2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

HFT: Lifting Perspective Representations via Hybrid Feature Transformation for BEV Perception

Jiayu Zou[1], Zheng Zhu[2], Junjie Huang[2], Tian Yang[2], Guan Huang[2], Xingang Wang[1]

***Abstract*— Restoring an accurate Bird's Eye View (BEV) map plays a crucial role in the perception of autonomous driving. The existing works of lifting representations from frontal view to BEV can be classified into two categories, *i.e.*, Camera model-Based Feature Transformation (CBFT) and Camera model-Free Feature Transformation (CFFT). We empirically analyze the significant differences between CBFT and CFFT. The former method lift perspective features based on the flat-world assumption, which often causes distortion of regions lying above the ground plane. The latter method is limited in the perception performance due to the absence of geometric priors and time-consuming computing. In this paper, we propose a novel framework with a Hybrid Feature Transformation module (HFT) to lift perspective representations. Furthermore, we design a mutual learning scheme to augment hybrid transformation. The deformable attention mechanism enables the model to pay more attention to relevant regions and capture features with more semantics. We illustrate the effectiveness of HFT in BEV perception tasks, such as segmentation and object detection. Notably, in the task of semantic segmentation, extensive experiments demonstrate that HFT outperforms the previous state-of-the-art method by relatively 17.9% on the Argoverse and 22.0% on the KITTI 3D Object dataset. With negligible computing budget, HFT outperforms existing image-based methods on 3D object detection. The code will be released soon.**

I. INTRODUCTION

In recent years, with the rapid development of autonomous driving technologies, researchers make extensive efforts in 3D object detections [1]–[3], vehicle behavior predictions [4]–[7], and scene perceptions [8]–[10]. Autonomous vehicles demand a detailed and compact representation of their surrounding scenes for path planning and obstacle avoidance [11]–[14]. Due to the compactness and information richness, researchers usually re-represent the real world in the form of Bird's Eye View (BEV) instead of reconstructing the entire 3D world [15]–[17].

To estimate scene layout in BEV accurately, recent industrially-led methods rely on expensive sensors such as LiDAR and radar [18], [19]. Considering the limited resolution and lack of sufficient semantic information of those sensors, we choose to perform the BEV reconstruction by a single monocular frontal view (FV) RGB camera. The existing methods used for lifting the outdoor scenes from FV to top-down view can be broadly divided into two categories, *i.e.*, 1) Camera model-Based Feature Transformation (CBFT) [20]–[22] and 2) Camera model-Free Feature Transformation (CFFT) [23]–[25]. The CBFT exploits the scene geometry to transform the coordinates from image-based space to BEV. They capture scene geometry by using Inverse Perspective Mapping (IPM) or estimating the pixel depth to unproject 2D pixels into 3D space, then refine the coarse perspective features by adopting deep Convolutional Neural Networks (CNN) in a top-down view. On the contrary, the CFFT simulates the global projection process without geometric prior, which can get rid of the limitation of the flat-world assumption. However, both of those methods fall into erroneous BEV maps due to their inherent drawbacks and bring great challenges to accurate perception in BEV space.

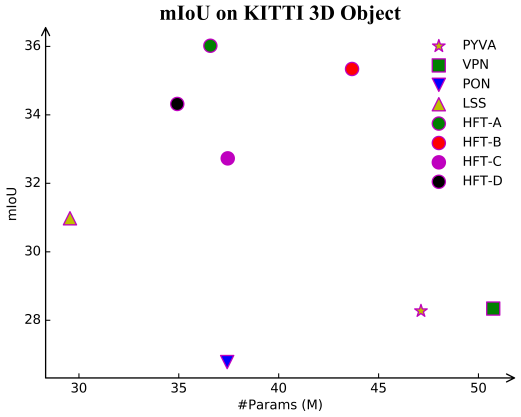


Fig. 1. The number of parameters and segmentation performance of different methods on the KITTI 3D Object dataset.

To analyze the differences and problems of the aforementioned methods, we first visualize the induced changes of the perspective transformation on the nuScenes dataset. As shown in Fig. 3, on the one hand, the CBFT only needs a few epochs to get an accurate representation of BEV space. However, there is a distortion in the CBFT for those regions above the ground, *e.g.*, the top view of the vehicle should be rectangular. On the other hand, CFFT requires more training epochs to learn a perspective transformation unconstrained by the flat-world assumption. These findings suggest that the geometric priors help the model converge, while the global information serves to get rid of the limitations of the flat-world assumption.

