DiffBEV: Conditional Diffusion Model for Bird's Eye View Perception

Jiayu Zou[1,3], Kun Tian[1,3], Zheng Zhu[2], Yun Ye[2], Xingang Wang[1]*

[1]Institute of Automation, Chinese Academy of Sciences
[2]PhiGent Robotics
[3]University of Chinese Academy of Sciences
zoujiayu2020@ia.ac.cn, kun.tian@phigent.ai, zhengzhu@ieee.org, yun.ye@phigent.ai, xingang.wang@ia.ac.cn

Abstract

BEV perception is of great importance in the field of autonomous driving, serving as the cornerstone of planning, controlling, and motion prediction. The quality of the BEV feature highly affects the performance of BEV perception. However, taking the noises in camera parameters and LiDAR scans into consideration, we usually obtain BEV representation with harmful noises. Diffusion models naturally have the ability to denoise noisy samples to the ideal data, which motivates us to utilize the diffusion model to get a better BEV representation. In this work, we propose an end-to-end framework, named DiffBEV, to exploit the potential of diffusion model to generate a more comprehensive BEV representation. To the best of our knowledge, we are the first to apply diffusion model to BEV perception. In practice, we design three types of conditions to guide the training of the diffusion model which denoises the coarse samples and refines the semantic feature in a progressive way. What's more, a cross-attention module is leveraged to fuse the context of BEV feature and the semantic content of conditional diffusion model. DiffBEV achieves a 25.9% mIoU on the nuScenes dataset, which is 6.2% higher than the best-performing existing approach. Quantitative and qualitative results on multiple benchmarks demonstrate the effectiveness of DiffBEV in BEV semantic segmentation and 3D object detection tasks.

Introduction

Bird's Eye View (BEV) perception plays a crucial role in autonomous driving tasks, which need a compact and accurate representation of the real world. One of the most important components of BEV perception is the quality of the BEV feature. Taking the classical LSS (Philion and Fidler 2020) as an illustration, it first extracts image features from the backbone encoder and then transforms them into BEV space along with depth estimation. However, the downstream perception results are often distorted, since the flat-world assumption is not always valid and the feature distribution in BEV is usually sparse. As shown in Fig. 1, when LSS (Philion and Fidler 2020) is utilized as the view transformer, the final segmentation results have three deficiencies: (1) The prediction of dynamic object boundaries is ambiguous, where pixels of different vehicles are connected; (2) The perception of static areas such as the pedestrian crossing and walkway is too rough. In particular, there are a lot of redundant predictions on the nuScenes benchmark; (3) LSS (Philion and Fidler 2020) has a poor discriminative ability for background and foreground pixels. In the last two rows of Fig. 1, the interested drivable area and vehicle objects are misclassified into the background.

The above observations intuitively motivate us to explore more fine-grained and highly detailed BEV feature for downstream perception tasks. Taking the noises in camera parameters and LiDAR scans into consideration, we usually obtain BEV representation with harmful noises. Diffusion models naturally have the ability to denoise noisy samples to the ideal data. Recently, the diffusion probability models (DPM) have illustrated their great power in generative tasks (Meng et al. 2021; Kim, Kwon, and Ye 2022; Bond-Taylor et al. 2022; Janner et al. 2022), but their potential in BEV perception tasks has not been fully explored. In this work, we propose DiffBEV, a novel framework that utilizes conditional DPM to improve quality of the BEV feature and push the boundary of BEV perception. In DiffBEV, the depth distribution or the BEV feature obtained from the view transformer is the input of conditional DPM. DiffBEV explores the potential of conditional diffusion model and progressively refines the noisy BEV feature. Then, the cross-attention module is proposed to fuse the fine-grained output of conditional diffusion model and the original BEV feature. This module adaptively builds the content relationship between the generated feature and the source BEV content, which helps to obtain a more precise and compact perception result.

DiffBEV is an end-to-end framework and can be easily extended by altering task-specific decoders. In this paper, we evaluate the performance of BEV semantic segmentation on standard benchmarks, *i.e.* nuScenes (Caesar et al. 2020), KITTI Raw (Geiger, Lenz, and Urtasun 2012), KITTI Odometry (Behley et al. 2019), and KITTI 3D Object (Geiger, Lenz, and Urtasun 2012). DiffBEV achieves a **25.9%** mIoU on the nuScenes benchmark, which is **6.2%** higher than previous best-performing approaches. DiffBEV outperforms other methods in the segmentation of drivable area, pedestrian crossing, walkway, and car by a substantial margin (**+5.0%**, **+10%**, **+6.7%**, and **+11.6%** IoU scores).

*Corresponding author.

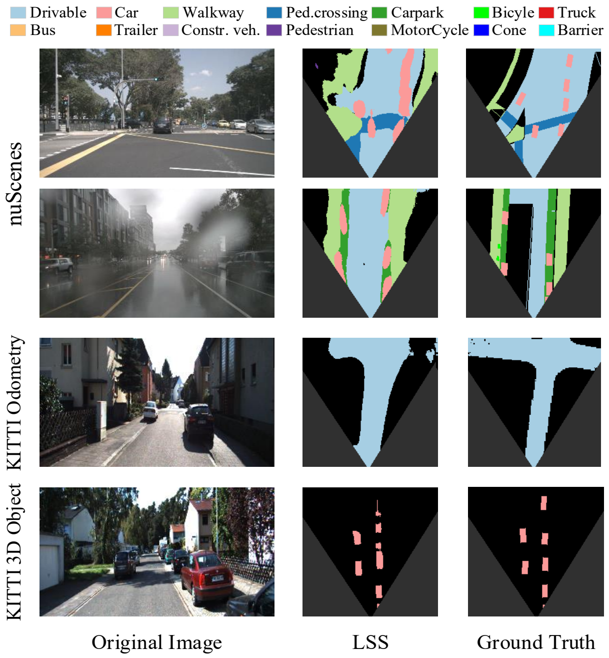

Figure 1: The poor segmentation results of LSS (Philion and Fidler 2020) model on the nuScenes and KITTI datasets.

