This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Spatiotemporal Decoupling for Efficient Vision-Based Occupancy Forecasting

Jingyi Xu[1*] Xieyuanli Chen[2*] Junyi Ma[3] Jiawei Huang[4] Jintao Xu[4] Yue Wang[5] Ling Pei[1†]
[1]Shanghai Key Laboratory of Navigation and Location Based Services, Shanghai Jiao Tong University
[2]College of Intelligence Science and Technology, National University of Defense Technology
[3]Department of Automation, Shanghai Jiao Tong University [4]HAOMO.AI
[5]State Key Laboratory of Industrial Control and Technology, Zhejiang University

Abstract

The task of occupancy forecasting (OCF) involves utilizing past and present perception data to predict future occupancy states of autonomous vehicle surrounding environments, which is critical for downstream tasks such as obstacle avoidance and path planning. Existing 3D OCF approaches struggle to predict plausible spatial details for movable objects and suffer from slow inference speeds due to neglecting the bias and uneven distribution of changing occupancy states in both space and time. In this paper, we propose a novel spatiotemporal decoupling vision-based paradigm to explicitly tackle the bias and achieve both effective and efficient 3D OCF. To tackle spatial bias in empty areas, we introduce a novel spatial representation that decouples the conventional dense 3D format into 2D bird's-eye view (BEV) occupancy with corresponding height values, enabling 3D OCF derived only from 2D predictions thus enhancing efficiency. To reduce temporal bias on static voxels, we design temporal decoupling to improve end-to-end OCF by temporally associating instances via predicted flows. We develop an efficient multi-head network EfficientOCF to achieve 3D OCF with our devised spatiotemporally decoupled representation. A new metric, conditional IoU (C-IoU), is also introduced to provide a robust 3D OCF performance assessment, especially in datasets with missing or incomplete annotations. The experimental results demonstrate that EfficientOCF surpasses existing baseline methods on accuracy and efficiency, achieving state-of-the-art performance with a fast inference time of 82.33 ms with a single GPU. Our code is released at: https://github.com/BIT-XJY/EfficientOCF.

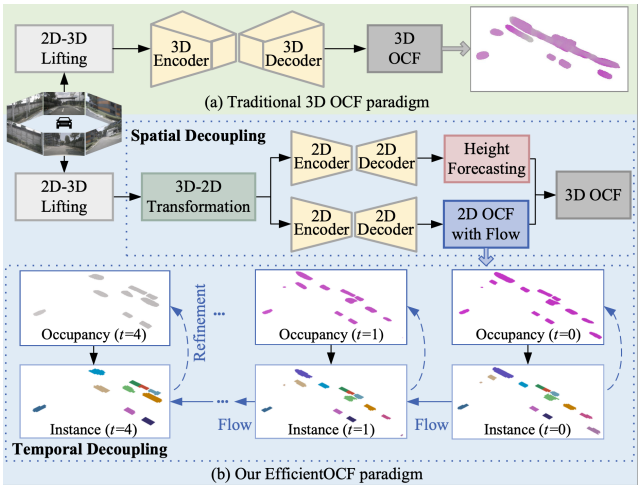

Figure 1. Unlike the traditional 3D vision-based OCF pipeline, our proposed paradigm implements spatiotemporal decoupling, thus achieving the efficiency of 2D OCF methods with the instance-aware capability of 3D OCF approaches.

1. Introduction

Accurately estimating occupancy states in the surrounding environment based on visual inputs is crucial for autonomous mobile systems, as it enables the identification of general obstacles, including objects absent from semantic databases. Most existing vision-based occupancy estimation methods [3, 7, 9, 10, 17, 34, 39] rely on complex model architectures but often overlook temporal changes in occupancy states. To address this, occupancy forecasting (OCF) approaches [1, 8, 12, 15, 25, 38] incorporate the prediction of surrounding object movements over time. These methods use past observations to forecast occupancy states in the present and near future, providing valuable information in advance for downstream tasks such as planning and control in autonomous driving.

Existing vision-based OCF methods predict future occupancy states for the entire 3D scene, as illustrated in Fig. 1(a). These methods process all voxel grids in each scene through end-to-end 3D feature encoding and decod-

*Equal contribution
†Corresponding author

This work was funded by Science and Technology Commission of Shanghai Municipality (24DZ3101300, 24TS1402600, 24TS1402800).

ing [21, 25]. However, this indiscriminate dense representation often leads to inefficient and less accurate OCF due to two main biases: (1) a spatial bias toward empty voxels, which dominate the 3D space; (2) a temporal changing bias toward static objects, as only a small subset of movable objects change positions in near future. These biases not only result in high computational costs and wasteful resource use but also increase uncertainty in predicting movable object occupancy, leading to rapid error accumulation in forecasts.

To tackle these issues, we propose a novel spatiotemporal decoupling representation and an efficient network EfficientOCF for vision-based OCF. As shown in Fig. 1(b), it decouples 3D OCF both spatially and temporally. We design a new spatial representation that decouples bird's-eye view (BEV) occupancy with height values. By storing only heights for 2D occupied grids, it achieves 3D OCF while avoiding the traditional memory-intensive 3D occupancy representation [20, 29, 34, 35]. For temporal changing forecasting, we decouple current instance segmentation from future occupancy estimation, leveraging instance movement understanding to enhance OCF rather than relying on end-to-end forecasting. To achieve these goals, we design EfficientOCF, a novel network with flow, segmentation, and height prediction heads. This network compresses 3D voxel features into 2D BEV features and employs lightweight 2D encoder-decoder structures to predict BEV occupancy. Simultaneously, it predicts 2D backward centripetal flow and height values for grids containing movable objects at each timestep. Flow vectors establish instance associations between current observations and future forecasts, refining OCF results at the instance level for enhanced accuracy and efficiency. The final 3D OCF is obtained by lifting refined 2D OCF with height estimation. We further introduce new metrics and conduct extensive experiments on the nuScenes [2], nuScenes-Occupancy [34], and Lyft-Level5 [11] datasets to demonstrate that our EfficientOCF generates high-quality OCF results and achieves superior real-time performance compared to the SOTA.

In summary, the contributions of our work are threefold:

- We propose a spatiotemporal decoupling framework for effective and efficient vision-based 3D OCF. Spatial decoupling is designed with a representation that combines BEV occupancy with height values to avoid 3D voxel redundancy. Temporal decoupling is achieved via instance-aware associations, obtaining enhanced 3D OCF results.
- We introduce EfficientOCF, a novel efficient vision-based 3D OCF network, that predicts the future occupancy states of movable objects using our proposed spatiotemporally decoupled representation. EfficientOCF employs lightweight 2D encoder-decoder structures rather than 3D counterparts and incorporates an adaptive dual pooling strategy for 3D-2D feature transformation, achieving SOTA 3D OCF performance and efficiency.
- We propose new evaluation metrics tailored for OCF on fine-grained general movable objects. These metrics address the shortcomings of traditional intersection over union (IoU) by minimizing penalties for false positives within the bounding box of movable objects, which enhances the rationality of assessments, especially in cases of missing or incomplete occupancy annotations.