Based on the aforementioned analysis, we propose the Hybrid Feature Transformation (HFT) module, which consists of two different branches: Geometric Transformer and

[1]Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, {zoujiayu2020, xingang.wang}@ia.ac.cn
[2]PhiGent Robotics, Beijing, China, {zhengzhu, junjie.huang}@ieee.org, {tian.yang, guan.huang}@phigent.ai

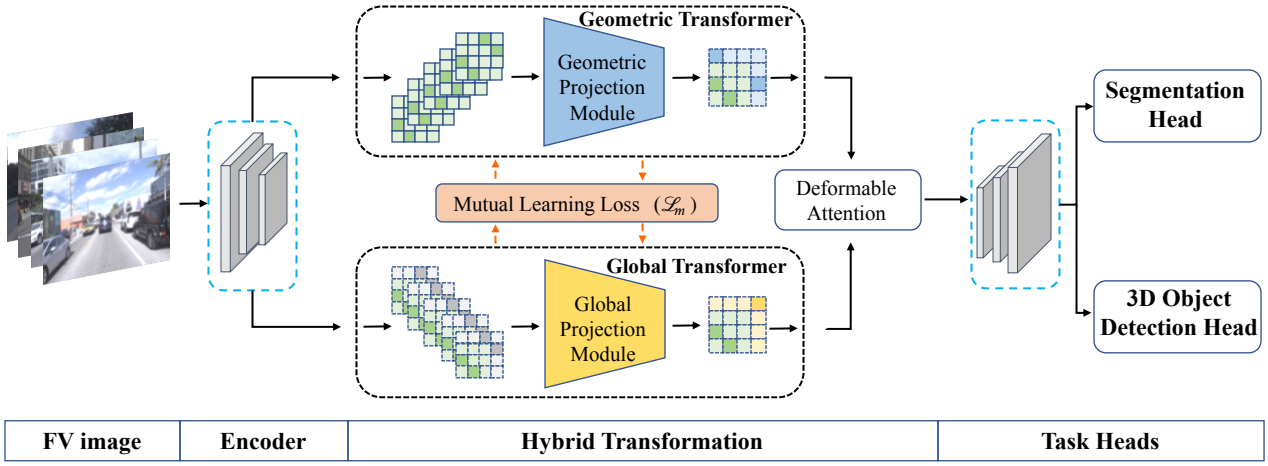


Fig. 2. Architecture diagram showing an overview of our proposed HFT framework.

Global Transformer, as shown in Fig. 2. Specifically, the Geometric Transformer utilizes an IPM-style layer to project views, which obtains a coarse BEV feature map in the early training phase. However, Geometric Transformer is limited by the flat-world assumption. Thus Global Transformer is designed to transform frontal views into top-down views with fully connected layers or attention mechanisms, which can globally model the relation between different views. To fully exploit the increased modeling capacity of HFT, we introduce a mutual learning [26], [27] scheme to encourage two different branches to learn potential representations from each other. To focus on the most relevant features, the deformable attention mechanism helps the model capture more informative features to refine feature maps in BEV. As illustrated in Fig. 1, with negligible extra overhead than CBFT and even less computation budget than CFFT, HFT obtains significant improvement in segmentation performance. Compared with View Parsing Network (VPN) [25], HFT reduces the number of parameters by 31.2% from 50.74M to 34.92M and still achieves a large relative 24.6% improvement of mIoU.

The contributions of our paper are summarized as follows:

- We empirically explore the significant differences between CBFT and CFFT. To the best of our knowledge, we are the first to point out that both the geometric priors in CBFT and the global contexts in CFFT are important for reconstructing BEV semantic maps.
- We propose a novel end-to-end learning framework, named HFT, to improve BEV perception performance using only monocular FV images. Mutual learning and deformable attention improve the model's performance to alleviate projection error and focus on the most informative regions.
- We illustrate the effectiveness of HFT in BEV perception tasks. HFT achieves state-of-the-art segmentation performance, *i.e.*, at least relative 17.9% and 22.0% improvement than previous methods on the Argoverse and KITTI 3D Object dataset respectively. With negligible computing budget, HFT outperforms existing image-based methods on 3D object detection.

II. Related Works

In autonomous driving tasks, BEV representation occupies a crucial position [28], [29]. Several works have been proposed to restore the BEV space from a single 2D FV image and these works can be broadly classified into CBFT and CFFT.

A. Camera model-Based Feature Transformation

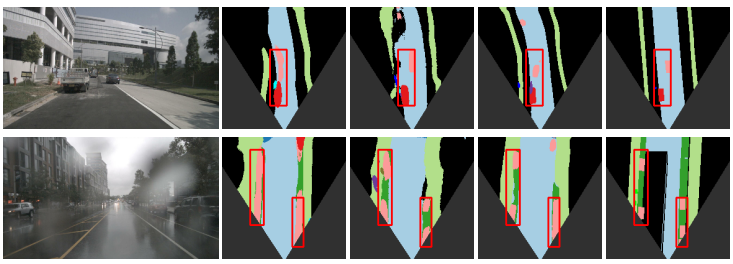
These methods account for the geometry of the scene to transform the original FV image into the BEV space. Pyramid Occupancy Networks (PON) [21] encodes the image onto the BEV space with pre-defined depth. Lift-Splat-Shoot (LSS) [22] estimates the probability distribution of the depth and unprojects FV features into a voxel grid for perspective transformation. CVT [30] leverages cross-view transformer to efficiently generate BEV semantic map. BEVDet [31], [32] series follow LSS [22] and use augmentation in both image-view and BEV space to improve 3D object detection performance. BEVerse [33] follows LSS [22] and jointly reasons about 3D object detection, map construction, and motion prediction. PETR [34] encodes the 3D positional embedding derived from camera parameters into multi-view features.

