

HybridOcc: NeRF Enhanced Transformer-based Multi-Camera 3D Occupancy Prediction

Xiao Zhao , Bo Chen, Mingyang Sun, Dingkang Yang, Youxing Wang, Xukun Zhang, Mingcheng Li, Dongliang Kou, Xiaoyi Wei, and Lihua Zhang , Member, IEEE

Abstract—Vision-based 3D semantic scene completion (SSC) describes autonomous driving scenes through 3D volume representations. However, the occlusion of invisible voxels by scene surfaces poses challenges to current SSC methods in hallucinating refined 3D geometry. This paper proposes HybridOcc, a hybrid 3D volume query proposal method generated by Transformer framework and NeRF representation and refined in a coarse-to-fine SSC prediction framework. HybridOcc aggregates contextual features through the Transformer paradigm based on hybrid query proposals while combining it with NeRF representation to obtain depth supervision. The Transformer branch contains multiple scales and uses spatial cross-attention for 2D to 3D transformation. The newly designed NeRF branch implicitly infers scene occupancy through volume rendering, including visible and invisible voxels, and explicitly captures scene depth rather than generating RGB color. Furthermore, we present an innovative occupancy-aware ray sampling method to orient the SSC task instead of focusing on the scene surface, further improving the overall performance. Extensive experiments on nuScenes and SemanticKITTI datasets demonstrate the effectiveness of our HybridOcc on the SSC task.

Index Terms—computer vision, autonomous driving, neural networks, semantic scene completion, 3D occupancy.

I. INTRODUCTION

CAMERA-BASED 3D scene understanding is a crucial component of the autonomous driving perception system. It involves acquiring accurate and comprehensive real-world 3D information, even with slight movement of the vehicle. In recent years, multi-camera systems have produced competitive results with Lidar in tasks such as depth estimation [1], [2] and 3D object detection [3], [4]. Semantic scene completion (SSC) [2], [5]–[9] has recently gained more attention than

Manuscript received: February 4, 2024; Revised: April 8, 2024; Accepted: June 7, 2024. This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by the National Key R&D Program of China under No.2021ZD0113503, Shanghai Municipal Science and Technology Major Project under No.2021SHZDZX0103. (Corresponding author: Lihua Zhang.)

Xiao Zhao, Mingyang Sun, Dingkang Yang, Youxing Wang, Xukun Zhang, Mingcheng Li, Dongliang Kou, and Xiaoyi Wei are with the Academy for Engineering and Technology, Fudan University, Shanghai 200000 China (e-mail: zhaox21@m.fudan.edu.cn)

Bo Chen and Youxing Wang are with the China FAW Group Corp., Ltd., Nanjing 211100, China.

Lihua Zhang is with the Academy for Engineering and Technology, Fudan University, Shanghai 200000 China, the Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, 200000 China, and also with the Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, 130000, China (e-mail: lihuazhang@fudan.edu.cn).

Digital Object Identifier (DOI): see top of this page.

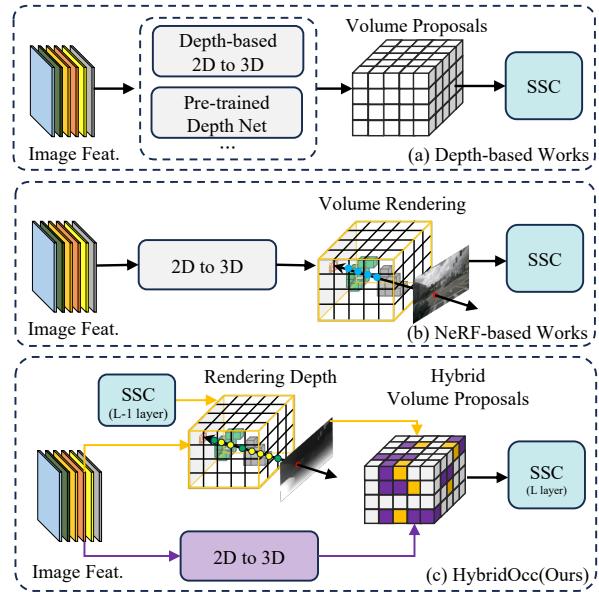


Fig. 1. Comparison of depth-based methods, NeRF-based and our HybridOcc. (a) Advanced methods such as FB-Occ [9] and VoxFormer [7] require additional depth prediction networks for generating 3D voxels. (b) NeRF-based methods [10], [11] only focus on the visible surface (blue sampling points) of the scene and render based on the transformed 3D voxel features. (c) In HybridOcc, the NeRF branch combined with the Transformer branch gradually refines SSC from coarse-to-fine. We propose a 3D occupancy-aware ray sampling (yellow sampling points) to enable the model to focus on occupied voxels of all scenes rather than visible surfaces.

3D object detection. SSC is more appropriate for autonomous driving downstream tasks as it can represent scenes of arbitrary shapes and categories. However, inferring the comprehensive semantic scene from limited observation views is challenging.

MonoScene [5] proposed directly lifting 2D images to 3D voxels through feature projection for the SSC task. Recently, some works [2], [8], [12] proposed to lift multi-view camera features to 3D representation based on spatial cross-attention [4]. In the coarse-to-fine framework proposed by Occ3D [8], the performance is limited by the lack of depth signals. Other studies [6], [7], [9] adopted additional depth estimation modules to improve the quality of 3D voxel representation, as shown in Fig. 1(a). FB-Occ [9] used a pre-trained depth prediction model and a depth-aware back-projection model to assist in generating 3D voxel features. However, most depth-based methods focus on the visible surface of the scene and lack inference of occluded regions. VoxFormer

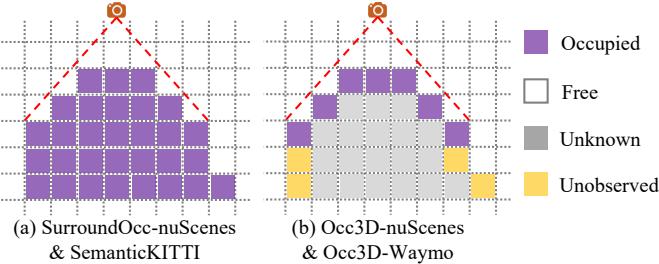


Fig. 2. Illustration of the 3D occupancy prediction data set. SurroundOcc-nuScenes [2] and SemanticKITTI [14] evaluate all occupied voxels, while Occ3D-nuScenes and Occ3D-Waymo [8] only evaluate visible surfaces.

