

Towards Transferable Multi-modal Perception Representation Learning for Autonomy: NeRF-Supervised Masked AutoEncoder

Xiaohao Xu*

Abstract—This work proposes a unified self-supervised pre-training framework for transferable multi-modal perception representation learning via masked multi-modal reconstruction in Neural Radiance Field (NeRF), namely NeRF-Supervised Masked AutoEncoder (NS-MAE). Specifically, conditioned on certain view directions and locations, multi-modal embeddings extracted from corrupted multi-modal input signals, *i.e.*, Lidar point clouds and images, are rendered into projected multi-modal feature maps via neural rendering. Then, original multi-modal signals serve as reconstruction targets for the rendered multi-modal feature maps to enable self-supervised representation learning. Extensive experiments show that the representation learned via NS-MAE shows promising transferability for diverse multi-modal and single-modal (camera-only and Lidar-only) perception models on diverse 3D perception downstream tasks (3D object detection and BEV map segmentation) with diverse amounts of fine-tuning labeled data. Moreover, we empirically find that NS-MAE enjoys the synergy of both the mechanism of masked autoencoder and neural radiance field. Our code shall be released upon acceptance.

I. INTRODUCTION

Toward robust autonomous driving, multi-modal perception [18], [25], which aims to sense the surrounding scene by extracting and fusing representations from diverse modalities, is a crucial research direction. Meanwhile, with the booming development of modern network architectures [76], [5], great efforts on transferable representation learning have been witnessed to relieve the appetite for data [26], [2]. However, **pre-training for multi-modal perception is in its infancy**.

We first review the widely-adopted multi-stage fully-supervised training paradigm for current advanced multi-modal perception models [45], [49], [92]. In this paradigm, the networks of Lidar and camera branches are first pre-trained separately in a fully-supervised learning way, and then the whole architecture is jointly fine-tuned. However, such a paradigm is not scalable due to the scarcity of labels for supervision. Specifically, annotating paired images and extremely sparse Lidar point clouds for 3D perception can be cumbersome [80], thus high-quality 3D labels are scarce. To relieve this problem, some self-supervised pre-training methods have been raised for single-modal perception models [59], [41], [29], [28], [55] to enable label-efficient transfer. However, their optimization formulations are not unified and there is no work to explore transferable representation learning for advanced multi-modal perception models.

X.H. Xu is with the Robotics Institute, University of Michigan, Ann Arbor, MI, USA. (xiaohaox@umich.edu).

The author would like to thank the suggestions from X.Z. Zhu, H. Tian, and L.W. Lu during the preliminary development stage of this work.

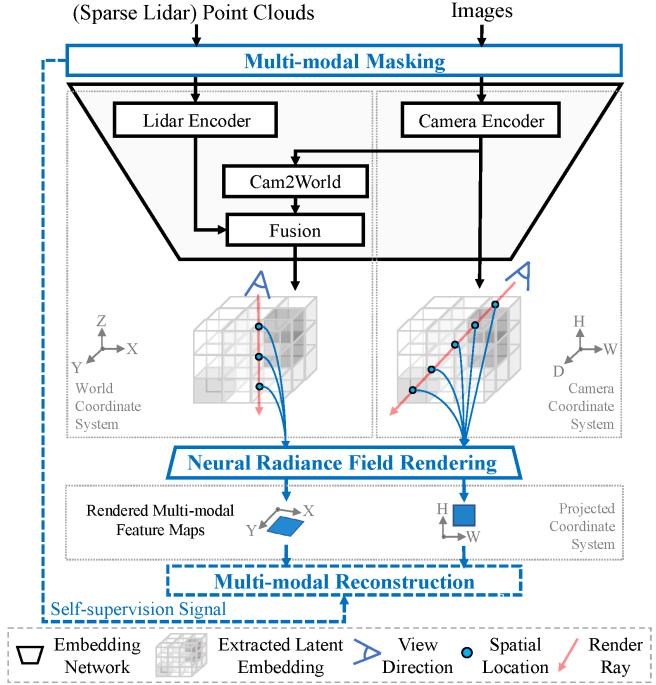


Fig. 1. Overview of our unified pre-training framework for transferable multi-modal perception representation learning. Blue parts denote the proposed plug-and-play components for representation learning. Multi-modal inputs, *i.e.*, sparse Lidar point clouds and images, are first partially masked out and sent to multi-modal encoders, *i.e.*, the embedding network, for latent representation encoding. Then, given specific view directions and locations, the extracted embeddings are rendered into projected feature maps of various modalities, *e.g.*, color and projected point cloud maps, via a differential neural volume rendering mechanism. Finally, the rendered feature maps are supervised by original inputs via self-supervised multi-modal reconstruction.

To this end, we go on to think about **what should the ideal multi-modal perception pre-training framework for transferable representation learning be like?** Inspired by successful pre-training frameworks for vision [26] and language [16], apart from the effectiveness to boost performance, we argue that the following features are crucial:

- **Unified:** it should be generic to various settings, *e.g.*, input modalities, architectures, and downstreams.
- **Scalable:** it should embrace the huge unlabeled yet valuable multi-modal data and potentially huge models.
- **Neat:** it should possess a simple formulation to be adapted to more diverse modalities for perception.

With these high-level goals in mind, we decouple the design of the pre-training framework for multi-modal perception into two sub-problems: (1) **learning** transferable multi-modal representation; (2) **unifying** the optimization formulation for multi-modal representation.

On one hand, to learn multi-modal representation in a self-supervised way, we get inspiration from the great success of Masked AutoEncoder [26], [91], [2] (MAE) paradigm. Specifically, MAE [26] follows a mask-then-reconstruct paradigm and shows inspiring transferability on various downstream tasks. To take a step forward, we would like to explore the possibility of adapting MAE to learn transferable multi-modal perception representation, thus empowering advanced multi-modal perception models [49], [45].

On the other hand, to unify the multi-modal representation optimization, Neural Radiance Field (NeRF) [1] provides a neat form to encode diverse physical properties, *e.g.*, color and geometry, of the scene via differential neural volume rendering procedure. We consider NeRF as a useful building block for multi-modal perception pre-training for two reasons. Firstly, the imaging process of optical systems for perception, *e.g.*, Lidar and camera, can be approximately modeled with rendering in the radiance field. Secondly, neural rendering unifies the multi-modal reconstruction via a neat and explainable physics formula. Thus, we introduce the rendering process of NeRF to perception pre-training, enabling unified multi-modal representation optimization.

Inspired by the remarkable attainments achieved by MAE and NeRF, we make a synergy of them and propose a unified self-supervised multi-modal perception pre-training framework (NS-MAE), which learns transferable multi-modal representations in a conceptually-neat formulation. As is shown in Fig. 1, we assemble the plug-and-play pre-training components (blue parts of Fig. 1) to a typical embedding network of multi-modal perception models [49] (black parts of Fig. 1). During the pre-training, we first conduct modality-specific masking operations for multi-modal inputs, *i.e.*, images and Lidar point clouds, separately. Then, we send the corrupted modalities to the multi-modal embedding network for embedding extraction. Specifically, we extract the embeddings that are generated and fused in two crucial coordination, *i.e.*, the world and the camera coordination, for perception models [43], [61], [69]. Afterward, according to specific view directions and spatial sampling locations, the embeddings are further rendered into diverse projected modality feature maps via differential volume rendering. Finally, the rendered feature maps are supervised by original images and point clouds via reconstruction-based self-supervised optimization.

At last, we evaluate the transferability of multi-modal representation learned via NS-MAE for diverse multi-modal and single-modal perception models.