B. Camera model-Free Feature Transformation

Contrary to the CBFT methods, CFFT fully relies on the model's representation abilities to lift perspective representations. VED [24] utilizes a variational autoencoder (VAE) to encode the FV features and decode them into BEV space directly. VPN [25] adopts a multi-layer perceptron to transform FV features into BEV space. Projecting Your View Attentively (PYVA) [23] ignores geometric priors and uses the attention mechanism to learn the warping from the FV to BEV.

III. Method

In this section, we first analyze the differences between CBFT and CFFT, and then present our approach for downstream tasks in BEV perception. As illustrated in Fig. 2, we take an image $I_i \in \mathbb{R}^{H\times W\times 3}$ with an intrinsic matrix



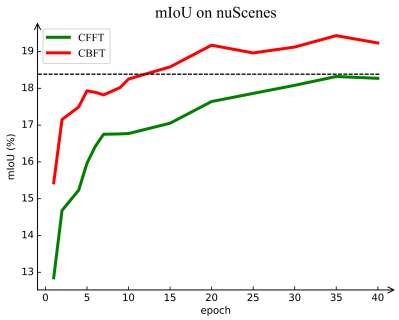


Fig. 3. The differences between CBFT and CFFT. The images in the 2nd, 3rd and 4th column are semantic maps in BEV produced by PYVA [23] trained for 20k iterations, PON [21] trained for 20k iterations and PYVA [23] trained for 200k iterations respectively.

$M_i \in \mathbb{R}^{3\times3}$ as input. We strive for a semantic representation of the scene in BEV map $B_i \in \mathbb{R}^{C\times X\times Y}$, where C is the number of categories and (X,Y) is the reference coordinates. It's worth noting that we use only a monocular FV image during the training or testing phase without any other sensors.

A. Difference Analysis

CBFT exploits the scene geometry explicitly by incorporating the camera projection model into the view transformer module. Such approaches follow the trivial IPM homography from image-based space to BEV. However, the flat-world assumption in IPM hinders the segmentation accuracy in areas above the ground, *e.g*, ramps and curbs with heights. The absence of geometric priors leads to estimating the geometry of layouts and the number of objects imprecisely. CFFT takes no geometric priors and fully relies on the network to learn the mapping relationship between FV and BEV. Compared with CBFT, CFFT costs more computing budget and longer inference time. Without geometric priors, CFFT tends to overfit the training data if given limited data. The flat-world assumption in CBFT and lack of geometric priors in CFFT limits the accuracy and effectiveness of the overall BEV space reconstruction.

We select PON [21] and PYVA [23] as the representative method of CBFT and CFFT, respectively. As shown in Fig. 3, we study the evolution of mean Intersection over Union (mIoU) scores on nuScenes set by applying PON [21] and PYVA [23]. We observe that PON [21] converges faster than PYVA [23], but obtains distortion in areas above the ground. Due to lacking of scene geometry, PYVA [23] converges slower than PON [21].

B. Network Overview

The differences between the two categories of methods inspire us to design a novel framework that follows the principle of reaping the benefits of both CBFT and CFFT as well as avoiding their drawbacks. As illustrated in Fig. 2, HFT is composed of a shared backbone (*i.e.*, encoder), a HFT module (*i.e.*, hybrid transformation), and task heads. Each component plays a different role in the task of BEV perception. The shared backbone adopts SwinTransformer [35] to extract FPN-style features with five different scales. Those features are fed into the HFT module which contains two different feature representation spaces and independently transforms both the static and dynamic elements into BEV maps. Furthermore, we propose a mutual learning strategy and deformable attention in HFT to unearth the potential of encouraging the CBFT branch (Geometric Transformer) and the CFFT branch (Global Transformer) to learn from each other. By leveraging task-specific heads, HFT can serve as a unified framework for different downstream tasks, such as segmentation and 3D object detection.

C. Hybrid Feature Transformation

As shown in Fig. 2, HFT consists of three parts: Global Transformer, Geometric Transformer and mutual learning scheme. Both the Global Transformer branch and the Geometric Transformer branch generate semantic probability occupancy grids in BEV space. Mutual learning measures the similarities between these two branches and seeks a better lifting perspective representation by feature mimicking.