Qualitative visualization results show that DiffBEV presents more clear edges than existing approaches. Furthermore, we compare the performance of 3D object detection on the popular nuScenes benchmark with other modern 3D detectors. Without bells and whistles, DiffBEV offers benefits to 3D object detection and provides approximately 1% NDS improvement on nuScenes. DiffBEV achieves leading performance both in BEV semantic segmentation and 3D object detection.

Our contributions can be summarized into three folds as follows.

(1) To the best of our knowledge, DiffBEV is the first work that utilizes conditional DPM to assist multiple autonomous driving perception tasks in BEV. Furthermore, DiffBEV needs no extra pre-training stage and is optimized in an end-to-end manner along with downstream tasks.

(2) The conditional DPM and the attentive fusion module are proposed to refine the original BEV feature in a progressive way, which can be seamlessly extended to different perspective view transformers, *e.g.* VPN (Pan et al. 2020), LSS (Philion and Fidler 2020), PON (Roddick and Cipolla 2020), and PYVA (Yang et al. 2021).

(3) Extensive experiments on multiple benchmarks demonstrate that DiffBEV achieves state-of-the-art performance and is effective in semantic segmentation and 3D object detection. DiffBEV achieves a **25.9%** mIoU on the nuScenes dataset, which outperforms previous best-performing approach (Philion and Fidler 2020) by a substantial margin, *i.e.* 6.2% mIoU.

Related Works

Diffusion Model

Diffusion models are widely used in Artificial Intelligence Generated Content (AIGC), which are of great importance in generative models. Diffusion models have illustrated their power in image generation (Rombach et al. 2022; Xiao, Kreis, and Vahdat 2021; Graikos et al. 2022; Huang, Lim, and Courville 2021), detection (Chen et al. 2022a), segmentation (Chen et al. 2022b; Amit et al. 2021; Baranchuk et al. 2021), image-to-image translation (Kawa et al. 2022; Jooyoung et al. 2021), super resolution (Saharia et al. 2022), image inpainting (Bond-Taylor et al. 2022), image editing (Meng et al. 2021), text-to-image (Kim, Kwon, and Ye 2022; Avrahami, Fried, and Lischinski 2022; Gu et al. 2022a), video generation (Singer et al. 2022; Ho et al. 2022), point cloud (Zeng et al. 2022; Zhou, Du, and Wu 2021; Luo and Hu 2021), and human motion synthesis (Janner et al. 2022; Shao et al. 2022).

DDPM-Segmentation (Baranchuk et al. 2021) is the first work to apply the diffusion model to semantic segmentation, which pre-trains a diffusion model and then trains classifiers for each pixel. But the two-stage paradigm, *i.e.* pre-training and fine-tuning, costs much training time, which is harmful to model efficiency. DiffusionInst (Gu et al. 2022b) applies the diffusion model to instance segmentation. A generalist framework (Chen et al. 2022b) leverages the diffusion model to generate results of panoptic segmentation. To this end, we are motivated to further explore the potential of employing the diffusion model to generate a high-quality representation for BEV perception tasks. Compared with DDPM-Segmentation (Baranchuk et al. 2021), DiffBEV is a generalist end-to-end framework, which can be optimized along with downstream tasks.

BEV Semantic Segmentation

BEV semantic segmentation is a fundamental and crucial vision task in BEV scene understanding and serves as the cornerstone of path planning and controlling. VPN (Pan et al. 2020) and PYVA (Yang et al. 2021) present the layout of static or dynamic objects through learnable fully connected layers and attention mechanisms, respectively. LSS (Philion and Fidler 2020) takes advantage of camera parameters to lift image-view features to BEV and is widely applied in modern 3D detectors. HFT (Zou et al. 2022) presents an approach to leverage the strengths of both camera parameter-free methods and camera parameter-based methods. CVT (Zhou and Krähenbühl 2022) extracts the content from surrounding-view images and achieves a simple yet effective design. GitNet (Gong et al. 2022) follows a two-stage paradigm, improving the segmentation performance by geometry-guided pre-alignment module and ray-based transformer. However, these works suffer from defective factors, such as distortion caused by inaccurate camera parameters. In DiffBEV, we propose a conditional diffusion model to refine the distorted features and improve the performance of previous methods for BEV semantic segmentation.