In summary, this work has the following contributions:

- We propose a **novel unified self-supervised pre-training framework for multi-modal perception representation learning**, *i.e.*, NS-MAE, which is generic to both multi-modal and single-modal perception models.
- We enable both the **self-supervised learning** and **optimization unification** of transferable multi-modal representation via plug-and-play designs in the spirit of multi-modal reconstruction in neural radiance field.
- We employ NS-MAE to various advanced single-modal and multi-modal perception models and **verify the transferability of multi-modal representation derived via NS-MAE** on diverse 3D perception tasks with diverse amounts of fine-tuning data.

II. RELATED WORKS

Masked AutoEncoder (MAE) paradigm [4], [9], [22], which originates from masked language modeling [16] in natural language processing, learns transferable visual representation with masked signal reconstruction from partial observation. BEiT [4] and its subsequent works [17], [60], [38], [21] leverage a pre-trained tokenizer and reconstruct masked image patches in the token space. MAE [26] and SimMIM [88] demonstrate that directly reconstructing masked patches in raw pixel space can also lead to good transferability and scalability. Other works perform reconstruction in a high-level feature space [99], [20], [12] or handcrafted feature space [86]. Enlightened by these methods, we would like to tap the potential of introducing MAE to multi-modal representation learning for perception in this work.

Neural Radiance Field (NeRF) [54] shows impressive view synthesis results by using implicit functions to encode volumetric density and color observations. To improve the few-shot generalization ability of NeRF, data-driven priors recovered from general training data [97], other tasks [33], [87], or mixed priors [8] are leveraged to fill in missing information of test scenes. For efficient NeRF training, some works [47], [57], [87], [15] attempt to regulate the 3D geometry with depth. Considering the neat formulation of the neural rendering mechanism to optimize both the appearance and geometry, we introduce NeRF in perception pre-training.

Perception Models for Autonomous Driving include the following three main categories: (1) **Camera-only** models [52], [68], [50], [35], [98], [100], [67], [79], [66], [78] initially focus on monocular 3D detection [24]. Thanks to the emergence of larger benchmarks [7], [18], researchers start to study perception with range-view images [81], [82], [85]. Later, LSS [62], which transforms perspective camera feature maps into 3D Ego-car coordinate, helps shift the perspective-view perception into bird's eye view (BEV) perception [31], [30], [67], thus largely boosting the performance. (2) Early **Lidar-only** methods either operate on raw Lidar point clouds [65], [64], [73], [93], [42] or transform original point clouds into voxel [102] or pillar representation [36], [84], [94]. Later, these two feature representations are unified in one single model [11], [101], [71]. (3) **Multi-modal** perception models [32], [74], [77], [95], [96], [3], [40], [92], which unleash the complementary power of multiple modalities, become the de-facto standard in 3D perception. Recent works [49], [45] propose more effective and robust multi-modal models via disentangled modality modeling. Despite some attempts to explore pre-training for single-modal perception [83], [53], [23], [59], [29], [41], [55], [44], there is no trial for multi-modal perception models. Thus, we hope to explore representation learning for perception models with both single-modality and multi-modality in a unified form.

III. OUR APPROACH

A. Overview

Problem Formulation of Perception Pre-training. Given a tuple of images $\mathbf{I} = (I_1, I_2, \dots, I_N \in \mathbb{R}^{H \times W \times 3})$ collected from N views with their corresponding camera parameters $(\mathbf{P}_1 \mathbf{K}_1, \mathbf{P}_2 \mathbf{K}_2, \dots, \mathbf{P}_N \mathbf{K}_N)$ (where \mathbf{P} and \mathbf{K} denote camera pose and intrinsic matrix) and sparse Lidar point clouds $\mathcal{P} = \{p_i = [x_i, y_i, z_i, r_i] \in \mathbb{R}^4\}_{i=1, \dots, t}$,¹ the goal is to design a proxy task to learn the parameter set, *i.e.*, transferable representation, of an embedding network ϕ_{emb} , which can be used to initialize the parameter set of the downstream perception model $\phi_{down}(\supset \phi_{emb})$ for further fine-tuning.

Pipeline of NS-MAE Pre-training (as is shown in Fig. 1) includes the following three key steps to enable the transferable multi-modal perception representation learning:

- (1) **Masking** (Sec. III-C): The inputs, *i.e.*, images, and voxelized Lidar point clouds, are separately masked;
- (2) **Rendering** (Sec. III-D): The embeddings encoded from masked images and point clouds are rendered into color and projected point cloud feature maps via neural rendering [1];
- (3) **Reconstruction** (Sec. III-E): The rendered results are supervised by the ground-truth images and point clouds via multi-modal reconstruction-based optimization.

B. Training Architecture

As is shown in Fig. 1, the pre-training architecture, *i.e.*, a typical embedding network of multi-modal perception models, comprises the following components:

Camera Encoder takes masked images as input and generates the perspective-view (PER) image embedding $\mathbf{e}_I^{PER} \in \mathbb{R}^{H/\kappa \times W/\kappa \times D \times C}$ (κ denotes the down-sampling ratio). It can be implemented with Transformer-based [76], [48] or convolution-neural-network-based [27] architectures.

Lidar Encoder takes masked Lidar voxels as input and generates BEV embedding of Lidar modality $\mathbf{e}_L^{BEV} \in \mathbb{R}^{X \times Y \times Z \times C_L}$. Following common practice of 3D perception [49], [45], it is typically implemented with VoxelNet [102].

Cam2World module is used to transform the perspective-view image embedding \mathbf{e}_I^{PER} of camera coordination to BEV embedding $\mathbf{e}_I^{BEV} \in \mathbb{R}^{X \times Y \times Z \times C_I}$ of world coordination. Thus, image and Lidar embeddings can be aligned. We follow the implementation of lift-splat-shoot [62].

Fusion block fuses the BEV embedding of camera branch \mathbf{e}_I^{BEV} and Lidar branch \mathbf{e}_L^{BEV} , to generate fused multi-modal BEV embedding $[\mathbf{e}_I^{BEV}; \mathbf{e}_L^{BEV}] \in \mathbb{R}^{X \times Y \times Z \times (C_I + C_L)}$ via a simple concatenation $([.;.])$.

C. Multi-modal Masking

Image Masking. Following the MAE schema [26], [88], the original unmasked image I is first divided into regular non-overlapping image patches. Then, a random binary mask $M \in \{0, 1\}^{H \times W}$

¹ x, y, z denotes the position of a point in the world coordinate space; the Lidar intensity r is optional for the input but is used as common practice for the Lidar-based encoder of perception models [102] to boost performance.

is applied to mask out a large portion of image patches by replacing them with a learnable [MASK] token ($\in \mathbb{R}^{s \times s \times 3}$, where $s \times s$ denotes the patch size). Afterward, the image that is partially masked out is sent to the camera encoder for embedding extraction.

Lidar Masking. Inspired by the great success of MAE-style visual pre-training, some recent works [55], [28], [58] extend it for point cloud pertaining. Similarly, after transforming the input Lidar point cloud into its voxelized form, we mask a large fraction of non-empty voxels (70% to 90%). Then, the partially-masked voxels are processed in the Lidar encoder to generate the Lidar embedding.

D. Neural Radiance Field Rendering

1) **Rendering Network: Vanilla Rendering Network** of NeRF [1] takes a set of posed images and encodes the scene with volume density and emitted radiance for the purpose of view synthesis. In NeRF, a rendering network f maps a given 3D point $\mathbf{x} \in \mathbb{R}^3$ (\mathbb{R}^3 denotes the scene’s world space) and a particular viewing direction $\omega \in \mathbb{S}^2$ (\mathbb{S}^2 denotes the sphere of directions) to the differential sigma field density $\sigma \in \mathbb{R}$ and RGB color $\mathbf{c} \in \mathbb{R}^3$, like so: $f(\mathbf{x}, \omega) = (\sigma, \mathbf{c})$.