Geometric Transformer. We design a Geometric Transformer by enhancing the IPM with the hallucinatory ability of the CNN. The IPM makes full use of camera intrinsic matrices M_i to produce perspective projection features F_i^{ipm} with rich geometric priors. However, due to the flat-world assumption, there is a spatial misalignment if pixels are above the ground plane. We resolve this error by using a resampling layer $C(x)$ to a certain extent. In other words, we first flatten the FV feature map F_i in height H dimension and warp it into different depth extents, and then estimate a coarse feature map with different depth ranges in top-down view space via adopting the IPM algorithm. Finally, by using $C(x)$ which is composed of a convolution layer and a bilinear sampling, we get the refined BEV features F_i^{geo} with geometric priors.

Global Transformer. Different from the Geometric Transformer, the Global Transformer ignores geometric priors and thus gets rid of the flat-world assumption, which can achieve better performance on those classes above ground. We use MLP and attention mechanism to model the perspective relation of FV and BEV spaces respectively. To this end, we first warp the 2D features F_i of the shape $C_i \times H_i \times W_i$ from the backbone into a volumetric grid of size $C_i \times Z_i \times W_i$, where the Z_i is various pre-defined depth. Subsequently, we generate the spatial occupancy relation $R(x)$ to get feature maps F_i^{glo} in BEV space.

Mutual Learning. With the aim of reaping the benefits of both the rich geometric priors of the Geometric Transformer and the strong global representation of the Global Trans-

former, we get ahead on the task of restoring BEV space. Inspired by Deep Mutual Learning (DML) [36] which uses feature maps to mutually distill knowledge from two models, we design a scheme to strike a balance between these two branches by the metric L_m. Finally, we conduct regularization on the semantic probability occupancy maps under different depth extents and concatenate feature maps to form the final BEV features F_i^{bev}.

Deformable Attention. We leverage the deformable attention mechanism [37] to enhance the location accuracy in BEV space. After gaining the BEV features from the Geometric Transformer and Global Transformer, we try to find a more accurate BEV representation. BEV features F_i^{geo} are linearly projected to query tokens Q, while F_i^{glo} is projected to key tokens K and value tokens V. We treat F_i^{geo} as the reference points and leverage a deformable attention mechanism between two kinds of BEV features. More specifically, we learn an offset $\Delta q = G_{offset}(F_i^{glo})$ for each reference point and use projector Γ to obtain the corresponding tokens.

$$\begin{aligned} Q = F_i^{geo} W_q, K = \hat{F}_i^{glo} W_k, V = \hat{F}_i^{glo} W_v \\ \text{where} \quad \hat{F}_i^{glo} = \Gamma(F_i^{glo} + \Delta q) \\ \text{Attn}\,(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \end{aligned} \tag{1}$$

D. Loss Function

For semantic segmentation, we design a loss function that consists of semantic loss and mutual learning loss. For 3D object detection, the total loss consists of classification loss, regression loss, and mutual learning loss. These losses are described as follows.

Mutual Learning Loss. We design the mutual learning loss L_m to measure the similarity of the CBFT branch and the CFFT branch, which can facilitate the flow of different lifted features, *i.e.*, F^{geo} and F^{glo}. We denote the i-th sub-feature as F_i^{geo} and F_i^{glo}, respectively. The mutual learning loss can be formulated as

$$L_m = \lambda_1 ||F^{geo} - F^{glo}||_2 + \sum_i^{N_{fea}} \lambda_2 ||F_i^{geo} - F_i^{glo}||_2 \tag{2}$$

where N_{fea} means the number of BEV semantic maps with respect to different depth extents and $\lambda_1 = 0.05$ and $\lambda_2 = 0.01$ are used to control the degree of re-weighting.

Semantic Loss. Inspired by focal loss [38], we modify the cross-entropy loss to address the class imbalance problem by increasing the weights of the infrequently occurring classes. For each prediction with a classification score c_i, it has a corresponding binary label y_i. The semantic loss and segmentation loss can be written as

$$L_s = \frac{1}{N_{pos}} \left[-\sum_i^{N_{pos}} w_i y_i \log c_i - \sum_i^{N_{neg}} (1-w_i)(1-y_i)\log(1-c_i) \right] \tag{3}$$

$$L_{seg} = L_s + L_m \tag{4}$$

where w_i is the weight of a class, which is computed as the inverse square root of its relative frequency.

Detection Loss. Following BEVDet [31], we adopt focal loss [38] for classification and L1 Loss for bounding box regression. We denote (c^*, box^*) and (c, box) for ground truths and predictions, respectively. The detection loss is formulated as

$$L_{det} = L_{cls}(c^*, c) + L_{reg}(box^*, box) + L_m \tag{5}$$

IV. Experiment

A. Dataset

We evaluate our approach on five different datasets, *i.e.*, nuScenes [39], Argoverse [40], KITTI Raw [41], KITTI Odometry [42], and KITTI 3D Object [41]. The nuScenes dataset includes 1000 autonomous driving scenes, whose source data is collected with six surround view cameras, five radar sensors and a LiDAR sensor. Argoverse [40] dataset includes images and point cloud data with seven surround view cameras, two stereo cameras and two LiDAR sensors. We only choose the images captured by the FV camera as our framework's input. The KITTI Odometry [42] provides 22442 well-annotated images both in FV and BEV from the Semantic KITTI dataset. The KITTI Raw [41] generates ground truth segmentation by registering the depth and semantic segmentation of LiDAR scans. For further comparison with existing 3D vehicle segmentation approaches, we also evaluate segmentation performance on the KITTI 3D object [41].