3D Object Detection

3D object detection (Duan et al. 2019; Wang et al. 2022b) is a prevailing research topic in autonomous driving. FCOS3D (Wang et al. 2021) proposes 3D centerness and learns the 3D attributes. PGD (Wang et al. 2022a) explores the geometric relationship of different objects and improves depth estimation. PETR (Liu et al. 2022) projects the camera parameters of multi-view images into 3D positional embeddings. BEVDet (Huang et al. 2021) shows the positive effects of data augmentation in image view and BEV. BEVDet4D (Huang and Huang 2022) explores both the spatial and temporal content to improve the performance. BEVDepth (Li et al. 2022) exploits the explicit depth supervision of multi-view images and further pushes the boundary of 3D object detection. BEVerse (Zhang et al. 2022) proposes a unified framework that jointly handles the tasks of 3D object detection, map construction, and motion prediction. In our work, we further exploit the ability of the conditional diffusion model to handle the task of 3D object detection.

Approach

Framework Overview

Fig. 2 shows the overall architecture of DiffBEV, which comprises of image view backbone, view transformer, conditional diffusion model, cross-attention module, and task-specific decoder. DiffBEV doesn't require an independent stage of pre-training and is trained in an end-to-end manner.

The image view backbone extracts the image features and the view transformer lifts the image-view features to BEV. Conditional diffusion model refines noisy samples and generates high-quality semantic feature. Cross-attention module is in charge of merging BEV feature and the output of conditional diffusion model. Finally, a task-specific decoder is applied for some downstream BEV perception tasks, such as segmentation and 3D object detection. In practice, LSS (Philion and Fidler 2020) is adopted as the default view transformer in our implementation.

Conditional Diffusion Probability Model

Diffusion Probability Model. We formulate the conditional diffusion probability model in this section. The feature generated by the view transformer is treated as the condition of diffusion model. Noise x_T obeys standard normal distribution $\mathcal{N}(0, I)$. Diffusion model transforms the noise x_T to the original sample x_0 in a progressive way. We denote the variance at step $t (0 \leqslant t \leqslant T)$ as β_t.

The forward process of the conditional diffusion model is presented as follows.

$$q(x_t|x_{t-1}) \sim \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

For convenience, we denote a series of constant.

$$\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \tag{2}$$

The noisy sample at step t is transformed from the input data x_0 by Eq. 3.

$$q(x_t|x_0) \sim \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_{t-1}, (1-\alpha)I) \tag{3}$$

$$x_t \sim \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{where} \quad \epsilon \sim \mathcal{N}(0, I) \tag{4}$$

$\Sigma_\theta(x_t, t)$ is the covariance predictor and $\epsilon_\theta(x_t, t)$ is the denoising model. In our experiments, a typical variance of UNet (Wu et al. 2022) is used as the denoising network. In the denoising process, the diffusion model progressively refines the noisy sample x_t. The reverse diffusion process is written as Eq. 5.

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

The Design of Condition. In practice, there are three types of conditions x_{cond} to choose: (1) The original BEV feature from the view transformer ($F^{O-BEV} \in \mathbb{R}^{C \times H \times W}$); (2) The semantic feature learned from the depth distribution ($F^{S-BEV} \in \mathbb{R}^{C \times H \times W}$); (3) The element-wise sum of F^{O-BEV} and F^{S-BEV}.

The view transformer lifts the image-view feature to BEV space, obtaining the original BEV feature F^{O-BEV}. For each point, the view transformer estimates the distribution on different predefined depth ranges and generates the corresponding depth distribution $F^d \in \mathbb{R}^{c \times h \times w}$. We employ a 1×1 convolutional layer to convert the channel and interpolate F^d into F^{S-BEV}, which has the same size as F^{O-BEV}.

The above three conditions are features in BEV space, where we add gaussian noise. By denoising samples progressively, we hope the conditional diffusion model helps to learn the fine-granularity content of objects, such as precise boundary and highly detailed shape. We strictly follow the standard DPM model to add BEV noise, while the difference is that we employ condition-modulated denoising, which is shown in Fig. 2.

Given noisy BEV feature x_t and condition x_{cond} at time step t, x_t is further encoded and interacts with x_{cond} through element-wise multiplication. To alleviate the computational burden, we set a flexible choice for the encoding mechanism of noisy BEV feature x_t, *i.e.* the self-attention mechanism or a simple convolutional layer, which will be discussed in Section . A UNet-style structure, whose components include an encoder and a decoder, serves as the denoising network $\epsilon_\theta(x_t, t)$.

Cross-Attention Module

After obtaining the output of conditional diffusion model, we design a cross-attention module (CA) to refine the original BEV feature, which is shown in Fig. 3.

Specifically, the output of the conditional diffusion model is treated as the source of K and V, while the original BEV feature from the perspective view transformer is projected into Q. The cross-attention process of the two-stream features is formulated as:

$$\begin{aligned} \mathrm{CA}(Q, K, V) &= \mathrm{Attn}(QW_i^Q, KW_i^K, VW_i^V)W^{Out}, \\ \mathrm{Attn}\,(Q, K, V) &= \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \end{aligned} \tag{6}$$

Q, K, and V are linearly mapped to calculate the attention matrix Attn, where W_i^Q, W_i^K, W_i^V are the projection layers with the shape of $\mathbb{R}^{d_{model} \times d_q}$, $\mathbb{R}^{d_{model} \times d_k}$, $\mathbb{R}^{d_{model} \times d_v}$.

