2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

CVFormer: Learning Circum-View Representation and Consistency for Vision-Based Occupancy Prediction via Transformers

Zhengqi Bai[#,1,3], Wenjun Shi[#,1], Dongchen Zhu[1,3], Hanlong Kang[2], Guanghui Zhang[1], Gang Ye[2], Yang Xiao[2], Lei Wang[1,3], Xiaolin Zhang[1,3,4], Bo Li[2], and Jiamao Li[1,3,*]

***Abstract*— With the increasing demands for perception accuracy in autonomous driving, there is a growing focus on fine-grained 3D semantic occupancy prediction. Effectively representing detailed three-dimensional scenes has become a significant challenge in the development of this task. In this paper, we present a novel transformer-based framework named CVFormer, which leverages two-dimensional circum-views from the ego to excavate three-dimensional features of the surrounding environment. Circum-views provide a novel solution for effectively addressing the representation of dense and fine-grained scenes. Specifically, a multi-attention module CTMA is designed for fusing temporal features from circum-views to fully exploit the spatiotemporal correlations between frames and capture more comprehensive clues. Furthermore, a novel 2D projection constraint is established by observing objects from different perspective directions, and multiple 3D constraints based on object invariance and semantic consistency are also conducted for supervising the network, which enhances its performance of understanding the scene. Experimental results on nuScenes dataset demonstrate that the proposed CVFormer obviously outperforms existing methods for occupancy prediction.**

I. INTRODUCTION

With the advancement of autonomous driving technology and the increasing demand for enhanced driving experiences, more comprehensive and precise 3D environmental perception algorithms have become increasingly crucial. Despite the effectiveness of LiDAR-based methods [29]–[33] in accurately determining depth, vision-based approaches [15], [18] are more commonly applied in vehicles due to their lower cost and better robustness. In order to achieve effective perception in all directions around the vehicle, a series of algorithms based on Bird's-Eye View (BEV) have been proposed and applied to tasks such as 3D object detection, semantic segmentation, and motion prediction. However, road scenes are often intricate, and object-centric techniques may struggle in open-world traffic scenarios where the shape or appearance of targets is ambiguous [34]. This is especially evident in cases of shape-shifting long-tail obstacles, irregular obstacles, and obstacles of unknown categories. Hence, this paper focuses on utilizing fine-grained voxels to represent the surrounding environmental information of vehicles, specifically addressing the task of 3D semantic occupancy prediction.

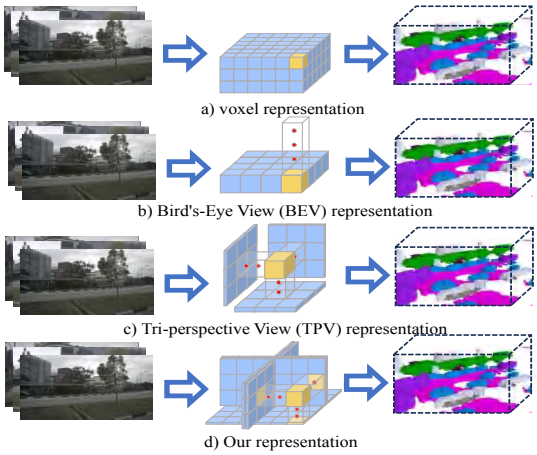

Fig. 1. Typical 3D environment feature representation of different semantic occupancy prediction methods and our approach.

The challenge in the 3D Occupancy Prediction task lies in effectively representing fine-grained 3D road environment information over a large area. The simplest and most straightforward approach involves using 3D features directly as intermediaries to predict the feature for each voxel through a series of operations [18], as shown in Fig.1 (a). However, this method's complexity results in excessive computational demands and poses challenges for deployment. Algorithms like BEVFormer [15] extend the BEV-based approach by incorporating height prediction to generate three-dimensional predictions, as illustrated in Fig.1 (b). Nevertheless, this method lacks optimization for shape-shifting long-tail obstacles and fails to achieve the desired outcomes. The TPVFormer [2] introduces the concept of using three orthogonal perspective views to reconstruct three-dimensional features, as depicted in Fig.1 (c). However, in both the side and front views, the method needs to describe features from different objects of opposite directions on the same coordinate, which leads to inaccuracies in predicting the 3D scene. Therefore, we propose a circum-vehicle surround view scheme, named circum-view, which obtains projections from the vehicle itself in five directions: forward, backward, left, right, and BEV, as shown in Fig.1 (d). This scheme aligns with the camera distribution and perceptual intuition.

#These authors contributed equally to this work.