[7] proposed an additional masked autoencoder-based module [13] to consider the occluded voxels, but its cumbersome two-stage structure is not conducive to end-to-end model training. Various current methods have shown the importance of depth signals to the SSC task. Notably, there are currently two types of 3D occupancy datasets for autonomous driving with different functions. One is only evaluating the visible surface (Fig. 2(b)) [8], while the other is for the complete occupation of the scene, that is, the SSC task (Fig. 2(a)) [2], [14]. This paper focuses more on the SSC task, which takes into account occluded objects or regions. Current SSC works [2], [6], [7] mostly suffer from occlusion, making per-voxel features contain many ambiguities. Consequently, the occupancy prediction of occluded voxels still faces challenges.

The introduction of neural radiance fields (NeRFs) [15], [16] greatly improved the 3D scene reconstruction performance. SceneRF [16] designed a probabilistic ray sampling method for radiance field and applied it to the 3D reconstruction of autonomous driving scenes. Recently, some methods [10], [11], [17] utilized the lifted 3D voxel features for depth and color rendering. Since NeRF-based 3D reconstruction methods focus on the visible surface of the scene, as shown in Fig. 1(b), the SSC task requires extra attention to voxel features in invisible regions. Therefore, a rough and direct application of the NeRF model on the SSC task may not be conducive to optimizing implicit function and the SSC task.

To address these challenges, we propose HybridOcc, a multi-camera semantic scene completion method. HybridOcc refines hybrid occupancy proposals generated by NeRF representation and Transformer architecture in a coarse-to-fine structure. As shown in Fig. 1(c), HybridOcc contains two branches. The Transformer branch, inspired by SurroundOcc [2] and Occ3D [8], uses learnable cross-attention to lift 2D images to 3D volume and gradually refine 3D volume queries from a coarse-to-fine structure. The NeRF branch innovatively adapts volume rendering with depth supervision to predict complete occupancy. Due to the NeRF optimization challenges posed by occlusions in autonomous driving scenes, we propose occupancy-aware ray sampling to optimize large radiance volumes. The implicit function is trained to serve the SSC task by taking occupancy-aware sampling points across visible and invisible voxels along the ray. The occupancy priors for each layer need to be carefully considered in the coarse-to-fine structure. Improved NeRF can hallucinate the occupancy

of occluded invisible regions. The binary occupancy predicted by NeRF and coarse-grained Transformer is hybridized as a new volume query set to refine the semantic occupancy. In summary, our contributions are threefold:

1) We propose a novel complementary combination of contextual feature aggregation of the Transformer and depth supervision of NeRF. The hybrid occupancy proposals generated by NeRF representation and Transformer framework are refined end-to-end in a coarse-to-fine framework.

2) We introduce a novel depth-supervised neural radiance field that takes into account all visible and occluded invisible voxels for the SSC task. It adds depth signals to the coarse-to-fine SSC prediction framework and includes an occupancy-aware ray sampling strategy.

3) Extensive experiments demonstrate the effectiveness of our HybridOcc, which outperforms methods based on depth prediction networks such as FB-Occ and VoxFormer.

II. RELATED WORKS

A. 3D Semantic Scene Completion.

3D semantic scene completion can provide a more detailed understanding of autonomous driving scenes. Some previous works [18], [19] are studied in small-scale indoor scenes. With the release of the SemanticKITTI dataset [14] and nuScenes dataset [20], the SSC benchmark [21]–[23] for large-scale autonomous driving scenes has been rapidly proposed recently. SurroundOcc [2] and Occ3D [8] constructs nuScenes-based 3D occupancy prediction datasets respectively, one is oriented to the dense SSC task, and the other only evaluates occupancy of visible surfaces. These occupancy methods can be simply categorized into building 3D voxel features based on depth prediction [6], [7], [9] and using Transformer-based learnable voxel feature aggregation [2], [8], [12], [24]. Some methods [9], [25], [26] introduce historical frame data to solve depth prediction and occlusion problems. OccFiner [26] proposes to implicitly capture and process multiple local frames. Additionally, some methods [10], [11], [17] use NeRF [15] representation to explore occupancy task, but they focus more on reconstruction rather than SSC. We propose a method that combines the advantages of the Transformer paradigm and NeRF representation to enhance the SSC task performance.

B. 3D Scene Reconstruction.

3D scene reconstruction aims to model 3D surface information from single- or multi-view 2D images. Early reconstruction methods focused on explicit representation of voxels [27] but now neural radiance fields (NeRF) [15], [28] and 3D Gaussian Splatting [29], [30] are becoming more popular for implicit reconstruction. Considering that NeRF has the problem of slow rendering, some methods [31]–[33] improve the rendering speed while maintaining the rendering quality. Implicit reconstruction works [34], [35] based on image features extends object-level reconstruction to indoor scenes and is committed to building a generalized implicit network. [34] and [35] adopt a coarse-to-fine approach fuse multi-scale features to obtain more accurate 3D reconstruction of indoor scenes. SceneRF [16] proposes spherical U-Net and