Conditional Rendering Network for Pre-training. As our goal (representation learning) is different from the goal of vanilla NeRF (view synthesis), we leverage a different rendering network formulation. In specific, we additionally introduce a latent multi-modal embedding \mathbf{e} from the embedding network of the perception model to the inputs of the rendering network f , like so: $f(\mathbf{x}, \omega, \mathbf{e}) = (\sigma, \mathbf{c})$. Thus, the gradient from differential rendering and the following reconstruction-based self-supervision stages can be back-propagated to the embedding network for end-to-end representation learning.

2) **Rendering Target: Vanilla Color Rendering.** Given the pose \mathbf{P} and intrinsic \mathbf{K} of a virtual camera, we shoot rays $\mathbf{r}(t) = \mathbf{o} + td$ originating from the \mathbf{P} ’s center of projection \mathbf{o} in direction ω derived from its intrinsic \mathbf{K} to render the RGB color $\hat{\mathbf{C}}(\mathbf{r})$ via standard volume rendering [34], which is formulated as:

$$\hat{\mathbf{C}}(t) = \int_0^\infty T(t)\sigma(t)\mathbf{c}(t)dt, \quad (1)$$

where $\mathbf{c}(t)$ and $\sigma(t)$ are the differential color radiance and density, and $T(t) = \exp(-\int_0^t \sigma(s)ds)$ checks for occlusions by integrating the differential density between 0 to t . Specifically, the discrete form can be approximated as:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^N T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i, \quad (2)$$

where N is the number of sampled points along the ray, $\delta_i = t_{i+1} - t_i$ is the distance between two adjacent ray samples and the accumulated transmittance T_i is $\exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$.

Multi-modal Rendering for Pre-training. For multi-modal pre-training, the rendering targets can be extended to unleash the power of multi-modal data. In specific, apart from the color that reflects the semantics of the scene, the Lidar ray

that captures 3D geometry in the form of point clouds is also a kind of radiance. Going beyond the differential RGB color radiance $\mathbf{c}(t)$ for color rendering in Eq. (1), we introduce the differential radiance of *any-modality* $\mathbf{a}(t)$ for multi-modal rendering. Specifically, the rendering of the projected feature map of *any-modality* $\hat{\mathbf{A}}(t)$ is formulated as:

$$\hat{\mathbf{A}}(t) = \int_0^\infty T(t)\sigma(t)\mathbf{a}(t)dt, \quad (3)$$

Typically, to render the projected 3D point cloud feature map, *i.e.*, 2D depth $\hat{\mathbf{D}}(\mathbf{r})$, the differential radiance of *any-modality* $\mathbf{a}(t)$ can be set as the integration of the distance distribution field $\int_0^t dt$. Then, the discrete form to render the feature map of the projected 3D point cloud can be expressed as:

$$\hat{\mathbf{D}}(\mathbf{r}) = \sum_{i=1}^N (T_i(1 - \exp(-\sigma_i \delta_i)) \sum_{j=1}^{i-1} \delta_j), \quad (4)$$

E. Multi-modal Reconstruction

Vanilla Color Reconstruction Objective. Given a set of rendering rays \mathcal{S}_r passing through the pixels of the original image, the goal is to minimize the square of the L_2 -norm of the difference between the ground truth color $\mathbf{C}(\mathbf{r})$ and the rendered color $\hat{\mathbf{C}}(\mathbf{r})$ of ray r :

$$\mathcal{L}_{\mathbf{C}} = \frac{1}{|\mathcal{S}_r|} \sum_{r \in \mathcal{S}_r} \|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2, \quad (5)$$

Multi-modal Reconstruction Objective for Pre-training. Given a set of rendering rays \mathcal{S}_r passing through the ground-truth target, *i.e.*, *pixel*, of *any-modality*, we minimize the L_p -norm of the difference between the ground-truth view-specific projected feature map of *any-modality* $\mathbf{A}(\mathbf{r})$ and the rendered result $\hat{\mathbf{A}}(\mathbf{r})$ of ray r to the p -th power:

$$\mathcal{L}_{\mathbf{A}} = \frac{1}{|\mathcal{S}_r|} \sum_{r \in \mathcal{S}_r} \|\hat{\mathbf{A}}(\mathbf{r}) - \mathbf{A}(\mathbf{r})\|_p^p, \quad (6)$$

The **overall objective function** jointly optimize the reconstruction for multiple (K) view-specific modalities:

$$\mathcal{L} = \sum_{k=1}^K (\lambda_k \cdot \mathcal{L}_{\mathbf{A}_k}) \quad (7)$$

λ_k indicates the coefficient to modulate the k -th sub-loss.

F. Training Setup and Details

Training Strategy. The embedding network (Sec. III-B) is pre-trained for 50 epochs. We use AdamW [51] optimizer with a learning rate of 1e-4 and a weight decay of 0.01. Following mainstream perception models [102], [49], [45], one-cycle scheduler [75] is adopted. Early-stopping strategy [63] is used to avoid over-fitting. The network is trained on 8 NVIDIA V100 with a total batch size of 16.

Masking. (1) For image: the masking patch size is set as 4×4 and 8×8 for images with resolutions of 128×352 and 256×704 ; the masking ratio is set as 50%. (2) For Lidar: the point cloud range is set as $-54 \sim 54$ (m), $-54 \sim 54$ (m), and $-5 \sim 3$ (m) for X , Y , and Z axes, respectively, and the voxel

size is set as [0.075, 0.075, 0.2](m); the masking ratio of non-empty Lidar voxel is set as 90% or in a range-aware manner [55] for single-sweep and multi-sweep point clouds.

Rendering. (1) For rendering view directions, we select two typical and critical views for 3D perception models, *i.e.*, bird’s eye view (BEV) ω^{BEV} and perspective view (PER) ω^{PER} . (2) For the rendering network, we implement it with *conv* layers. Specifically, the rendering network first transforms the embedding $\mathbf{e} \in \mathbb{R}^{D_1 \times D_2 \times D_3 \times D_C}$ ² extracted from the embedding network into sigma-field feature volume $V_\sigma \in \mathbb{R}^{D_1 \times D_2 \times D'_3 \times 1}$ and color feature volume $V_c \in \mathbb{R}^{D_1 \times D_2 \times D'_3 \times 3}$, which are then processed via rendering to derive the projected feature maps for further reconstruction-based optimization. (3) For discretized rendering functions, the parameter δ is approximately set as 0.2 and 0.8 for the rendering in BEV and perspective view.

Reconstruction. (1) For reconstruction targets, we leverage color field map \mathbf{C} which corresponds to multi-view camera-collected images, perspective-view depth \mathbf{D}^{PER} which is generated by projecting Lidar point clouds on perspective-view image planes, and BEV depth \mathbf{D}^{BEV} which is generated by projecting the voxelized Lidar point cloud on the BEV plane. (2) For the set of rendering rays \mathcal{S}_r , it is constructed with rays emitted orthogonally to the BEV plane to render depth in BEV (\mathbf{D}^{BEV}), and it is constructed with rays passing through image plane to render color and depth in perspective view (\mathbf{C} , \mathbf{D}^{PER}). (3) For the parameter p in Eq. (6), we set it as 2 and 1 for color and depth, respectively. (4) For coefficients of the color, perspective-view depth, and BEV depth, we set them as 1e4, 1e-2, and 1e-2, respectively, to normalize the numerical value of diverse modalities.

