UniPAD: A Universal Pre-training Paradigm for Autonomous Driving

Honghui Yang[1,2*], Sha Zhang[2,6], Di Huang[2,7], Xiaoyang Wu[2,5], Haoyi Zhu[2,6], Tong He[2†]
Shixiang Tang[2], Hengshuang Zhao[5], Qibo Qiu[8], Binbin Lin[3,4†], Xiaofei He[1], Wanli Ouyang[2]
[1]State Key Lab of CAD&CG, Zhejiang University [2]Shanghai Artificial Intelligence Laboratory
[3]School of Software Technology, Zhejiang University [4]Fullong Inc. [5]HongKong University
[6]University of Science and Technology of China [7]The University of Sydney [8]Zhejiang Lab

Abstract

In the context of autonomous driving, the significance of effective feature learning is widely acknowledged. While conventional 3D self-supervised pre-training methods have shown widespread success, most methods follow the ideas originally designed for 2D images. In this paper, we present UniPAD, a novel self-supervised learning paradigm applying 3D volumetric differentiable rendering. UniPAD implicitly encodes 3D space, facilitating the reconstruction of continuous 3D shape structures and the intricate appearance characteristics of their 2D projections. The flexibility of our method enables seamless integration into both 2D and 3D frameworks, enabling a more holistic comprehension of the scenes. We manifest the feasibility and effectiveness of UniPAD by conducting extensive experiments on various 3D perception tasks. Our method significantly improves lidar-, camera-, and lidar-camera-based baseline by 9.1, 7.7, and 6.9 NDS, respectively. Notably, our pre-training pipeline achieves 73.2 NDS for 3D object detection and 79.4 mIoU for 3D semantic segmentation on the nuScenes validation set, achieving state-of-the-art results in comparison with previous methods.

1. Introduction

Self-supervised learning for 3D point cloud data is of great significance as it is able to use vast amounts of unlabeled data efficiently, enhancing their utility for various downstream tasks like 3D object detection [20, 51, 63, 64, 89, 92] and semantic segmentation [16, 47, 48, 52, 76, 104]. Although significant advances have been made in self-supervised learning for 2D images [9, 10, 26, 27], extending these approaches to 3D point clouds have presented considerably more significant challenges. This is partly caused by the inherent sparsity of the data, and the variability in point distribution due to sensor placement and occlusions by other scene elements. Previous pre-training paradigms for 3D scene understanding adapted the idea from the 2D image domain and can be roughly categorized into two groups: contrastive-based and MAE-based.

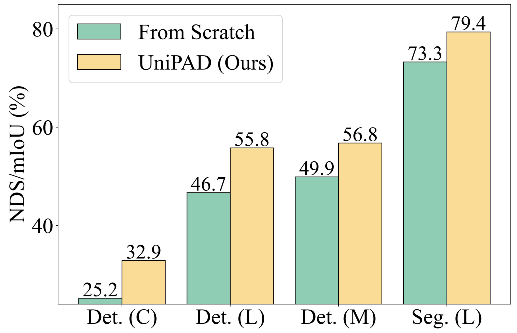

Figure 1. Effect of our pre-training for 3D detection and segmentation on the nuScenes [5] dataset, where C, L, and M denote camera, LiDAR, and fusion modality, respectively.

Contrastive-based methods [12, 102] explore pulling similar 3D points closer together while pushing dissimilar points apart in feature space through a contrastive loss function. For example, PointContrast [80] directly operates on each point and has demonstrated its effectiveness on various downstream tasks. Nonetheless, the sensitivity to positive/negative sample selection and the associated increased latency often impose constraints on the practical applications of these approaches. Masked AutoEncoding (MAE) [27], which encourages the model to learn a holistic understanding of the input beyond low-level statistics, has been widely applied in the autonomous driving field. Yet, such a pretext task has its challenges in 3D point clouds due to the inherent irregularity and sparsity of the data. Voxel-MAE [28] proposed to divide irregular points into discrete voxels and predict the masked 3D structure using voxel-wise supervision. The coarse supervision may lead to insufficient representation capability.

In this paper, we come up with a novel pre-training paradigm tailored for effective 3D representation learn-

[*] This work was done during his internship at Shanghai Artificial Intelligence Laboratory.

[†] Corresponding author

ing, which not only eliminates the need for complex positive/negative sample assignments but also implicitly provides continuous supervision signals to learn 3D shape structures. The whole framework, as illustrated in Figure 2, takes the masked point cloud as input and aims to reconstruct the missing geometry on the projected 2D depth image via 3D differentiable neural rendering.

Specifically, when provided with a masked LiDAR point cloud, our approach employs a 3D encoder to extract hierarchical features. Then, the 3D features are transformed into the voxel space via voxelization. We further apply a differentiable volumetric rendering method to reconstruct the complete geometric representation. The flexibility of our approach facilitates its seamless integration for pre-training 2D backbones. Multi-view image features construct the 3D volume via lift-split-shoot (LSS) [61]. To maintain efficiency during the training phase, we propose a memory-efficient ray sampling strategy designed specifically for autonomous driving applications, which can greatly reduce training costs and memory consumption. Compared with the conventional methods, the novel sampling strategy boosts the accuracy significantly.

Extensive experiments conducted on the competitive nuScenes [5] dataset demonstrate the superiority and generalization of the proposed method. For pre-training on the 3D backbone, our method yields significant improvements over the baseline, as shown in Figure 1, achieving enhancements of **9.1** NDS for 3D object detection and **6.1** mIoU for 3D semantic segmentation, surpassing the performance of both contrastive- and MAE-based methods. Notably, our method achieves the state-of-the-art mIoU of **79.4** for segmentation on nuScenes dataset. Furthermore, our pre-training framework can be seamlessly applied to 2D image backbones, resulting in a remarkable improvement of **7.7** NDS for multi-view camera-based 3D detectors. We directly utilize the pre-trained 2D and 3D backbones to a multi-modal framework. Our method achieves **73.2** NDS for detection, reaching the level of existing state-of-the-art methods. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to explore the 3D differentiable rendering for self-supervised learning in the context of autonomous driving.
- The flexibility of the method makes it easy to be extended to pre-train a 2D backbone. With a novel sampling strategy, our approach exhibits superiority in both effectiveness and efficiency.
- We conduct comprehensive experiments on the nuScenes dataset, wherein our method surpasses the performance of six pre-training strategies. Experimentation involving seven backbones and two perception tasks provides convincing evidence for the effectiveness of our approach.

