TO VIEW TRANSFORM OR NOT TO VIEW TRANSFORM: NERF-BASED PRE-TRAINING PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural radiance fields (NeRFs) have emerged as a prominent pre-training paradigm for vision-centric autonomous driving, which enhances 3D geometry and appearance understanding in a fully self-supervised manner. To apply NeRF-based pre-training to 3D perception models, recent approaches have simply applied NeRFs to volumetric features obtained from view transformation. However, coupling NeRFs with view transformation inherits conflicting priors; view transformation imposes discrete and rigid representations, whereas radiance fields assume continuous and adaptive functions. When these opposing assumptions are forced into a single pipeline, the misalignment surfaces as blurry and ambiguous 3D representations that ultimately limit 3D scene understanding. Moreover, the NeRF network for pre-training is discarded during downstream tasks, resulting in inefficient utilization of enhanced 3D representations through NeRF. In this paper, we propose a novel NeRF-Resembled Point-based 3D detector that can learn continuous 3D representation and thus avoid the misaligned priors from view transformation. NeRP3D preserves the pre-trained NeRF network regardless of the tasks, inheriting the principle of continuous 3D representation learning and leading to greater potentials for both scene reconstruction and detection tasks. Experiments on nuScenes dataset demonstrate that our proposed approach significantly improves previous state-of-the-art methods, outperforming not only pretext scene reconstruction tasks but also downstream detection tasks.

1 INTRODUCTION

Accurate and fine-grained 3D scene understanding is essential for autonomous driving, supporting critical tasks such as 3D object detection Reading et al. (2021); Li et al. (2023; 2024), high-definition (HD) map construction Liao et al. (2023); Shin et al. (2025), and occupancy prediction Tong et al. (2023); Tian et al. (2023). To facilitate these open-world perceptions, view transformation backbones Li et al. (2023; 2024; 2022) have drawn great attention, which project multi-view 2D image features into a unified 3D representation on bird's-eye-view (BEV) or voxel space. A unified 3D representation, aligning various modalities Liu et al. (2023b); Li et al. (2022); Yan et al. (2023); Kim et al. (2023) in a common metric frame, provides a single 3D canvas that can be leveraged across diverse downstream tasks Hu et al. (2023); Jiang et al. (2023); Weng et al. (2024).

In parallel, neural fields, such as NeRFs Mildenhall et al. (2021) and 3DGS Kerbl et al. (2023), have emerged as a dominant paradigm for reconstructing 3D representation and synthesizing novel views by learning a continuous field of color and volume density in a self-supervised manner. Sharing the goal of understanding the 3D environment, recent studies Yang et al. (2024); Huang et al. (2024); Xu et al. (2024) proposed combining NeRFs or 3DGS with view transformation, enabling self-supervised pre-training through photometric and depth reconstruction without the need for expensive manual annotations.

Although both view transformation and NeRFs ultimately aim to reconstruct a 3D representation of the world from 2D signals, they embody conflicting priors. Existing approaches Yang et al. (2024); Huang et al. (2024) extract point features for radiance fields by interpolating discretized and fixed voxel features from a view transformation backbone, and then pre-train the backbone through photometric and depth errors rendered from those point features. However, this pipeline inevitably leads to NeRF inheriting the discrete and rigid priors of the view transformation, which conflicts with the continuous radiance fields and restricts the fidelity of the reconstructed 3D representation.