B. Implementation details

Models are trained by AdamW [43] optimizer with learning rate 2e-4, and gradient clip is exploited. The total batch size is 48 on 4 NVIDIA GeForce RTX 3090 GPUs. For the shared backbone, we initialize the SwinTransformer [35] backbone with weights pre-trained on the ImageNet [44] and apply a step learning rate policy which drops the learning rate at 17 and 20 epochs by a factor of 0.1. The total schedule is terminated within 40 and 20 epochs for segmentation and 3D object detection, respectively.

C. Approaches evaluation and benchmark results

For segmentation, we compare the HFT framework with previous methods on five public datasets. For 3D object detection, we evaluate the performance of the existing vision-based methods on nuScenes dataset. As illustrated in Fig. 1, we compare the perception performance of CBFT, CFFT, and HFT methods. As listed in Table II, we construct HFT networks with different geometric transformers and global transformers. We adopt HFT-A as the default setting.

Static Scene Layout Estimation. We compare HFT against Monocular semantic Occupancy (MonoOcc) [24], Monocular 3D (Mono3D) [45], PYVA [23] and PON [21]. Table III summarizes the performance of existing approaches on the KITTI Raw and KITTI Odometry benchmarks. We compare all methods under the same training protocol and manually densify the sparse semantic labels. As observed, HFT model outperforms all the existing baselines by large

TABLE I

INTERSECTION OVER UNION SCORES (%) OF HYBRID SCENE LAYOUT ESTIMATION ON THE ARGOVERSE DATASET.

Method	#param.	GFLOPs	Drivable	Vehicle	Pedest.	Large veh.	Bicycle	Bus	Trailer	Motorcy.	mIoU
IPM	-	-	43.7	7.5	1.5	-	0.4	7.4	-	0.8	-
Unproj.	-	-	33.0	12.7	3.3	-	1.1	20.6	-	1.6	-
VED [24]	-	-	23.9	6.2	9.7	0.9	3.0	0.4	1.9	13.9	
PYVA [23]	47.12M	127.1	78.95	33.91	6.87	18.29	6.1	32.5	32.39	1.01	26.25
PON [21]	37.42M	135.6	65.70	27.72	6.56	8.08	0.25	19.87	16.49	0.16	18.10
Ours	**36.58M**	**121.54**	**79.82**	**41.31**	**9.24**	**26.52**	**8.43**	**36.45**	**39.27**	**7.63**	**30.95**

TABLE II

THE COMPONENTS OF GEOMETRIC TRANSFORMER AND GLOBAL TRANSFORMER IN HFT.

Setting	Geometric Transformer	Global Transformer
HFT-A (default)	PON [21]	PYVA [23]
HFT-B	PON [21]	VPN [25]
HFT-C	LSS [22]	PYVA [23]
HFT-D	LSS [22]	VPN [25]

TABLE III

SEGMENTATION PERFORMANCE OF STATIC SCENE LAYOUT ESTIMATION ON KITTI RAW AND KITTI ODOMETRY AND DYNAMIC SCENE LAYOUT ESTIMATION ON KITTI 3D OBJECT.

KITTI	Raw		Odometry		3D Object	
Method	mIoU	mAP	mIoU	mAP	mIoU	mAP
MonoOcc [24]	58.41	66.01	65.74	67.84	20.45	22.29
Mono3D [45]	59.58	79.07	66.81	81.79	17.11	26.62
PYVA [23]	65.70	81.62	78.19	85.55	29.52	36.86
PON [21]	60.47	77.45	70.92	76.27	26.78	44.50
OFT [46]	-	-	-	-	25.34	34.69
VPN [25]	-	-	-	-	26.52	35.54
Ours	**66.42**	**81.84**	**79.49**	**88.92**	**36.02**	**55.84**

TABLE IV

ABLATION STUDY OF MUTUAL LEARNING SCHEMES ON THE NUSCENES DATASET. WE ADOPT MIOU AS THE SEGMENTATION METRIC.

schemes	baseline	CFFT-Teacher	output sim.	sub-feature sim.
mIoU	16.4	17.3	18.4	18.1

margins on both datasets. HFT achieves the highest mIoU of 66.42% and mAP of 81.84% than coexisting models in the KITTI Raw benchmark. Moreover, we observe a significant improvement with the KITTI Odometry dataset in the mIoU and mAP when compared to both CBFT and CFFT baselines.