*National Science and Technology Major Project from Minister of Science and Technology, China(2021ZD0201403), National Natural Science Foundation of China(62103399), Youth Innovation Promotion Association, Chinese Academy of Sciences(2021233,2023242), Shanghai Academic Research Leader(22XD1424500)

[1]Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China. *Corresponding author: Jiamao Li jmli@mail.sim.ac.cn)

[2]Lotus Robotics Ltd., Hangzhou 310056, China.

[3]University of Chinese Academy of Sciences, Beijing 100049, China.

[4]ShanghaiTech University, Shanghai 201210, China.

This approach overcomes the issues present in Fig.1 (c) and improves algorithm accuracy without significantly increasing computational cost. Additionally, because each view aligns with the camera distribution, it effectively reduces retrieval costs compared to other methods.

In addition, we noticed that many methods regard 3D semantic occupancy prediction as a voxel-level fine-grained segmentation task, and it inevitably leads to some outliers. The fact is that the semantic labels of the internal or surface voxels of the same object in the 3D space of a driving scene should be the same unchanged, and the geometric shape or volume composed of these voxels should also remain unchanged and show continuity in space (the moving object only changes its position over time). To mitigate the risks posed by outliers during driving, we introduce a consistency constraint based on object invariance. This module utilizes the features of surrounding voxels to influence the classification of the current point's category. Furthermore, to guide the view in acquiring semantic and depth features in 3D space, we utilize the characteristics of the circum-views to design a 2D projection constraint based on perspective consistency.

In summary, our main contributions are as follows:

- We pull out a circum-view representation for describing the 3D scene efficiently and construct a multi-attention module to fuse features of long-term range and wide-space perspective frames. It successfully generalizes object-centric perception to fine-grained 3D voxel perception, avoiding redundant computations for dense voxel information.
- We develop a 3D semantic perception framework, named CVFormer. It integrates a novel 2D constraint and multiple 3D constraints based on perspective consistency and object invariance, improving the prediction accuracy by observing objects from different directions.
- We evaluate the proposed CVFormer on nuScenes benchmark for the 3D semantic occupancy prediction task. Under comparable parameters and computational overhead, CVFormer outperforms existing techniques, achieving a mIoU of 43.09%, particularly excelling at small moving objects that are challenging to observe.

II. Related work

A. 3D Semantic Occupancy Prediction

In recent years, with the advancement of the autonomous driving field, there has been an increasing demand for higher perception accuracy and comprehensiveness, leading to a shift towards multi-view perception. As a result, several solutions [15]–[28] based on Bird's Eye View (BEV) perception have been proposed. These solutions include those utilizing the Lift-Splat-Shoot (LSS) [24] technique for forward inference [18]–[22] and others employing Transformers [25] for reverse inference [15]–[17], [28]. This series of approaches typically revolves around object-centric tasks such as 3D object detection. However, there is a continuous increase in the requirements for perception accuracy and granularity.

For a more accurate perception of the surrounding environment and safer autonomous driving, occupancy prediction was introduced as a solution and received increasing attention. It involves partitioning the 3D space into voxels and describing environmental information by classifying the occupancy probability of each voxel. This subsequently paved the way for the development of various algorithms for the occupancy prediction task [3]–[11]. MonoScene [1] introduced the first monocular method for completing semantic scenes. This method employs a 3D UNet [42] to process voxel features. Occformer [4], on the other hand, handles 3D features through a multi-scale approach.

The most significant challenge in the occupancy prediction task lies in efficiently describing the 3D scene. Given the substantial number of empty voxels present in the scene, EsscNet [12] proposed combining non-empty voxels before performing 3D convolutions. DDRNet [13] and AIC-Net [14] altered the convolution approach to reduce algorithmic complexity. BevFormer [15] employs BEV features to predict height and describe the 3D space. TPVFormer [2] introduced a three-view representation to depict the 3D scene. While the three-plane representation simplifies computation, it has not achieved optimal performance.

B. Temporal Modeling for Multi-view Perception

In camera-only BEV-based methods, temporal information is often utilized to optimize motion and occlusion information in autonomous driving scenes. Temporal features can effectively enhance the detection performance of 3D perception algorithms [26]. Methods like BEVDet [19] introduce temporal modeling, while BEVFormer [27] proposes the utilization of attention mechanisms to integrate previous BEV features. Algorithms like SoloFusion [27] and Hop have also explored fusion through various approaches. Similarly, in lidar-based 3D perception, many temporal fusion approaches have been proposed, incorporating methods such as RNN [36] and LSTM [35] into the framework. However, in the context of the occupancy prediction task, there has been relatively less exploration of utilizing temporal information. The FB-occ [7] method attempts to leverage temporal information to smooth the predicted results.