2. Related Work

3D occupancy prediction. 3D occupancy prediction aims to estimate whether each voxel in the surrounding environment is currently occupied, representing general obstacles and enhancing the expressiveness of complex scenes. MonoScene, proposed by Cao *et al*. [3], first introduces a 3D semantic scene completion framework using only camera inputs, but it primarily focuses on front-view scenarios. To address this, TPVFormer by Huang *et al*. [9] proposes a tri-perspective view (TPV) representation to lift image features, achieving comparable performance with other LiDAR methods. Building on the TPV representation, they further propose SelfOcc [10], employing a self-supervised scheme to learn 3D occupancy states. Similar to TPVFormer, transformer modules [32] are also utilized in Occ3D [31], OccFormer [39], and SurroundOcc [35]. Occ3D [31] proposes CTF-Occ, a novel transformer-based course-to-fine occupancy network, achieving 3D occupancy predictions with high spatial resolution. OccFormer [39] proposes a dual-path transformer network to fully integrate 2D global features and 3D local features, while SurroundOcc proposed by Wei *et al*. [35] adopts spatial 2D-3D attention to lift images to the 3D volume space. Wang *et al*. [34] propose OpenOccupancy, the first surrounding semantic occupancy perception benchmark, and they introduce nuScenes-Occupancy dataset extending the large-scale nuScenes dataset [2] with fine-grained occupancy annotations. These approaches consider the complex challenge of inferring the spatial distribution of obstacles but ignore their temporal evolution. However, understanding the detailed and dynamic environment is crucial for planning and decision-making in autonomous driving, ensuring robust and reliable performance under various driving conditions.

BEV occupancy forecasting. BEV representation has been widely used in autonomous mobile systems due to its lightweight 2D format [16, 18, 19, 22, 26]. To overcome the challenge of capturing the temporal evolution in 2D format, some occupancy perception methods are proposed to achieve occupancy forecasting and instance prediction in BEV. FIERY, proposed by Hu *et al*. [8], is the first to use surround-view camera inputs to directly predict future instance segmentation and the motion of dynamic agents using BEV representation. Based on FIERY, StretchBEV by Akan *et al*. [1] incorporates a stochastic temporal model to refine residual updates. Zhang *et al*. [38] propose BE-

Verse, which also uses BEV occupancy to jointly reason about multiple tasks, including object detection, semantic map construction, and motion prediction. Similarly, MotionNet by Wu *et al*. [36] encodes object category and motion information in each BEV grid to predict scene flow and detect objects. Compared to MotionNet, MP3 proposed by Casas *et al*. [4] enhances occupancy flow parameterization with probabilistic modeling for multi-modal behavior. In contrast, PowerBEV by Li *et al*. [15] relies sorely on a semantic segmentation head alongside a backward centripetal flow head to achieve future instance prediction. Despite their effectiveness, these BEV-based methods lack a comprehensive understanding of spatial structures along the z-axis, which is crucial for accurately capturing the full 3D states of obstacles.

3D occupancy forecasting. Recently, researchers have increasingly explored end-to-end occupancy forecasting in 3D space. As depicted in the previous work [25], occupancy states can be directly generated from point cloud prediction [5, 23, 24, 27, 28] by trivial voxelization, but cannot achieve acceptable occupancy forecasting accuracy. OccWorld [40] and OccSora [33] use generation models to simultaneously predict the movement of the ego car and the evolution of the surrounding scenes. Ma *et al*. [25] introduce Cam4DOcc, the first benchmark to standardize evaluation protocol of vision-only occupancy forecasting in 3D space, and propose OCFNet, an end-to-end network to forecast dense 3D occupancy states. Compared to these 3D occupancy forecasting methods, in this work, we decouple spatiotemporal representation, where BEV occupancy with height values is utilized for spatial decoupling and instance-aware refinement in consecutive frames is designed for temporal decoupling. Based on the spatiotemporally decoupled representation, we propose a novel lightweight network EfficientOCF. Our experiments show that our proposed method achieves the SOTA performance with fast inference speed.

3. Proposed Method

3.1. Task Definition

Given N_p past and present surrounding visual images $\{\mathcal{I}|I_t \in \mathbb{R}^{N_\text{cam}\times 3\times H_\text{v}\times W_\text{v}}\}$, $t \in [-N_\text{p}, 0]$ as input, the vision-based OCF task aims to predict the occupancy states (occupied or free) of general movable objects in the present frame and N_f future frames $\{\mathcal{O}^\text{3D}|O_t^\text{3D} \in \mathbb{R}^{1\times H\times W\times L}\}$, $t \in [0, N_\text{f}]$. N_cam represents surrounding camera numbers. H_v, W_v represent the height and width of input images, and H, W, L represent the height, width, and length of the specific range defined in the present ($t = 0$) coordinate system.

Compared to traditional prediction methods, our work first spatially decouples OCF in 2D BEV occupancy $\{\mathcal{O}^\text{2D}|O_t^\text{2D} \in \mathbb{R}^{1\times H\times W}\}$ and corresponding height prediction $\{\mathcal{O}^\text{height}|O_t^\text{height} \in \mathbb{R}^{1\times H\times W}\}$, $t \in [0, N_\text{f}]$. It predicts the future height values of each voxel column and then lifts $\mathcal{O}^\text{2D}$ to $\mathcal{O}^\text{3D}$. Unlike static objects that remain unchanged over short periods, general movable objects exhibit more dynamic motion characteristics. Therefore, following [25], we focus on forecasting the occupancy states of general movable objects, e.g., bicycles, buses, cars, etc., in the nuScenes dataset [2]. This emphasis on movable objects enhances the relevance and applicability of the forecasting task to real-world autonomous driving scenarios, where accurately predicting the motion of these objects is vital for both safety and efficiency. While existing 3D OCF methods [25] forecast movable objects end-to-end, we propose temporal decoupling to propagate current instance segmentation through flow association and achieve refined 3D OCF.

3.2. Network Architecture

The pipeline of our EfficientOCF is shown in Fig. 2, which consists of four main components: a perception module, an aggregation module, a prediction module, and a refinement module. The perception module extracts features from the input surrounding images and lifts them from 2D to 3D space following Lift-Splat-Shoot [30] to obtain 3D spatial voxel features. The aggregation module then aggregates voxel features from sequential frames and transforms them into 2D BEV features, which are fed into the following prediction module. Next, the prediction module, composed of multiple 2D branches, decodes the aggregated 2D BEV features to initial 2D OCF, backward centripetal flow, and height values respectively. Then, the refinement module propagates instance segmentation into each future timestep, which is further used to refine initial OCF results in 2D BEV. Ultimately, we use height values predicted by the prediction module to lift the refined 2D occupancy to 3D space.