2. Related Work

Self-supervised learning in point clouds has gained remarkable progress in recent years [12, 28, 30, 37, 42, 45, 57, 59, 68, 77, 85, 96, 102, 106]. PointContrast [80] contrasts point-level features from two transformed views to learn discriminative 3D representations. Point-BERT [99] introduces a BERT-style pre-training strategy with standard transformer networks. OcCo [71] occludes point clouds based on different viewpoints and learns to complete them. PointContrast [80] contrasts point-level features from two transformed views to learn discriminative 3D representations. MSC [78] incorporates a mask point modeling strategy into a contrastive learning framework. PointM2AE [101] utilizes a multiscale strategy to capture both high-level semantic and fine-grained details. STRL [32] explores the rich spatial-temporal cues to learn invariant representation in point clouds. GD-MAE [90] applies a generative decoder for hierarchical MAE-style pre-training. ALSO [4] regards the surface reconstruction as the pretext task for representation learning. Unlike previous works primarily designed for point clouds, our pre-training framework is applicable to both image-based and point-based models.

Representation learning in image has been well-developed [1, 3, 8, 69, 74, 75], and has shown its capabilities in all kinds of downstream tasks as the backbone initialization. Contrastive-based methods, such as MoCo [26] and MoCov2 [11], learn the representations of images by discriminating the similarities between different augmented samples. MAE-based methods [24, 67] obtain the promising generalization ability by recovering masked patches. In autonomous driving, models pre-trained on ImageNet [19] are widely utilized in image-related tasks [29, 38, 40, 43, 46, 50, 86]. For example, to compensate for the insufficiency of 3D priors in tasks like 3D object detection, depth estimation [60] and monocular 3D detection [73] are usually exploited as the additional pre-training techniques.

Neural rendering for autonomous driving utilizes neural networks to differentially render images from 3D scene representation [7, 56, 58, 82, 84, 94]. Those methods can be roughly divided into two categories: perception and simulation. Being capable of capturing semantic and accurate geometry, NeRFs are gradually utilized to do different perception tasks including panoptic segmentation [23], object detection [82, 83], segmentation [35], and instance segmentation [103]. For simulation, MARS [79] models the foreground objects and background environments separately based on NeRF, making it flexible for scene controlling in autonomous driving simulation. Considering the limited labeled LiDAR point clouds data, NeRF-LiDAR [100] proposes to generate realistic point clouds along with semantic

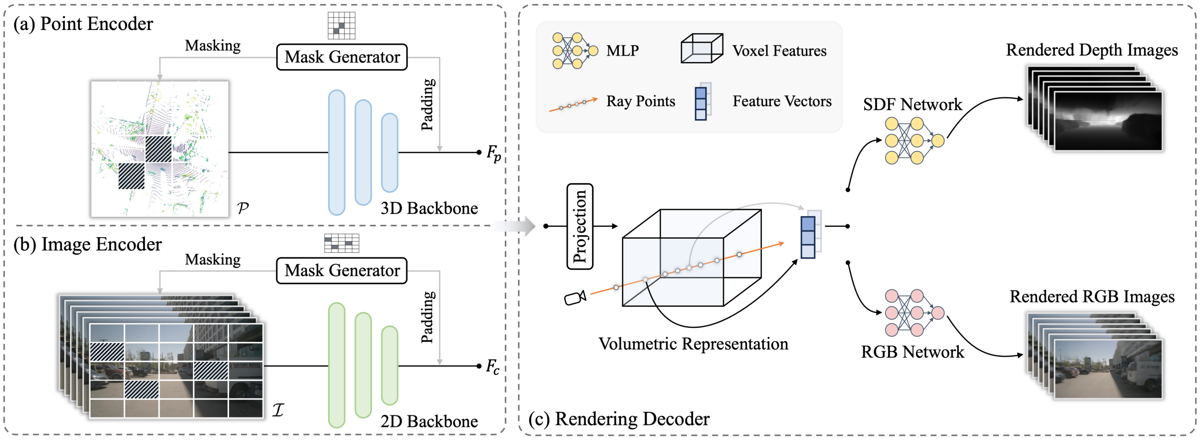

Figure 2. The overall architecture. Our framework takes LiDAR point clouds or multi-view images as input. We first propose the mask generator to partially mask the input. Next, the modal-specific encoder is adapted to extract sparse visible features, which are then converted to dense features with masked regions padded as zeros. The modality-specific features are subsequently transformed into the voxel space, followed by a projection layer to enhance voxel features. Finally, volume-based neural rendering produces RGB or depth prediction for both visible and masked regions.

labels for the LiDAR simulation. Besides, READ [41] explores multiple sampling strategies to make it possible to synthesize large-scale driving scenarios. Inspired by them, we make novel use of NeRF, with the purpose of universal pre-training, rather than of novel view synthesis.

3. Methodology

The UniPAD framework is a universal pre-training paradigm that can be easily adapted to different modalities, e.g., 3D LiDAR point and multi-view images. Our framework is shown in Figure 2, which contains two parts, i.e., a modality-specific encoder and a volumetric rendering decoder. For processing point cloud data, we employ a 3D backbone for feature extraction. In the case of multi-view image data, we leverage a 2D backbone to extract image features, which are then mapped into 3D space to form the voxel representation. Similar to MAE [27], a masking strategy is applied for the input data to learn effective representation. For decoders, we propose to leverage off-the-shelf neural rendering with a well-designed memory-efficient ray sampling. By minimizing the discrepancy between rendered 2D projections and the input, our approach encourages the model to learn a continuous representation of the geometric or appearance characteristics of the input data.

3.1. Modal-specific Encoder

UniPAD takes LiDAR point clouds $\mathcal{P}$ or multi-view images $\mathcal{I}$ as input. The input is first masked out by the mask generator (detailed in the following) and the visible parts are then fed into the modal-specific encoder. For the point cloud $\mathcal{P}$, a point encoder, e.g., VoxelNet [87], is adopted to extract hierarchical features F_p, as shown in Figure 2(a). For images, features F_c are extracted from $\mathcal{I}$ with a classic convolutional network, as illustrated in Figure 2(b). To capture both high-level information and fine-grained details in data, we employ additional modality-specific FPN [44] to efficiently aggregate multi-scale features in practice.