Implementation. We implement the network in PyTorch using the open-sourced MMDetection3D [13]. Data augmentations mainly follow official Lidar and image augmentations for 3D perception models [49], [45] except the ones that require ground-truth labels, *e.g.*, database-sampler [103].

IV. EXPERIMENT

On multi-modal 3D perception benchmarks (Sec. IV-A), we first verify the transferability of the representation learned via NS-MAE (Sec. IV-B and Sec. IV-C). Then, we study the component effectiveness of NS-MAE (Sec. IV-D).

A. Dataset and Evaluation Metric

nuScenes [7] is a large-scale autonomous driving dataset for 3D perception. Each frame in nuScenes contains six cameras with surrounding views and Lidar point clouds. For 3D object detection, there are up to 1.4 million annotated 3D bounding boxes for 10 classes; detection score (NDS) and mean average precision (mAP) across 10 foreground classes are used for evaluation. For BEV map segmentation, the Intersection-over-Union (IoU) on 6 background classes and the class-averaged mean IoU (mIoU) are used for evaluation.

²In the special case, when $\omega = \omega^{PER}$, dimensions D_1, D_2, D_3 refer to H, W, D axes of the camera coordination; when $\omega = \omega^{BEV}$, D_1, D_2, D_3 refer to X, Y, Z axes of the world coordination.

TABLE I

3D OBJECT DETECTION RESULTS FOR MULTI-MODAL PERCEPTION MODEL (BEVFUSION [49]) ON nUSCENES [7] val. THE NOTION OF MODALITY: CAMERA (C), LIDAR (L). #SWEEP DENOTES THE NUMBER OF LIDAR SWEEPS. #IMGSIZE DENOTES THE RESOLUTION OF IMAGES. THE NOTION OF CLASS: CONSTRUCTION VEHICLE (C.V.), TRAILER (TRAIL.), BARRIER (BARR.), MOTORCYCLE (MOTO.), PEDESTRIAN (PED.), TRAFFIC CONE (T.C.).

| Method | Modality | #Sweep | #ImgSize | Per-class mAP | | | | | | | | | | mAP | NDS | |
|----------------------------|----------|--------|-----------|---------------|-------|------|------|--------|-------|-------|------|------|------|------|------|------|
| | | | | Car | Truck | C.V. | Bus | Trail. | Barr. | Moto. | Bike | Ped. | T.C. | | | |
| BEVFusion [49] + NS-MAE | LC | 1 | 128 × 352 | 81.1 | 37.4 | 12.3 | 59.0 | 31.5 | 64.1 | 46.7 | 28.9 | 80.3 | 63.1 | 50.5 | 53.3 | |
| BEVFusion [49] + NS-MAE | LC | 9 | 256 × 704 | 87.4 | 40.4 | 25.7 | 67.0 | 38.8 | 71.6 | 68.2 | 48.6 | 85.5 | 74.5 | 60.8 | 64.1 | |
| | | | | 256 × 704 | 88.1 | 45.9 | 25.1 | 68.8 | 37.2 | 73.8 | 70.8 | 56.6 | 86.9 | 77.4 | 63.0 | 65.5 |

TABLE II

3D OBJECT DETECTION RESULTS OF MULTI-MODAL PERCEPTION MODELS (VFF [39] WITH VARIOUS 3D DETECTION HEADS, INCLUDING SECOND [89], PVRCNN [72], AND VOXELRCNN [14]) ON KITTI-3D [24] val. RESULTS ARE FOR THE CAR CATEGORY AND REPORTED IN AP_{R40} @0.7, 0.7, 0.7. DB INDICATES DATABASE-SAMPLER [103] DATA AUGMENTATION IS USED.

| Method | DB | AP _{BEV} | | | AP _{3D} | | | mAP | NDS |
|---------------------------|--------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----|-----|
| | | Easy | Moderate | Hard | Easy | Moderate | Hard | | |
| VFF-SECOND + NS-MAE | | 91.71 92.65 | 85.77 88.24 | 83.54 85.83 | 87.25 88.25 | 76.44 78.40 | 74.02 74.37 | | |
| VFF-SECOND + NS-MAE | ✓ ✓ | 92.83 93.04 | 88.92 90.43 | 88.22 88.46 | 89.45 91.48 | 82.32 82.58 | 79.39 79.77 | | |
| VFF-PVRCNN + NS-MAE | ✓ ✓ | 92.65 92.94 | 90.86 91.00 | 88.55 90.49 | 91.75 92.03 | 85.09 85.31 | 82.68 83.05 | | |
| VFF-VoxelRCNN + NS-MAE | ✓ ✓ | 95.67 95.57 | 91.56 91.69 | 89.20 89.23 | 92.46 92.51 | 85.25 85.59 | 82.93 82.95 | | |

KITTI-3D [24] is one of the most widely-used benchmarks for 3D object detection, which consists of 7,481 and 7,518 image-Lidar pairs for training and testing, respectively. The training set is commonly divided into *train* split with 3712 samples and *val* split with 3769 samples [10]. As KITTI-3D is a small-scale benchmark existed for a long time, the performance of multi-modal perception models on it is close to saturation. 3D IoU (AP_{3D}) and BEV IoU (AP_{BEV}) with the average precision metric are used for evaluation.

B. Main Transfer Results

In this section, we evaluate the representation quality of pre-training by directly transferring to various perception models on two mainstream 3D perception tasks.

We follow the fine-tuning setups of baseline models and fine-tune the whole framework of them in an end-to-end manner and without using any extra data. Concretely, we fine-tune models on nuScenes and KITTI-3D for 20 and 80 epochs, respectively. By default, CBGS [103] trick is not used during fine-tuning. Then, we compare the performance between models that are *without pre-training* and models whose embedding networks are *pre-trained via NS-MAE*.

1) *Transfer to 3D Object Detection Task: Transfer to Multi-modal Perception Models.* For representation transferability evaluation, we select two baseline models, *i.e.*, BEVFusion [49] and VFF [39], which represent the state-of-the-art performance for multi-modal 3D perception on nuScenes [6] and KITTI-3D benchmarks, respectively. In Table I, we show that NS-MAE can effectively boost the performance for the multi-modal model BEVFusion under

various input settings, including varied Lidar sweeps and image resolutions. Notably, for the multi-modal version of the BEVFusion model with multi-sweep Lidar and higher image resolution (#Sweep:9, #ImgSize:256×704), **NS-MAE brings more than 2% improvement (+2.2%) in mAP and 1.4% improvement in NDS**. Moreover, in Table II, for VFF [39] framework with various 3D detection heads, including SECOND [89], PVRCNN [72], and VoxelRCNN [14], the representation pre-trained via NS-MAE also improves the BEV and 3D detection performance in nearly all sub-metrics than training from scratch.

Transfer to Single-modal Perception Models. In Table III, we show the generality of pre-trained representation via NS-MAE to be transferred to single-modal, *i.e.*, camera-only and Lidar-only, perception models. For the camera-only perception model BEVDet³ [31], **NS-MAE brings more than 2% NDS improvement** for the settings of both the default (20) and doubled (2×: 40) training epochs. For the Lidar-only baselines [3], [94], the transferable representation learned via our NS-MAE consistently shows its effectiveness. Moreover, compared to a concurrent work (Voxel-MAE [55]) that targets pre-training for the Lidar-only perception models, our NS-MAE that targets unified pre-training for multi-modal perception models also shows better performance (62.1% v.s. 61.4% in NDS) for the Lidar-only baseline, CenterPoint [94].

2) *Transfer to BEV Map Segmentation Task:* As is shown in Table IV, NS-MAE largely boosts the performance for BEV map segmentation. In specific, **our pre-training schema (NS-MAE) largely helps improve the baseline setting with about 4% mIoU improvement** for both the multi-modal and camera-only perception models. Notably, NS-MAE brings better segmentation quality for all classes.