Perception module. To extract 3D voxel features from the input surrounding images, we adopt the approach used in previous studies [15, 25], applying a shared ResNet [6] backbone to derive 2D features from both past and present images. These features are represented as $\{\mathcal{F}^\text{2D}|F_t^\text{2D} \in \mathbb{R}^{c_\text{v}\times h_\text{v}\times w_\text{v}}\}$, where $t \in [-N_\text{p}, 0]$, c_v represents the 2D feature channel, and $h_\text{v} \times w_\text{v}$ denotes the 2D spatial resolution. These 2D features are then lifted and integrated into 3D voxel features $\{\mathcal{F}^\text{3D}|F_t^\text{3D} \in \mathbb{R}^{c\times h\times w\times l}\}$ using a 2D-3D lifting module [30], where c is the 3D feature channel, and $h \times w \times l$ represents the 3D spatial resolution.

Aggregation module. All the obtained 3D voxel features are projected into the present ($t = 0$) coordinate system exploiting 6-DOF ego-car poses. Here we propose an adaptive dual pooling strategy to further transform 3D voxel features into 2D BEV features. We first focus on the holistic information of each voxel column and employ average pooling to generate 2D features $\{\mathcal{F}^\text{avg}|F_t^\text{avg} \in \mathbb{R}^{c\times h\times w}\}$, where $t \in [-N_\text{p}, 0]$. Additionally, we use max pooling to generate the concurrent 2D features $\{\mathcal{F}^\text{max}|F_t^\text{max} \in \mathbb{R}^{c\times h\times w}\}$, cap-

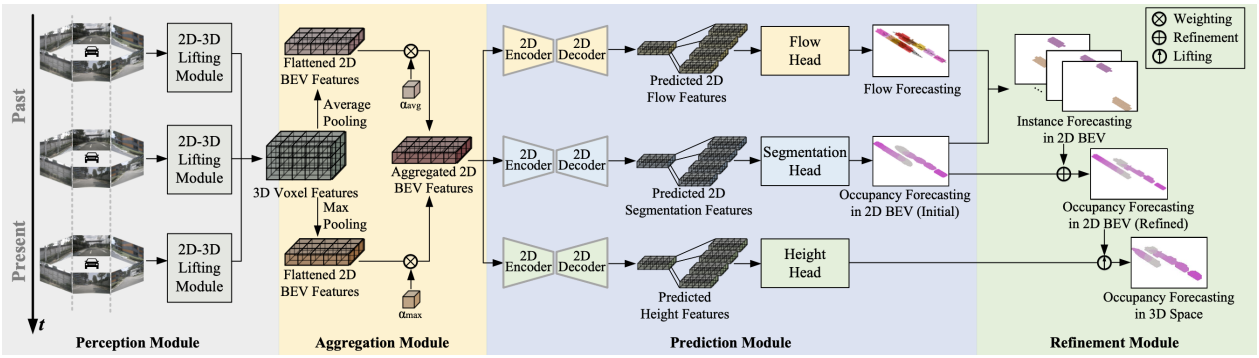

Figure 2. Architecture of our proposed EfficientOCF. The perception module (gray area) extracts 2D visual features from the input surrounding images and lifts them to 3D voxel features. Subsequently, the voxel features are further compressed into BEV features and aggregated by the aggregation module (yellow area). Next, the prediction module (blue area) with three branches sharing the same 2D structure, achieves initial OCF in 2D BEV and predicts backward centripetal flow and height values for BEV occupancy. The refinement module (green area) forecasts instances using estimated flow to refine initial 2D OCF results, which are ultimately lifted to 3D space.

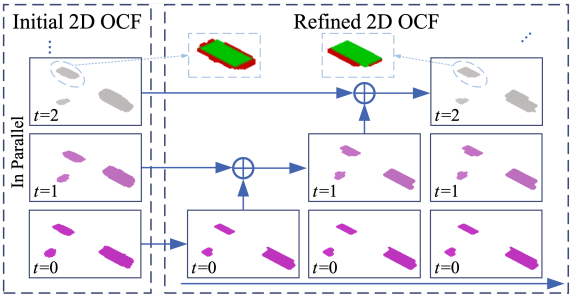

Figure 3. The process of refining initial OCF results following temporal decoupling. We exploit step-wise association along the time axis to improve forecasting quality.

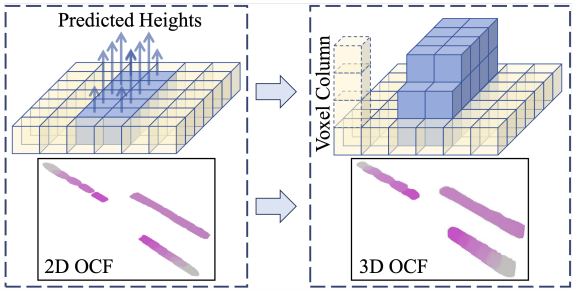

Figure 4. EfficientOCF achieves 3D OCF by assigning 2D grids with refined occupancy states and heights changing with time.

turing the voxel grid with prominent occupancy features implicitly corresponding to height values. Learnable weights, α_{avg} and α_{max}, are designed to adaptively adjust the contribution of these two types of 2D BEV features $\mathcal{F}^{\text{avg}}$ and $\mathcal{F}^{\text{max}}$. Finally, we aggregate features from both N_{p} past and present inputs to obtain the aggregated 2D BEV features $F^{\text{agg}} \in \mathbb{R}^{(N_{\text{p}}+1)\times c\times h\times w}$.

Prediction module. The proposed prediction module consists of three 2D convolution-based branches for forecasting initial 2D occupancy, height values, and backward centripetal flow vectors respectively. The aggregated 2D BEV features are fed into 2D encoder-decoder structures, and then processed by three predictive heads. The segmentation head estimates initial present and future occupancy states $\{\mathcal{O}^{\text{2D}}|O_t^{\text{2D}} \in \mathbb{R}^{1\times H\times W}\}$, $t \in [0, N_{\text{f}}]$ in 2D space. Concurrently, the height head determines the height values of 2D voxel grids, denoted as $\{\mathcal{O}^{\text{height}}|O_t^{\text{height}} \in \mathbb{R}^{1\times H\times W}\}$ for $t \in [0, N_{\text{f}}]$, based on our proposed representation with spatial decoupling. In addition, the flow head generates 2D backward centripetal flow [15] for each voxel grid between consecutive frames, denoted as $\{\mathcal{O}^{\text{flow}}|O_t^{\text{flow}} \in \mathbb{R}^{2\times H\times W}\}$, $t \in [0, N_{\text{f}}]$, which points from a voxel at time t to its corresponding potential 2D instance center at $t-1$.