Mask Generator Prior self-supervised approaches, as exemplified by He et al. [27], have demonstrated that strategically increasing training difficulty can enhance model representation and generalization. Motivated by this, we introduce a mask generator as a means of data augmentation, selectively removing portions of the input. Given points $\mathcal{P}$ or images $\mathcal{I}$, we adopt block-wise masking [90] to obscure certain regions. Specifically, we first generate the mask according to the size of the output feature map, which is subsequently upsampled to the original input resolution. For points, the visible areas are obtained by removing the information within the masked regions. For images, we replace the traditional convolution with the sparse convolution as in [67], which only computes at visible places. After the encoder, masked regions are padded with zeros and combined with visible features to form regular dense feature maps.

3.2. Unified 3D Volumetric Representation

To make the pre-training method suitable for various modalities, it is crucial to find a unified representation. Transposing 3D points into the image plane would result in a loss of depth information, whereas merging them into the bird's eye view would lead to the omission of height-related details. In this paper, we propose to convert both modalities into the 3D volumetric space, as shown in Figure 2(c), preserving as much of the original information from their corresponding views as possible. For multi-view images, the view transformation [61] is adopted to transform 2D features into the 3D ego-car coordinate system to obtain the volume features. Specifically, we first predefine the 3D voxel coordinates $X_p \in \mathbb{R}^{X \times Y \times Z \times 3}$, where $X \times Y \times Z$ is

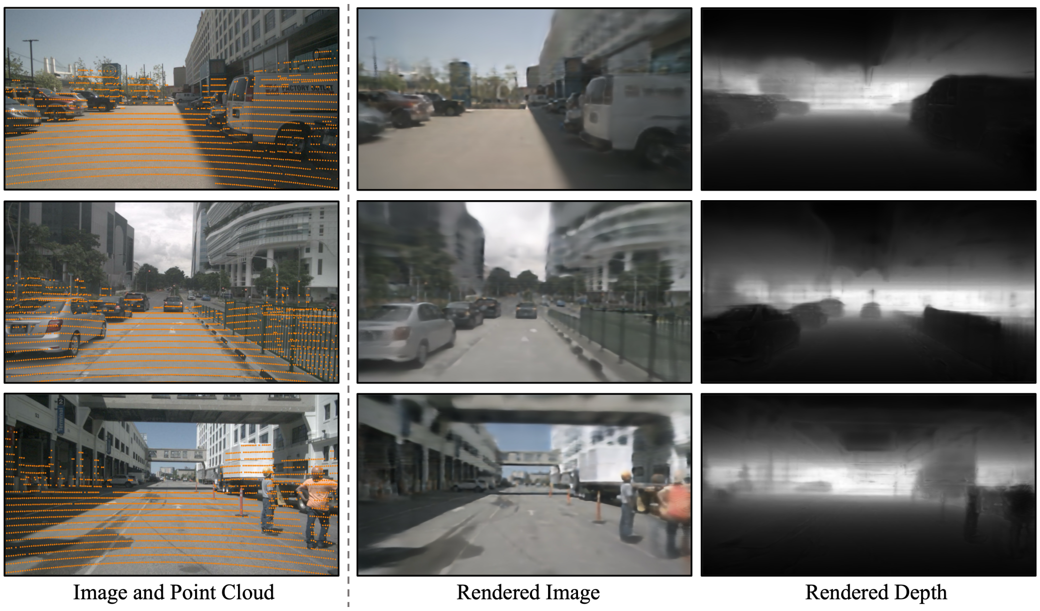

Figure 3. Illustration of the rendering results, where the ground truth RGB and projected point clouds, rendered RGB, and rendered depth are shown on the left, middle, and right, respectively.

the voxel resolution. Subsequently, the X_p is projected on multi-view images to index the corresponding 2D features, which are then multiplied by a learnable scaling factor. The process can be calculated by:

$$X'_p = T_{\mathrm{c2i}} T_{\mathrm{l2c}} X_p, \quad \mathcal{V} = \mathcal{B}(X'_p, F_c)\mathcal{T}(X'_p, \phi(F_c)), \quad (1)$$

where X'_p is the projected coordinates in the image plane, and T_{l2c} and T_{c2i} denote the transformation matrices from the LiDAR coordinate system to the camera frame and from the camera frame to image coordinates, respectively. $\mathcal{V}$ is the constructed volumetric feature, F_c is the image features, and ϕ is determined by a convolutional layer with a Softmax function. $\mathcal{B}$ and $\mathcal{T}$ represent the bilinear and trilinear interpolation to retrieve the corresponding 2D features and scaling factor, respectively. For the 3D point modality, we follow [38] to directly retain the height dimension in the point encoder. Finally, we leverage a projection layer involving L conv-layers to enhance the voxel representation.

3.3. Neural Rendering Decoder

Differentiable Rendering We represent a novel use of neural rendering to flexibly incorporate geometry or textural clues into learned voxel features with a unified pretraining architecture, as shown in Figure 2(c). Specifically, when provided the volumetric features, we sample some rays $\{\mathbf{r}_i\}_{i=1}^K$ from multi-view images or point clouds and use differentiable volume rendering to render the color or depth for each ray. The flexibility further facilitates the incorporation of 3D priors into the acquired image features, achieved via supplementary depth rendering supervision. This capability ensures effortless integration into both 2D and 3D frameworks. Figure 3 shows the rendered RGB images and depth images based on our rendering decoder.