C. Label-Efficient Transfer Results

Efficient transfer (*e.g.*, fewer annotations) on downstream tasks brings about great application value to real-world scenarios. Thus, we evaluate both label-efficient regime 3D detection and BEV segmentation on nuScenes. Specifically, nuScenes training split is sampled with different ratios (1% to 10%) to generate label-efficient fine-tuning datasets. For fair comparisons, all models are end-to-end fine-tuned for the same iterations as the default fine-tuning setting (100%).

In Table V, **our pre-training schema shows the best performance under all settings of sampling ratios of labeled**

³We use an efficient implementation of BEVDet [31] in [49].

TABLE III

3D OBJECT DETECTION RESULTS FOR CAMERA-ONLY (TOP) AND LIDAR-ONLY (DOWN) PERCEPTION MODELS ON NUSCENES [7] val. $2\times$ DENOTES DOUBLED TRAINING EPOCHS. #SWEEP DENOTES THE NUMBER OF LIDAR SWEEPS/SCANS. #IMGSIZE DENOTES THE SPATIAL RESOLUTION OF IMAGES.

| Method | Modality | #Sweep | #ImgSize | Per-class mAP | | | | | | | | | | mAP | NDS |
|---|----------|--------|-----------|---------------|-------|------|------|--------|-------|-------|------|------|------|------|------|
| | | | | Car | Truck | C.V. | Bus | Trail. | Barr. | Moto. | Bike | Ped. | T.C. | | |
| BEVDet [31] + NS-MAE | C | / | 256 × 704 | 35.3 | 15.7 | 2.7 | 18.2 | 5.5 | 31.0 | 19.8 | 18.3 | 26.9 | 46.0 | 21.9 | 29.4 |
| | | / | 256 × 704 | 37.1 | 18.0 | 3.6 | 18.0 | 7.0 | 41.3 | 20.3 | 19.8 | 28.1 | 47.6 | 24.1 | 32.1 |
| BEVDet ^{2×} [31] + NS-MAE | C | / | 256 × 704 | 37.1 | 16.4 | 2.7 | 18.9 | 5.8 | 42.0 | 20.7 | 18.3 | 28.3 | 49.0 | 23.9 | 31.8 |
| | | / | 256 × 704 | 38.3 | 17.9 | 4.5 | 18.3 | 8.1 | 47.1 | 21.5 | 18.8 | 29.4 | 49.8 | 25.4 | 33.9 |
| CenterPoint [94] + VoxelMAE [55] + NS-MAE | L | 9 | / | 80.9 | 52.4 | 14.4 | 64.0 | 29.6 | 58.7 | 59.4 | 45.6 | 80.4 | 60.8 | 54.6 | 61.3 |
| | | 9 | / | 80.6 | 53.7 | 13.7 | 63.2 | 29.2 | 61.1 | 60.5 | 45.4 | 80.4 | 61.1 | 54.9 | 61.4 |
| | | 9 | / | 81.2 | 53.0 | 14.7 | 63.7 | 30.2 | 60.0 | 60.1 | 47.1 | 81.6 | 61.3 | 55.3 | 62.1 |

TABLE IV

BEV-MAP SEGMENTATION RESULTS FOR MULTI-MODAL AND CAMERA-ONLY PERCEPTION MODELS ON NUSCENES [7] val. THE NOTION OF CLASS: DRIVABLE (DRI.), PEDESTRIAN CROSSING (P.C.), WALKAWAY (WALK.), STOP LINE (S.L.), CARPARK (CAR.), DIVIDER (DIV.). IMAGES OF SIZE 256 × 704 AND MULTI-SWEEP (9) LIDAR POINTS ARE USED AS INPUT.

| Method | Modality | Per-class IoU | | | | | mIoU | |
|----------------------------|----------|---------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | Dri. | P.C. | Walk. | S.L. | Car. | | |
| BEVFusion [49] + NS-MAE | LC | 75.0 | 42.6 | 52.6 | 24.4 | 26.6 | 36.0 | 42.9 |
| | | 78.0 | 45.9 | 55.5 | 26.1 | 35.4 | 38.9 | 46.6 |

| Method | Modality | mIoU |
|-------------------------|----------|-------------|
| BEVDet [31] + NS-MAE | C | 72.7 |
| | | 76.1 |
| | | 35.6 |
| | | 39.9 |
| | | 44.7 |
| | | 49.0 |
| | | 21.1 |
| | | 23.5 |
| | | 34.0 |
| | | 41.6 |
| | | 32.3 |
| | | 35.6 |
| | | 40.1 |
| | | 44.3 |

TABLE V

LABEL-EFFICIENT REGIME 3D OBJECT DETECTION (TOP) AND BEV MAP SEGMENTATION (DOWN) RESULTS FOR MULTI-MODAL PERCEPTION MODEL ON NUSCENES [7] val. MODELS ARE FINE-TUNED WITH VARIED RATIOS OF ANNOTATIONS. IMAGES OF SIZE 256 × 704 AND MULTI-SWEEP (9) LIDAR POINTS ARE USED AS INPUT.

| Method | Metric | Sampling Ratio | | | | |
|---|--------|----------------|-------------|-------------|-------------|--|
| | | 1% | 5% | 10% | 100% | |
| Label-Efficient regime 3D object detection | | | | | | |
| BEVFusion [49] + NS-MAE | mAP | 26.2 | 46.1 | 54.2 | 60.8 | |
| | | 30.2 | 47.6 | 55.9 | 63.0 | |
| BEVFusion [49] + NS-MAE | NDS | 44.2 | 55.4 | 60.3 | 64.1 | |
| | | 45.4 | 57.0 | 61.4 | 65.5 | |
| Label-Efficient regime BEV map segmentation | | | | | | |
| BEVFusion [49] + NS-MAE | mIoU | 29.7 | 39.4 | 41.3 | 42.9 | |
| | | 31.1 | 41.6 | 45.1 | 46.6 | |

fine-tuning data for both 3D object detection and BEV-map segmentation tasks, demonstrating good label-efficient transferability of the representation pre-trained via NS-MAE.

D. Ablation Study

We ablate the components of NS-MAE in Table VI and provide the following insights about how to learn transferable multi-modal representation for perception via NS-MAE.

Masking of input modalities is critical for effective representation learning. *Without masking:* When rendering the feature maps of color and depth without masking (settings (A1) and (B3)) for representation learning, the transferred performance could be even worse than the baseline setting (Baseline). *With masking:* When enabling the multi-modal masking for the inputs, the comparison between settings (A2)(B1)(B2) with the baseline (Baseline) verifies the effectiveness of masked reconstruction of the feature map of color

TABLE III

3D OBJECT DETECTION RESULTS FOR CAMERA-ONLY (TOP) AND LIDAR-ONLY (DOWN) PERCEPTION MODELS ON NUSCENES [7] val. $2\times$ DENOTES DOUBLED TRAINING EPOCHS. #SWEEP DENOTES THE NUMBER OF LIDAR SWEEPS/SCANS. #IMGSIZE DENOTES THE SPATIAL RESOLUTION OF IMAGES.

