained by images of size 768 $\times$ 448 pixels on nuScenes for 30 epochs and images of size 1408 $\times$ 384 pixels on KITTI-360 for 20 epochs. Random horizontal flips, and random perturbations of the image attributes (brightness, contrast and saturation) are used for dataset augmentation. The batch size is set to 4 for nuScenes and 2 for KITTI-360. The EfficientDet [33] backbone is initialized with weights pre-trained on the COCO dataset, and we use SGD as the optimizer with multi-step learning rate from 0.005 to 0.001. All experiments are conducted on a NVIDIA A6000 GPU.

TABLE I

INTERSECTION OVER UNION SCORES [%] OF BEV SEMANTIC SEGMENTATION PERFORMANCE ON NUSCENES DATASET. BOLD INDICATES THE BEST AND UNDERLINED INDICATES THE SECOND BEST.

Method	Drivable	Car	Pedestrian	Vegetation	Terrain	Invisible	Manmade	Sidewalk	Truck	Motorcycle	Mean
IPM [1]	50.56	6.00	0.12	21.42	10.44	0.00	18.79	8.69	1.21	0.09	11.73
VED [15]	73.68	29.67	1.58	33.47	29.28	32.14	34.07	23.20	22.74	1.95	28.18
VPN [16]	73.16	30.72	2.54	32.27	29.47	31.01	33.03	23.82	23.55	6.25	28.58
PON [18]	74.07	32.21	2.94	34.40	29.03	32.21	31.56	23.25	27.56	5.56	29.28
PoBEV [31]	77.32	40.53	4.98	**35.06**	33.56	36.65	**36.72**	**28.55**	**33.47**	9.63	33.65
DVT (Ours)	**77.33**	**42.10**	**5.29**	34.54	**33.80**	**37.31**	36.49	28.45	32.50	**10.42**	**33.82**

TABLE II

INTERSECTION OVER UNION SCORES [%] OF BEV SEMANTIC SEGMENTATION PERFORMANCE ON KITTI-360 DATASET. BOLD INDICATES THE BEST AND UNDERLINED INDICATES THE SECOND BEST.

Method	TIIM [23]	PoBEV [31]	SkyEye [21]	Ours
Road	63.08	70.14	71.39	**74.60**
Sidewalk	28.66	35.23	**37.62**	37.08
Building	13.70	34.68	**37.48**	28.69
Car	33.31	39.77	32.73	**42.12**
Truck	8.52	**14.38**	10.48	13.32
2-Wheeler	**6.45**	5.63	4.72	5.16
Mean	25.62	33.31	32.46	**35.16**

C. Metrics

Our evaluation metric for BEV semantic segmentation is the Intersection over Union (IoU) score, which we compute by binarizing the output probability maps with the threshold of 0.5 following [18]. Invisible areas are considered on nuScenes and ignored on KITTI-360 following [31] and [21]. To benchmark our auxiliary task, we fuse the output of semantic segmentation and instance segmentation according to the method of [32] to obtain panoptic segmentation, and use the panoptic quality (PQ), recognition quality (RQ), and segmentation quality (SQ) metric as the evaluation criteria.

D. Experimental Results

We evaluate the performance of our proposed DVT model in comparison with IPM [1], VED [15], VPN [16], PON [18], PoBEV [31], TTIM [23] and SkyEye [21]. Specially, the IPM baseline is combined with the EfficientPS [32] to yield BEV semantic map. In Table I, we can observe that our proposed DVT outperforms all previous methods on the nuScenes dataset, especially in the classes of drivale area, car, pedestrian, invisible area and motorcycle. Our DVT also achieves a leading performance on the KITTI-360 Dataset, as shown in Table II. We outperform poBEV [31] by 1.85 pp on the mean IoU metric and improve the segmentation performance of drivable area and car by 4.46 pp and 2.35 pp on the IoU metric, respectively.

To obtain the panoptic segmentation results, we combine IPM [1], VPN [16], PON [18] the instance head and the panoptic fusion module from two panoptic segmentation networks EfficientPS (EPS) [32] and Panoptic-DeepLab (PDL) [40]. We adopt the same panoptic fusion module proposed in EPS [32]. Table III shows that our performance of auxiliary task exceeds previous (state-of-the-art) SOTA PoBEV [31] on almost all metrics, which means that our depth-aware weight and class-aware weight designed for the main task are also beneficial for the performance on auxiliary tasks. The improvement of our main task by introducing auxiliary tasks is specifically described in the section IV-E.