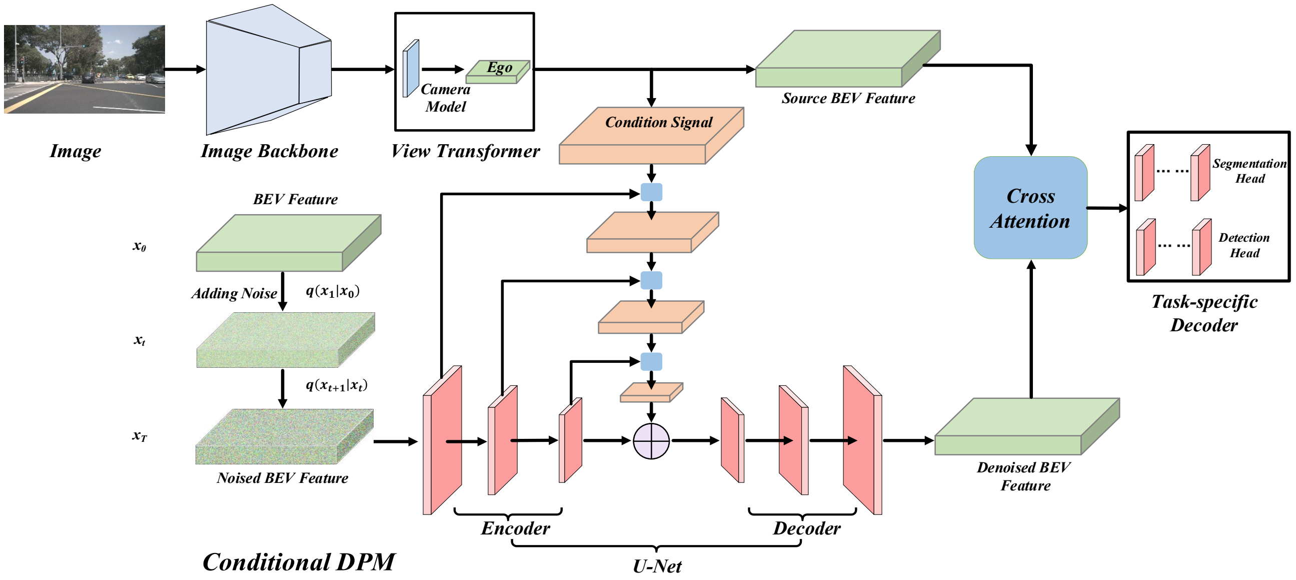

Figure 2: Overall architecture of DiffBEV. DiffBEV is comprised of the image backbone, view transformer, conditional diffusion model, cross-attention module, and task-specific decoder. By flexibly changing the task-specific decoder, DiffBEV can be easily extended to different downstream tasks, such as segmentation and 3D object detection.

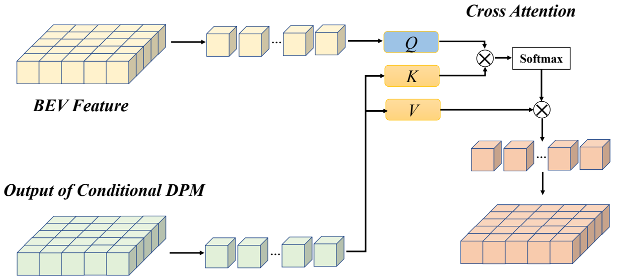

Figure 3: Overall structure of the cross-attention module.

Then, the refined BEV feature is obtained from the output layer $W^{Out} \in \mathbb{R}^{d_v \times d_{model}}$, which aims to facilitate the downstream tasks to learn better.

Training Loss

Depth Loss. Given the intrinsic parameter matrix $K_i \in \mathbb{R}^{3\times3}$, rotation matrix $R_i \in \mathbb{R}^{3\times3}$, and translation matrix $t_i \in \mathbb{R}^3$, we introduce a depth loss $\mathcal{L}_{depth}$ to assist model training. The depth loss is defined as the binary cross entropy (BCE) between the predicted depth map D_i and D_i^*. The specific process is expressed as:

$$\begin{aligned} P_i = K_i\left(R_i P + t_i\right), D_i^* = one\_hot(P_i), \\ \mathcal{L}_{depth} = \mathrm{BCE}(D_i^*, D_i) \end{aligned} \tag{7}$$

Diffusion Loss. We denote the gaussian noise at time step t as $\bar{z}_t$. Please refer to Section for the meaning of the rest symbols. The diffusion loss $\mathcal{L}_{diff}$ is defined as:

$$\mathcal{L}_{diff} = \mathbb{E}[||\bar{z}_t - \Sigma_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{z}_t, t)||^2] \tag{8}$$

Task-specific Training Loss. The training loss for segmentation and detection can be written as Eq. 9. In practice, we empirically set the loss weights $\lambda_1 = 10$ and $\lambda_2 = 1$. We introduce the details of segmentation loss $\mathcal{L}_{wce}$ and detection loss $\mathcal{L}_{detect}$ in the supplementary material.

$$\begin{aligned} \mathcal{L}_{seg} &= \mathcal{L}_{wce} + \lambda_1\mathcal{L}_{depth} + \lambda_2\mathcal{L}_{diff} \\ \mathcal{L}_{det} &= \mathcal{L}_{detect} + \lambda_1\mathcal{L}_{depth} + \lambda_2\mathcal{L}_{diff} \end{aligned} \tag{9}$$

Task-specific Decoder

As a general framework for BEV perception, DiffBEV can reason about different downstream tasks by altering the task-specific decoder. We adopt a residual-style decoding head for the semantic segmentation task, which consists of 8 convolutional blocks and a fully connected (FC) layer. Each convolutional block has a convolution layer, followed by batch normalization (BN) and a rectified linear unit (ReLU) layer. As for the 3D object detection task, the classification and regression heads are composed of several convolution layers respectively. Please refer to CenterPoint (Yin, Zhou, and Krahenbuhl 2021) for more structure details.