III. Method

To perform fine-grained occupancy prediction of the scene, we propose a transformer-based circum-view perception scheme called CVFormer. Fig.2 illustrates the overall algorithmic pipeline. After extracting image features, we utilize a Circum-view and Temporal-view Multi-Attention module (abbreviated as CTMA) to construct and fuse features of spatial-temporal views to obtain fine-grained 3D information. Subsequently, the multiple spatial and semantic consistency constraints on object invariance are employed to enhance contextual features across voxels, ultimately resulting in the prediction of a 3D voxel grid with classification labels. Follow-up to introduce their implementation details.

A. Circum-view and Temporal-view Multi-Attention

Circum-View. Taking full account of the surface information observation of 3D space by vision, we first construct

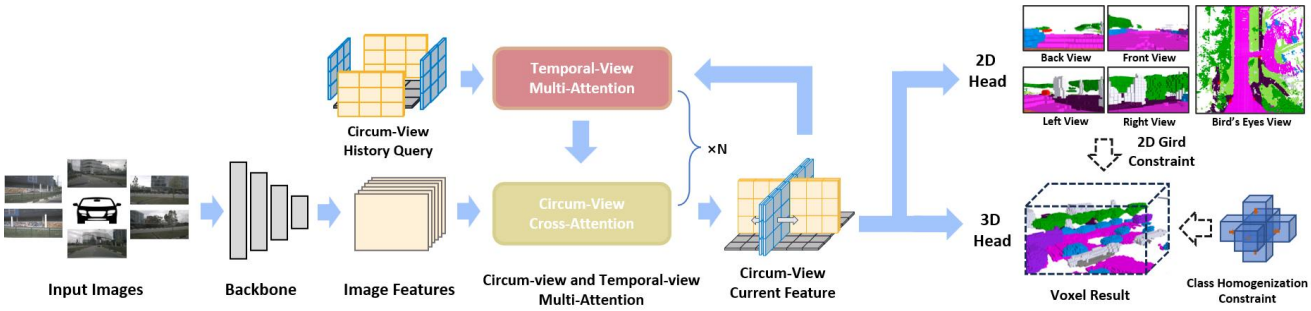

Fig. 2. CVFormer first extracts image features using a pre-trained backbone. Then, it generates 2D circum-view features with historical information through multiple layers of Circum-view and Temporal-view Multi-Attention. Subsequently, it obtains 3D voxel features using a Decoder, which is supervised by 2D grid constraint and the class homogenization constraint.

a concept of circum-view. It consists of five projection view planes originating from the vehicle's coordinate system and extending forward, backward, left, right, and in the BEV direction. These projections include visible surface image features. The circum-view aligns with sensor configuration schemes and adheres to common visual perception practices. Each view v corresponds to a certain number of camera images I_i, which are associated into a sequence S using the correspondence relation f_{rex}. The Query $Q_{h,w}$ directly within the sequence of images formed by these projections significantly reduces the computational complexity:

$$S = f_{rex}(I_i, V^v), \tag{1}$$

where $i = 1, \ldots, k$, and k depends on the field of observation corresponding to the view. This approach, while simplifying representation, effectively captures 3D features, especially those of objects near the vehicle. It is more efficient than methods that attempt to recover the features of all voxels from several images, including occluded voxels. It is more accurate than TPVFormer [2], which ignores the difference between the left and right sides of the ego as well as the front and rear.

Based on the circum-view, the multi-attention module CTMA is implemented by two modules: the Circum-View Cross-Attention (CVCA) module and the Temporal-View Multi-Attention (TVMA) module. The former fuses the perceived six image features through cross-attention structures to obtain circum-view feature slices. The latter combines historical frame feature information with the current frame to enhance the temporal features in pursuit of higher accuracy and robustness to occlusions.

Circum-View Cross-Attention. Specifically, We construct 3D pillar regions orthogonal to the view plane at the scene scale and generate feature sampling points in them. We then build a circum-view query in the corresponding view and refine the data using the camera image. After that, we add learnable positional embeddings to circum-view query $Q_{h,w}$. For efficiency, we use deformable attention $\mathcal{DA}$ to achieve cross-attention between images [25]. The construction of the reference point and F_{CVCA} can refer to the formula:

$$F_{CVCA}\left(Q_{h,w}, I\right) = \frac{1}{|S|} \sum_{pos}^{S} \mathcal{DA}\left(Q_{h,w}, Ref_{h,w}^{pos}, I_{pos}\right), \tag{2}$$

where $Ref_{h,w}^{S} = ((h - \frac{H}{2}) \times s, (w - \frac{W}{2}) \times s, z)$, it represents the corresponding reference point in the (h, w) position within the sequence S. We sample and map the image features by projecting them into pixel coordinates. Each view mapping uses a 2D grid representation. Each grid cell corresponds to the direction information of the region from the coordinate center of the ego car.