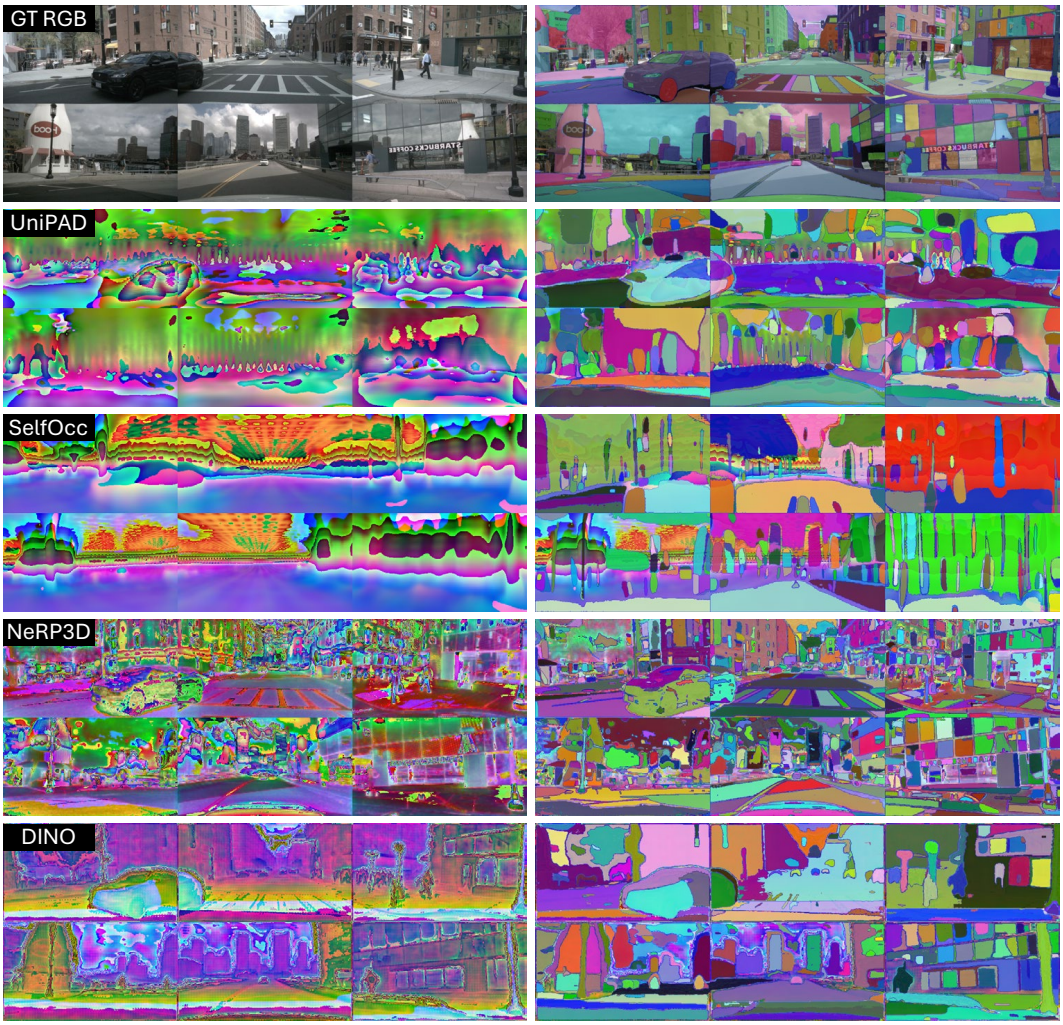

Figure 1: Comparison of 2D feature maps (left) and their instance segmentation (right) results using SAM Kirillov et al. (2023); Ren et al. (2024); Ravi et al. (2024) across different methods. All 2D feature maps, except for ground truth RGB (row 1) and DINO Caron et al. (2021); Oquab et al. (2023) feature (row 5), are obtained by accumulating 3D point-wise representations along each ray onto the image plane with predicted density. They are extracted directly after radiance field pre-training without any task-specific fine-tuning. UniPAD Yang et al. (2024) (row 2) and SelfOcc Huang et al. (2024) (row 3) produce blurry and inaccurate features that fail to separate nearby or crowded objects, resulting in under-segmented instances. In contrast, NeRP3D (row 4) produces precise and well-localized features with distinct object boundaries without any distillation or fine-tuning from 2D foundation models, comparable to those from DINO features. Consequently, we observe the potential for the enhancement of 3D representation to be reflected in the improved instance segmentation quality.

Moreover, the pre-trained NeRF is discarded during downstream tasks, preventing effective transfer of NeRF knowledge and limiting the exploitation of enhanced 3D representations from pre-training. As a result, distinct objects can be collapsed into a single blurry blob, as shown in Fig. 1.