Experiment

Datasets

We compare the performance of DiffBEV with the existing methods on four different benchmarks, *i.e.* nuScenes (Caesar et al. 2020), KITTI Raw (Geiger, Lenz, and Urtasun 2012), KITTI Odometry (Behley et al. 2019), and KITTI 3D Object (Geiger, Lenz, and Urtasun 2012). As a popular benchmark in autonomous driving, nuScenes (Caesar et al. 2020) dataset is collected by six surrounding cameras and one LiDAR, which includes multi-view images and point cloud of 1,000 scenes. KITTI Raw (Geiger, Lenz, and Urtasun 2012) and KITTI Odometry (Behley et al. 2019) provide the images and BEV ground truth of the static road layout, while KITTI 3D Object (Geiger, Lenz, and Urtasun 2012) provides the images and labels for dynamic vehicles.

By flexibly leveraging different task-specific decoders, DiffBEV can be extended to various downstream tasks. In this work, extensive experiments are conducted on the BEV semantic segmentation and 3D object detection tasks.

Implementation Details

We train all semantic segmentation models using the AdamW optimizer (Ilya and Frank 2017) with learning rate and weight decay as 2e-4 and 0.01. Two NVIDIA GeForce RTX 3090 are utilized and the mini-batch per GPU is set to 4 images. The input resolution is 800 × 600 for nuScenes and 1024 × 1024 for KITTI datasets. The total training schedule includes 20,000 iterations (200, 000 iterations for nuScenes) and the warm-up strategy (Goyal et al. 2017) gradually increases the learning rate for the first 1,500 iterations. Then, a cyclic policy (Yan, Mao, and Li 2018) linearly decreases the learning rate from 2e-4 to 0 during the remainder training process. For 3D object detection, we follow the implementation details of BEVDet (Huang et al. 2021).

For the image backbone, the SwinTransformer (Liu et al. 2021) is initialized with the weights pre-trained on the ImageNet (Russakovsky et al. 2015) dataset. The model structures of VPN (Pan et al. 2020), PON (Roddick and Cipolla 2020), LSS (Philion and Fidler 2020), and PYVA (Yang et al. 2021) are the same as the original paper. In addition, we mainly follow the methods of the BEVDet (Huang et al. 2021) family to achieve 3D object detection. The training and testing details are consistent with (Huang et al. 2021) and (Huang and Huang 2022). Last but not least, there is no extra pre-training stage for the conditional diffusion probability model, which can be optimized in an end-to-end manner along with the downstream tasks.

BEV Semantic Segmentation

Evaluation on the nuScenes benchmark. In this part, we compare the effectiveness of DiffBEV with other approaches on the pixel-wise segmentation task. Both the layout of static objects and dynamic objects are estimated on the nuScenes benchmark.

As illustrated in Tab. 1, we report the segmentation performance of DiffBEV and some advanced methods described in Section . It can be seen that the previous state-of-the-art method LSS (Philion and Fidler 2020) is good at predicting static objects with wide coverage, such as the drivable area, walkway, and pedestrian crossing, compared to the car, pedestrian, bicycle, *etc.* This is because dynamic objects usually occupy fewer pixels and appear less frequently in BEV. A similar performance can also be observed from PYVA (Yang et al. 2021) and PON (Roddick and Cipolla 2020), which achieve a comparable accuracy in the drivable area class but perform worse in the rare class, such as truck, bus, and trailer.

In contrast, DiffBEV has a remarkable improvement in the Intersection over Union (IoU) score of both static and dynamic objects. As listed in Tab. 1, we design three varieties according to the condition. The condition of DiffBEV-B, DiffBEV-D, and DiffBEV-DB comes from the original BEV feature (F^{O-BEV}), conditional features learned from the depth distribution (F^{S-BEV}), and the element-wise sum of F^{O-BEV} and F^{S-BEV}, respectively. DiffBEV-D leads the performance in most classes and achieves a **25.9%** mIoU score, which is **6.2%** higher than previous best-performing approach (Philion and Fidler 2020). In particular, DiffBEV improves the segmentation accuracy of the drivable area, pedestrian crossing, walkway, and car by a substantial margin (**+5.0%**, **+10.0%**, **+6.7%**, and **+11.6%** IoU scores), which are crucial classes for the safety of autonomous driving systems. We attribute this improvement to that the conditional DPM reduces noises and complements more spatial information about objects of interest. DiffBEV significantly improves the pixel-wise perception accuracy of the model in both high-frequency classes and sparsely distributed classes. Please refer to visualization results for a more intuitive analysis and explanation.

Evaluation on KITTI Raw, KITTI Odometry, and KITTI 3D Object benchmark. Tab. 2 reports the quantitative results of static scene layout estimation on KITTI Raw and KITTI Odometry datasets. The performance comparison on KITTI 3D Object dataset shows the segmentation results for dynamic vehicles. Three varieties of DiffBEV obtain higher mIoU and mAP scores than existing methods. For example, DiffBEV-Dep surpasses the second-best model PYVA (Yang et al. 2021) by 0.71%, 1.51%, and 7.97% mIoU on KITTI Raw, KITTI Odometry, and KITTI 3D Object dataset, which achieves state-of-the-art perception accuracy consistently on all evaluation benchmarks.