Inspired by [72], we represent a scene as an implicit signed distance function (SDF) field to be capable of representing high-quality geometry details. The SDF symbolizes the 3D distance between a query point and the nearest surface, thereby implicitly portraying the 3D geometry. For ray $\mathbf{r}_i$ with camera origin $\mathbf{o}$ and viewing direction $\mathbf{d}_i$, we sample D ray points $\{\mathbf{p}_j = \mathbf{o} + t_j\mathbf{d}_i \mid j = 1, ..., D, t_j < t_{j+1}\}$, where $\mathbf{p}_j$ is the 3D coordinates of sampled points, and t_j is the corresponding depth along the ray. For each ray point $\mathbf{p}_j$, the feature embedding $\mathbf{f}_j$ can be extracted from the volumetric representation by trilinear interpolation. Then, the SDF value s_j is predicted by $\phi_{\mathrm{SDF}}(\mathbf{p}_j, \mathbf{f}_j)$, where ϕ_{SDF} represents a shallow MLP. For the color value, we follow [58] to condition the color field on the surface normal $\mathbf{n}_j$ (i.e., the gradient of the SDF value at ray point $\mathbf{p}_j$) and a geometry feature vector $\mathbf{h}_i$ from ϕ_{SDF}. Thus, the color representation is denoted as $c_j = \phi_{\mathrm{RGB}}(\mathbf{p}_j, \mathbf{f}_j, \mathbf{d}_i, \mathbf{n}_j, \mathbf{h}_j)$, where ϕ_{RGB} is parameterized by a MLP. Finally, we render RGB value $\hat{Y}_i^{\mathrm{RGB}}$ and depth $\hat{Y}_i^{\mathrm{depth}}$ by integrating predicted colors and sampled depth along rays:

$$\hat{Y}_i^{\mathrm{RGB}} = \sum_{j=1}^{D} w_j c_j, \quad \hat{Y}_i^{\mathrm{depth}} = \sum_{j=1}^{D} w_j t_j, \quad (2)$$

where w_j represents an unbiased and occlusion-aware weight [72] given by $w_j = T_j\alpha_j$. $T_j = \prod_{k=1}^{j-1}(1-\alpha_k)$ is the accumulated transmittance, and α_j is the opacity value computed by:

$$\alpha_j = \max\left(\frac{\sigma_s(s_j) - \sigma_s(s_{j+1})}{\sigma_s(s_j)}, 0\right), \quad (3)$$

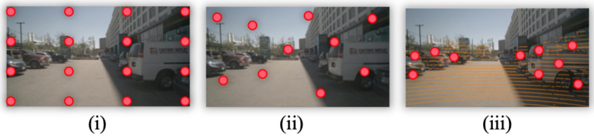

Figure 4. Illustration of ray sampling strategies: i) dilation, ii) random, and iii) depth-aware sampling.

where $\sigma_s(x) = (1 + e^{-sx})^{-1}$ is a Sigmoid function modulated by a learnable parameter s.

Memory-friendly Ray Sampling Previous novel view synthesis methods prioritize dense supervision to enhance image quality. However, rendering a complete set of $S \times H \times W$ rays — where S represents the number of camera views and $H \times W$ is the image resolution — presents substantial computational challenges, especially in the context of autonomous driving scenes.

To alleviate computational challenges, we devise three memory-friendly ray sampling strategies to render a reduced subset of rays: *Dilation Sampling*, *Random Sampling*, and *Depth-aware Sampling*, illustrated in Figure 4. 1) *Dilation Sampling* traverses the image at intervals of I, thereby reducing the ray count to $\frac{S \times H \times W}{I^2}$. 2) In contrast, *Random Sampling* selects K rays indiscriminately from all available pixels. 3) Although both dilation and random sampling are straightforward and significantly cut computation, they overlook the subtle prior information that is inherent to the 3D environment. For example, instances on the road generally contain more relevant information over distant backgrounds like sky and buildings. Therefore, we introduce *depth-aware sampling* to selectively sample rays informed by available LiDAR information, bypassing the need for a full pixel set. To implement this, we project point clouds onto the multi-view images and acquire the set of projection pixels with a depth less than the τ threshold. Subsequently, rays are selectively sampled from this refined pixel set as opposed to the entire array of image pixels. In doing so, our approach not only alleviates the computational burden but also enhances the learned representation by concentrating on the most relevant segments within the scene.

Pre-training Loss The overall pre-training loss consists of the color loss and depth loss:

$$\begin{aligned} L = & \frac{\lambda_{\text{RGB}}}{K} \sum_{i=1}^{K} |\hat{Y}_i^{\text{RGB}} - Y_i^{\text{RGB}}| \\ & + \frac{\lambda_{\text{depth}}}{K^+} \sum_{i=1}^{K^+} |\hat{Y}_i^{\text{depth}} - Y_i^{\text{depth}}|, \end{aligned} \tag{4}$$

where Y_i^{RGB} and Y_i^{depth} are the ground-truth color and depth for each ray, respectively. $\hat{Y}_i^{\text{RGB}}$ and $\hat{Y}_i^{\text{depth}}$ are the corresponding rendered ones in Eq. 2. K^+ is the count of rays with available depth.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct the experiments on the NuScenes [5] dataset, which is a challenging dataset for autonomous driving. It consists of 700 scenes for training, 150 scenes for validation, and 150 scenes for testing. Each scene is captured through six different cameras, providing images with surrounding views, and is accompanied by a point cloud from LiDAR. The dataset comes with diverse annotations, supporting tasks like 3D object detection and 3D semantic segmentation. For detection evaluation, we employ nuScenes detection score (NDS) and mean average precision (mAP), and for segmentation assessment, we use mean intersection-over-union (mIoU).

4.2. Implementation Details

We base our code on the MMDetection3D [17] toolkit and train all models on 4 NVIDIA A100 GPUs. The input image is configured to 1600 × 900 pixels, while the voxel dimensions for point cloud voxelization are $[0.075, 0.075, 0.2]$. During the pre-training phase, we implemented several data augmentation strategies, such as random scaling and rotation. Additionally, we partially mask the inputs, focusing only on visible regions for feature extraction. The masking size and ratio for images are configured to 32 and 0.3, and for points to 8 and 0.8, respectively. ConvNeXt-small [53] and VoxelNet [87] are adopted as the default image and point encoders, respectively. A uniform voxel representation with the shape of $180 \times 180 \times 5$ is constructed across modalities. The feature projection layer reduces the voxel feature dimensions to 32 via a 3-kernel size convolution. For the decoders, we utilize a 6-layer MLP for SDF and a 4-layer MLP for RGB. In the rendering phase, 512 rays per image view and 96 points per ray are randomly selected. We maintain the loss scale factors for λ_{RGB} and λ_{depth} at 10. The model undergoes training for 12 epochs using the AdamW optimizer with initial learning rates of $2e^{-5}$ and $1e^{-4}$ for point and image modalities, respectively. In the ablation studies, unless explicitly stated, fine-tuning is conducted for 12 epochs on 50% of the image data and for 20 epochs on 20% of the point data, without the use of CBGS [105] strategy and cut-and-paste [87] augmentation.