In this paper, we introduce NeRP3D, a novel NeRF-Resembled Point-based 3D detector that fully inherits the continuous function of neural radiance fields Mildenhall et al. (2021); Wang et al. (2021), effectively overcoming the inherent discrepancy with view transformation.Unlike methods relying on rigidly discretized voxel-based representations, NeRP3D directly models 3D scenes as continuous 3D features, geometry, and appearance from any continuous 3D location in a feedforward manner, as illustrated in Fig. 2. Experiments on the nuScenes Caesar et al. (2020) benchmark demonstrate that

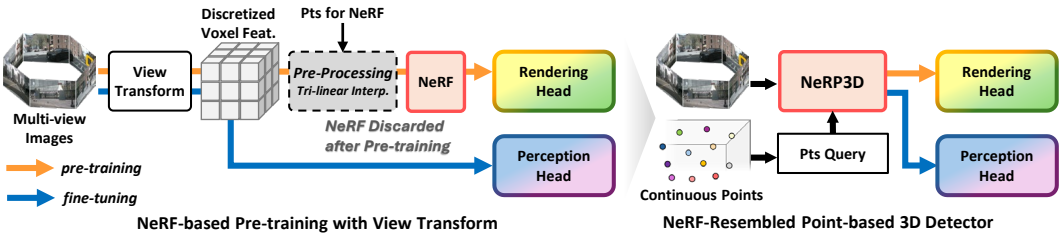

Figure 2: Comparison of the previous NeRF-based pre-training methods and our NeRP3D pipeline.

our approach significantly improves not only the rendering quality but also the downstream perception tasks for autonomous driving compared to previous approaches that simply incorporate NeRF-based pre-training into view transformation frameworks. These findings highlight the importance of aligning the 3D backbone with the pre-training model as well as continuous 3D representation learning in advancing NeRF-based pre-training for enhanced 3D scene understanding.

In summary, our contributions are:

- NeRP3D preserves the full knowledge from pre-training, since the NeRF-resembled design makes it effectively inherit and utilize continuous and fine-grained representations for both pretext and downstream tasks.
- Regardless of tasks, NeRP3D provides a unified framework allowing for consistent feature extraction with adaptive sampling, ray-wise and uniform spatial sampling, available through our proposed continuous function.

2 Related Work

Neural Radiance Fields Neural radiance fields (NeRFs) Mildenhall et al. (2021) and their variants Wang et al. (2021); Fridovich-Keil et al. (2022); Müller et al. (2022); Barron et al. (2022; 2023) have established a powerful paradigm for 3D scene reconstruction by learning continuous volumetric functions from posed multi-view images. NeRFs are typically trained in a self-supervised manner, minimizing photometric reconstruction loss across multiple views. These prior works have demonstrated their ability to understand and enhance fine 3D geometry and appearance through high-fidelity novel view synthesis and 3D reconstruction. To move from dense toward sparse image sets, conditioning the radiance fields with image features Yu et al. (2021); Chen et al. (2021); Liu et al. (2022b) shows reliable novel view synthesis results, demonstrating that generic 2D representations can guide NeRF training. Moreover, depth supervision Roessle et al. (2022); Wei et al. (2023a); Deng et al. (2022); Wei et al. (2021) is incorporated to understand more accurate geometry. NeRF's enhanced 3D understanding is increasingly being extended to autonomous driving applications, and NeRP3D aims to fully leverage these capabilities.

Neural Radiance Fields with Autonomous Driving The inherent ability of NeRFs Mildenhall et al. (2021); Wang et al. (2021); Barron et al. (2022; 2023) to capture 3D scene structure from multi-view 2D observations in a self-supervised manner has positioned them as a promising foundation for various autonomous driving applications. For sensor simulation in driving environments, offline scene reconstruction methods Yang et al. (2023c); Tonderski et al. (2024); Yang et al. (2023b) have demonstrated NeRFs' capability to synthesize realistic camera images, generate scenarios through object manipulation, and decompose static-dynamic scenes. Moreover, DistillNeRF Wang et al. (2024) builds upon EmerNeRF Yang et al. (2023b) by extending it into a feed-forward model, while feature distillation from 2D foundation models Radford et al. (2021); Oquab et al. (2023) further enhances 3D scene understanding.

The most relevant branch of this paper is the integration of NeRFs in pre-training to improve downstream perception tasks. UniPAD Yang et al. (2024) introduces a universal NeRF-based pre-training framework to enhance the 3D object detection downstream task. Occupancy predictions Huang et al. (2024); Zhang et al. (2023) are also integrated with NeRF, which is optimized through multi-view consistency Godard et al. (2017; 2019); Zhou et al. (2017). GaussianPretrain Xu et al. (2024) has demonstrated the feasibility of 3D Gaussian Splatting Kerbl et al. (2023) for pre-training 3D scene