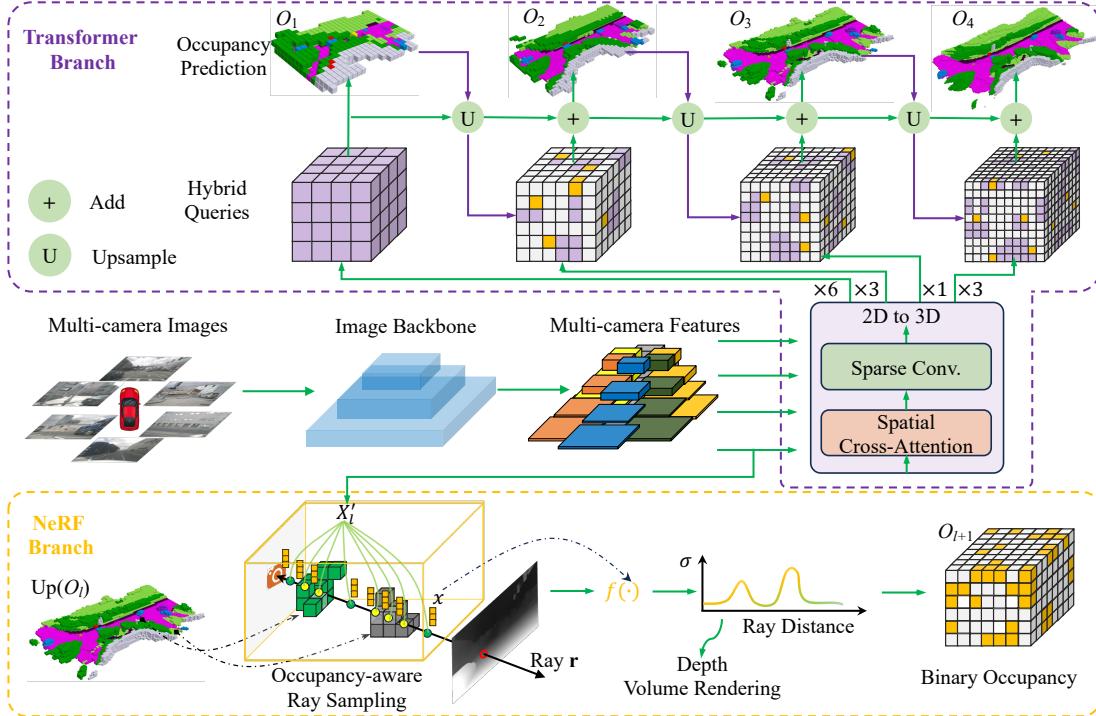


Fig. 3. The pipeline of the proposed HybridOcc for multi-camera 3D semantic occupancy prediction. It consists of the image backbone for extracting multi-scale features and the dual branch composed of Transformer and NeRF to learn a sparse 3D feature volume from coarse to fine. The Transformer branch contains a 2D to 3D transformation module for lifting the 2D features to 3D volumes, and the NeRF branch obtains supervision from depth signals to enhance the Transformer branch.

probabilistic ray sampling to expand NeRF for large-scale outdoor scenes. It is worth noting that 3D reconstruction under the NeRF paradigm requires the sampling points along the ray to be concentrated near the 3D surface for better rendering of color or semantics. However, for the SSC task, it makes more sense to concentrate the radiation field on occupied voxels.

III. APPROACH

A. Overall Architecture

The overall pipeline of HybridOcc is shown in Fig. 3. Taking the multi-camera images as the inputs, we use an image backbone to extract multi-scale camera features. Then we learn the sparse 3D volume features through a dual branch composed of Transformer framework and NeRF representation. Specifically, the Transformer branch learns 3D volume-shaped queries from multi-camera features via a 2D to 3D transformation module. The hybrid 3D query proposals are derived from Transformer and NeRF respectively, and gradually refined in a coarse-to-fine manner (see Sec. III-B). In the NeRF branch, the vanilla NeRF paradigm is replaced by the new autonomous driving scene occupancy prediction NeRF module. Volume rendering occupancy prediction models are directly supervised by depth rather than RGB color (see Sec. III-C). The semantic occupancy ground truth supervises multi-scale volume semantic occupancy prediction.

B. Transformer Branch

Coarse-to-fine Approach. Unlike the dense 3D volume obtained in SurroundOcc [2], inspired by Occ3D [8], we

adopt a coarse-to-fine approach to gradually refine the sparse volume, as shown in the upper part of Fig. 3. Specifically, the semantic occupancy O_l of the 3D volume space $V_l \in \mathbb{R}^{H_l \times W_l \times Z_l \times C}$ at each scale is predicted by an MLP following SurroundOcc [2]. Voxels with an occupancy value lower than the occupancy threshold θ are defined as empty voxels. The l -th volume occupancy O_l serves as part of the query prior position distribution in the higher resolution of the volume V_{l+1} , as shown in the purple arrow and purple square in Fig. 3. Sparse voxels of V_{l+1} are recorded as sparse query proposals $Q_{l+1,s} \in \mathbb{R}^{N_{l+1} \times C}$, $Q_{l+1,s} \subseteq V_{l+1}$, and $Q_{l+1,s}$ is learned from multi-camera features at each scale via the 2D to 3D module. Finally, $Q_{l+1,s}$ is skip-connected with upsampled $Q_{l,s}$ and fed to the MLP to predict $l+1$ -th layer semantic occupancy. The semantic occupancy prediction can be expressed as:

$$O_{l+1} = h(Q_{l+1,s} + \text{up}(Q_{l,s})), \quad (1)$$

where up is $2 \times$ upsample, $h(\cdot)$ represent MLP.

It is worth noting that the initial query proposals of the coarse-grained volume V_1 are densely constructed. The prior spatial distribution of query proposals of the fine-grained V_2 , V_3 , and V_4 are composed of the hybrid of binary occupancies of the Transformer branch and the NeRF branch at each scale, respectively (see Sec. III-C).

2D to 3D Transformation. Inspired by recent Transformer-based multi-camera 3D perception methods [2], [4], we project the 3D reference points of the volume onto the 2D camera to aggregate features. Specifically, each 3D reference point corresponding to query $q \subseteq Q_l$, is projected to the 2D feature map according to the given camera intrinsic and

extrinsic parameters and performs deformable cross-attention (DeformAtt) to learn features:

$$\text{DeformAtt}(q, X) = \sum_{m=1}^M W_m \sum_{k=1}^K A_{mk} \cdot W'_m X(p + \Delta p_{mk}), \quad (2)$$

where X is the multi-camera features, W_m and W'_m are the weights obtained by linear projection, A_{mk} is the attention weights and $A_{mk} \in [0, 1]$, $X(p + \Delta p_{mk})$ are the sampled features corresponding to the 2D reference point p , Δp_{mk} is the learned position offsets corresponding to p . Other settings follow SurroundOcc [2] and BEVFormer [4]. Finally, the volume-shaped query Q is further optimized through 3D sparse convolution, so that each voxel query subset pays attention to local information of each other.