4.3. Comparison with State-of-the-Art Methods

3D Object Detection. In Table 1, we compare UniPAD with previous detection approaches on the nuScenes validation set. We adopt UVTR [38] as our baselines for point-modality (UVTR-L), camera-modality (UVTR-C), Camera-Sweep-modality (UVTR-CS), and fusion-modality (UVTR-M). Benefits from the effective pre-training, UniPAD consistently improves the baselines, namely, UVTR-L, UVTR-C, and UVTR-M, by 2.9, 2.4, and 3.0 NDS,

Table 1. Comparisons of different methods with a single model on the nuScenes *val* set. We compare with classic methods on different modalities *without* test-time augmentation. †: denotes our reproduced results based on MMDetection3D [17]. L, C, CS, and M indicate the LiDAR, Camera, Camera Sweep, and Multi-modality input, respectively.

Methods	Present at	Modality	CS	CBGS	NDS↑	mAP↑
PVT-SSD [91]	CVPR'23	L		✓	65.0	53.6
CenterPoint [97]	CVPR'21	L		✓	66.8	59.6
FSD [22]	NeurIPS'22	L		✓	68.7	62.5
VoxelNeXt [14]	CVPR'23	L		✓	68.7	63.5
LargeKernel3D [13]	CVPR'23	L		✓	69.1	63.3
TransFusion-L [2]	CVPR'22	L		✓	70.1	65.1
CMT-L [86]	ICCV'23	L		✓	68.6	62.1
UVTR-L [38]	NeurIPS'22	L		✓	67.7	60.9
UVTR-L+UniPAD (Ours)	-	L		✓	**70.6**	**65.0**
BEVFormer-S [40]	ECCV'22	C		✓	44.8	37.5
SpatialDETR [21]	ECCV'22	C			42.5	35.1
PETR [50]	ECCV'22	C		✓	44.2	37.0
Ego3RT [55]	ECCV'22	C			45.0	37.5
3DPPE [65]	ICCV'23	C		✓	45.8	39.1
BEVFormerV2 [88]	CVPR'23	C			46.7	39.6
CMT-C [86]	ICCV'23	C		✓	46.0	40.6
FCOS3D† [73]	ICCVW'21	C			38.4	31.1
FCOS3D+UniPAD (Ours)	-	C			**40.1**	**33.2**
UVTR-C [38]	NeurIPS'22	C			45.0	37.2
UVTR-C+UniPAD (Ours)	-	C			**47.4**	**41.5**
UVTR-CS [38]	NeurIPS'22	C	✓		48.8	39.2
UVTR-CS+UniPAD (Ours)	-	C	✓		**50.2**	**42.8**
PointPainting [70]	CVPR'20	C+L		✓	69.6	65.8
MVP [98]	NeurIPS'21	C+L		✓	70.8	67.1
TransFusion [2]	CVPR'22	C+L		✓	71.3	67.5
AutoAlignV2 [15]	ECCV'22	C+L		✓	71.2	67.1
BEVFusion [43]	NeurIPS'22	C+L		✓	71.0	67.9
BEVFusion [54]	ICRA'23	C+L		✓	71.4	68.5
ObjectFusion [6]	ICCV'23	C+L		✓	72.3	69.8
DeepInteraction [93]	NeurIPS'22	C+L		✓	72.6	69.9
SparseFusion [81]	ICCV'23	C+L		✓	72.8	70.4
CMT-M [86]	ICCV'23	C+L		✓	72.9	70.3
UVTR-M [38]	NeurIPS'22	C+L		✓	70.2	65.4
UVTR-M+UniPAD (Ours)	-	C+L		✓	**73.2**	**69.9**

Table 2. Comparisons of different methods with a single model on the nuScenes segmentation dataset.

Methods	Modality	Backbone	Split	
			val	*test*
RangeFormer [34]	L	Transformer	78.1	80.1
SphereFormer [36]	L	Transformer	78.4	81.9
WaffleIron [62]	L	Conv2D	79.1	-
SPVNAS [66]	L	SpConv	-	77.4
Cylinder3D [107]	L	SpConv	76.1	77.2
SpUNet [16]	L	SpConv	73.3	-
SpUNet+UniPAD (Ours)	L	SpConv	**79.4**	**81.1**

respectively. When taking multi-frame cameras as inputs, UniPAD-CS brings 1.4 NDS and 3.6 mAP gains over UVTR-CS. Our pre-training technique also achieves 1.7 NDS and 2.1 mAP improvements over the monocular-based baseline FCOS3D [73]. Without any test time augmentation or model ensemble, our single-modal and multi-modal methods, UniPAD-L, UniPAD-C, and UniPAD-M, achieve impressive NDS of 70.6, 47.4, and 73.2, respectively, reaching the level of existing state-of-the-art methods.

3D Semantic Segmentation. In Table 2, we compare UniPAD with previous point cloud semantic segmentation approaches on the nuScenes Lidar-Seg dataset. We adopt SpUNet [16] implemented by Pointcept [18] as our baseline. Benefiting from effective pre-training, UniPAD improves the baselines by 6.1 mIoU, achieving state-of-the-art performance on the validation set. Meanwhile, UniPAD achieves an impressive mIoU of 81.1 on the *test* set, which is comparable with existing state-of-the-art methods.

4.4. Comparisons with Pre-training Methods.

Image-based Pre-training. In Table 3, we conduct comparisons between UniPAD and several other image-based pre-training approaches: 1) Depth Estimator: we follow [60] to inject 3D priors into 2D learned features via depth estimation; 2) Detector: the image encoder is initialized using pre-trained weights from MaskRCNN [25] on the nuImages dataset [5]; 3) 3D Detector: the weights from the widely used monocular 3D detector [73] is used for

Table 3. Comparison with different image-based pre-training.