3D Object Detection

We conduct 3D object detection experiments on the nuScenes benchmark and Tab. 3 reports the official evaluation metrics: mean Average Precision (mAP), Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), Average Attribute Error (AAE), and NuScenes Detection Score (NDS). Note that we select LSS (Philion and Fidler 2020) as the default view transformer, and use the semantic feature learned from the depth distribution (F^{S-BEV}) as the condition of DiffBEV. The data augmentations in image view and BEV are strictly consistent with that of the BEVDet (Huang et al. 2021) and BEVDet4D (Huang and Huang 2022).

After applying the conditional diffusion model, it can be observed that all evaluation metrics for 3D object detection are improved. This is because DiffBEV progressively refines the original BEV feature and interactively exchanges the semantic context through the cross-attention mechanism. Without bells and whistles, BEVDet (Huang et al. 2021) with DiffBEV raises the NDS score from 38.7% to 39.8%, while BEVDet4D (Huang and Huang 2022) with DiffBEV raises the NDS score from 47.6% to 48.6%.

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	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	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
VPN	58.0	27.3	29.4	12.9	25.5	17.3	20.0	16.6	4.9	7.1	5.6	4.4	4.6	10.8	17.5
PON	60.4	28.0	31.0	18.4	24.7	16.8	20.8	16.6	12.3	8.2	7.0	9.4	5.7	8.1	19.1
LSS	55.9	31.3	34.4	23.7	27.3	16.8	27.3	17.0	9.2	6.8	6.6	6.3	4.2	9.6	19.7
DiffBEV-BEV	65.3	40.2	41.0	27.2	37.9	21.3	32.9	20.5	7.6	9.2	13.7	13.1	7.2	16.0	25.2
DiffBEV-DepBEV	64.9	39.7	40.7	27.7	37.7	22.3	32.5	21.4	**12.7**	9.2	13.3	12.8	6.6	15.9	25.5
DiffBEV-Dep	**65.4**	**41.3**	**41.1**	**28.4**	**38.9**	**23.1**	**33.7**	**21.1**	8.4	**9.6**	**14.4**	**13.2**	**7.5**	**16.7**	**25.9**

Table 1: Intersection over Union scores (%) of hybrid scene layout estimation on the nuScenes **val** dataset.

KITTI	Raw		Odometry		3D Object	
Method	mIoU	mAP	mIoU	mAP	mIoU	mAP
OFT	-	-	-	-	25.34	34.69
MonoOcc	58.41	66.01	65.74	67.84	20.45	22.29
Mono3D	59.58	79.07	66.81	81.79	17.11	26.62
VPN	64.65	78.20	78.16	84.73	26.52	35.54
PYVA	65.70	81.62	78.19	85.55	29.11	36.86
PON	60.47	77.45	70.92	76.27	26.78	44.50
DiffBEV-B	66.19	81.08	79.48	88.30	36.76	52.81
DiffBEV-DB	66.40	81.89	79.58	88.44	**37.08**	**53.96**
DiffBEV-D	**66.41**	**81.91**	**79.70**	**89.68**	36.99	53.61

Table 2: Segmentation performance of static scene layout estimation on KITTI Raw and KITTI Odometry, and dynamic scene layout estimation on KITTI 3D Object.

Ablation Study

Condition Design. In order to exploit the advantages of the conditional diffusion model, we conduct ablation experiments for different DPM conditions on the KITTI Raw dataset to estimate the layout of static roads. Specifically, there are three DPM conditions to choose, *i.e.* the original BEV feature (F^{O-BEV}), the semantic feature learned from the depth distribution (F^{S-BEV}), and the element-wise sum of F^{O-BEV} and F^{S-BEV} ($both$).

As shown in Tab. 4, no matter which condition is used, three conditions can guide the DPM to learn discriminative BEV feature. F^{S-BEV} and F^{S-BEV} & F^{O-BEV} achieve better modulation effects than the F^{O-BEV}, while the best segmentation result comes from F^{S-BEV}. This observation demonstrates the effectiveness of semantic feature learned from the depth distribution.

Feature Interaction Mechanism. Another ablation study is to explore the most effective way for feature interaction.

As shown in each row of Tab. 4, regardless of which feature interaction mechanism is employed, DiffBEV achieves better segmentation results than the baseline model with 63.38% mIoU. It can be seen that cross-attention can learn better BEV feature than the other two simple feature interactions, which is beneficial for the downstream perception tasks. In summary, the combination of F^{S-BEV} and the cross-attention feature interaction mechanism achieves the best segmentation results, which improves 2.48% mIoU based on LSS (Philion and Fidler 2020) model. If not specified, the DiffBEV model corresponds to the setting of F^{S-BEV} with the cross-attention mechanism.

Encoding Mechanism for Noisy BEV Samples. For the noisy BEV sample x_t, we calculate the self-attention semantic map or obtain the refined affinity map through a simple convolutional layer. Tab. 5 shows the comparison between the computational burden and segmentation performance. The DiffBEV model using self-attention mechanism achieves a higher 65.86% mIoU and an 80.62% mAP. By simplifying self-attention to a simple convolutional layer, the DiffBEV model achieves a 64.23% mIoU and a 78.34% mAP while decreases the GFLOPs from 446.81 to 433.72.