C. Neural Radiance Field Branch

Depth Rendering Supervision. Vanilla NeRF [15], [32] optimizes a continuous radiance field $f(\cdot) = (c, \rho)$ based on the density ρ of sample points along the ray, and supervises volume rendering with RGB. The difference is that we design a new radiance field based on SceneRF [16], and the new NeRF model has depth supervision to predict 3D occupancy.

The NeRF branch is shown at the bottom of Fig. 3. The NeRF branch performs occupancy prediction and depth rendering based on l -th level multi-camera features $X_l, l = 2, 3$, and 4 from the image backbone. We uniformly sampled I pixels from the pixel coordinates of each camera, and sample N points along each ray passing through these pixels. The uniform sampling strategy is consistent with SceneRF [16]. Then following SceneRF converts X_l into a spherical space to obtain X'_l , so that each sampled point x can be projected on the spherical space for retrieving the image feature vector $X'_l(x)$ through bilinear interpolation. Finally, the feature $X'_l(x)$ of the point x and the 3D position encoding $\gamma(x)$ are fed to the implicit expression function MLP to predict the binary occupancy σ_l of volume V_l . Note that the NeRF branch only needs to provide prior spatial distribution information of query for the Transformer-based coarse-to-fine structure, therefore, we only predict class-agnostic occupancy. The binary occupancy prediction implicit radiance field is defined as:

$$f(\gamma(x), X'_l(x)) = (d_l, \sigma_l), \quad (3)$$

where d_l is the l -th scale depth.

Unlike most NeRFs [15], [36] that use volume rendering from density to predict colors, we attempt to reveal depth explicitly from the radiance volume. Depth volume rendering is performed on multi-scale image features X_2 , X_3 , and X_4 respectively, so that multi-scale features can obtain depth supervision. For l -th scale feature, we define depth volume rendering as:

$$D_{r,l} = \sum_{i=1}^N w_{i,l} d_{i,l}, \quad (4)$$

$$w_{i,l} = T_{i,l} (1 - \exp(-\sigma_{i,l} \delta_{i,l})),$$

where $D_{r,l}$ is l -th ray depth, $T_{i,l}$ is the accumulated transmittance, $T_{i,l} = -\sum_{i=1}^{n-1} \sigma_{i,l} \delta_{i,l}$, and $\delta_{i,l}$ is the distance to the previous adjacent point.

Occupancy-aware Ray Sampling. Previous research [15], [16], [37] has shown that sampling points along the ray near the surface of the scene can effectively increase rendering efficiency. Fig. 4(a) illustrates the hierarchical volume sampling [15] method produces a probability density function (PDF) focused on the surface along the ray to optimize the sampling points. In the NeRF module of the SSC task, the objective is to estimate the depth based on features of visible surfaces and invisible voxels, rather than the color or category. The implicit function of the NeRF branch needs to complete the 3D volume binary occupancy prediction of the entire scene. As for the occupancy prediction NeRF model, intuitively, the sampling points along the ray falling on non-empty voxels can improve the rendering. Therefore, we propose an occupancy-aware ray sampling strategy in which the occupancy prediction result O_l of volume V_l explicitly guides each sampling point along the ray in the volume V_{l+1} , as shown in Fig. 4(b).

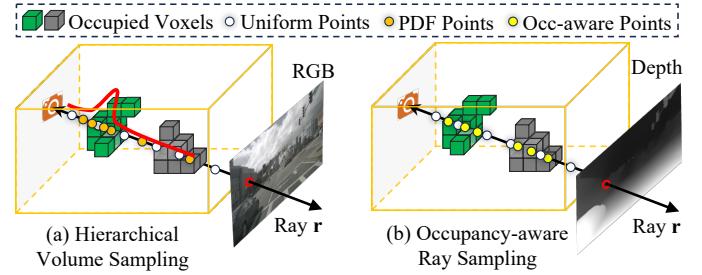


Fig. 4. Comparisons of the proposed occupancy-aware ray sampling with hierarchical volume sampling. Our ray sampling strategy focuses on sampling all visible or invisible occupied voxels passing through the ray.

Specifically, for each ray, we first sample 128 points uniformly between the near and far boundaries. These points are then projected into 3D volume V_3 to query the occupancy state O_3 , and we sample 32 points based on the occupancy state. If more than 32 points are occupied, 32 points are randomly sampled from them. Otherwise, we accept all occupied points and randomly sample the remaining points. The occupancy-aware ray sampling strategy concentrates on sampling occupied voxels within the scene, optimizing binary occupancy prediction and volume rendering of depth.

Hybrid Query Proposals. The aforementioned NeRF branch has the 3D occupancy prediction capability. Take the nuScenes [20] dataset as an example, we first independently partition each camera’s features X_2 , X_3 , and X_4 into the 3D voxel space, and predict the binary occupation of the 3D volume by the NeRF module. Then, we fuse the multi-camera results in volume coordinate $V_{\text{imp},l}^{H_l \times W_l \times Z_l}$ with camera extrinsics and obtain the occupancy distribution $O_{\text{imp},l} \in \{0; 1\}^{H_l \times W_l \times Z_l}$. Meanwhile, the depth supervision signal also updates the image features, making the model sensitive to depth. Finally, we fuse the implicitly predicted occupancy $O_{\text{imp},l}$ from the NeRF branch with the explicitly estimated $O_{\text{exp},l}$ from the coarse-grained Transformer branch in volume V_{l-1} . The hybrid query proposals serve as the l -th level query