Temporal-View Multi-Attention. Building upon the circum-view features obtained through CVCA, we further introduce the Temporal-View Multi-Attention(TVMA) to fully exploit the inter-frame correlations and historical information. Fig.3 shows its architecture in detail.

Since the features from different historical frames are in different coordinate systems, it's necessary to unify them into one coordinate system. To align the features, we transform the current frame from ego coordinate $Loc_{ego}^{(v,t)}$ to the global coordinate $Loc_{global}^{(v,t)}$ through the matrix $T_{global}^{ego(t,v)}$ in view v at time t:

$$Loc_{global}^{(v,t)} = T_{global}^{ego(t,v)} Loc_{ego}^{(v,t)} - T_{t-1}^{t} T_{ego(t-1,v)}^{global} Loc_{ego}^{(v,t-1)}. \tag{3}$$

After alignment, we use deformable attention [25] $\mathcal{DA}$ to construct temporal-attention $\mathcal{TA}$, utilizing the preceding circum-view feature $V_{(t-1)}^{v}$ as the value which is from the view v on time t . This process yields and stores view features for each time step. And recursing from time $t-n$ to the current time t in this way, we ultimately obtain the view feature V_t^v for the current time with temporal information. This process can be represented as:

$$V_t^v = \mathcal{TA}(Q_t^v, V_{t-1}^v) = \mathcal{DA}(Q_t^v, V_{t-1}^v, Loc_{global}^{(v,t)}), \tag{4}$$

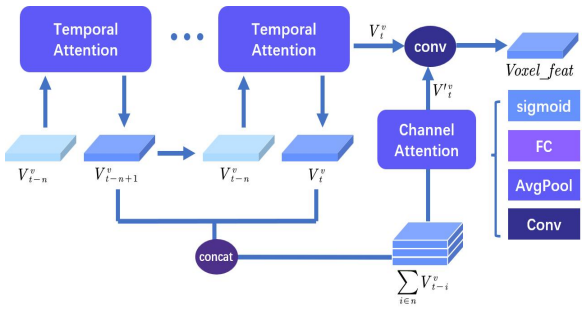

Fig. 3. The architecture of Temporal-View Multi-Attention module.

where Q_t^v represents the query at $Loc_{global}^{(v,t)}$. At the same time, we stack the saved temporal features V_{t-i}^v to increase the feature density along the view. Then, we construct a channel attention $\mathcal{CA}$ to filter out useful geometric and semantic features V'^v_t from historical information. Finally, we fuse V'^v_t with V_t^v those obtained through $\mathcal{TA}$, resulting in fine-grained circum-view features $Voxel\_feats$ enhanced with historical information.

$$Voxel\_feat = Conv(\mathcal{CA}(\sum_{i \in n} V_{t-i}^v) + V_t^v). \tag{5}$$

B. Consistency Constraints with Object Immutability

For the 3D occupancy prediction task in autonomous driving applications, objects in the scene can basically be regarded as rigid objects undergoing rigid motion, relative to the perceptual granularity of voxels. Therefore, the geometry of the scene is fixed at the current moment, and objects show continuity in space and are not isolated. Here, we use these properties to construct more constraints between the occupancy output and the semantic labels of each 3D voxel as well as the projection 2D prediction between the corresponding ground truth, driving the model to perceive scene semantics and geometric information better.

Projected perspective grids have consistent categories. Currently, occupancy perception methods focus on the constraints between predictions and ground truth in a unified 3D space, including both geometric and semantic aspects. However, image features from the six views cannot fully correspond to all locations in 3D space. The prediction of many voxels does not have good enough feature support, so the 3D supervision is difficult for these occluded voxels, and may even cause the network to fall into a deadlock. Therefore, we add a 2D head branch to the framework based on the circum-view projection and introduce the 2D grid constraints to enhance supervision.

This 2D constraint is implemented in a similar way as TSDF (Truncated Signed Distance Function) [40]. We obtain the 2D prediction label maps Sur^{pre} corresponding to five different views through projection, which represent the semantic labels of the 3D surface viewed from circum-views:

$$Sur_{v_{front}}^{pre} = argmax(Pre_{bi}[H/2 : H-1, :, :], 1), \tag{6}$$

$$Sur_{v_{back}}^{pre} = argmax(Pre_{bi}[0 : H/2-1, :, :], 1), \tag{7}$$

$$Sur_{v_{left}}^{pre} = argmax(Pre_{bi}[:, 0 : W/2-1, :], 0), \tag{8}$$

$$Sur_{v_{right}}^{pre} = argmax(Pre_{bi}[:, W/2 : W-1, :], 0), \tag{9}$$

$$Sur_{v_{bev}}^{pre} = argmax(Pre_{bi}[:, :, 0 : Z-1], 2), \tag{10}$$

where Pre_{bi} is the binarization of the output prediction Pre whose dimension is $W \times H \times Z$. $Pre_{bi}(x, y, z) = 0$ when its label is $free$, and otherwise $Pre_{bi}(x, y, z) = 1$. In the same way, the 2D ground truth maps Sur_v^{tar} can be constructed.