Refinement module. We find that existing OCF methods [15, 25] suffer from shape divergence over time due to insufficient instance consistency across temporal frames. To address this, we propose decoupling current instance segmentation from future occupancy forecasts and refining 3D OCF through instance association, rather than learning 3D OCF end-to-end. The motivation is that instance segmentation based on current observations is often more accurate than end-to-end forecasts for future unobserved frames. By separating segmentation from occupancy forecasting, we simplify the learning process, making it easier compared to end-to-end 3D OCF. This allows us to leverage more accurate instance shapes to improve 3D OCF results. To this end, we design a refinement module that first performs instance segmentation in 2D BEV and then uses associated instances to mask inaccurate occupancy predictions for movable objects in the initial 2D OCF results, which are then lifted by height information to obtain refined 3D OCF. Specifically, we extract the centers of instances by non-maximum suppression (NMS) at $t=-1$ and associate pixel-wise instance ID between adjacent frames iteratively over time $t \in [0, N_{\text{f}}]$ using the predicted 2D backward centripetal flow. Here we denote the instance forecasting results as $\{\mathcal{M}^{\text{2D}}|M_t^{\text{2D}} \in \mathbb{R}^{1\times H\times W}\}$, $t \in [0, N_{\text{f}}]$. For each timestep t, we convert M_t^{2D} to the instance mask $\bar{M}_t^{\text{2D}}$ by $\texttt{CLIP}(M_t^{\text{2D}})$, which converts voxels of instances to occupied. Ultimately, we generate the final refined 2D OCF results $\bar{\mathcal{O}}^{\text{2D}}(i,j) = 1$, if $O_t^{\text{2D}}(i,j) = \bar{M}_t^{\text{2D}}(i,j) = 1$, else 0, where (i,j) is the

voxel index in BEV, as Fig. 3 shows. We therefore achieve temporal decoupling by splitting the end-to-end OCF into step-wise instance associations along the time axis. By assigning each 2D voxel grid with corresponding refined occupancy states and height values over time, we achieve lifting 2D OCF $\bar{\mathcal{O}}^{\text{2D}}$ to 3D OCF $\{\bar{\mathcal{O}}^{\text{3D}}|\bar{O}_t^{\text{3D}} \in \mathbb{R}^{1\times H\times W\times L}\}$, $t \in [0, N_{\text{f}}]$, as shown in Fig. 4.

3.3. Training and Inference

Our approach is trained and evaluated on three publicly available datasets, including the nuScenes [2], nuScenes-Occupancy [34], and Lyft-Level5 [11] datasets.

Data generation. We develop a novel data generation method based on three publicly available datasets. Initially, we split these datasets into sequences with a time length of $N = N_{\text{p}}+N_{\text{f}}+1$. Sequential semantic and instance annotations of general movable objects are then extracted for each sequence. Based on the extracted semantic information, we set the voxel grids corresponding to the general movable objects to 1 and the other positions to 0. We thereby obtain the bounding-box-aware 3D occupancy states $\{\hat{\mathcal{O}}^{\text{bb}}|\hat{O}_t^{\text{bb}} \in \mathbb{R}^{1\times H\times W\times L}\}$ (where 1 indicates occupied and 0 indicates free), which corresponds to the inflated annotations mentioned in [25]. In this case, we assign the heights of ground-truth bounding boxes to the voxel grids within them. Subsequently, we compress 3D voxel annotations into 2D BEV counterparts $\{\hat{\mathcal{O}}^{\text{bev}}|\hat{O}_t^{\text{bev}} \in \mathbb{R}^{1\times H\times W}\}$. This is achieved under the principle that if any voxel in a vertical voxel column has a value of 1, the corresponding position in the BEV is set to 1. We extract ground-truth height values of each voxel column $\{\hat{\mathcal{O}}^{\text{height}}|\hat{O}_t^{\text{height}} \in \mathbb{R}^{1\times H\times W}\}$ by selecting the highest grid with the semantics of general movable objects. We then conduct fine-grained 3D occupancy states $\{\hat{\mathcal{O}}^{\text{fg}}|\hat{O}_t^{\text{fg}} \in \mathbb{R}^{1\times H\times W\times L}\}$ by assigning height values to 2D BEV occupancy annotations.

Loss functions. We employ the cross-entropy loss as the OCF loss $\mathcal{L}_{\text{occ}}$, and use the smooth L1 loss for both the height prediction loss $\mathcal{L}_{\text{height}}$ and the flow prediction loss $\mathcal{L}_{\text{flow}}$. The total loss used to train our EfficientOCF is the weighted sum of these three losses, computed as follows:

$$\begin{aligned}\mathcal{L}_{\text{all}} = \frac{1}{N_{\text{f}}+1}\sum_{t=0}^{N_{\text{f}}}(\lambda_1\mathcal{L}_{\text{occ}}(\hat{O}_t^{\text{bev}}, \bar{O}_t^{\text{2D}}) \\ + \lambda_2\mathcal{L}_{\text{height}}(\hat{O}_t^{\text{height}}, O_t^{\text{height}}) + \lambda_3\mathcal{L}_{\text{flow}}(\hat{O}_t^{\text{flow}}, O_t^{\text{flow}})),\end{aligned} \tag{1}$$

where λ_1, λ_2, and λ_3 are the weighting parameters to balance the supervision for different heads.

Evaluation metrics. We employ intersection-over-union (IoU), and conditional IoU (C-IoU) to evaluate OCF performance. We also use video panoptic quality (VPQ) [13] to assess the performance of 3D instance forecasting. We utilize IoU_{c}, IoU_{f}, and $\tilde{\text{IoU}}_{\text{f}}$ to separately evaluate the occupancy estimation at the present moment ($t = 0$), the forecasting results for future timesteps ($t \in [1, N_{\text{f}}]$), and the occupancy estimation over the entire time horizon ($t \in [0, N_{\text{f}}]$). These evaluation metrics are computed by

$$\text{IoU} = \frac{1}{N_{\text{all}}}\sum_{t=t_{\text{s}}}^{t_{\text{e}}}\frac{\sum \hat{O}_t^{\text{occ}}\cdot\bar{O}_t^{\text{occ}}}{\sum \hat{O}_t^{\text{occ}} + \bar{O}_t^{\text{occ}} - \hat{O}_t^{\text{occ}}\cdot\bar{O}_t^{\text{occ}}}. \tag{2}$$

We set $\hat{O}_t^{\text{occ}}/\bar{O}_t^{\text{occ}}$ as $\hat{O}_t^{\text{bev}}/\bar{O}_t^{\text{2D}}$ to access 2D OCF performance. To evaluate the performance of forecasting 3D inflated and fine-grained occupancy respectively, we calculate IoU by setting $\hat{O}_t^{\text{occ}}/\bar{O}_t^{\text{occ}}$ as $\hat{O}_t^{\text{bb}}/\bar{O}_t^{\text{3D}}$ and $\hat{O}_t^{\text{fg}}/\bar{O}_t^{\text{3D}}$. We set $t_{\text{s}} = t_{\text{e}} = 0$ and $N_{\text{all}} = 1$ to calculate IoU_{c}, $t_{\text{s}} = 1$, $t_{\text{e}} = N_{\text{f}}$, and $N_{\text{all}} = N_{\text{f}}$ for IoU_{f}, and $t_{\text{s}} = 0$, $t_{\text{e}} = N_{\text{f}}$, and $N_{\text{all}} = N_{\text{f}} + 1$ for $\tilde{\text{IoU}}$.