Dynamic Scene Layout Estimation. Considering the variability of scales and mobility, dynamic scene layout estimation is a more challenging task than static scene layout estimation. As shown in Table III, HFT peaks both mIoU and mAP by a large margin than previous approaches, reaching a mIoU score of 36.02% and a mAP score of 55.84%. Note that, Mono3D [45] is a two-stage method and OFT [46] employs a parameter-heavy transformer which slows the speed down significantly. HFT avoids the limitations of the drawback for coexisting models to a large extent.

Hybrid Scene Layout Estimation. We present results for complex outdoor layout estimation on the Argoverse and nuScenes datasets. We follow the train-validation splits and ground-truth generation schemes provided by [21], [40]. We reimplement the experiments of PYVA [23] and PON [21]. As shown in Table I, HFT achieves the highest IoU score in all categories of objects on the Argoverse dataset, gaining a 30.95% mean IoU score, which is at least a relative 14.6% improvement from previous methods. Especially for dynamic classes such as vehicles and motorcycles, HFT shows a fabulous ability of semantic representation in BEV space and significantly improves performance for these kinds of objects. Furthermore, as listed in Table V, HFT scores 22.1% over 14 classes on the nuScenes dataset, which outperforms previous models by a relative improvement of 14.5%. HFT achieves state-of-the-art IoU scores in most of the classes and comparative results in drivable areas.

3D Object Detection. HFT can be used as a plugin-in method and can be easily extended to 3D object detection. In this part, we compare HFT with other baseline methods. Without bells and whistles, we use HFT to replace the FV2BEV module (*i.e.* LSS [22]) in BEVDet [31] series and conduct experiments on 3D object detection. The metrics are reported in Table VI on the nuScenes val set. With a negligible computing budget, HFT offers about 1% improvement on mAP and NDS metrics.

D. Ablation studies

The Effectiveness of HFT Framework. To explore the effectiveness of the HFT framework, we study the performances under different configurations on the KITTI 3D Object. As illustrated in Fig. 1, HFT achieves better semantic segmentation performance than both CBFT and CFFT. With a negligible computing budget, HFT models achieve much higher mIoU scores. By properly simplifying the channel or hidden size of image backbone or decoder, the HFT-D model reduces the number of parameters by 31.2% from 50.74M to 34.92M compared with VPN [25], but it still obtains an amazingly relative 24.6% improvement of mIoU.

Mutual Learning Schemes. In order to exploit the influence of different mutual learning schemes, we construct three different ways described as follows. As listed in Table IV, we show the positive effects of mutual learning schemes. 1) The model which uses the CFFT model as distillation teacher (CFFT-teacher) outperforms baseline method by 0.9%, reaching 17.3% score of mIoU. 2) Setting regularization on the output (output sim.) of two modules offers a positive impact on small and rare objects in autonomous driving scenarios. 3) Furthermore, by setting regularization on sub-features (sub-feature sim.), which means the semantic maps in BEV

TABLE V

INTERSECTION OVER UNION SCORES (%) OF HYBRID SCENE LAYOUT ESTIMATION ON THE NUSCENES DATASET.

Method	Drivable	Ped. crossing	Walkway	Carpark	Car	Truck	Bus	Trailer	Constr. veh.	Pedestrian	Motorcycle	Bicycle	Traf. Cone	Barrier	Mean
IPM	40.1	-	14.0	-	4.9	-	3.0	-	-	0.6	0.8	0.2	-	-	-
Unproj.	27.1	-	14.1	-	11.3	-	6.7	-	-	2.2	2.8	1.3	-	-	-
VED [24]	54.7	12.0	20.7	13.5	8.8	0.2	0.0	7.4	0.0	0.0	0.0	0.0	0.0	4.0	8.7
PYVA [23]	56.2	26.4	32.2	21.3	19.3	13.2	21.4	12.5	**7.4**	4.2	3.5	4.3	2.0	6.3	16.4
PON [21]	54.6	32.0	33.7	19.4	30.2	16.0	23.5	16.0	3.5	6.8	9.1	9.8	4.9	10.9	19.3
Ours	**56.3**	**36.2**	**35.8**	**23.8**	**31.2**	**19.7**	**28.5**	**19.1**	6.7	**8.5**	**12.0**	**12.4**	**6.7**	**11.4**	**22.1**

TABLE VI

3D OBJECT DETECTION PERFORMANCE OF DIFFERENT PARADIGMS ON THE NUSCENES VAL SET. TINY AND BASE MEANS TINY AND BASE SWIN TRANSFORMER RESPECTIVELY.