TABLE VI

ABLATION STUDY ON NUSCENES [7] val WITH THE MULTI-MODAL PERCEPTION BASELINE MODEL, BEVFUSION [49]. RESULTS OF 3D OBJECT DETECTION ARE REPORTED. \mathbf{C} , \mathbf{D}^{PER} , AND \mathbf{D}^{BEV} DENOTE COLOR, PERSPECTIVE-VIEW DEPTH, AND BEV DEPTH. IMAGES OF SIZE 128×352 AND SINGLE-SWEEP LIDAR POINT CLOUDS ARE USED.

| Setting | Masking | Rendering Targets | | | mAP | NDS |
|----------|---------|-------------------|--------------------|--------------------|------|------|
| | | \mathbf{C} | \mathbf{D}^{PER} | \mathbf{D}^{BEV} | | |
| Baseline | | | | | 50.5 | 53.3 |
| (A1) | | ✓ | | | 50.2 | 53.9 |
| (A2) | ✓ | ✓ | | | 50.7 | 54.2 |
| (B1) | ✓ | | ✓ | | 50.9 | 54.2 |
| (B2) | ✓ | | | ✓ | 51.0 | 54.3 |
| (B3) | | | ✓ | ✓ | 49.6 | 52.8 |
| (B4) | ✓ | | ✓ | ✓ | 51.3 | 54.4 |
| Default | ✓ | ✓ | ✓ | ✓ | 51.5 | 54.7 |

and projected point cloud, *i.e.*, depth.

Rendering with more modalities and more view directions encourages more transferable representation. *More modalities:* When multi-modal masking is enabled, the comparisons between each setting of (A2)(B2)(B3) and the baseline (Baseline) verify the effectiveness of each rendering target (\mathbf{C} , \mathbf{D}^{PER} , and \mathbf{D}^{BEV}). *More view directions:* As point clouds lie in the 3D space, rendered feature maps of projected point clouds (\mathbf{D}) in diverse views can be simultaneously optimized in a self-supervised manner. When rendering the projected feature maps of point clouds with more views for reconstruction-based optimization (B4), the performance can be further boosted compared to using only a single view direction (B1) or (B2)). Finally, by incorporating all rendering targets above with multi-modal masking used, the default setting of NS-MAE (Default) achieves the best performance with 1.5% higher NDS and 1.0% higher mAP than the baseline setting due to their mutual benefits.

E. Analysis

In this section, we analyze the emerging properties of NS-MAE for multi-modal perception pre-training.

NS-MAE Induces Robust Multi-modal Reconstruction. In Fig. 2, given the input images with diverse masking ratios, we show the rendering results of perspective-view images and projected point cloud feature maps, *i.e.*, depth maps. Case (a.1)&(a.2) of Fig. 2 shows that the multi-modal embedding network pre-trained with NS-MAE can render visually-high-quality color and depth of the scene. In Case (b) of Fig. 2,

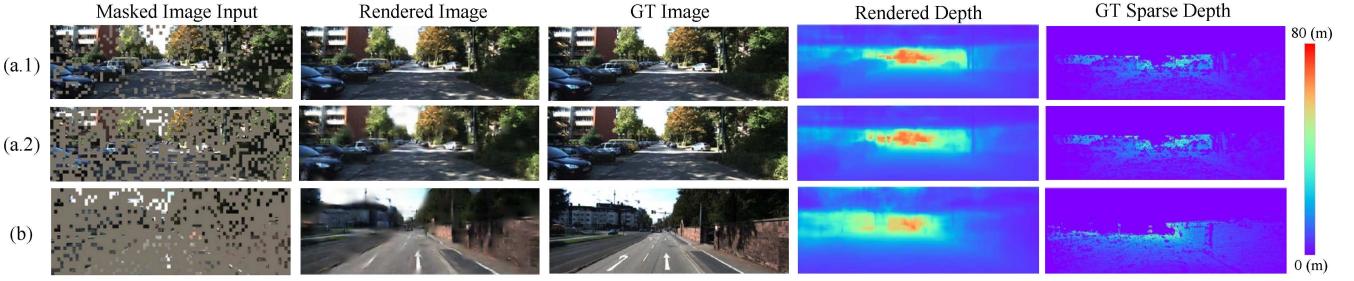


Fig. 2. Qualitative results of perspective-view image and depth, rendered with front-view camera parameters, on KITTI [24] val set.

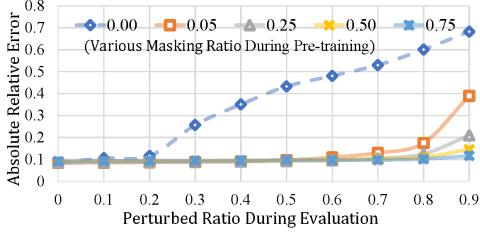


Fig. 3. Quantitative comparisons on depth reconstruction quality (measured by absolute relative error) on KITTI [24] val set with varied perturbed ratios of the input image during evaluation. Here, perturbation is implemented with random masking.

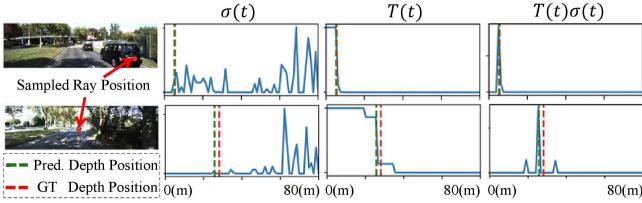


Fig. 4. Qualitative visualizations of the radiance terms in rendering equations (c.f. Sec. III-D), including volume density $\sigma(t)$, accumulated transmittance $T(t)$, and their product $T(t)\sigma(t)$, of sampled rendering rays on KITTI [24] val. Here, the rendering view direction is orthogonal to the perspective-view image plane.

even if the input image is masked out with an extremely high masking ratio (80%), the rendered image and depth can still reconstruct coarse-level color and geometry well. To quantitatively study the robustness of the multi-modal representation pre-trained via NS-MAE, we provide the depth reconstruction results of models trained and evaluated with various masking ratios on KITTI [24]. As is shown in Fig. 3, the models trained with masked modeling exhibit better robustness to perturbations thanks to the learned modality completion and generalization ability.

NS-MAE Is Physics-informed Unknown-region Filter. In Fig. 4, we visualize the radiance terms in rendering equations (Eq. (2) and Eq. (4) in Sec. III-D) of sampled rendering rays to understand the property of the rendering equations better. During reconstruction-based optimization (Sec. III-E), the accumulated transmittance $T(t)$ term in the prior rendering functions serves as a *band-pass filter* along the ray direction, e.g., the direction from the origin (0m) to the farthest (80m) of Fig. 4. Specifically, $T(t)$ helps selectively filter out the *unknown* region (from the position of an object to the farthest, i.e., *approximate infinity*), and only optimize *non-occupied*

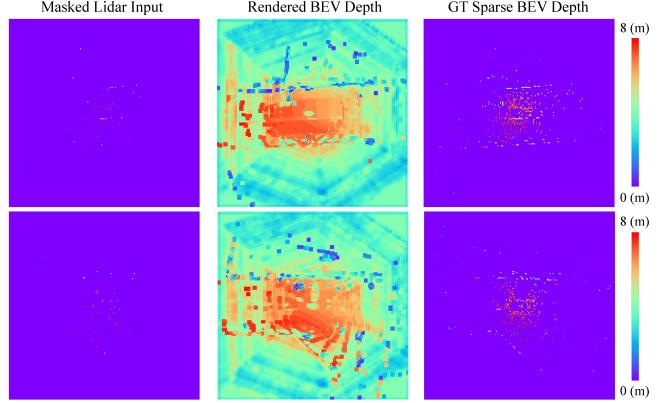


Fig. 5. Qualitative results of bird's-eye-view (BEV) depth, i.e., BEV occupancy, rendered from a viewpoint that is orthogonal to the road plane, on nuScenes [6] val. Zoom in to view better.

regions (from the origin of the ray to the object) and *occupied* regions (where objects exist), respectively. Besides, the sigma field $\sigma(t)$ term is learned to predict zero for *non-occupied* regions between the origin of the ray and the object, deriving a focused impulse in the product of accumulated transmittance and sigma field $T(t)\sigma(t)$.