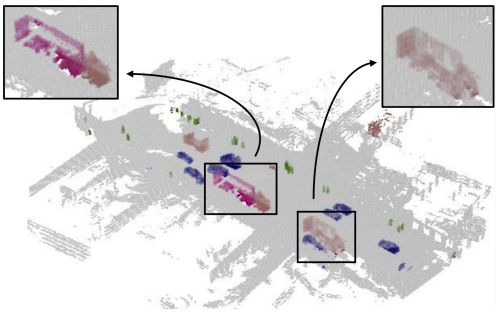

Figure 5. Examples of incomplete 3D occupancy labels due to LiDAR sparsity in nuScenes-Occupancy [34], which cannot comprehensively assess occupancy forecasting methods.

While conventional IoU metrics can evaluate forecasting results within bounding boxes or fine-grained voxels [25], they struggle with scenarios involving incomplete 3D annotations due to LiDAR data sparsity, as illustrated in Fig. 5. The nuScenes-Occupancy dataset [34] extends the original nuScenes dataset [2] with dense semantic occupancy annotation using the proposed augmenting and purifying (AAP) pipeline. However, during the AAP's automatic labeling process, some annotation omissions are inevitable, resulting in missed 3D occupancy labels due to factors such as occlusion or sparse LiDAR observations, and even introducing pseudo occupancy labels with self-training [37]. Additionally, during the manual purifying phase, while noise removal is effective, high-precision leak detection remains challenging. Therefore, we propose a new evaluation metric for 3D occupancy forecasting, conditional IoU (C-IoU):

$$\text{C-IoU} = \frac{1}{N_{\text{all}}}\sum_{t=t_{\text{s}}}^{t_{\text{e}}}\frac{|TP_t| + |\widetilde{FP}_t|}{|TP_t| + |FN_t| + (|FP_t| - |\widetilde{FP}_t|)}, \tag{3}$$

where TP_t, FN_t, and FP_t represent true positives, false negatives, and false positives at timestep t, computed in the 3D fine-grained format. $\widetilde{FP}_t$ represents partial false positives of FP_t, which lies in the corresponding annotated bounding boxes. Note that we only use this metric for nuScenes and nuScenes-Occupancy datasets because Lyft-Level5 lacks fine-grained voxel-wise annotations. C-IoU comprehensively considers predictions within

Table 1. Comparison of performance on 3D occupancy forecasting

Approach	nuScenes						Lyft-Level5						nuScenes & nuScenes-Occupancy					
	2D			3D			2D			3D			3D					
	IoU_c	IoU_f	$\widetilde{\text{IoU}}$	IoU_c	IoU_f	$\widetilde{\text{IoU}}$	IoU_c	IoU_f	$\widetilde{\text{IoU}}$	IoU_c	IoU_f	$\widetilde{\text{IoU}}$	IoU_c	IoU_f	$\widetilde{\text{IoU}}$	C-IoU_c	C-IoU_f	C-$\widetilde{\text{IoU}}_\text{f}$
PowerBEV	33.62	30.07	30.77	28.34	25.47	26.05	45.78	43.44	43.91	36.15	34.18	34.58	-	-	-	-	-	-
OpenOccupancy	28.77	25.93	26.50	24.26	22.69	23.15	38.16	35.11	35.79	32.95	29.73	30.40	14.99	13.06	13.47	27.11	24.68	25.17
OccFormer	38.98	31.05	32.76	34.75	28.08	29.56	47.00	32.10	35.02	41.68	28.55	31.00	15.80	11.30	12.18	44.56	35.98	37.67
OCFNet^-	30.34	26.27	27.47	27.32	24.40	24.98	36.02	33.12	33.69	31.04	28.92	29.36	12.66	10.65	11.05	29.64	27.47	27.91
OCFNet	35.50	32.02	32.71	33.24	29.36	29.94	45.74	43.67	44.08	39.71	38.14	38.46	14.73	12.57	13.00	34.83	33.14	33.49
EfficientOCF^- (ours)	31.69	29.00	29.56	30.24	28.17	28.60	36.39	33.24	33.83	33.95	31.12	31.70	15.98	14.05	14.46	35.89	34.20	34.56
EfficientOCF (ours)	**39.93**	**36.15**	**36.93**	**35.60**	**32.73**	**33.32**	**47.42**	**44.96**	**45.46**	**44.27**	**41.90**	**42.39**	**21.28**	**19.02**	**19.50**	**47.53**	**45.57**	**45.99**

bonding boxes and fine-grained occupancy predictions, accounting for the discrepancies between the nuScenes and nuScenes-Occupancy datasets to avoid penalizing false positives in bounding boxes unfairly. This mitigates the effect of missing and incomplete occupancy annotations within bounding boxes of movable objects. Additionally, we use VPQ [13] to evaluate 3D instance forecasting lifted from 2D counterparts by height values, which is calculated by

$$\text{VPQ}(\hat{\mathcal{O}}^{\text{3D}}, \bar{\mathcal{O}}^{\text{3D}}) = \sum_{t=0}^{N_\text{f}} \frac{\sum_{TP_t} \text{IoU}_t}{|TP_t| + \frac{1}{2}|FP_t| + \frac{1}{2}|FN_t|}. \quad (4)$$

We set $\hat{\mathcal{O}}^{\text{3D}}$ as $\hat{\mathcal{O}}^{\text{bb}}$ and $\hat{\mathcal{O}}^{\text{fg}}$ for inflated and fine-grained formats, denoted as VPQ^{bb} and VPQ^{fg} respectively.

4. Experiments

4.1. Experimental Setups

Dataset details. We evaluate our proposed EfficientOCF alongside the SOTA baselines using the nuScenes [2], nuScenes-Occupancy [34] and Lyft-Level5 [11] datasets. Following existing works [15, 25, 34], we split the nuScenes and nuScenes-Occupancy datasets into training and test sets with 700 and 150 scenes respectively, and the Lyft-Level5 dataset into 130 training scenes and 50 test scenes. Each sequence has a time length N of 7 frames ($N_\text{p} = 2$, $N_\text{f} = 4$), meaning that we use 2 past frames and 1 present observation to estimate the present occupancy states and forecast the counterparts of the next 4 frames. The nuScenes and nuScenes-Occupancy datasets are labeled at a frequency of 2 Hz, and the Lyft-Level5 dataset at 5 Hz, resulting in different forecasting time durations across these datasets.