to attend to the 2D to 3D process. Hybrid query proposals $Q_{l,s}$ can be expressed as:

$$Q_{l,s} = Q[O_{\text{imp},l}] \cup Q[O_{\text{exp},l}]. \quad (5)$$

D. Loss Fusion

Our dual-branch occupancy prediction network is an end-to-end optimization model. The overall loss of the model $L_{\text{total}} = L_{\text{exp}} + \beta L_{\text{imp}}$, where L_{exp} is the explicit loss of the Transformer branch, L_{imp} is the implicit loss of the NeRF branch, and β is set to 0.5. The supervision of multi-scale 3D volumes is inspired by SurroundOcc [2]. We also supervise each scale volume to get coarse-grained and fine-grained 3D features. We adopt cross-entropy loss for 3D semantic occupancy prediction. L_{exp} is expressed as:

$$L_{\text{exp}} = \sum_{i=1}^L \alpha_i L_i(\sigma_i), \quad (6)$$

where α_i is the decayed loss weight for the i -th scale supervision. Hybrid queries of high-resolution volume V_4 require sufficient supervision signals.

For the loss of the NeRF branch, we use binary cross-entropy loss for class-agnostic occupancy prediction and utilize SILog loss [38] to optimize the depth, the depth is supervised by the projection of the LiDAR points. Also includes the decayed loss weight α_i . L_{imp} is expressed as:

$$L_{\text{imp}} = \sum_{i=2}^L \alpha_i (L_{\text{depth},i}(D_r, \hat{D}_r) + L_i(\sigma)). \quad (7)$$

IV. EXPERIMENT

A. Datasets

We conduct multi-camera semantic scene completion experiments on the nuScenes dataset [20], which contains surround RGB image data from 6 cameras and Lidar sweeps covering the full 360-degree field of view. The 1 000 multi-modal data are split into train/val/test splits with 700/150/150. SurroundOcc [2] proposed a 3D SSC benchmark based on the nuScenes dataset, and there are 17 categories of 3D occupancy. The perception range is clipped into $[-50m, 50m]$ for X, Y axis and $[-5m, 3m]$ for Z axis. The ground truth volume dimension of semantic occupancy is $200 \times 200 \times 16$ with $0.5m$ voxel size. For Occ3D-nuScenes, The perception range is clipped into $[-40m, 40m]$ for X, Y axis and $[-1m, 5, 4m]$ for Z axis. The final output occupancy shape is $200 \times 200 \times 16$ with $0.4m$ voxel size.

To further demonstrate the effectiveness of our method, we conduct monocular semantic scene completion experiment on the SemanticKITTI dataset [14], which annotates autonomous driving scene with 21 semantic classes (19 semantics, 1 free, and 1 unknown). The dataset contains 22 sequences and is split into 10/1/11 for train/val/test. The perception range is clipped into $[-25.6m, 25.6m]$ for X axis, $[0, 51.2m]$ for Y axis, and $[-2m, 4.4m]$ for Z axis. The ground truth semantic occupancy has a dimension of $256 \times 256 \times 32$ with $0.2m$ voxel size.

B. Metrics

For both SurroundOcc-nuScenes [2] and SemanticKITTI [14] datasets, we report the intersection over union (IoU) of occupied voxels as the evaluation metric of the class-agnostic scene completion (SC) task and the mIoU of all semantic classes for the SSC task following SurroundOcc [2]. For Occ3D-nuScenes [8], we report the mIoU following FB-Occ [9] and Occ3D [8]. It is worth noting that Occ3D-nuScenes only evaluates visible regions, as shown in Fig. 2. We refer the readers to previous papers [2], [8], [14] for more details.

C. Implementation Details

For the SurroundOcc-nuScenes [2], [20] dataset, the input image resolution is 900×1600 . We following SurroundOcc [2] adopt ResNet-101 [39], [40] initialized from the FCOS3D [41] checkpoint as the image backbone. The image backbone yields 3 level feature maps, and employs FPN [42] following the backbone to produce 4 level feature maps with hidden dimensions 256. For the SemanticKITTI dataset, we crop image of cam2 to size 370×1220 , and following [2], [6] use EfficientNetB7 [5] as the image backbone for fair comparison. For the Occ3D-nuScenes [2], [20] dataset, we follow FB-Occ [9] resize the input image resolution to 256×704 and adopt ResNet-50 as image backbone.

For both SurroundOcc-nuScenes and SemanticKITTI datasets, we set the number of 2D to 3D spatial cross-attention layers as 6, 3, 1, and 3 respectively. Each level of spatial cross-attention uses 8, 4, 4, and 4 sampling points around each reference point, respectively. The MLP structure of the implicit function $f(\cdot)$ in NeRF module is consistent with that of SceneRF [16]. We use 32 points per ray in occupancy-aware ray sampling. The occupancy threshold θ is set to 0.5. The Occ3D-nuScenes dataset evaluates the visible surfaces of a scene, and we simply set our HybridOcc sampling strategy to probabilistic ray sampling [16]. We train our model 24 epochs on nuScenes dataset and 30 epochs on SemanticKITTI dataset with a learning rate 2×10^{-4} by default. All models are trained with a batch size of 4 on 4 NVIDIA A800 GPUs.

D. Main Results

In Table I, we report the multi-camera semantic scene completion task on SurroundOcc-nuScenes [2] val set. We compare our HybridOcc with several vision-based methods [2], [6], [9]. Compared with end-to-end OccFormer [6], HybridOcc achieves a 1.53% IoU lead. OccFormer has a learning-based deep prediction network and contains two Transformer modules for aggregating contextual features. For the Transformer-based multi-scale SurroundOcc [2], our hybrid query strategy brings a remarkable boost of 1.58% IoU and 1.06% mIoU. Compared with the single-frame variant of FB-Occ [9], HybridOcc achieves 1.51% IoU and 1.19% mIoU lead. FB-Occ has a pre-trained depth prediction network, but lacks consideration of occluded invisible voxels. Our HybridOcc can achieve end-to-end training while the NeRF branch additionally considers invisible voxels.