Methods	Image Size	#param.	GFLOPs	**mAP**↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	**NDS**↑	FPS
CenterNet [47]	-	-	-	0.306	0.716	0.264	0.609	1.426	0.658	0.328	-
FCOS3D [48]	1600×900	52.5M	2,008.2	0.295	0.806	0.268	0.511	1.315	**0.170**	0.372	1.7
DETR3D [49]	1600×900	51.3M	1,016.8	0.303	0.860	0.278	0.437	0.967	0.235	0.374	2.0
PGD [50]	1600×900	53.6M	2,223.0	0.335	0.732	0.263	0.423	1.285	0.172	0.409	1.4
PETR-R50 [34]	1056×384	-	-	0.313	0.768	0.278	0.564	0.923	0.225	0.381	10.7
PETR-Tiny [34]	1408×512	-	-	0.361	0.732	0.273	0.497	0.808	0.185	0.431	-
BEVDet-Tiny [31]	704×256	52.6M	215.3	0.312	0.691	0.272	0.523	0.909	0.247	0.392	**15.6**
BEVDet-Base [31]	1600×640	126.6M	2,962.6	0.393	0.608	0.259	0.366	0.822	0.191	0.472	1.9
BEVFormer [51]	1600×900	-	-	0.416	0.673	0.274	0.372	0.394	0.198	0.517	1.7
BEVDet4D-Tiny [32]	704×256	53.6M	222.0	0.338	0.672	0.274	0.460	0.337	0.185	0.476	15.5
BEVDet4D-Base [32]	1600×640	127.6M	2,989.2	0.421	0.579	0.258	0.329	0.301	0.191	0.545	1.9
BEVDet-Tiny [31] + **HFT**	704×256	52.8M	217.3	0.323	0.661	0.270	0.493	0.793	0.215	0.401	15.5
BEVDet-Base [31] + **HFT**	1600×640	126.9M	2,964.7	0.399	0.596	0.248	0.364	0.781	0.185	0.478	1.8
BEVDet4D-Tiny [32] + **HFT**	704×256	53.8M	224.1	0.349	0.614	0.260	0.355	0.321	0.175	0.485	15.3
BEVDet4D-Base [32] + **HFT**	1600×640	127.9M	2992.3	**0.431**	**0.562**	**0.217**	**0.289**	**0.279**	0.171	**0.556**	1.8

TABLE VII

ABLATION STUDY OF DEFORMABLE ATTENTION TRANSFORMATION ON DIFFERENT DATASETS. WE ADOPT MIOU AS THE SEGMENTATION METRIC.

deformable	nuScenes	Argoverse	RAW	Odometry	Object
✗	21.3	30.31	66.29	79.42	34.49
✓	**22.1**	**30.95**	**66.42**	**79.49**	**36.02**

of different depth ranges. HFT framework achieves similar performance as the second mutual learning scheme.

Deformable Attention Feature Interaction. We conduct an ablation study on the deformable attention mechanism and report the segmentation performance in Table VII. It can be illustrated that deformable attention between BEV features offers approximately 1% improvement on mIoU scores.

E. Qualitative Results

In this part, we show some qualitative segmentation results using different methods on nuScenes [39] and Argoverse [40]. As we can see in the visualization figures listed in Fig. 4, the HFT can produce more accurate semantic segmentation results with a large improvement in visualization than PYVA [23] and PON [21]. It's worth noting that HFT can predict the layout of roads, vehicles and other kinds of objects, especially in uphill, downhill and uneven scenes, which are usually the hard cases for the existing CBFT and CFFT. Another significant promotion is that HFT can produce much more clear edges in dense scenes, which is helpful for crowded autonomous driving scenarios.

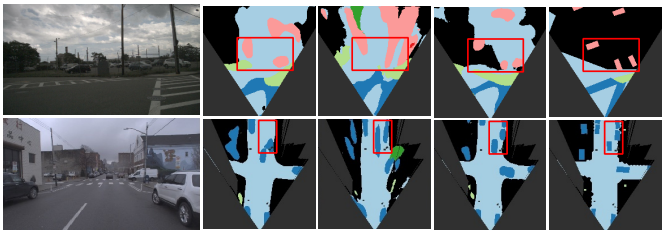


Fig. 4. Qualitative segmentation results on the nuScenes (the 1st row) and Argoverse (the 2nd row) datasets. For each grid location i, we visualise the class with the largest index c which has occupancy probability $p_i > 0.5$. Black regions (outside field of view or no lidar returns) are ignored during evaluation.

V. CONCLUSION

In this paper, we have proposed a novel framework to lift perspective representations from monocular images. We first explore the significant differences between CBFT and CFFT, concluding that scene geometry from the former method and global spatial context from the latter method are both crucial for BEV perception. Inspired by both factors, we propose the HFT framework, which employs a mutual learning scheme and deformable attention mechanism to learn better BEV representations. HFT efficiently offers an improvement on BEV perception tasks, such as semantic segmentation and 3D object detection. Extensive experiments demonstrate that HFT outperforms the previous state-of-the-art segmentation method by relatively 17.9% on the Argoverse and 22.0% on the KITTI 3D Object. With negligible computing budget, HFT outperforms existing image-based methods on 3D object detection.