Implementation details. Following [25], we set the evaluation range to $[-51.2\,\text{m}, 51.2\,\text{m}]$ for both the x-axis and y-axis, and $[-5\,\text{m}, 3\,\text{m}]$ for the z-axis. The voxel resolution is set to $0.2\,\text{m}$, resulting in 3D occupancy voxels of size $512 \times 512 \times 40$. We compare our EfficientOCF against four SOTA methods: PowerBEV [15], OpenOccupancy [34], OccFormer [39], and OCFNet [25]. All models are trained for 15 epochs using AdamW optimizer [14] with an initial learning rate 3×10^{-4} and a weight decay of 0.01. Eight A100 GPUs are used for training with a batch size of 8.

4.2. Occupancy Forecasting Assessment

We first assess the occupancy forecasting performance of EfficientOCF and baseline methods, as shown in Tab. 1. The nuScenes-Occupancy dataset provides fine-grained labels for the nuScenes dataset, allowing us to conduct evaluations based on fine-grained voxels. In contrast, the Lyft-Level5 dataset only contains bounding box annotations for movable objects, and thus the evaluation on this dataset is based on inflated objects where the C-IoU metric is not available. Notably, the vanilla PowerBEV only forecasts occupancy states in 2D space. Therefore, we lift its 2D results to 3D space by assigning a fixed height value to compute the 3D metrics. This fixed height for each test scene is determined by averaging the heights of all general movable objects. For OpenOccupancy and OccFormer, we estimate the occupancy states of the present frame and then use it as the prediction of all future timesteps based on the static-world model [25]. Besides, for OpenOccupancy, OccFormer, and OCFNet, which are end-to-end 3D occupancy perception methods, we first calculate 3D metrics and then compress the 3D results to BEV space to calculate the 2D metrics. Following [25], we additionally report the performance of OCFNet and our EfficientOCF when trained on limited $1/6$ of the training sequences, denoted as OCFNet^- and EfficientOCF^-. Tab. 1 shows that our proposed EfficientOCF estimates the present occupancy states and forecasts future counterparts better than all the other baselines on the nuScenes, nuScenes-Occupancy, and Lyft-Level5 datasets in both 2D and 3D evaluation metrics. In particular, EfficientOCF improves the proposed C-IoU of the baselines more significantly than the other existing metrics. It is also notable that EfficientOCF^- trained on limited data even outperforms the baseline OCFNet supervised by the holistic training set on all 3D fine-grained metrics. This validates that our method more effectively captures diverse motion patterns, even with only partial scenario perception, making it practically useful when training data is limited.

We further report the forecasting performance with different future time horizons. As shown in Tab. 2, our EfficientOCF remains the best performance within the future time horizons $\{0.5\,\text{s}, 1.0\,\text{s}, 1.5\,\text{s}, 2.0\,\text{s}\}$ on the nuScenes and nuScenes-Occupancy datasets, and $\{0.2\,\text{s}, 0.4\,\text{s}, 0.6\,\text{s}, 0.8\,\text{s}\}$ on the Lyft-Level5 dataset. Our

Table 2. Comparison of performance on 3D OCF in different future time horizons

Approach	nuScenes								Lyft-Level5								nuScenes & nuScenes-Occupancy							
	IoU_f (2D)				IoU_f (3D)				IoU_f (2D)				IoU_f (3D)				IoU_f (3D)				C-IoU_f (3D)			
	0.5s	1.0s	1.5s	2.0s	0.5s	1.0s	1.5s	2.0s	0.2s	0.4s	0.6s	0.8s	0.2s	0.4s	0.6s	0.8s	0.5s	1.0s	1.5s	2.0s	0.5s	1.0s	1.5s	2.0s
PowerBEV	32.13	31.31	30.61	30.07	27.17	26.49	25.92	25.47	45.04	44.64	44.05	43.44	35.52	35.18	34.68	34.18	-	-	-	-	-	-	-	-
OpenOccupancy	27.89	27.11	26.46	25.93	24.20	23.79	23.26	22.69	37.24	36.62	35.86	35.11	32.03	31.19	30.45	29.73	14.08	13.71	13.36	13.06	26.62	25.83	25.16	24.68
OccFormer	34.35	32.52	31.60	31.05	31.09	29.49	28.62	28.08	39.77	35.67	33.48	32.10	34.92	31.56	29.74	28.55	12.85	12.07	11.61	11.30	39.35	37.52	36.56	35.98
OCFNet$^-$	29.11	28.11	27.38	26.77	26.31	25.51	24.90	24.40	35.05	34.46	33.80	33.12	30.39	29.97	29.45	28.92	11.65	11.25	10.94	10.65	29.09	28.39	27.88	27.47
OCFNet	34.45	33.65	32.84	32.02	31.40	30.73	30.04	29.36	45.18	44.77	44.27	43.67	39.40	39.02	38.59	38.14	13.64	13.34	12.97	12.57	34.64	34.17	33.66	33.14
EfficientOCF$^-$	30.90	30.16	29.52	29.00	29.74	29.13	28.60	28.17	35.35	34.64	33.93	33.24	32.88	32.43	31.78	31.12	15.07	14.70	14.35	14.05	35.86	35.16	34.61	34.20
EfficientOCF	**38.80**	**37.85**	**36.98**	**36.15**	**34.81**	**34.07**	**33.37**	**32.73**	**46.70**	**46.20**	**45.59**	**44.96**	**43.54**	**43.07**	**42.49**	**41.90**	**20.21**	**19.86**	**19.46**	**19.02**	**47.63**	**46.91**	**46.24**	**45.57**