NS-MAE Generates Dense BEV Occupancy. Given an *extremely sparse* masked Lidar input, the pre-training schema of NS-MAE includes the sub-task of reconstructing the projected BEV point cloud feature maps, i.e., BEV depth, which indicates the occupancy of objects on the BEV plane. In Fig. 5, we show that NS-MAE pre-trained with only image and Lidar pairs can help the multi-modal perception model learn to complete *denser* BEV occupancy maps which decouple objects in the scene along the height axis of the world coordination. Such a good zero-shot BEV occupancy completion ability can further elevate the performance on BEV perception downstream tasks, especially for BEV map segmentation (as is verified in the experiments of Sec. IV-B).

To sum up, NS-MAE naturally inherits useful properties from MAE and NeRF, thus learning transferable representation for multi-modal perception models effectively.

V. CONCLUSIONS AND FUTURE WORK

We have proposed a unified self-supervised pre-training paradigm (NS-MAE) for multi-modal perception models. Specifically, NS-MAE conducts masked multi-modal reconstruction in NeRF to enable transferable multi-modal representation learning in a unified and neat fashion. Extensive

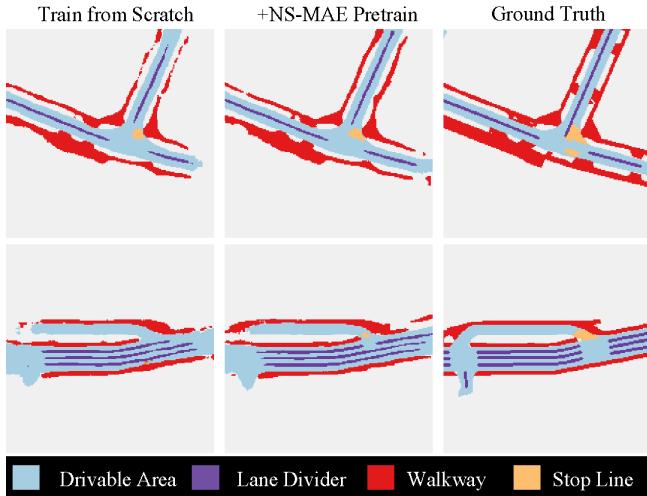


Fig. A. **Qualitative comparisons of BEV map segmentation results** on nuScenes [6] *val* with multi-modal perception model BEVFusion [49]. Specifically, we compare the qualitative results of the model which is trained from scratch (Train from Scratch) and the results of the model which is pre-trained via NS-MAE (+NS-MAE Pretrain). The ground-truth BEV map segmentation result (Ground Truth) is provided for reference.

experiments demonstrate the encouraging transferability and generalization of the learned representation via NS-MAE for both multi-modal and single-model perception models.

Future Work. Due to computation and time limitations, we currently do not explore NS-MAE with larger models and data. Besides, we consider it meaningful to study the generality of NS-MAE to perception models with more modalities, *e.g.*, Radar [56], [37], [90], or joint multi-modal encoders [92].

VI. ACKNOWLEDGEMENTS

The authors would like to thank Shusheng, Zhaoyang, Baixin, Huaxin, and Yiheng for their kind help with proof reading. Also, the authors would like to thank Yangyi, Chenyu, Jinguo, and Tianyu for their valuable viewpoints and help during the development of this project.

A. MORE QUALITATIVE COMPARISONS

As is shown in Fig. A, we compare the BEV map segmentation results predicted by the multi-modal perception model [49] when *trained from scratch* and the model *pre-trained via the proposed NS-MAE* self-supervised representation learning schema, which qualitatively demonstrates that the transferred representation learned via NS-MAE can help improve the segmentation quality for 3D perception.

B. MORE DETAILS FOR BASELINE MODELS

A. Multi-modal Perception Model

BEVFusion [49] is an advanced fusion-based/multi-modal perception model for both 3D object detection and BEV map segmentation. (1) For the embedding network part, BEVFusion uses Swin-T [48] with FPN [46] as the Camera Encoder, VoxelNet [89] as Lidar Encoder, and LSS [61] as Cam2World view-transformation module. (2) In the original implementation of BEVFusion, the Lidar point cloud is

voxelized with different resolutions for diverse downstream tasks, *i.e.*, 0.075(m) (for 3D object detection) and 0.1(m) (for BEV map segmentation). Differently, to unify and align the multi-modal pre-training and the multi-modal transferring, we voxelize the Lidar point cloud with the same resolution, *i.e.*, 0.075(m), for both detection and segmentation tasks. (3) Due to computation and time limitations, although CBGS [103] data augmentation is an effective trick to improve the performance for downstream tasks, we do not use it for all the experiments in the main paper. (4) For the implementation of the rendering network, we use separate *conv* layers for the bird’s eye view and perspective views. (5) We refer to the detailed implementations in the official repo of BEVFusion, which is under the license of Apache 2.0, for the pre-training, fine-tuning, and evaluation.

VFF [39], *namely* Voxel Field Fusion, is a perception model for cross-modality 3D object detection. (1) For the embedding network part, VFF uses ResNet-50 [27] as the Camera Encoder and the proposed voxel fusion mechanism for Cam2World transition and fusion. For the Lidar Encoder, VFF is instantiated with various voxel-based Lidar backbones, *e.g.*, PV-RCNN [70] and VoxelRCNN [14]. (2) In the VFF-SECOND configuration that we set up for transferability evaluation of the representation pre-trained via NS-MAE in the main paper, we refer to the baseline model setting of their paper for re-implementation. Specifically, we remove the voxel-field-fusion mechanism from the default setting of VFF and use the 3D detection head of SECOND [89]. (3) For the rendering network, we implement it with separate *conv* layers for the bird’s eye view and perspective views. Specifically, we empirically leverage FPN-like [46] convolution layers for the perspective-view rendering network and simple *conv3* × 3 layer for the bird’s-eye-view rendering network. (4) We refer to the detailed implementations in the official repo of VFF, which is under the license of Apache 2.0, for the pre-training, fine-tuning, and evaluation.

B. Camera-only Perception Model

BEVDet [31] is a sophisticated camera-only perception model for 3D object detection. We leverage an efficient implementation of BEVDet in [49] to align the setting between the multi-modal representation learning during pre-training and the fine-tuning stage. (1) For the embedding network part, BEVDet uses Swin-T [48] with FPN [46] as the Camera Encoder and leverages LSS [61] as the Cam2World view-transformation module. (2) During the pre-training of BEVDet, we additionally leverage the CBGS [103] data augmentation, which elongates the training iterations to 4.5 times than the default setting without it, for the convergence of the model. (3) We refer to an improved implementation of BEVDet in the repo of BEVFusion, which is under the license of Apache 2.0, for the pre-training, fine-tuning, and evaluation.

C. Lidar-only Perception Model

CenterPoint [94] is a widely-used Lidar-only perception model for 3D object detection. (1) For the embedding network,

CenterPoint uses a standard Lidar-based backbone network, *i.e.*, VoxelNet [102] to build a representation of the input point cloud. (2) We refer to the implementation in the open-sourced MMDetection3D [13] library, which is under the license of Apache 2.0, and the official repo of CenterPoint, which is under the license of MIT, for the pre-training, fine-tuning, and evaluation.