More View Transformers with DiffBEV

In the main experiments, we adopt LSS (Philion and Fidler 2020) as the view transformer. To investigate the generality of DiffBEV, we conduct experiments on more view transformers. As shown in Tab. 6, the model equipped with DiffBEV outperforms the version without DPM on both mIoU and mAP metrics by a significant margin. Benefited from DiffBEV, the models of VPN (Pan et al. 2020), PYVA (Yang et al. 2021), and PON (Roddick and Cipolla 2020) raise their performances on mIoU scores (**+1.19%**, **+1.61%**, **+0.59%**, respectively) and mAP scores (**+10.14%**, **+7.01%**, **+10.11%**, respectively). This observation illustrates that DiffBEV is not only effective for a specific view transformer.

Visualization Analysis

As indicated in Fig. 4, previous state-of-the-art methods tend to output relatively rough predictions. For instance, cars that should be independent individuals are connected into a strip region and the drivable area is misclassified as background.

Despite the complex and challenging street layouts on the nuScenes dataset, DiffBEV produces more accurate semantic maps and is able to resolve fine-grained details such as the spatial separation between neighboring vehicles, especially in the crowded autonomous driving scenarios.

Methods	Image Size	**mAP**↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	**NDS**↑
CenterNet	-	0.306	0.716	0.264	0.609	1.426	0.658	0.328
FCOS3D	1600×900	0.295	0.806	0.268	0.511	1.315	**0.170**	0.372
DETR3D	1600×900	0.303	0.860	0.278	0.437	0.967	0.235	0.374
PGD	1600×900	0.335	0.732	0.263	**0.423**	1.285	0.172	0.409
PETR-R50	1056×384	0.313	0.768	0.278	0.564	0.923	0.225	0.381
PETR-R101	1408×512	0.357	0.710	0.270	0.490	0.885	0.224	0.421
PETR-Tiny	1408×512	**0.361**	0.732	0.273	0.497	0.808	0.185	0.431
BEVDet-Tiny	704×256	0.310	0.681	0.273	0.570	0.933	0.223	0.387
BEVDet-Tiny+DiffBEV	704×256	0.315	0.660	0.265	0.567	0.878	0.219	0.398
BEVDet4D-Tiny	704×256	0.338	0.672	0.274	0.460	0.337	0.185	0.476
BEVDet4D-Tiny+DiffBEV	704×256	0.344	**0.652**	**0.262**	0.453	**0.312**	0.176	**0.486**

Table 3: 3D object detection performance of different paradigms on the nuScenes **val** set. Tiny means tiny Swin Transformer.

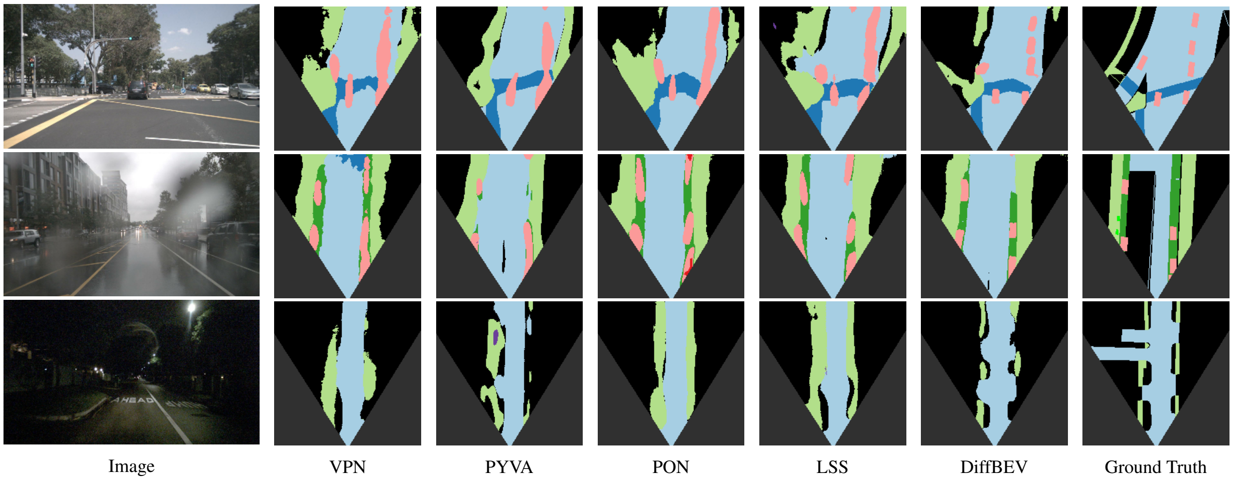

Figure 4: Qualitative segmentation results on the nuScenes benchmark. We visualize the class with the largest index c which has occupancy probability $p_i > 0.5$. Black regions (outside field of view or no LiDAR returns) are ignored during evaluation.

Interaction Mechanism	F^{S-BEV}	F^{O-BEV}	*both*
Concat	65.03	64.81	64.95
Add	64.85	64.11	64.50
Cross-Attention	**65.86**	64.33	65.16

Table 4: Ablation study on condition design and feature fusion mechanism.