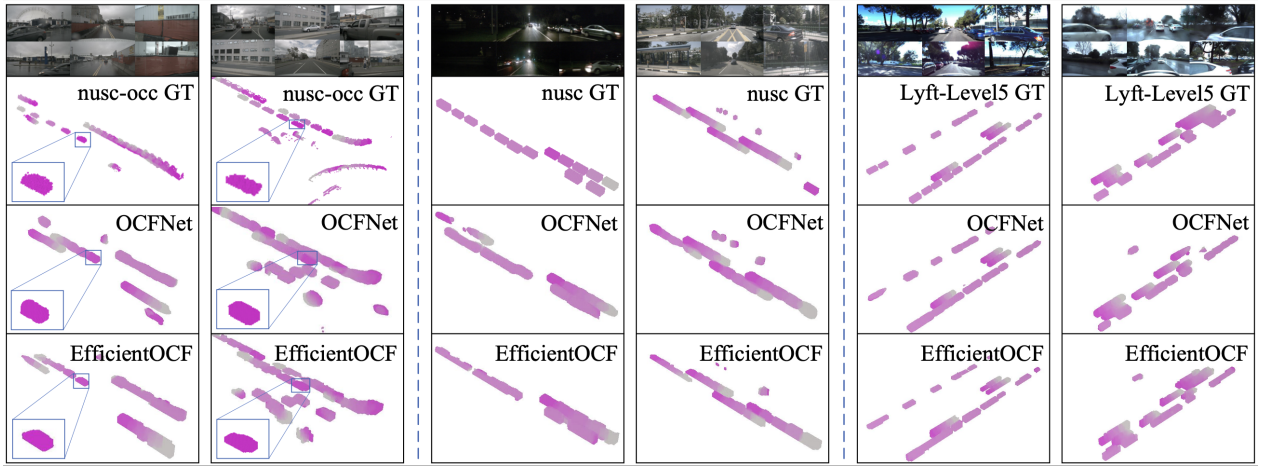

Figure 6. Visualization of OCF results of EfficientOCF and OCFNet, and ground-truth from nuScenes (nusc), nuScenes-Occupancy (nusc-occ), and Lyft-Level5 datasets. The OCF results and ground-truth labels from timesteps 0 to N_f are assigned colors from dark to light.

temporal decoupling leads to more accurate OCF through instance forecasting associated with predicted flow at each timestep. As can be seen in Fig. 6, EfficientOCF captures better contours of movable objects (e.g., the first and the second samples) compared to the SOTA baseline OCFNet which basically overlooks shape details. This is achieved by spatial decoupling, which provides more detailed 3D structures inherent in the combination of BEV occupancy and heights. Moreover, we demonstrate that EfficientOCF can generate more TP and fewer FP predictions in 2D space by Fig. 7, which also contributes to the higher forecasting accuracy of our EfficientOCF. Besides, it can be seen from the nusc-occ ground-truth (GT) of Fig. 6 that the publicly available fine-grained GT annotations are not always accurate due to LiDAR sparsity. This reflects the need for our new metrics C-IoU.

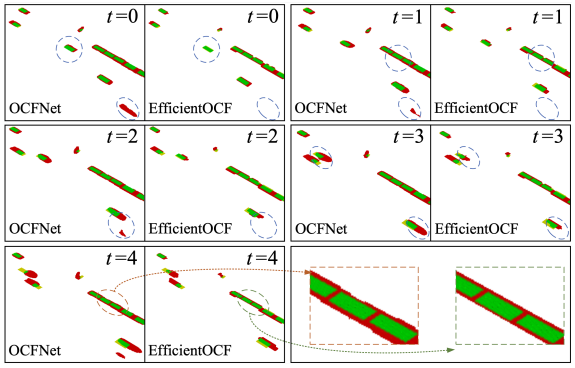

Figure 7. Visualization of TP (green), FP (red), and FN (yellow) results against GT at continuous timesteps.

4.3. Ablation Studies

We conduct several ablation studies to validate the effectiveness of different components in EfficientOCF. All the 3D evaluation metrics are derived by comparing fine-grained predictions with corresponding ground-truth labels.

Spatiotemporal decoupling. We refer the ablation study on spatial decoupling to Tab. 1, where EfficientOCF outperforms OCFNet that directly implements occupancy forecasting in dense 3D representation. We further ablate temporal decoupling by removing the instance-aware refinement in EfficientOCF shown in Tab. 3. We conduct a baseline EfficientOCF‡ that directly lifts the initial 2D OCF results with height values. As shown, our instance-aware temporal decoupling improves the initial OCF performance of the prediction module by absolute 1.64 and 2.55 percentage points in 3D $\widetilde{\text{IoU}}_\text{f}$ and C-$\widetilde{\text{IoU}}_\text{f}$ on the nuScenes-Occupancy dataset. This demonstrates that instance forecasting encompassing flow-based association provides the more accurate occupancy estimation for each decoupled timestep.

Adaptive dual pooling. We compare our adaptive dual pooling (ADP) to different pooling strategies that compress

Table 3. Ablation on temporal decoupling

Approach	2D			3D					
	IoU_c	IoU_f	$I\tilde{o}U_f$	IoU_c	IoU_f	$I\tilde{o}U_f$	C-IoU_c	C-IoU_f	C-$\tilde{I}oU_f$
EfficientOCF‡	39.44	35.89	36.61	19.98	17.31	17.86	45.52	42.90	43.44
EfficientOCF	**39.93**	**36.15**	**36.93**	**21.28**	**19.02**	**19.50**	**47.53**	**45.57**	**45.99**

Table 4. Ablation on adaptive dual pooling

Approach	2D			3D					
	IoU_c	IoU_f	$I\tilde{o}U_f$	IoU_c	IoU_f	$I\tilde{o}U_f$	C-IoU_c	C-IoU_f	C-$\tilde{I}oU_f$
Average Pooling	38.04	34.79	35.45	19.29	16.77	17.29	44.02	41.71	42.18
Max Pooling	38.84	35.53	36.20	19.42	16.90	17.41	44.69	42.33	42.81
ADP	**39.93**	**36.15**	**36.93**	**21.28**	**19.02**	**19.50**	**47.53**	**45.57**	**45.99**

Table 5. Ablation on segmentation/flow/height heads

Seg. Head	Height Head	Flow Head	2D			3D					
			IoU_c	IoU_f	$I\tilde{o}U_f$	IoU_c	IoU_f	$I\tilde{o}U_f$	C-IoU_c	C-IoU_f	C-$\tilde{I}oU_f$
✓			38.48	35.26	35.93	-	-	-	-	-	-
✓	✓		38.55	35.33	35.98	19.41	16.89	17.41	44.09	41.83	42.29
✓	✓	✓	**39.93**	**36.15**	**36.93**	**21.28**	**19.02**	**19.50**	**47.53**	**45.57**	**45.99**

3D features to 2D space. Tab. 4 shows that EfficientOCF with adaptive weighting for the combination of average and max pooling outperforms the baseline with only one single pooling strategy for 3D-2D transformation. The max pooling contributes more to forecasting accuracy than the average pooling, likely because height values and max-pooled features both capture prominent structural changes along the z-axis. We also deploy our ADP strategy as the 3D-2D transformation module in PowerBEV, and it enhances the performance of the vanilla PowerBEV in 2D IoU_c and IoU_f (2.0 s) by absolute 5.16 and 5.42 percentage points. This validates that other OCF methods can also benefit from our proposed ADP strategy.

Multiple heads. We ablate the devised heads in the prediction module of EfficientOCF in Tab. 5. The inclusion of the height head allows the transformation of forecasting results from 2D to 3D space, and provides more structure-aware supervision signals. It thus significantly improves OCF performance on all the metrics. Furthermore, the introduction of the flow head also provides additional enhancements because flow prediction enables the model to be aware of the concrete motion pattern of each voxel grid. Besides, the flow head helps to refine the initial occupancy following our proposed temporal decoupling.