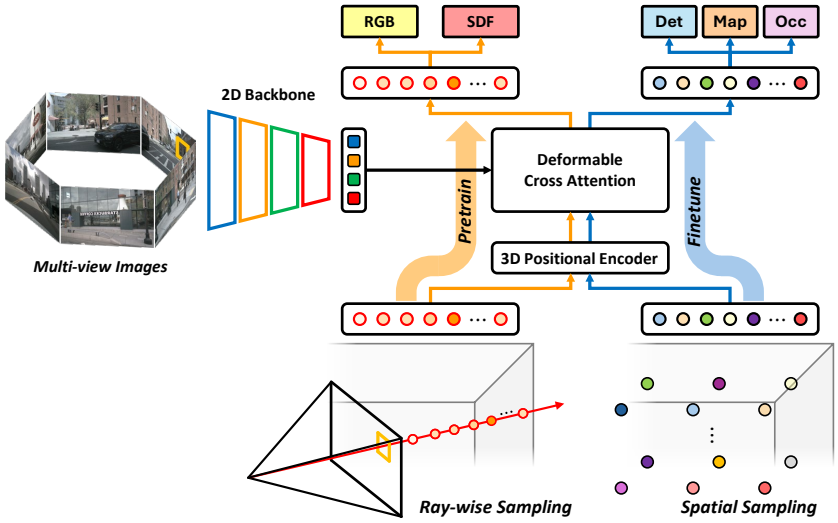

Figure 3: Overview of NeRP3D, illustrating both pre-training for rendering (orange) and fine-tuning for downstream (blue) pipelines. Through NeRF-resembled design, our method maintains a coherent 3D understanding from scattered points across diverse tasks while accommodating task-specific point sampling strategies, enabling the model to effectively leverage underlying geometric and appearance information while allowing for task-dependent feature specialization.

representations in driving environments. However, existing methods Yang et al. (2024); Huang et al. (2024); Tian et al. (2023); Xu et al. (2024), which rely on view transformation, have inherent constraints that diminish NeRF's capacity for continuous and fine-grained 3D representation. Moreover, pre-trained NeRF is discarded during downstream tasks, resulting in suboptimal 3D representations enhancement from pre-training. In contrast, NeRP3D fully inherits pre-trained NeRF knowledge and utilizes continuous and fine-grained representations through its NeRF-resembled design.

3 Method

NeRP3D is a simple and effective NeRF-resembled architecture that unifies scene reconstruction and perception tasks from single-timestep multi-view images. As illustrated in Fig. 3, our framework operates in two distinct stages within a unified architecture, without discarding or adding modules depending on stage or task requirements. This unified architecture enables adaptive exploration of regions of interest tailored to specific processing efficiency, while maintaining a coherent 3D understanding across diverse tasks.

3.1 Adaptive Sampling & Representation of Point

To reconstruct accurate 3D representations from sparse and dynamic multi-view inputs, NeRP3D directly samples 3D points of interest at arbitrary spatial locations and predicts the representation of sampled points with 2D image features to cope with dynamic driving scenes, without processing voxelized feature grids or any interpolation from them.

NeRP3D first samples 3D points $\mathbf{x} \in \mathbb{R}^3$ using one of two distinct strategies tailored to different processing phases, view-dependent ray-wise sampling and uniform spatial sampling. For volumetric rendering, we follow the standard NeRF. Specifically, for each pixel in the multi-view images, we define a camera ray $\mathbf{r}_i$ based on its origin $\mathbf{o}_i$ and direction $\mathbf{d}_i$, which are derived from camera intrinsics and extrinsics. Along each ray, we sample a set of points $\{\mathbf{x}_{ij} = \mathbf{o}_i + t_j\mathbf{d}_i\}$ at regular or stratified distances within a defined range $\{t_j | j = 1, ..., D, t_j < t_{j+1}\}$. These sampled points are then integrated into rendered color and depth along the ray for differentiable volumetric rendering. In contrast, for downstream tasks, where the goal is to utilize the learned 3D representation for autonomous driving tasks such as 3D object detection or occupancy prediction, we sample points across the scene volume rather than following camera rays. We sample points $\mathbf{x}_{xyz}$ uniformly in 3D space around the vehicle, covering regions relevant to perception tasks.