We can observe from Fig. 2 (a, b, e, f) that both DVT and PoBEV [31] are able to capture the characteristics of close regions and also can localize nearby vehicles with high accuracy. In addition, as evident from the improvement/error maps, our method precisely segments the drivable areas and vehicles in the distance as well. A similar observation can be made in Fig. 2 (c) where our method accurately predicts the curve of the road over long distances, but PoBEV [31] fails to do so and instead confuses the road with the vegetation. As shown in Fig. 2 (a, d, g), our method also performs better on reasoning about invisible areas and recognizing small objects. Fig. 2 (f) demonstrates that our method can localize a large number of vehicles simultaneously with higher accuracy, especially for distant vehicles. Fig. 2 (e, g, h) show the generalization of our method in challenging scenarios such as rainy days and dark nights.

E. Ablation Study

We begin with a simple baseline, consisting of a backbone network, a bottleneck network to map features from PV to BEV, and a segmentation head to predict final semantic maps. We then incrementally reintroduce each of the components of our approach: the dual-branch view transformation (T), the depth-aware weighting method (D), the auxiliary task head (A), and the class-aware weighting method (C). Each successive component improves the performance of mean IoU, as shown in Table IV. Particularly, the addition of the the dual-branch view transformation has a pronounced effect on the performance. The auxiliary task head significantly improves performance for the prediction on invisible area. The depth-aware and class-aware weighting method provide no advantage for large classes such as drivable area, but significantly improve performance for small, rare classes such as pedestrian and motorcycle.

V. CONCLUSIONS

In this paper, we propose DVT, a novel dual-branch view transformation for monocular BEV semantic segmentation.