Methods	Label		NDS	mAP
	2D	3D		
UVTR-C (Baseline)			25.2	23.0
+Depth Estimator			$26.9^{\uparrow 1.7}$	$25.1^{\uparrow 2.1}$
+Detector	✓		$29.4^{\uparrow 4.2}$	$27.7^{\uparrow 4.7}$
+3D Detector		✓	$31.7^{\uparrow 6.5}$	$29.0^{\uparrow 6.0}$
+UniPAD			$32.9^{\uparrow 7.7}$	$32.6^{\uparrow 9.6}$

Table 4. Comparison with different point-based pre-training.

Methods	Support		NDS	mAP
	2D	3D		
UVTR-L (Baseline)			46.7	39.0
+Occupancy-based		✓	$48.2^{\uparrow 1.5}$	$41.2^{\uparrow 2.2}$
+MAE-based		✓	$48.8^{\uparrow 2.1}$	$42.6^{\uparrow 3.6}$
+Contrast-based	✓	✓	$49.2^{\uparrow 2.5}$	$48.8^{\uparrow 9.8}$
+UniPAD	✓	✓	$55.8^{\uparrow 9.1}$	$48.1^{\uparrow 9.1}$

model initialization, which relies on 3D labels for supervision. UniPAD demonstrates superior knowledge transfer capabilities compared to previous unsupervised or supervised pre-training methods, showcasing the efficacy of our rendering-based pretext task.

Point-based Pre-training. For point modality, we also present comparisons with recently proposed self-supervised methods in Table 4: 1) Occupancy-based: we implement ALSO [4] in our framework to train the point encoder; 2) MAE-based: the leading-performing method [90] is adopted, which reconstructs masked point clouds using the chamfer distance. 3) Contrast-based: [49] is used for comparisons, which employs pixel-to-point contrastive learning to integrate 2D knowledge into 3D points. Among these methods, UniPAD achieves the best NDS performance. While UniPAD has a slightly lower mAP compared to the contrast-based method, it avoids the need for complex positive-negative sample assignments in contrastive learning. More implementation details will be provided in the supplementary material.

4.5. Effectiveness on Various Backbones.

Different View Transformations. In Table 5, we investigate different view transformation strategies for converting 2D features into 3D space, including BEVDet [31], BEVDepth [39], and BEVformer [40]. Due to the prevalent use of BEV representation, we integrate these methods into our framework by transforming features into volumetric representations. Consistent improvements ranging from 5.2 to 6.3 NDS can be observed across different transformation techniques, which demonstrates the strong generalization ability of the proposed approach.

Different Modalities. Unlike most previous pre-training methods, our framework can be seamlessly applied to various modalities. To verify the effectiveness of our approach, we set UVTR as our baseline, which contains detectors with point, camera, and fusion modalities. Table 6 shows the impact of UniPAD on different modalities. UniPAD consistently improves the UVTR-L, UVTR-C, and UVTR-M by 9.1, 7.7, and 6.9 NDS, respectively.

Table 5. Pre-training effect on different view transformations.

Methods	View Transform	NDS	mAP
BEVDet	Pooling	27.1	24.6
+UniPAD	Pooling	$32.7^{\uparrow 5.6}$	$32.8^{\uparrow 8.2}$
BEVDepth	Pooling & Depth	28.9	28.1
+UniPAD	Pooling & Depth	$34.1^{\uparrow 5.2}$	$33.9^{\uparrow 5.8}$
BEVformer	Transformer	26.8	24.5
+UniPAD	Transformer	$33.1^{\uparrow 6.3}$	$31.9^{\uparrow 7.4}$

Table 6. Pre-training effectiveness on different input modalities.

Methods	Modality	NDS	mAP
UVTR-L	LiDAR	46.7	39.0
+UniPAD	LiDAR	$55.8^{\uparrow 9.1}$	$48.1^{\uparrow 9.1}$
UVTR-C	Camera	25.2	23.0
+UniPAD	Camera	$32.9^{\uparrow 7.7}$	$32.6^{\uparrow 9.6}$
UVTR-M	LiDAR-Camera	49.9	52.7
+UniPAD	LiDAR-Camera	$56.8^{\uparrow 6.9}$	$57.0^{\uparrow 4.3}$

Scaling up Backbones. To test UniPAD across different backbone scales, we adopt an off-the-shelf model, ConvNeXt, and its variants with different numbers of learnable parameters. As shown in Table 7, one can observe that with our UniPAD pre-training, all baselines are improved by large margins of +6.0∼7.7 NDS and +8.2∼10.3 mAP. The steady gains suggest that UniPAD has the potential to boost various state-of-the-art networks.

4.6. Ablation Studies

Masking Ratio. Table 8 shows the influence of the masking ratio for the camera modality. We discover that a masking ratio of 0.3, which is lower than the ratios used in previous MAE-based methods, is optimal for our method. This discrepancy could be attributed to the challenge of rendering the original image from the volume representation, which is more complex compared to image-to-image reconstruction. For the point modality, we adopt a mask ratio of 0.8, as suggested in [90], considering the spatial redundancy inherent in point clouds.

Rendering Design. Our examinations in Tables 9, 10, and 11 illustrate the flexible design of our differentiable rendering. In Table 9, we vary the depth $(D_{\text{SDF}}, D_{\text{RGB}})$ of the SDF and RGB decoders, revealing the importance of suf-

Table 7. Pre-training effectiveness on different backbone scales.

Methods	Backbone		
	ConvNeXt-S	ConvNeXt-B	ConvNeXt-L
UVTR-C	25.2/23.0	26.9/24.4	29.1/27.7
+UniPAD	$32.9^{\uparrow 7.7}/32.6^{\uparrow 9.6}$	$34.1^{\uparrow 7.2}/34.7^{\uparrow 10.3}$	$35.1^{\uparrow 6.0}/35.9^{\uparrow 8.2}$

Table 8. Ablation studies of the masking ratio.

Ratio	0.1	0.3	0.5	0.7	0.9
NDS	31.9	**32.9**	32.3	32.1	31.4

Table 9. Ablation studies of the decoder depth.