In addition, we conduct semantic scene completion experiments on the monocular SemanticKITTI dataset, as shown in

TABLE I

SEMANTIC SCENE COMPLETION RESULTS ON SURROUND OCC-NUSCENES VAL SET. * INDICATES SURROUND OCC [2] REPORTS THESE MODELS. † REPRESENTS TRAINED ON SURROUND OCC-NUSCENES. FB-OCC(D) [9] IS THE SINGLE-FRAME VERSION.

Method	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	mammal	vegetation
BEVFormer* [4]	30.50	16.75	14.22	6.58	23.46	28.28	8.66	10.77	6.64	4.05	11.20	17.78	37.28	18.00	22.88	22.17	13.80	22.21
TPVFormer* [12]	30.86	17.10	15.96	5.31	23.86	27.32	9.79	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
FB-Occ(D)† [9]	31.56	20.17	20.31	12.29	26.33	31.07	10.78	15.95	13.31	11.14	13.24	22.13	39.56	22.26	25.14	23.59	13.92	21.64
SurroundOcc* [2]	31.49	20.30	20.59	11.68	28.06	30.86	10.70	15.14	14.09	12.06	14.38	22.26	37.29	23.70	24.49	22.77	14.89	21.86
OccFormer† [6]	31.54	20.97	21.98	11.92	28.77	31.91	11.62	14.92	14.26	11.57	15.38	23.60	40.01	22.93	25.74	24.14	14.51	22.29
HybridOcc(Ours)	33.07	21.36	22.29	12.13	29.78	32.34	10.94	16.33	14.07	12.69	14.63	23.98	40.43	23.69	26.15	24.53	15.23	22.60

TABLE II

MONOCULAR SEMANTIC SCENE COMPLETION RESULTS ON SEMANTICKITTI TEST SET. OUR METHOD SURPASSES VOXFORMER [7], WHICH USES AN ADDITIONAL DEPTH PREDICTION NETWORK [43].

Method	IoU	mIoU	road	sidewalk	parking	other-grnd	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traf.-sign
MonoScene [5]	34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
TPVFormer [12]	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
SurroundOcc [2]	34.72	11.86	56.90	28.30	30.20	6.80	15.20	20.60	1.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40
OccFormer [6]	34.53	12.32	55.90	30.30	31.50	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
VoxFormer [7]	44.02	12.35	54.76	26.35	15.50	0.70	17.65	25.79	5.08	0.59	0.51	3.77	24.39	5.63	29.96	1.78	3.32	0.00	7.64	7.11	4.18
HybridOCC(Ours)	36.34	12.68	57.38	31.10	30.92	7.20	16.20	21.70	1.60	1.70	1.80	4.70	16.20	4.20	21.50	2.10	2.20	0.20	12.20	4.20	3.90

TABLE III

3D OCCUPANCY PREDICTION PERFORMANCE ON OCC3D-NUSCENES VAL SET. * MEANS REPORTED BY CTF-OCC [8], † MEANS ONLY 2D SUPERVISION. FB-OCC(D) [9] IS THE SINGLE-FRAME VERSION.

Method	Backbone	Revolution	mIoU
SelfOcc† [17]	ResNet-50	768×1600	9.30
OccNeRF† [10]	ResNet-101	900×1600	10.81
OccFormer* [6]	ResNet-101	900 × 1600	21.93
RenderOcc [11]	ResNet-101	512×1408	23.95
BEVFormer* [4]	ResNet-101	900 × 1600	26.88
TPVFormer* [12]	ResNet-101	900 × 1600	27.83
CTF-Occ* [8]	ResNet-101	900 × 1600	28.53
SurroundOcc [2]	ResNet-50	256 × 704	36.32
FB-Occ(D) [9]	ResNet-50	256 × 704	37.39
HybridOcc	ResNet-50	256 × 704	37.82

Table II. HybridOcc outperforms several existing competitors [2], [6], [7]. It achieves approximately 12.68% mIoU on the SSC task, which achieves an improvement of 0.33% compared to two-stage VoxFormer [7]. VoxFormer uses an additional depth prediction network [43] to provide strong position priors for 3D voxel query proposals, achieving a remarkable 44.02% IoU. However, VoxFormer’s ablation experiments show that without adding a one-stage depth prediction network, it only has 34.64% IoU. Compared to this variant, our HybridOcc achieves a 1.7% IoU lead. Compared with the end-to-end OccFormer [6], which also has a depth prediction network, HybridOcc shows a 1.81% IoU lead. Monocular SSC experiments also demonstrate the effectiveness of our method.

In Table III, we further experiment on Occ3D-nuScenes [8], which only evaluates visible voxels. HybridOcc achieves

TABLE IV
ABLATION STUDY FOR ARCHITECTURAL COMPONENTS. W.O. DENOTES WITHOUT.

Method	SC IoU	SSC mIoU
w.o. Transformer branch	28.84	15.49
w.o. NeRF branch	32.19	20.61
w.o. Sparse convolution	32.87	21.16
Ours	33.07	21.36

37.82% mIoU, surpassing FB-Occ [9] with the depth network by 0.43% mIoU. Furthermore, HybridOcc achieves 1.50% mIoU lead compared to SurroundOcc [2], which uses an attention-based dense multi-scale supervision framework.

E. Ablation Studies

Architectural components. In Table IV, we ablate the architectural components in our HybridOcc. Since our proposed NeRF branch cannot operate independently of the Transformer structure, we use the probabilistic ray sampling method [16] in the NeRF model to ablate the case without the Transformer branch. The NeRF model shows certain performance on the SSC task. The 1st and 2nd rows indicate that the NeRF-based method performs worse than the Transformer-based method. In 3rd row, 3D sparse convolution positively contributes to performance. Finally, Hybrid query proposals include depth and invisible information supplementation from the NeRF branch, which improves the occupancy prediction precision.