Encoding	#param.	GFLOPs	mIoU	mAP
Conv	78.16M	433.72	64.23	78.34
Self-Attention	78.80M	446.81	65.86	80.62

Table 5: Ablation study on encoding mechanism in conditional diffusion model. The mIoU and mAP (%) of the basic LSS (Philion and Fidler 2020) on the KITTI Raw dataset are 63.38% and 77.52%, respectively.

Model	mIoU		mAP	
DiffBEV	✗	✓	✗	✓
VPN	27.02	28.21 (+1.19)	35.63	45.77 (**+10.14**)
PYVA	29.22	30.83 (**+1.61**)	36.97	43.98 (+7.01)
PON	36.49	37.08 (+ 0.59)	45.51	55.62 (+10.11)

Table 6: Extension experiments of more view transformers with DiffBEV on the KITTI 3D Object dataset. The metric (%) in the middle and right columns represent the performance without and with DiffBEV, respectively.

Conclusion

In this work, we propose a novel framework, namely DiffBEV, which first applies the conditional diffusion model to BEV perception tasks. DiffBEV utilizes BEV feature and semantic feature learned from the depth distribution as the condition of diffusion model, which progressively refines the noisy samples to generate highly detailed information. Then, a cross-attention module is proposed to attentively learn the interactive relationship between the output of conditional DPM and the BEV feature. Extensive experiments on multiple benchmarks illustrate that DiffBEV achieves favorable performance in both semantic segmentation and 3D object detection. DiffBEV obtains a 25.9% mIoU on the nuScenes, outperforming the previous state-of-the-art method by a substantial margin. The extension studies on different view transformers confirm the generality of DiffBEV. We hope to further explore the potential of DiffBEV and broaden its application ranges to more perception tasks.

References

Amit, T.; Nachmani, E.; Shaharbany, T.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Avrahami, O.; Fried, O.; and Lischinski, D. 2022. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*.

Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; and Babenko, A. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Bond-Taylor, S.; Hessey, P.; Sasaki, H.; Breckon, T. P.; and Willcocks, C. G. 2022. Unleashing transformers: parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *Proceedings of the European Conference on Computer Vision*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2022a. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2022b. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gong, S.; Ye, X.; Tan, X.; Wang, J.; Ding, E.; Zhou, Y.; and Bai, X. 2022. GitNet: Geometric Prior-based Transformation for Birds-Eye-View Segmentation. *arXiv preprint arXiv:2204.07733*.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Graikos, A.; Malkin, N.; Jojic, N.; and Samaras, D. 2022. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022a. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gu, Z.; Chen, H.; Xu, Z.; Lan, J.; Meng, C.; and Wang, W. 2022b. DiffusionInst: Diffusion Model for Instance Segmentation. *arXiv preprint arXiv:2212.02773*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *arXiv preprint arXiv:2204.03458*.

Huang, C.-W.; Lim, J. H.; and Courville, A. C. 2021. A variational perspective on diffusion-based generative models and score matching. In *Advances in Neural Information Processing Systems*.

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

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

Ilya, L.; and Frank, H. 2017. Fixing Weight Decay Regularization in Adam. *arxiv preprint arXiv:1711.05101*.

Janner, M.; Du, Y.; Tenenbaum, J. B.; and Levine, S. 2022. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*.

Jooyoung, C.; Sungwon; Kim, Y.; Jeong, Y.; Gwon, S.; and YoonJ. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2108.02938*.

Kawa, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2022. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*.

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

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

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Meng, C.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Pan, B.; Sun, J.; Leung, H. Y. T.; Andonian, A.; and Zhou, B. 2020. Cross-view semantic segmentation for sensing surroundings. In *IEEE Robotics and Automation Letters*.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*.

Roddick, T.; and Cipolla, R. 2020. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision*.

Saharia; C, H.; J, C.; and W. 2022. Image Super-Resolution via Iterative Refinement. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shao, R.; Zheng, Z.; Zhang, H.; Sun, J.; and Liu, Y. 2022. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *Proceedings of the European Conference on Computer Vision*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Wang, T.; Xinge, Z.; Pang, J.; and Lin, D. 2022a. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022b. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*.

Wu, J.; Fang, H.; Zhang, Y.; Yang, Y.; and Xu, Y. 2022. MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model. *arXiv preprint arXiv:2211.00611*.

Xiao, Z.; Kreis, K.; and Vahdat, A. 2021. Tackling the generative learning trilemma with denoising diffusion GANs. *arXiv preprint arXiv:2112.07804*.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. In *Sensors*.

Yang, W.; Li, Q.; Liu, W.; Yu, Y.; Ma, Y.; He, S.; and Pan, J. 2021. 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*.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zeng, X.; Vahdat, A.; Williams, F.; Gojcic, Z.; Litany, O.; Fidler, S.; and Kreis, K. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. *arXiv preprint arXiv:2210.06978*.

Zhang, Y.; Zhu, Z.; Zheng, W.; Huang, J.; Huang, G.; Zhou, J.; and Lu, J. 2022. BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving. *arXiv preprint arXiv:2205.09743*.

Zhou, B.; and Krähenbühl, P. 2022. Cross-view Transformers for real-time Map-view Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhou, L.; Du, Y.; and Wu, J. 2021. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zou, J.; Xiao, J.; Zhu, Z.; Huang, J.; Huang, G.; Du, D.; and Wang, X. 2022. HFT: Lifting perspective representations via hybrid feature transformation. *arXiv preprint arXiv:2204.05068*.