Layer	(2, 2)	(4, 3)	(5, 4)	(6, 4)	(7, 5)
NDS	31.3	31.9	32.1	**32.9**	32.7

Table 10. Ablation studies of the decoder width.

Dim.	16	32	64	128	256
NDS	32.1	**32.9**	32.5	**32.9**	32.4

ficient decoder depth for succeeding in downstream detection tasks. This is because deeper ones may have the ability to adequately integrate geometry or appearance cues during pre-training. Conversely, as reflected in Table 10, the width of the decoder has a relatively minimal impact on performance. Thus, the default dimension is set to 32 for efficiency. Additionally, we explore the effect of various rendering techniques in Table 11, which employ different ways for ray point sampling and accumulation. Using NeuS [72] for rendering records a 0.4 and 0.1 NDS improvement compared to UniSurf [58] and VolSDF [95], respectively, showcasing the learned representation can be improved by utilizing well-designed rendering methods and benefiting from the advancements in neural rendering.

Memory-friendly Ray Sampling. Instead of rendering the entire set of multi-view images, we sample only a subset of rays to provide supervision signals. Table 12 outlines the different strategies explored to minimize memory usage and computational costs during pre-training. Our observations indicate that depth-aware sampling holds a substantial advantage, improving scores by 0.4 and 1.0 NDS compared to random sampling ($K = 512$) and dilation sampling ($I = 16$), respectively. The sampling excludes regions without well-defined depth, like the sky, from contributing to the loss. This allows the representation learning to focus more on the objects in the scene, which is beneficial for downstream tasks. Meanwhile, it costs less memory usage than dilation sampling.

Feature Projection. The significance of feature projection is shown in Table 13. Removing projection from pre-training and fine-tuning leads to drops of 1.8 and 2.7 NDS, respectively, underscoring the essential role it plays in enhancing voxel representation. Concurrently, utilizing shared parameters for the projection during pre-training and fine-tuning induces reductions of 0.8 NDS and 0.6 mAP. This phenomenon is likely due to the disparity between the rendering and recognition tasks, with the final layers being more tailored for extracting features specific to each task.

Table 11. Ablation studies of the rendering technique.

Methods	NDS	mAP
UniSurf [58]	32.5	32.1
VolSDF [95]	32.8	32.4
NeuS [72]	**32.9**	**32.6**

Table 12. Ablation studies of the sampling strategy.

Methods	Memory	NDS	mAP
Dilation Sampling	1.4×	31.9	32.4
Random Sampling	**1**×	32.5	32.1
Depth-aware Sampling	**1**×	**32.9**	**32.6**

Table 13. Ablation studies of the feature projection.

Methods	NDS	mAP
Baseline	**32.9**	**32.6**
w/o Projection$_{\mathrm{FT}}$	$30.2^{\downarrow 2.7}$	$29.7^{\downarrow 2.9}$
w/o Projection$_{\mathrm{PT}}$	$31.1^{\downarrow 1.8}$	$30.5^{\downarrow 2.1}$
Shared Projection	$32.1^{\downarrow 0.8}$	$32.0^{\downarrow 0.6}$

Table 14. Ablation studies of the pre-trained components.

Methods	NDS	mAP
Baseline	25.2	23.0
+Encoder	$32.0^{\uparrow 6.8}$	$31.8^{\uparrow 8.8}$
+Encoder & FPN	$32.2^{\uparrow 0.2}$	$32.2^{\uparrow 0.4}$
+Encoder & FPN & VT	$\mathbf{32.9}^{\uparrow 0.7}$	$\mathbf{32.6}^{\uparrow 0.4}$

Pre-trained Components. In Table 14, the influence of pre-trained parameters on each component is investigated. Replacing the pre-trained weights of the FPN and view transformation (VT) with those from a random initialization induces declines of 0.2 and 0.7 NDS, respectively, thereby highlighting the crucial roles of these components.

5. Conclusion

We present UniPAD, an innovative self-supervised learning paradigm that excels in various 3D perception tasks. UniPAD stands out for its ingenious adaptation of NeRF as a unified rendering decoder, enabling seamless integration into both 2D and 3D frameworks. The inherent adaptability of our approach bridges the 2D and 3D domains, which could facilitate representation learning through advancements in the other domain. For instance, semantic knowledge can be infused into point clouds via additional semantic supervision, leveraging the outputs of well-developed models like SAM [33] in the 2D domain as learning targets.

Limitation. There are still certain limitations to the approach. For instance, we need to explicitly transform point and image features into volumetric representations, which would increase memory usage as voxel resolution rises.

Acknowledgement This work was supported in part by The National Nature Science Foundation of China (Grant Nos: 62273303), in part by Yongjiang Talent Introduction Programme (Grant No: 2022A-240-G), in part by Ningbo Key R&D Program (No. 2023Z231, 2023Z229), in part by the National Key R&D Program of China (NO. 2022ZD0160101).

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 6

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 2

[4] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. ALSO: automotive lidar self-supervision by occupancy estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 7

[5] 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 Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 5, 6

[6] Qi Cai, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Objectfusion: Multi-modal 3d object detection with object-centric fusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 6

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[8] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 2023. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020. 1

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[11] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 2

[12] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2

[13] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6

[14] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6

[15] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Deformable feature aggregation for dynamic multi-modal 3d object detection. In *Proceedings of the European Conference on Computer Vision*, 2022. 6

[16] Christopher B. Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 6

[17] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d`, 2020. 5, 6

[18] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. `https://github.com/Pointcept/Pointcept`, 2023. 6

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009. 2

[20] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 1

[21] Simon Doll, Richard Schulz, Lukas Schneider, Viviane Benzin, Markus Enzweiler, and Hendrik P. A. Lensch. Spatialdetr: Robust scalable transformer-based 3d object detection from multi-view camera images with global cross-sensor attention. In *Proceedings of the European Conference on Computer Vision*, 2022. 6

[22] Lue Fan, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Fully sparse 3d object detection. In *Advances in Neural Information Processing Systems*, 2022. 6

[23] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *International Conference on 3D Vision*, 2022. 2

[24] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *CoRR*, abs/2205.03892, 2022. 2