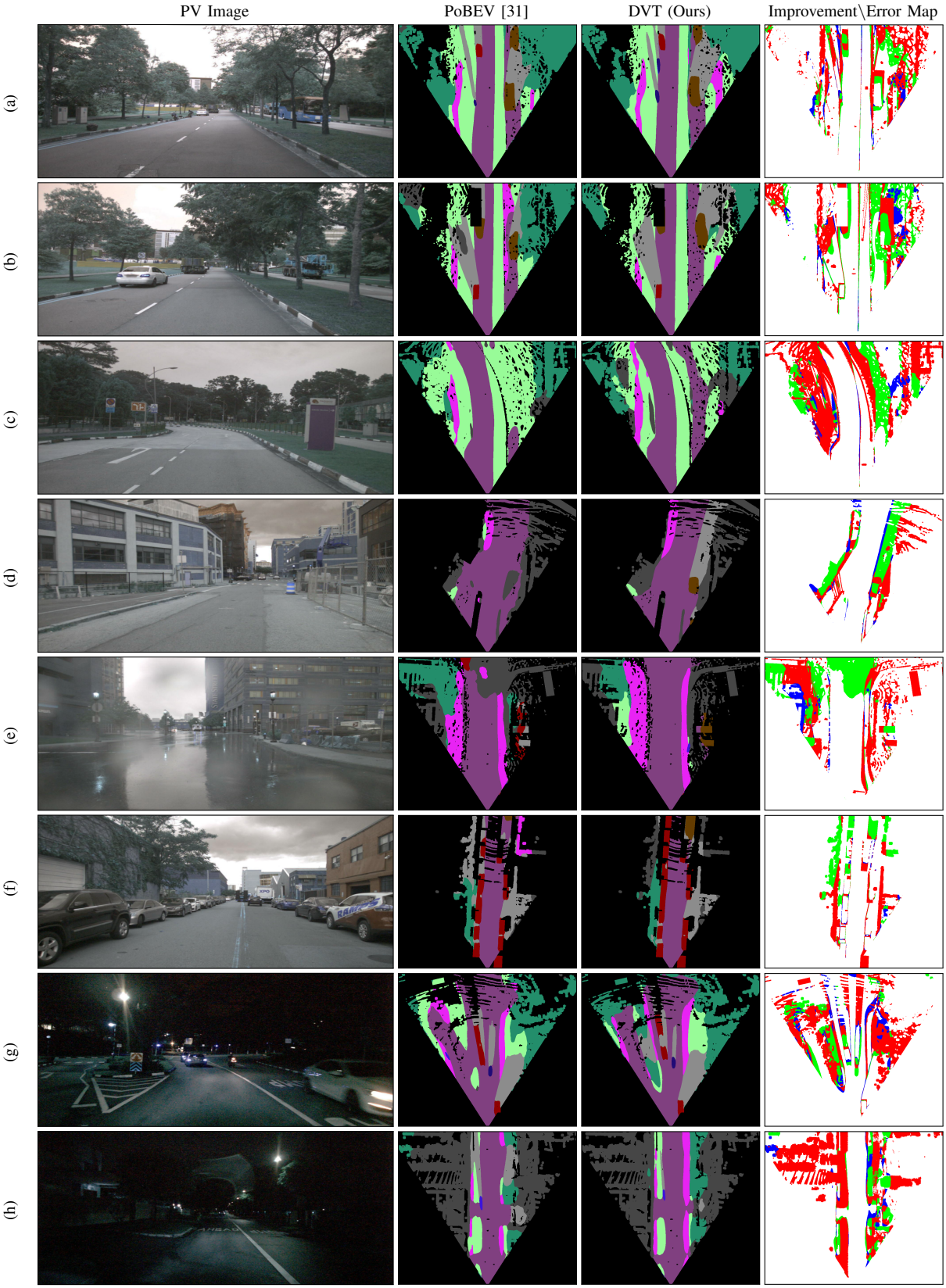

Fig. 2. Qualitative semantic segmentation results of our proposed DVT network in comparison to the state-of-the-art architecture PoBEV [31] on nuScenes benchmark datasets. In addition to the semantic segmentation output, we also show the improvement\error map which denotes the pixels that are misclassified by the PoBEV model but correctly predicted by the DVT model in green, the pixels that are misclassified by the DVT model but correctly predicted by the PoBEV model in blue, and the pixels that are misclassified by both the DVT model and the PoBEV model in red.

TABLE III

EVALUATION OF BEV PANOPTIC SEGMENTATION PERFORMANCE ON NUSCENES DATASET. ALL VALUES ARE IN [%].

Method	Overall			Thing			Stuff		
	PQ ↑	SQ ↑	RQ ↑	PQ (Thing)	SQ (Thing)	RQ (Thing)	PQ (Stuff)	SQ (Stuff)	RQ (Stuff)
IPM [1] + EPS [32]	5.63	35.13	8.62	0.04	13.87	0.07	9.35	49.29	14.32
VPN [16] + EPS [32]	14.35	63.67	21.16	6.35	66.16	9.52	19.69	62.00	28.92
VPN [16] + PDL [40]	14.91	64.44	22.01	7.76	**68.62**	11.39	19.67	61.64	29.08
PON [18] + EPS [32]	14.52	61.91	21.06	9.28	62.69	13.50	18.01	61.39	26.11
PON [18] + PDL [40]	14.72	63.04	21.21	8.98	65.40	12.78	18.54	61.46	26.33
PoBEV [31]	19.84	64.38	28.44	14.64	66.37	20.39	23.30	63.05	33.81
DVT (Ours)	**20.62**	**64.84**	**29.71**	**15.61**	67.36	**21.95**	**23.96**	**63.17**	**34.89**

TABLE IV

ABLATION STUDY ON THE VARIOUS ARCHITECTURAL COMPONENTS PROPOSED IN OUR MODEL. T: DUAL-BRANCH VIEW TRANSFORMATION. D: DEPTH-AWARE WEIGHTING METHOD. A: AUXILIARY TASK HEAD. C: CLASS-AWARE WEIGHTING METHOD. THE INTERSECTION OVER UNION SCORES [%] ARE REPORTED ON THE NUSCENES DATASET.

Model	Mean	Drivable	Car	Pedestrian	Vegetation	Terrain	Invisible	Manmade	Sidewalk	Truck	Motorcycle
Baseline	31.62	75.73	38.18	4.12	32.95	32.12	35.83	35.44	26.14	29.39	6.35
T	32.44	76.45	42.07	4.06	35.87	33.52	34.63	34.68	26.78	28.86	7.45
T+D	33.32	76.67	40.33	5.11	35.88	33.92	35.51	35.93	27.98	32.73	9.08
T+D+A	33.72	77.40	42.06	5.78	34.48	33.45	37.07	36.20	28.12	31.62	10.97
T+D+C	33.76	77.04	42.24	5.74	36.22	33.61	36.12	36.27	28.19	32.26	9.91
T+D+A+C	**33.82**	77.33	42.10	5.29	34.54	33.80	37.31	36.49	28.45	32.50	10.42

Our network includes several modules for addressing the problems in non-flat region distortion, distant depth ambiguity, and visual occlusion. Experimental results on two challenging datasets demonstrate the effectiveness of our approach. The limitation of our model is its reliance on large-scale labeled datasets, which are arduous to obtain. In the future, we will explore online automatic labeling of datasets and incremental learning policy on the BEV semantic segmentation.