By projecting the 3D prediction onto different view planes, the loss function of the 2D head branch is calculated as:

$$\begin{aligned} Loss_{2d} = -\sum [&Sur_v^{tar} \log Sur_v^{pre} \\ &+ (1 - Sur_v^{tar}) \log(1 - Sur_v^{pre})], \end{aligned} \tag{11}$$

where v belongs to the five circum-views.

Connected voxels have Consistent categories. Using the influence of surrounding voxels on the processed current voxel for semantic labeling, we can solve the problem of outliers in the predicted 3D results. Specifically, we proposed a class homogenization constraint branch after circum-view feature fusion to adjust the feature of the current voxel using the class prediction probabilities of surrounding voxels because the category within an object should be the same. This branch has two functions. One is to output a 3D occupancy result and calculate the spatial continuity of each voxel based on the result for constructing the loss. The other is to selectively integrate the learned high-dimensional features into the 3D label prediction head to enhance the final semantic occupancy estimation. For the latter, we filter out features that favor the prediction of category continuity using channel attention.

As for the former spatial continuity loss, it is assumed that two adjacent voxels are connected if they have the same category label. To simplify the calculation, we first use the neighboring elements of the six faces of the voxel to approximate. Within the object's scope, the category of a voxel is consistent with other voxels in the 6-connected direction. Then, we predict the number of connected voxels C_{num} with the same category as the current voxel ν_i:

$$C_{num}(\nu_i, d) = Count(\nu_i, d), \tag{12}$$

where $d = 1, 2, ..., 6$ represents the specified direction and $Count$ is the connectivity prediction function. To ensure efficient execution speed, this function is implemented using CUDA code programming. Finally, we build the constraint between C_{num}^{pre} of the predicted result and C_{num}^{tar} of the ground truth via smoothing $L1$:

$$Loss_{sc} = \sum smooth_{L_1}(C_{num}^{tar}(\nu_i, d) - C_{num}^{pre}(\nu_i, d)), \tag{13}$$

Through this constraint, the network further learns the spatial semantic context information of voxels, effectively expanding the receptive field and addressing the issue of outlier voxels in the prediction.

In addition, our 3D head not only has the semantic label cross-entropy loss but also introduces semantic and geometric affinity losses according to MonoScene [1]. Therefore, the entire loss function can be expressed as:

$$\begin{aligned} Loss_{all} = Loss_{ce} + \alpha Loss_{sca}^{geo} + \beta Loss_{sca}^{sem} \\ + \gamma Loss_{2d} + \delta Loss_{sc}, \end{aligned} \tag{14}$$

where α, β, γ, δ are hyperparameters. In experiments, we found that their small changes have little impact on the final results of the model. Therefore, we follow the principle of controlling all loss values to be of the same magnitude and set them as: $\alpha = 0.2$, $\beta = 0.2$, $\gamma = 1$, $\delta = 3$.

IV. Experiments

In this section, our experimental settings and the performances on occupancy prediction are presented to demonstrate the effectiveness of the proposed framework. We conducted comprehensive experiments on the Occ3D-nuScenes

TABLE I

OCCUPANCY PREDICTION RESULTS ON NUSCENES VAL SET. * REPRESENTS METHOD WITH OTHER BACKBONE

Method	backbone	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	dirve. suf.	other flat	sidewalk	terrain	manmade	vegetation
BEVFormer [15]	ResNet101	26.88	5.85	37.83	17.87	40.44	42.43	7.36	23.88	21.88	20.98	22.38	30.70	55.35	28.36	36.00	28.06	20.04	17.69
TPVFormer [2]	ResNet101	21.29	4.67	19.24	7.54	31.93	31.28	11.01	10.08	10.16	6.43	10.72	23.51	59.14	32.1	31.51	26.80	11.96	17.17
CTF-Occ [5]	ResNet101	28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
BEVDet* [19]	ResNet50	36.01	8.22	44.21	10.34	42.08	49.63	23.37	17.41	21.49	19.70	31.33	37.09	80.13	37.37	50.41	54.29	45.56	39.59
BEVDet [19]	SwinTrans	42.02	12.15	49.63	25.10	**52.02**	54.46	27.87	27.99	28.94	27.23	36.43	42.22	82.31	43.29	54.62	57.9	**48.61**	**43.55**
PanoOcc [6]	ResNet101	42.13	11.67	50.48	29.64	49.44	**55.52**	23.29	**33.26**	30.55	30.99	34.43	42.57	83.31	44.23	54.4	56.04	45.94	40.40
FB-Occ [7]	ResNet50	42.06	**14.30**	49.71	**30.00**	46.62	51.54	**29.3**	29.13	29.35	30.48	34.97	39.36	**83.07**	**47.16**	**55.62**	**59.88**	44.89	39.58
ours	ResNet101	**43.09**	14.01	**50.89**	29.82	50.94	54.57	28.63	32.70	**31.5**	**32.72**	**37.36**	**43.17**	81.84	43.94	53.02	57.78	47.42	42.25