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 6

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

[28] Georg Hess, Johan Jaxing, Elias Svensson, David Hagerman, Christoffer Petersson, and Lennart Svensson. Masked autoencoders for self-supervised learning on automotive point clouds. *CoRR*, abs/2207.00531, 2022. 1, 2

[29] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[30] Di Huang, Sida Peng, Tong He, Honghui Yang, Xiaowei Zhou, and Wanli Ouyang. Ponder: Point cloud pre-training via neural rendering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 2

[31] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *CoRR*, abs/2112.11790, 2021. 7

[32] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 8

[34] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 6

[35] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[36] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 6

[37] Lanxiao Li and Michael Heizmann. A closer look at invariances in self-supervised pre-training for 3d vision. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[38] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 5, 6

[39] 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*, 2023. 7

[40] 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 *Proceedings of the European Conference on Computer Vision*, 2022. 2, 6, 7

[41] Zhuopeng Li, Lu Li, and Jianke Zhu. Read: Large-scale neural scene rendering for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 3

[42] Hanxue Liang, Chenhan Jiang, Dapeng Feng, Xin Chen, Hang Xu, Xiaodan Liang, Wei Zhang, Zhenguo Li, and Luc Van Gool. Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2

[43] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In *Advances in Neural Information Processing Systems*, 2022. 2, 6

[44] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[45] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[46] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 2

[47] Jiaheng Liu, Tong He, Honghui Yang, Rui Su, Jiayi Tian, Junran Wu, Hongcheng Guo, Ke Xu, and Wanli Ouyang. 3d-queryis: A query-based framework for 3d instance segmentation. *CoRR*, abs/2211.09375, 2022. 1

[48] Jiahui Liu, Chirui Chang, Jianhui Liu, Xiaoyang Wu, Lan Ma, and Xiaojuan Qi. Mars3d: A plug-and-play motion-aware model for semantic segmentation on multi-scan 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[49] Yueh-Cheng Liu, Yu-Kai Huang, HungYueh Chiang, Hung-Ting Su, Zhe Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2d: Pixel-to-point knowledge transfer for 3d pretraining. *CoRR*, abs/2104.04687, 2021. 7

[50] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: position embedding transformation for multi-view 3d object detection. In *Proceedings of the European Conference on Computer Vision*, 2022. 2, 6

[51] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2022. 1

[52] Youquan Liu, Runnan Chen, Xin Li, Lingdong Kong, Yuchen Yang, Zhaoyang Xia, Yeqi Bai, Xinge Zhu, Yuexin

Ma, Yikang Li, Yu Qiao, and Yuenan Hou. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 1

[53] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 5

[54] 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 *International Conference on Robotics and Automation*, 2023. 6

[55] Jiachen Lu, Zheyuan Zhou, Xiatian Zhu, Hang Xu, and Li Zhang. Learning ego 3d representation as ray tracing. In *Proceedings of the European Conference on Computer Vision*, 2022. 6

[56] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[57] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *CoRR*, abs/2206.09900, 2022. 2

[58] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2, 4, 8

[59] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[60] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2, 6

[61] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 3

[62] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Using a waffle iron for automotive point cloud semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 6

[63] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[64] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3d object detection. *CoRR*, abs/2102.00463, 2021. 1

[65] Changyong Shu, JIajun Deng, Fisher Yu, and Yifan Liu. 3dppe: 3d point positional encoding for multi-camera 3d object detection transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 6

[66] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *Proceedings of the European Conference on Computer Vision*, 2020. 6

[67] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. In *International Conference on Learning Representations*, 2023. 2, 3

[68] Xiaoyu Tian, Haoxi Ran, Yue Wang, and Hang Zhao. Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[69] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 2

[70] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 6

[71] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2

[72] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, 2021. 4, 8

[73] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2021. 2, 6

[74] Wei-Chien Wang, Euijoon Ahn, Dagan Feng, and Jinman Kim. A review of predictive and contrastive self-supervised learning for medical images. *Machine Intelligence Research*, 2023. 2

[75] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 2023. 2

[76] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer V2: grouped vector attention and partition-based pooling. In *Advances in Neural Information Processing Systems*, 2022. 1

[77] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. *CoRR*, abs/2308.09718, 2023. 2

[78] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, 2023. 2

[79] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, Yuxin Huang, Xiaoyu Ye, Zike Yan, Yongliang Shi, Yiyi Liao, and Hao Zhao. MARS: an instance-aware, modular and realistic simulator for autonomous driving. *CoRR*, abs/2307.15058, 2023. 2

[80] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 2

[81] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6

[82] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 2

[83] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 2

[84] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[85] Runsen Xu, Tai Wang, Wenwei Zhang, Runjian Chen, Jinkun Cao, Jiangmiao Pang, and Dahua Lin. MV-JAR: masked voxel jigsaw and reconstruction for lidar-based self-supervised pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[86] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer via coordinates encoding for 3d object dectection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 2, 6

[87] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 3, 5

[88] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6

[89] Honghui Yang, Zili Liu, Xiaopei Wu, Wenxiao Wang, Wei Qian, Xiaofei He, and Deng Cai. Graph R-CNN: towards accurate 3d object detection with semantic-decorated local graph. In *Proceedings of the European Conference on Computer Vision*, 2022. 1

[90] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. GD-MAE: generative decoder for MAE pre-training on lidar point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 7

[91] Honghui Yang, Wenxiao Wang, Minghao Chen, Binbin Lin, Tong He, Hua Chen, Xiaofei He, and Wanli Ouyang. PVT-SSD: single-stage 3d object detector with point-voxel transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6

[92] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[93] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. In *Advances in Neural Information Processing Systems*, 2022. 6

[94] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[95] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems*, 2021. 8

[96] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[97] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 6

[98] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. In *Advances in Neural Information Processing Systems*, 2021. 6

[99] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[100] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. *CoRR*, abs/2304.14811, 2023. 2

[101] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *Advances in Neural Information Processing Systems*, 2022. 2

[102] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any

point-cloud. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1, 2

[103] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[104] Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[105] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *CoRR*, abs/1908.09492, 2019. 5

[106] Haoyi Zhu, Haoshu Fang, and Cewu Lu. X-nerf: Explicit neural radiance field for multi-scene 360° insufficient RGB-D views. *CoRR*, abs/2210.05135, 2022. 2

[107] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6