REFERENCES

[1] H. A. Mallot, H. H. Bülthoff, J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological cybernetics*, vol. 64, no. 3, pp. 177–185, 1991.

[2] S. Sengupta *et al.*, "Automatic dense visual semantic mapping from street-level imagery," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2012, pp. 857–862.

[3] A. Loukkal *et al.*, "Driving among flatmobiles: Bird-eye-view occupancy grids from a monocular camera for holistic trajectory planning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 51–60.

[4] Y. B. Can *et al.*, "Understanding bird's-eye view of road semantics using an onboard camera," *IEEE Robotics and Automation Letters (RAL)*, vol. 7, no. 2, pp. 3302–3309, 2022.

[5] L. Reiher *et al.*, "A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–7.

[6] S. Srivastava *et al.*, "Learning 2D to 3D lifting for object detection in 3D for autonomous vehicles," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4504–4511.

[7] T. Bruls *et al.*, "The right (angled) perspective: Improving the understanding of road scenes using boosted inverse perspective mapping," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 302–309.

[8] Y. Wang *et al.*, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8445–8453.

[9] Y. You *et al.*, "Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving," *Computing Research Repository (CoRR)*, vol. abs/1906.06310, 2019. [Online]. Available: https://arxiv.org/abs/1906.06310

[10] R. Qian *et al.*, "End-to-end Pseudo-LiDAR for image-based 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5881–5890.

[11] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 194–210.

[12] C. Reading *et al.*, "Categorical depth distribution network for monocular 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8555–8564.

[13] A. Hu *et al.*, "FIERY: Future Instance Prediction in Bird's-Eye View From Surround Monocular Cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 273–15 282.

[14] T. Roddick *et al.*, "Orthographic feature transform for monocular 3d object detection," *Computing Research Repository (CoRR)*, vol. abs/1811.08188, 2018. [Online]. Available: https://arxiv.org/abs/1811.08188

[15] C. Lu, M. J. G. van de Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, 2019.

[16] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[17] A. Saha *et al.*, "Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5133–5139.

[18] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 138–11 147.

[19] W. Yang *et al.*, "Projecting your view attentively: Monocular road

scene layout estimation via cross-view transformation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 536–15 545.
[20] J. Zou *et al.*, "HFT: Lifting Perspective Representations via Hybrid Feature Transformation," *Computing Research Repository (CoRR)*, vol. abs/2204.05068, 2022. [Online]. Available: https://arxiv.org/abs/2204.05068
[21] N. Gosala, K. Petek, P. L. Drews-Jr, W. Burgard, and A. Valada, "Skyeye: Self-supervised bird's-eye-view semantic mapping using monocular frontal view images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 901–14 910.
[22] S. Chen *et al.*, "Efficient and Robust 2D-to-BEV Representation Learning via Geometry-guided Kernel Transformer," *Computing Research Repository (CoRR)*, vol. abs/2206.04584, 2022. [Online]. Available: https://arxiv.org/abs/2206.04584
[23] A. Saha *et al.*, "Translating images into maps," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9200–9206.
[24] S. Gong *et al.*, "GitNet: Geometric prior-based transformation for birds-eye-view segmentation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*. Springer, 2022, pp. 396–411.
[25] H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.
[26] Y. Liao *et al.*, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[27] Y. Ma *et al.*, "Vision-centric BEV perception: A survey," *Computing Research Repository (CoRR)*, vol. abs/2208.02797, 2022. [Online]. Available: https://arxiv.org/abs/2208.02797
[28] H. Li *et al.*, "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[29] X. Ma *et al.*, "Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6851–6860.
[30] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.
[31] N. Gosala and A. Valada, "Bird's-eye-view panoptic segmentation using monocular frontal view images," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1968–1975, 2022.
[32] R. Mohan and A. Valada, "Efficientps: Efficient panoptic segmentation," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1551–1579, 2021.
[33] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.
[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
[35] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens, "Searching for efficient multi-scale architectures for dense image prediction," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
[37] Y.-F. Zhang *et al.*, "Focal and efficient iou loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, 2022.
[38] Z. Wang *et al.*, "Revisiting evaluation metrics for semantic segmentation: Optimization and evaluation of fine-grained intersection over union," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[39] Y. Cui *et al.*, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9268–9277.
[40] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. 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*, 2020, pp. 12 475–12 485.