REFERENCES

[1] Z. Wu, W. Zhang, J. Wang, M. Wang, Y.-Z. Gan, X. Gou, M. Fang, and J.-Y. Song, "Disentangling and vectorization: A 3d visual perception approach for autonomous driving based on surround-view fisheye cameras," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5576–5582, 2021.

[2] D. Feng, Y. Zhou, C. Xu, M. Tomizuka, and W. Zhan, "A simple and efficient multi-task network for 3d object detection and road understanding," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 7067–7074, 2021.

[3] M. Zhu, S. Zhang, Y. Zhong, P. Lu, H. Peng, and J. K. Lenneman, "Monocular 3d vehicle detection using uncalibrated traffic cameras through homography," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3814–3821, 2021.

[4] L. Wang, T. Wu, H. Fu, L. Xiao, Z. Wang, and B. Dai, "Multiple contextual cues integrated trajectory prediction for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 6, pp. 6844–6851, 2021.

[5] X. Xie, K. Shao, Y. Wang, F. Miao, and D. Zhang, "Automated type-aware traffic speed prediction based on sparse intelligent camera system," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4869–4874, 2021.

[6] Z. Sui, Y. Zhou, X. Zhao, A. Chen, and Y. L. Ni, "Joint intention and trajectory prediction based on transformer," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 7082–7088, 2021.

[7] E. M. Rella, J.-N. Zaech, A. Liniger, and L. V. Gool, "Decoder fusion rnn: Context and interaction aware decoders for trajectory prediction," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5937–5943, 2021.

[8] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1012–1025, 2014.

[9] S. Gupta, V. Tolani, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," *International Journal of Computer Vision*, vol. 128, pp. 1311–1330, 2019.

[10] T. Clunie, M. DeFilippo, M. Sacarny, and P. Robinette, "Development of a perception system for an autonomous surface vehicle using monocular camera, lidar, and marine radar," in *IEEE International Conference on Robotics and Automation*. IEEE, 2021, pp. 14 112–14 119.

[11] B. Li, Y. Zhang, T. Acarman, Y. Ouyang, C. Yaman, and Y. Wang, "Lane-free autonomous intersection management: A batch-processing framework integrating reservation-based and planning-based methods," in *International Conference on Robotics and Automation*. IEEE, 2021, pp. 7915–7921.

[12] S. V. Albrecht, C. Brewitt, J. Wilhelm, B. Gyevnar, F. Eiras, M. Dobre, and S. Ramamoorthy, "Interpretable goal-based prediction and planning for autonomous driving," in *International Conference on Robotics and Automation*. IEEE, 2021, pp. 1043–1049.

[13] J. Philion, "Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 574–11 583, 2019.

[14] Y. Luo, M. Meghjani, Q. H. Ho, D. Hsu, and D. Rus, "Interactive planning for autonomous urban driving in adversarial scenarios," in *International Conference on Robotics and Automation*. IEEE, 2021, pp. 5261–5267.

[15] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "Learning to look around objects for top-view representations of outdoor scenes," in *Proceedings of the European Conference on Computer Vision*, 2018.

[16] S. Sengupta, P. Sturgess, L. Ladicky, and P. H. S. Torr, "Automatic dense visual semantic mapping from street-level imagery," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 857–862, 2012.

[17] Z. Wang, B. Liu, S. Schulter, and M. Chandraker, "A parametric top-view representation of complex road scenes," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10 317–10 325, 2019.

[18] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. De La Escalera, "Birdnet: a 3d object detection framework from lidar information," in *International Conference on Intelligent Transportation Systems*. IEEE, 2018, pp. 3517–3523.

[19] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," *arXiv preprint arXiv:1608.07916*, 2016.

[20] L. Reiher, B. Lampe, and L. Eckstein, "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 *International Conference on Intelligent Transportation Systems*. IEEE, 2020, pp. 1–7.

[21] 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.

[22] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 194–210.

[23] W. Yang, Q. Li, W. Liu, Y. Yu, Y. Ma, S. He, and J. Pan, "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*, 2021, pp. 15 536–15 545.

[24] 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.

[25] 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.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[27] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," *Advances in neural information processing systems*, vol. 29, 2016.

[28] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 273–15 282.

[29] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," *arXiv e-prints*, pp. arXiv–2107, 2021.

[30] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[31] J. Huang, G. Huang, Z. Zhu, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[32] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[33] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[34] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," *arXiv preprint arXiv:2203.05625*, 2022.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[36] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4328, 2018.

[37] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4794–4803.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[39] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[40] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2019, pp. 8748–8757.

[41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[42] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9297–9307.

[43] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[45] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[46] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," in *British Machine Vision Conference*, 2019.

[47] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6569–6578.

[48] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 913–922.

[49] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[50] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Conference on Robot Learning*. PMLR, 2022, pp. 1475–1485.

[51] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.