Ablation on Ray Sampling. In Table V, we conduct a comprehensive study on the ray sampling strategy of the neural radiance field branch. Probabilistic ray sampling [16]

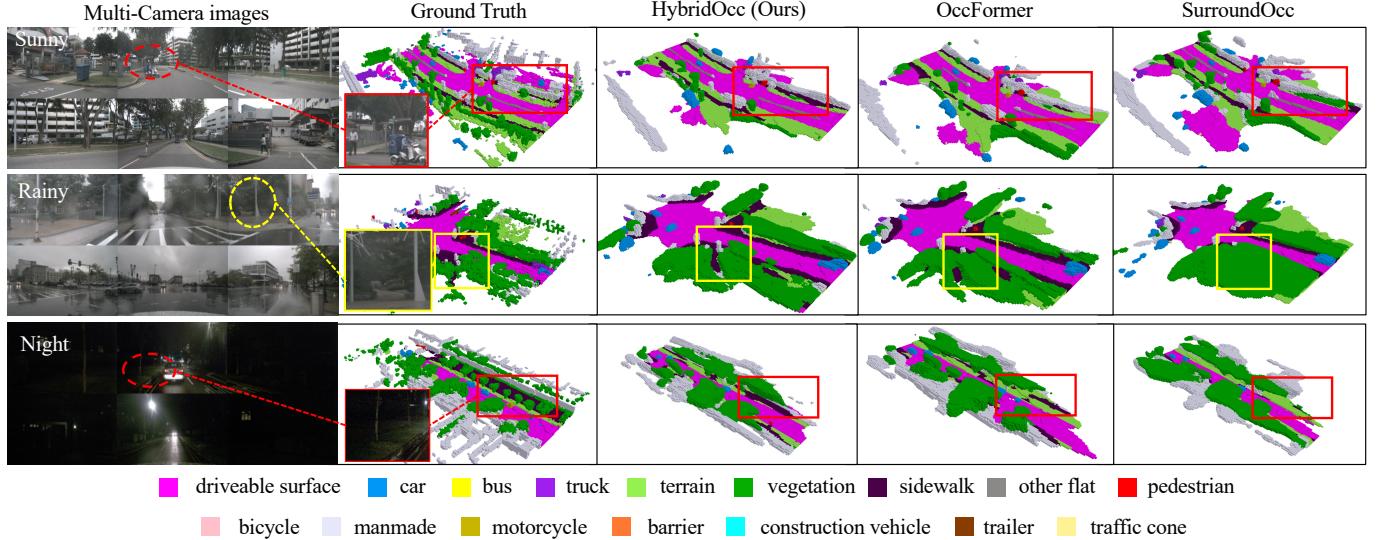


Fig. 5. Visualizations on nuScenes validation set. The leftmost column is the input multi-camera surround image, and the next four columns show the ground truth semantic occupancy, the 3D semantic occupancy predicted by our HybridOcc, OccFormer [6] and SurroundOcc [2].

TABLE V
ABLATION STUDY FOR RAY SAMPLING, WE VARY POINT NUMBER AND SAMPLING METHOD.

Method	Point	SC IoU	SSC mIoU
Hierarchical volume sampling [15]	64	32.51	20.79
Probabilistic ray sampling [16]	32	32.70	20.82
Occupancy-aware ray sampling	16	32.56	20.85
	32	33.07	21.36
	64	33.10	21.28

TABLE VI
THE MODEL EFFICIENCY AND PERFORMANCE COMPARISON ON SURROUNDOCC-NUSCENES [2] DATASET. THE LATENCY OF ALL METHODS IS MEASURED ON A SINGLE A800 GPU.

Method	mIoU	Memory(G)	Latency(ms)
FB-Occ(D) [9]	20.17	5.40	337
SurroundOcc [2]	20.30	5.90	413
HybridOcc(NeRF [16])	21.36	7.40	522
HybridOcc(Instant-NGP [31])	21.29	6.30	426

concentrates sampling points near the geometric surface. This type of approach performs worse than occupancy-aware ray sampling. The potential reason is that voxel features of visible surfaces and occluded regions are needed to jointly optimize the radiance field in the SC task. Additionally, due to the sparsity of the scene, 32 sampling points per ray are sufficient to optimize a single view with a sensing range of 50m.

Ablation on efficiency of NeRF branch. In Table VI, we study the impact of the NeRF variant on model efficiency and compare it with SurroundOcc [2] and FB-Occ [9]. There are two potential possibilities for accelerating NeRF on the SSC task. One is to reduce sampling points, and the other is to reduce the number and dimensions of MLP layers in NeRF. We adopt the lightweight MLP proposed by Instant-NGP [31] to accelerate our model. The results show that Instant-NGP reduces the inference memory usage and the

inference latency (Instant-NGP variant (55 ms) *v.s.* NeRF (151 ms)). The introduction of the NeRF branch inevitably makes its inference speed slower than FB-Occ, but our model does not require an additional pre-trained depth network and has slightly higher performance.

F. Visualization

Fig. 5 shows a visualization of the SSC prediction results of proposed HybridOcc, OccFormer [6] and SurroundOcc [2] on the SurroundOcc-nuScenes [2] dataset. We show sunny day, rainy day, and night visualization. Compared to OccFormer and SurroundOcc, which have large and continuous occupancies, our HybridOcc results are much more refined. Such as objects of different categories have more independent and complete boundaries, which also reduces the predicted false positive phenomenon (yellow marks). In addition, HybridOcc shows better semantic refinement capabilities, such as better distinction between terrain and sidewalk (red marks) that are approximately the same plane.

V. CONCLUSION

In this paper, we have presented HybridOcc, which generates sparse hybrid 3D volume query proposals by Transformer and NeRF for semantic scene completion prediction. To more precisely lift the 2D camera features to 3D volumes, the proposed hybrid query from coarse to fine gradually refines autonomous driving scenes. The newly designed NeRF branch implicitly infers the visible and invisible scene occupancy, explicitly obtains depth signal supervision, and combines it with the Transformer branch to hallucinate the overall volume semantic layout of the scene with higher accuracy. Considering that the occupancy prediction NeRF branch relies on non-empty voxel features, we have designed an occupancy-aware ray sampling method to optimize the hybrid model significantly. The superiority of our HybridOcc is demonstrated by comparing the nuScenes and SemanticKITTI datasets.