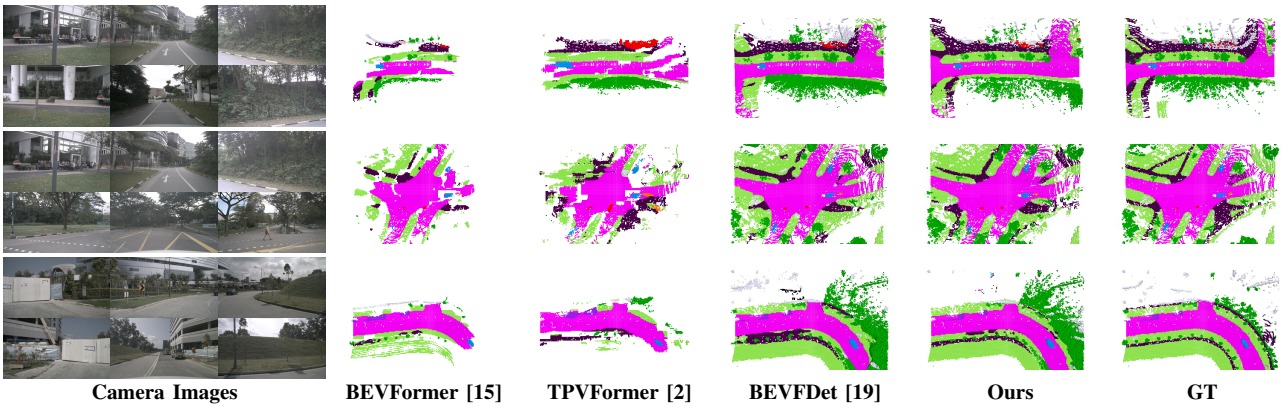

Fig. 4. Qualitative results on the dense annotated Occ3D-nuScenes validation set for 3D semantic occupancy prediction. With circum-view representation and multiple constraints, our CVFormer can predict better and denser occupancy.

[5] dataset, which generates dense annotations for 3D semantic occupancy prediction from nuScenes [39].

A. Implementation Details

Our algorithm uses ResNet101 [38] as the backbone network, which is commonly used in this task. Then using FPN [41] as neck, multi-scale features are obtained, which are 1/16, 1/32 and 1/64 of the original feature size respectively. For 3D semantic occupancy prediction, the default size of queries is 200×200×16 and the number of sampled points is 4, following the standard baseline BEVFormer-occ [15]. The perception ranges for the X and Y axes are [-40m, 40m], and for the Z axis, it is [-1m, 5.4m]. The resolution of the voxel grid is set to 0.4m. Our models are trained with AdamW optimizer, in which gradient clip is exploited with learning rate 4×10^{-4}, a total batch size of 8. For the following experimental results of our models, the maximum number of epochs for training is set to 24.

B. 3D Semantic Occupancy Prediction

We compare our CVFormer with other algorithms for the occupancy prediction task on the nuScene validation set. As shown in TableI, compared to TPVFormer [2] and BEVFormer [15], which serve as our baselines and has no temporal feature fusion, our method shows significant performance improvements (as $21.29\%, 26.88\%$ *vs.* 43.09%), and outperform the PanoOcc [6] and VoxFormer [3] (with mIoU 40.7% reported in OccTransformer [11]) with the same backbone ResNet101. At the same time, our model performs better than BEVDet-occ [19] and SurroundOcc [10](with mIoU 40.7% reported in OccTransformer [11]) even if it uses a more sufficient backbone. It also achieves better accuracy than the current SOTA method FB-occ [7] on ResNet50, even though FB-occ has been trained with a range of larger models and pre-trained networks. Comparing the results of different categories, our algorithm has advantages in locating and identifying small moving objects such as pedestrian, traffic cone and trailer. It can also achieve slightly better or more competitive results than the SOTA method in other categories. This indicates that our effective utilization of temporal information and object surface consistency significantly enhances the accuracy of sparsely annotated objects, particularly those around vehicles, which are critical elements in driving scenarios.