4.4. Instance Prediction Assessment

This experiment presents the 3D instance prediction performance of SOTA OCFNet and our EfficientOCF with the above-mentioned VPQ metric. As shown in Tab. 6, EfficientOCF improves the VPQ of OCFNet forecasting by absolute 1.42 and 4.08 percentage points in VPQ^{bb} and VPQ^{fg} respectively on nuScenes, and by 8.40 percentage points on Lyft-Level5. EfficientOCF^{-} still outperforms OCFNet^{-} for both datasets. EfficientOCF forecasts instances in 2D space which are further lifted to 3D counterparts following our proposed spatiotemporal decoupling. OCFNet directly implements instance association in dense 3D formats, leading to more prediction artifacts and uncertainties.

Table 6. Evaluation on 3D instance prediction

Approach	nuScenes		Lyft-Level5
	VPQ^{bb}	VPQ^{fg}	VPQ^{bb}
OCFNet^{-}	13.68	4.25	24.11
OCFNet	18.53	6.61	28.35
EfficientOCF^{-} (ours)	14.11	6.50	27.06
EfficientOCF (ours)	**19.95**	**10.69**	**36.75**

Table 7. Comparison of inference time

Approach	Inference time [ms]			
	Perception	Processing	Prediction	Total
OCFNet	107.05	**4.18**	13.15	124.38
EfficientOCF (ours)	**64.78**	5.05	**12.50**	**82.33**

4.5. Inference Time

We compare the runtime of OCFNet and our EfficientOCF on 3D OCF shown in Tab. 7. As seen, the processing time of EfficientOCF is slightly higher than OCFNet due to the additional 3D-2D transformation, while EfficientOCF achieves a faster overall inference speed (82.33 ms/12.15 Hz) compared to OCFNet. Moreover, we find that EfficientOCF reduces memory consumption by 478 MB compared to OCFNet. This is because EfficientOCF utilizes lightweight 2D encoders and decoders in the perception module to achieve 3D occupancy forecasting by lifting 2D results to 3D space using height values. In contrast, OCFNet uses heavy 3D encoders and decoders with dense 3D representation, leading to more inference time.

5. Conclusion

This paper proposes spatiotemporal decoupling for efficient vision-based occupancy forecasting (OCF). We achieve spatial decoupling by transforming 3D dense representations into compact 2D BEV occupancy with height values. We also propose using temporal decoupling to refine the end-to-end OCF by step-wise instance forecasting along the time axis. Based on our proposed spatiotemporally decoupled representation, we further develop EfficientOCF, a lightweight network that concurrently forecasts future 2D occupancy states and height values of general movable objects, effectively bridging 2D predictions into the 3D space. It also refines initial OCF results with instance forecasting at each individual timestep. Additionally, we introduce new evaluation metrics providing a comprehensive assessment of OCF performance for datasets with incomplete annotations. Our experiments on extensive evaluation metrics demonstrate that EfficientOCF outperforms existing methods in both accuracy and efficiency, achieving SOTA results in both 2D and 3D spaces. We hope that the spatiotemporal decoupling pipeline proposed in this work can be used as a foundation paradigm in the field of vision-based OCF.

References

[1] Adil Kaan Akan and Fatma Güney. Stretchbev: Stretching future instance prediction spatially and temporally. In *European Conference on Computer Vision*, pages 444–460. Springer, 2022. 1, 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 3, 5, 6

[3] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2

[4] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14403–14412, 2021. 3

[5] Hehe Fan and Yi Yang. Pointrnn: Point recurrent neural network for moving point cloud processing. *arXiv preprint arXiv:1910.08287*, 2019. 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] Yu Hong, Qian Liu, Huayuan Cheng, Danjiao Ma, Hang Dai, Yu Wang, Guangzhi Cao, and Yong Ding. Univision: A unified framework for vision-centric 3d perception. *arXiv preprint arXiv:2401.06994*, 2024. 1

[8] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex 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*, pages 15273–15282, 2021. 1, 2

[9] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 1, 2

[10] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19946–19956, 2024. 1, 2

[11] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Lyft level 5 perception dataset 2020, 2019. 2, 5, 6

[12] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1116–1124, 2023. 1

[13] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020. 5, 6

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Peizheng Li, Shuxiao Ding, Xieyuanli Chen, Niklas Hanselmann, Marius Cordts, and Juergen Gall. Powerbev: a powerful yet lightweight framework for instance prediction in bird's-eye view. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1080–1088, 2023. 1, 3, 4, 6

[16] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2

[17] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 1

[18] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 2

[19] Haochen Liu, Zhiyu Huang, and Chen Lv. Multi-modal hierarchical transformer for occupancy flow field prediction in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1449–1455. IEEE, 2023. 2

[20] Jian Liu, Sipeng Zhang, Chuixin Kong, Wenyuan Zhang, Yuhang Wu, Yikang Ding, Borun Xu, Ruibo Ming, Donglai Wei, and Xianming Liu. Occtransformer: Improving bevformer for 3d camera-only occupancy prediction. *arXiv preprint arXiv:2402.18140*, 2024. 2

[21] Xinhao Liu, Moonjun Gong, Qi Fang, Haoyu Xie, Yiming Li, Hang Zhao, and Chen Feng. Lidar-based 4d occupancy completion and forecasting. *arXiv preprint arXiv:2310.11239*, 2023. 2

[22] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 2

[23] Fan Lu, Guang Chen, Zhijun Li, Lijun Zhang, Yinlong Liu, Sanqing Qu, and Alois Knoll. Monet: Motion-based point cloud prediction network. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13794–13804, 2021. 3

[24] Zhen Luo, Junyi Ma, Zijie Zhou, and Guangming Xiong. Pcpnet: An efficient and semantic-enhanced transformer network for point cloud prediction. *IEEE Robotics and Automation Letters*, 8(7):4267–4274, 2023. 3

[25] Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang.

Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21486–21495, 2024. 1, 2, 3, 4, 5, 6

[26] Reza Mahjourian, Jinkyu Kim, Yuning Chai, Mingxing Tan, Ben Sapp, and Dragomir Anguelov. Occupancy flow fields for motion forecasting in autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):5639–5646, 2022. 2

[27] Benedikt Mersch, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In *Conference on Robot Learning*, pages 1444–1454. PMLR, 2022. 3

[28] Kaustab Pal, Aditya Sharma, Avinash Sharma, and K Madhava Krishna. Atppnet: Attention based temporal point cloud prediction network. *arXiv preprint arXiv:2401.17399*, 2024. 3

[29] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai, and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023. 2

[30] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 3

[31] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[32] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[33] Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*, 2024. 3

[34] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 1, 2, 5, 6

[35] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 2

[36] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11385–11395, 2020. 3

[37] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 5

[38] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 1, 2

[39] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 1, 2, 6

[40] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023. 3