REFERENCES

- [1] V. Guizilini, I. Vasiljevic, R. Ambrus, G. Shakhnarovich, and A. Gaidon, “Full surround monodepth from multiple cameras,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5397–5404, 2022.
- [2] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, “Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.
- [3] J. Philion and S. 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*. Springer, 2020, pp. 194–210.
- [4] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, “Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers,” in *European conference on computer vision*. Springer, 2022, pp. 1–18.
- [5] A.-Q. Cao and R. de Charette, “Monoscene: Monocular 3d semantic scene completion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3991–4001.
- [6] Y. Zhang, Z. Zhu, and D. Du, “Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction,” *ArXiv*, vol. abs/2304.05316, 2023.
- [7] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. 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*, 2023, pp. 9087–9098.
- [8] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, “Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [9] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, “Fb-occ: 3d occupancy prediction based on forward-backward view transformation,” *arXiv preprint arXiv:2307.01492*, 2023.
- [10] C. Zhang, J. Yan, Y. Wei, J. Li, L. Liu, Y. Tang, Y. Duan, and J. Lu, “Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields,” *arXiv preprint arXiv:2312.09243*, 2023.
- [11] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, L. Liu, and S. Zhang, “Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision,” *arXiv preprint arXiv:2309.09502*, 2023.
- [12] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, “Tri-perspective view for vision-based 3d semantic occupancy prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9223–9232.
- [13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [14] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, “Semantickitti: A dataset for semantic scene understanding of lidar sequences,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.
- [15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [16] A.-Q. Cao and R. de Charette, “Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9387–9398.
- [17] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu, “Selfocc: Self-supervised vision-based 3d occupancy prediction,” *arXiv preprint arXiv:2311.12754*, 2023.
- [18] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, “Semantic scene completion from a single depth image,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1746–1754.
- [19] Y. Cai, X. Chen, C. Zhang, K.-Y. Lin, X. Wang, and H. Li, “Semantic scene completion via integrating instances and scene in-the-loop,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 324–333.
- [20] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Lioung, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [21] Y. Li, S. Li, X. Liu, M. Gong, K. Li, N. Chen, Z. Wang, Z. Li, T. Jiang, F. Yu *et al.*, “Ssbench: Monocular 3d semantic scene completion benchmark in street views,” 2023.
- [22] X. Liu, M. Gong, Q. Fang, H. Xie, Y. Li, H. Zhao, and C. Feng, “Lidar-based 4d occupancy completion and forecasting,” *arXiv preprint arXiv:2310.11239*, 2023.
- [23] A.-Q. Cao, A. Dai, and R. de Charette, “Pasco: Urban 3d panoptic scene completion with uncertainty awareness,” *arXiv preprint arXiv:2312.02158*, 2024.
- [24] Z. Ming, J. S. Berrio, M. Shan, and S. Worrall, “Occfusion: A straightforward and effective multi-sensor fusion framework for 3d occupancy prediction,” *arXiv preprint arXiv:2403.01644*, 2024.
- [25] H. Liu, H. Wang, Y. Chen, Z. Yang, J. Zeng, L. Chen, and L. Wang, “Fully sparse 3d panoptic occupancy prediction,” *arXiv preprint arXiv:2312.17118*, 2023.
- [26] H. Shi, S. Wang, J. Zhang, X. Yin, Z. Wang, Z. Zhao, G. Wang, J. Zhu, K. Yang, and K. Wang, “Occfiner: Offboard occupancy refinement with hybrid propagation,” *arXiv preprint arXiv:2403.08504*, 2024.
- [27] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, “Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images,” *International Journal of Computer Vision*, vol. 128, no. 12, pp. 2919–2935, 2020.
- [28] J. Yang, B. Ivanovic, O. Litany, X. Weng, S. W. Kim, B. Li, T. Che, D. Xu, S. Fidler, M. Pavone *et al.*, “Emernerf: Emergent spatial-temporal scene decomposition via self-supervision,” *arXiv preprint arXiv:2311.02077*, 2023.
- [29] Y. Li, Z. Wang, Y. Wang, Z. Yu, Z. Gojcic, M. Pavone, C. Feng, and J. M. Alvarez, “Memorize what matters: Emergent scene decomposition from multitraverse,” *arXiv preprint arXiv:2405.17187*, 2024.
- [30] K. Cheng, X. Long, K. Yang, Y. Yao, W. Yin, Y. Ma, W. Wang, and X. Chen, “Gaussianpro: 3d gaussian splatting with progressive propagation,” *arXiv preprint arXiv:2402.14650*, 2024.
- [31] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.
- [32] T. Hu, S. Liu, Y. Chen, T. Shen, and J. Jia, “Efficientnerf efficient neural radiance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 902–12 911.
- [33] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, “Zip-nerf: Anti-aliased grid-based neural radiance fields,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 697–19 705.
- [34] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, “Neuralrecon: Real-time coherent 3d reconstruction from monocular video,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 593–15 602, 2021.
- [35] A. Bovzic, P. R. Palafox, J. Thies, A. Dai, and M. Nießner, “Transformerfusion: Monocular rgb scene reconstruction using transformers,” pp. 1403–1414, 2021.
- [36] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, “pixelnerf: Neural radiance fields from one or few images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4578–4587.
- [37] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao, “Efficient semantic scene completion network with spatial group convolution,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 733–749.
- [38] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” pp. 2366–2374, 2014.
- [39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [40] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, 2017.
- [41] T. Wang, X. Zhu, J. Pang, and D. Lin, “Fcoss3d: Fully convolutional one-stage monocular 3d object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913–922.
- [42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [43] F. Shamsafar, S. Woerz, R. Rahim, and A. Zell, “Mobilestereonet: Towards lightweight deep networks for stereo matching,” in *Proceedings of the ieee/cvf winter conference on applications of computer vision*, 2022, pp. 2417–2426.