In Fig.4, we visualize the results of TPVFormer [2], BEVFormer [15], BEVDet-occ [19], and our model on 3D semantic occupancy prediction. Our method produces denser and more accurate results compared to TPVFormer and BEVFormer, especially for distant objects. Additionally, compared with BEVDet using dense 3D features, our method provides smoother predictions, and multiple emphasis on circum-views solves the problem of missed and false detection of small-sized targets to a certain extent.

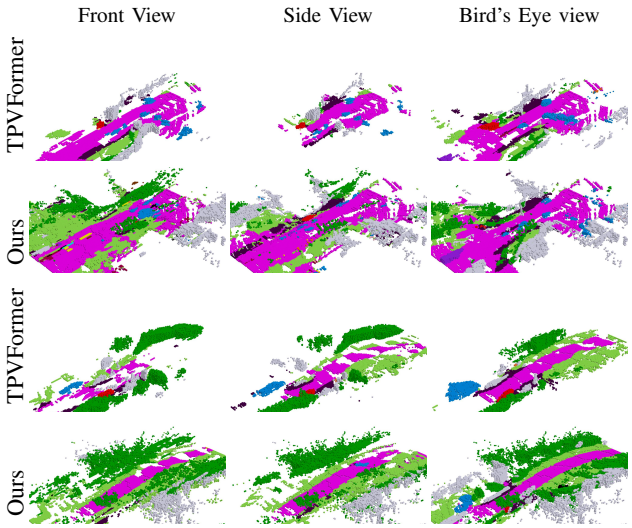

Fig. 5. Qualitative comparison results of TPVFormer and our method from different views perception.

C. Ablation Studies

To delve into the effect of different modules, we conduct numerous ablation experiments here.

Effectiveness of the proposed components. We first study the improvement due to the various components that we propose in our CVFormer architecture. Results from this experiment are shown in Table II. We adopt TPVFormer [2] as the original setting at baseline and reproduce according to the TPVFormer using our CTMA consisting of circum-view and temporal-view attention modules. The results in the second and third rows illustrate that the proposed representation achieves a larger improvement compared to the baseline regardless of whether the temporal fusion module has channel attention or not. Of course, the result is better with "ac", which can selectively fuse the temporal feature series The fourth row shows the improvement brought by multiple constraints from our 2D and 3D head when the reference frame sequence length is 0. Finally, we incorporate all proposed modules into the model in the last row that fuses features and learns consistency effectively which enables it to obtain the best mIoU 43.09%.

Effectiveness of circum-view representation. To observe the effect of the circum-view representation with different perspective supervision, we visualize the semantic prediction of different views. We combined predictions from the left and right views as well as the front and rear views, aligning our approach with TPVFormer's [2] scheme, and then compared

TABLE II

ABLATION STUDY ON THE PROPOSED NETWORK. "CA" REPRESENTS THE CHANNEL ATTENTION IN MODULE TVMA.

methods	Multi-Attention	Multi-Constraints	mIoU
Baseline			21.29
+CTMA	✓(w/o ca)		38.28
+CTMA	✓(w/ ca)		40.16
+L_{2d}&L_{sc}		✓	35.93
All	✓(w/ ca)	✓	43.09

TABLE III

THE EFFECTIVENESS OF THE GLANCING PERSPECTIVE.

methods	Front View	Side View	Bird's Eye View	Final Result
TPVFormer [2]	8.36	11.78	19.88	21.29
ours(w/o temporal)	23.89	25.96	33.21	35.93
ours(w/ temporal)	28.17	31.28	38.47	43.09

the results. Qualitative and quantitative results are shown in Table III and Fig. 5. The parameters for different models are adjusted to be the same except for the view. The first three columns in Table III illustrate an interesting phenomenon, that is, the results of the front and side view are far worse than the results of BEV. This is related to the characteristics of the driving task environment, and once again demonstrates the importance of BEV perception. Among the three views of TPVFormer, the perception accuracy of the front view and the side view are almost only half of that of BEV, and the final 3D occupancy result combined with three views is improved by less than 3% compared with the result of BEV. However, our circum-view representation and the corresponding consistency constraint in the 2D head have increased the mIoU of the front view by nearly three times and the mIoU of the side view by more than twice, greatly shortening the gap between them and that of BEV. Moreover, with the support of temporal features, the final result integrated of the three views is nearly 5% higher than the result of the corresponding BEV. It is also obvious from Fig.5 that the three views' results of our CVFormer are significantly better than the results of TPVFormer, whether it belongs to the small moving targets or the large static background.