C. MORE DETAILS FOR EVALUATION PROTOCOL

3D object detection on nuScenes: We leverage the official overall metrics: mean Average Precision (mAP) [19] and nuScenes Detection Score (NDS), and report these two metrics in the main paper for performance comparisons. Here, we detail the calculation of them as follows. In specific, mAP is computed by averaging over distance thresholds of 0.5(m), 1(m), 2(m), and 4(m) across ten sub-classes: car, truck, bus, trailer, construction vehicle, pedestrian, motorcycle, bicycle, barrier, and traffic cone. Moreover, nuScenes Detection Score (NDS) is a consolidated metric of mAP and all other indicators (*e.g.*, translation, scale, orientation, velocity, and attribute) to comprehensively judge the 3D detection quality on nuScenes dataset.

BEV map segmentation on nuScenes: Following the evaluation protocol on the BEV map segmentation task [61], [43], [49], [45], the region for evaluation is limited in the [-50m, 50m] \times [-50m, 50m] region around the ego car.

D. LISCENSES OF DATASETS

nuScenes [6] is a large-scale 3D perception dataset released under the CC BY-NC-SA 4.0 license.

KITTI-3D [24] is a small-scale 3D perception dataset released under the the CC BY-NC-SA 3.0 license.

REFERENCES

- [1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020.
- [2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022.
- [3] 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. *arXiv preprint arXiv:2203.11496*, 2022.
- [4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [6] Holger Caesar, Varun Bankiti, Alex H Lang, and et al. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.
- [7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Lioung, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.
- [8] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, 2021.
- [9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.
- [10] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, et al. 3d object proposals for accurate object class detection. In *NeurIPS*, 2015.
- [11] Xiaozhi Chen, Huimin Ma, Jixiang Wan, B. Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017.
- [12] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *ECCV*, 2022.
- [13] MMDetection3D Contributors. Mmdetection3d: Open-mmlab next-generation platform for general 3d object detection. <https://github.com/open-mmlab/mmdetection3d>, 2020.
- [14] Jiajun Deng, Shaoshuai Shi, Peiwei Li, et al. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, 2021.
- [15] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022.
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [17] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.
- [18] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021.
- [19] M. Everingham, L. Gool, Christopher K. I. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2009.
- [20] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). *arXiv preprint arXiv:2205.01397*, 2022.
- [21] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022.
- [22] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.
- [23] Huan Fu, Mingming Gong, Chaohui Wang, and others. Deep Ordinal Regression Network for Monocular Depth Estimation. In *CVPR*, 2018.
- [24] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.
- [25] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *ICRA*, 2023.
- [26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [28] Georg Hess, Johan Jaxing, Elias Svensson, David Hagerman, Christopher Petersson, and Lennart Svensson. Masked autoencoder for self-supervised pre-training on lidar point clouds. In *WACV*, 2023.
- [29] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *ECCV*, 2022.
- [30] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.
- [31] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.
- [32] Tengteng Huang, Zhe Liu, Xiwu Chen, and X. Bai. EPNet: Enhancing point features with image semantics for 3d object detection. In *ECCV*, 2020.
- [33] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021.
- [34] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 1984.
- [35] Abhinav Kumar, Garrick Brazil, and Xiaoming Liu. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *CVPR*, 2021.
- [36] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019.

- [37] Peizhao Li, Pu Wang, Karl Berntorp, and Hongfu Liu. Exploiting temporal relations on radar perception for autonomous driving. In *CVPR*, 2022.
- [38] Xiaotong Li, Yixiao Ge, Kun Yi, Zixuan Hu, Ying Shan, and Ling-Yu Duan. mc-beit: Multi-choice discretization for image bert pre-training. In *ECCV*, 2022.
- [39] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *CVPR*, 2022.
- [40] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Bo Wu, Yifeng Lu, Denny Zhou, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. *arXiv preprint arXiv:2203.08195*, 2022.
- [41] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *AAAI*, 2022.
- [42] Zhichao Li, Feng Wang, and Naiyan Wang. Lidar r-cnn: An efficient and universal 3d object detector. In *CVPR*, 2021.
- [43] 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 *ECCV*, 2022.
- [44] 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 *ICCV*, 2021.
- [45] 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 *NeurIPS*, 2022.
- [46] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.
- [47] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020.
- [48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [49] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huiyi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023.
- [50] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *ICCV*, 2021.
- [51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [52] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *ICCV*, 2021.
- [53] Xinzhu Ma, Zhihui Wang, Haojie Li, et al. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *ICCV*, 2019.
- [54] 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 *ECCV*, 2020.
- [55] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxelmae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022.
- [56] Ramin Nabati and H CenterFusion Qi. Center-based radar and camera fusion for 3d object detection. In *WACV*, 2021.
- [57] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H Mueller, Chakravarty R Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In *Computer Graphics Forum*, 2021.
- [58] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022.
- [59] Dennis Park, Rares Ambrus, and Vitor others Guizilini. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 2021.
- [60] Zhiiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.
- [61] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020.
- [62] Jonah Philion and S. Fidler. Lift, Splat, Shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020.
- [63] Lutz Prechelt. Early stopping—but when? *Neural networks: tricks of the trade: second edition*, 2012.
- [64] C. Qi, W. Liu, Chenxia Wu, Hao Su, and L. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018.
- [65] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.
- [66] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *AAAI*, 2019.
- [67] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, 2021.
- [68] Thomas Roddick, Alex Kendall, and R. Cipolla. Orthographic feature transform for monocular 3d object detection. In *BMVC*, 2019.
- [69] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *ICRA*, 2022.
- [70] Shaoshuai Shi, Chaoxu Guo, Li Jiang, et al. Pv-rnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, 2020.
- [71] 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 *CVPR*, 2020.
- [72] 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 *CVPR*, 2020.
- [73] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3d object proposal generation and detection from point cloud. In *CVPR*, 2019.
- [74] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *ICRA*, 2019.
- [75] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 2017.
- [76] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
- [77] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. PointAugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, 2021.
- [78] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, 2021.
- [79] Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, and Xiangyang Xue. Progressive coordinate transforms for monocular 3d object detection. In *NeurIPS*, 2021.
- [80] Tai Wang, Conghui He, Zhe Wang, Jianping Shi, and Dahua Lin. Flava: Find, localize, adjust and verify to annotate lidar-based point clouds. In *UIST*, 2020.
- [81] Tai Wang, Zhu Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *CoRL*, 2022.
- [82] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *CVPR*, 2021.
- [83] Yan Wang, Wei-Lun Chao, Divyansh Garg, et al. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019.
- [84] Yue Wang, Alireza Fathi, Abhijit Kundu, David A. Ross, Caroline Pantofaru, Thomas A. Funkhouser, and Justin M. Solomon. Pillar-based object detection for autonomous driving. In *ECCV*, 2020.
- [85] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, 2022.
- [86] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.
- [87] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021.
- [88] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhiliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.
- [89] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 2018.
- [90] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic

- objects. In *ECCV*, 2020.
- [91] Shusheng Yang, Yixiao Ge, Kun Yi, Dian Li, Ying Shan, Xiaohu Qie, and Xinggang Wang. Masked visual reconstruction in language semantic space. *arXiv preprint arXiv:2301.06958*, 2023.
- [92] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. In *NeurIPS*, 2022.
- [93] Zetong Yang, Y. Sun, Shu Liu, and Jiaya Jia. 3DSSD: Point-based 3d single stage object detector. In *CVPR*, 2020.
- [94] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *CVPR*, 2021.
- [95] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. In *NeurIPS*, 2021.
- [96] Jin Hyeok Yoo, Yeocheol Kim, Ji Song Kim, and J. Choi. 3D-CVF: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *ECCV*, 2020.
- [97] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021.
- [98] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021.
- [99] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.
- [100] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *CVPR*, 2021.
- [101] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *CoRL*, 2020.
- [102] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018.
- [103] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019.