V. CONCLUSIONS

In this paper, we presented our CVFormer architecture for 3D semantic occupancy prediction that achieves state-of-the-art performance among equivalent models. Different from the previous works, our method explores a circum-view representation that aligns with sensor distribution and perceptual habits, efficiently describing the 3D environment. First, we introduced the multi-attention module based on circum-view and temporal-view, which adequately leverages the temporal and spatial correlations between frames to meet the requirements of fine-grained predictions. Further, we investigated the continuity of the semantic target in 3D space and the consistency of observation from different views and proposed multiple constraints to supervise the network in training through our 2D head as well as the 3D head. Qualitative and quantitative results on the nuScenes dataset demonstrate the effectiveness of CVFormer on semantic perception from 3D voxel occupancy and 2D view projection.

REFERENCES

[1] Cao, Anh-Quan, and Raoul de Charette. "Monoscene: Monocular 3d semantic scene completion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Huang, Yuanhui, et al. "Tri-perspective view for vision-based 3d semantic occupancy prediction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[3] Li, Yiming, et al. "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[4] Zhang, Yunpeng, Zheng Zhu, and Dalong Du. "OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction." arXiv preprint arXiv:2304.05316 (2023).

[5] Tian, Xiaoyu, et al. "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving." arXiv preprint arXiv:2304.14365 (2023).

[6] Wang, Yuqi, et al. "PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentation." arXiv preprint arXiv:2306.10013 (2023).

[7] Li, Zhiqi, et al. "FB-OCC: 3D Occupancy Prediction based on Forward-Backward View Transformation." arXiv preprint arXiv:2307.01492 (2023).

[8] Miao, Ruihang, et al. "Occdepth: A depth-aware method for 3d semantic scene completion." arXiv preprint arXiv:2302.13540 (2023).

[9] Wang, Xiaofeng, et al. "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception." arXiv preprint arXiv:2303.03991 (2023).

[10] Wei, Yi, et al. "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving." arXiv preprint arXiv:2303.09551 (2023).

[11] Liu, Jian, et al. "OccTransformer: Improving BEVFormer for 3D camera-only occupancy prediction."

[12] Zhang, Jiahui, et al. "Efficient semantic scene completion network with spatial group convolution." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[13] Li, Jie, et al. "Rgbd based dimensional decomposition residual network for 3d semantic scene completion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[14] Li, Jie, et al. "Anisotropic convolutional networks for 3d semantic scene completion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[15] Li, Zhiqi, et al. "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[16] Zhang, Tianyuan, et al. "Mutr3d: A multi-camera tracking framework via 3d-to-2d queries." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[17] Yang, Chenyu, et al. "BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[18] Huang, Junjie, et al. "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view." arXiv preprint arXiv:2112.11790 (2021).

[19] Huang, Junjie, and Guan Huang. "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection." arXiv preprint arXiv:2203.17054 (2022).

[20] Huang, Junjie, and Guan Huang. "Bevpoolv2: A cutting-edge implementation of bevdet toward deployment." arXiv preprint arXiv:2211.17111 (2022).

[21] Li, Yinhao, et al. "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 2. 2023.

[22] Zhang, Yunpeng, et al. "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving." arXiv preprint arXiv:2205.09743 (2022).

[23] Liu, Yingfei, et al. "Petr: Position embedding transformation for multi-view 3d object detection." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

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

[25] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." arXiv preprint arXiv:2010.04159 (2020).

[26] Li, Yinhao, et al. "Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo." arXiv preprint arXiv:2209.10248 (2022).

[27] Park, Jinhyung, et al. "Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection." arXiv preprint arXiv:2210.02443 (2022).

[28] Liu, Yingfei, et al. "Petrv2: A unified framework for 3d perception from multi-camera images." arXiv preprint arXiv:2206.01256 (2022).

[29] Lang, Alex H., et al. "Pointpillars: Fast encoders for object detection from point clouds." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[30] Shi, Shaoshuai, Xiaogang Wang, and Hongsheng Li. "Pointrcnn: 3d object proposal generation and detection from point cloud." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[31] Zhou, Yin, and Oncel Tuzel. "Voxelnet: End-to-end learning for point cloud based 3d object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[32] Yan, Yan, Yuxing Mao, and Bo Li. "Second: Sparsely embedded convolutional detection." Sensors 18.10 (2018): 3337.

[33] Deng, Jiajun, et al. "Voxel r-cnn: Towards high performance voxel-based 3d object detection." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 2. 2021.

[34] Shi, Yining, et al. "Grid-centric traffic scenario perception for autonomous driving: A comprehensive review." arXiv preprint arXiv:2303.01212 (2023).

[35] Marinello, Nicola, Marc Proesmans, and Luc Van Gool. "TripletTrack: 3D object tracking using triplet embeddings and LSTM." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[36] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[37] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[38] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[39] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[40] Werner, Diana, Ayoub Al-Hamadi, and Philipp Werner. "Truncated signed distance function: experiments on voxel size." Image Analysis and Recognition: 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part II 11. Springer International Publishing, 2014.

[41] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[42] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.