Despite the difference in sampling methods, all 3D points, whether sampled along camera rays or spatially, are represented in the identical system, ensuring consistency across tasks and sharing a unified spatial understanding. In addition, we parameterize 3D coordinates to account for unbounded environments, inspired by Barron et al. (2022):

$$p(\mathbf{x}') = \begin{cases} \alpha \mathbf{x}' & |\mathbf{x}'| \leq 1 \\ \left(1 - \frac{(1-\alpha)}{|\mathbf{x}'|}\right) \frac{\mathbf{x}'}{|\mathbf{x}'|} & |\mathbf{x}'| > 1 \end{cases}, \tag{1}$$

where $p(\cdot)$ denotes a parameterized function that preserves real-scale coordinates for points within the inner range, while distributing distant points proportionally to disparity, including those at infinite distance. $\mathbf{x}'$ denotes normalized $\mathbf{x}$ to the range $[0, 1]$ and $\alpha \in [0, 1]$ denotes the contraction ratio.

After sampling 3D points, a set of 3D points $\{\mathbf{x}\}$ is conditioned with sparse 2D observations to represent 3D dynamic environments in a feed-forward manner. Given *N* multi-view images $\{I_i\}_{i=1}^{N}$, we feed each image to the image backbone to obtain 2D image features $\mathbf{F} \in \mathbb{R}^{N \times H \times W \times C}$. Then, to enhance 3D point representations with image-aligned context, we adopt a deformable cross-attention Zhu et al. (2020) with 2D image features $\mathbf{F}$. We first encode each 3D query point $\mathbf{x}$ by $\gamma(\cdot)$ and learn a set of N_s sampling offsets $\{\Delta\pi_s \mid s = 1, ..., N_s\}$ relative to its projected 2D location $\pi(\mathbf{x})$, focusing interaction with relevant image regions. The final representation $\mathbf{z}$ of 3D point $\mathbf{x}$ is defined as:

$$\mathbf{z} = \sum_{h=1}^{N_h} \mathbf{W}_h \sum_{s=1}^{N_s} \mathbf{A}_{h,s} \mathbf{W}'_s \mathbf{F}(\pi(\mathbf{x}) + \Delta\pi_{h,s}(\gamma(p(\mathbf{x}')))), \tag{2}$$

where $\mathbf{N_h}$ denotes the number of heads for multi-head attention. $\mathbf{W_h} \in \mathbb{R}^{C \times (C/N_h)}$ and $\mathbf{W}'_s \in \mathbb{R}^{(C/N_h) \times C)}$ denotes learnable weights and $\mathbf{A}_{hs}$ denotes the attention weights which are normalized as $\sum_s \mathbf{A}_{h,s} = 1$. The resulting point embedding $\mathbf{z}$ serves as input to both rendering heads and detection heads described in the following sections.

3.2 Point-based 3D Scene Reconstruction & Perception

Volumetric Rendering To support 3D scene understanding for downstream tasks in autonomous driving, we first optimize radiance fields in a self-supervised manner Yang et al. (2024), using the signed distance function (SDF) and RGB reconstruction to represent 3D geometry and appearance. Given a set of sampled points along each ray and its embedded features $\{\mathbf{z}_{ij}\}$, RGB color values of 3D points $\mathbf{x}_j$ are predicted by $c_j = \phi_{rgb}(\mathbf{z}_j, \mathbf{d}_i)$, and its signed distance s_j extracted by signed distance function $\phi_{sdf}(\mathbf{z}_j)$ is transformed into opacity α_j derived with:

$$\alpha_j = \max\left(\frac{\Phi_\omega(\phi_{sdf}(\mathbf{z_j})) - \Phi_\omega(\phi_{sdf}(\mathbf{z}_{j+1}))}{\Phi_\omega(\phi_{sdf}(\mathbf{z}_j))}, 0\right), \tag{3}$$

where $\Phi_\omega(x) = (1 + e^{-\omega x})^{-1}$ is the sigmoid function with a learnable parameter ω. Then, the unbiased and occlusion-aware weights Wang et al. (2021) $w_j = T_j \alpha_j$ is computed from α_j, where $T_j = \prod_{k=1}^{j-1}(1 - \alpha_k)$ is the accumulated transmittance. The final color and depth values are computed by accumulating the contributions of 3D points sampled along ray $\mathbf{r}_i$, weighted by the probability distribution $\{w_j\}$:

$$\hat{\mathbf{C}}(\mathbf{r}_i) = \sum_{j=1}^{D} w_j \mathbf{c}_j, \quad \hat{D}(\mathbf{r}_i) = \sum_{j=1}^{D} w_j t_j, \tag{4}$$

where $\hat{\mathbf{C}}(\mathbf{r}_i)$ and $\hat{D}(\mathbf{r}_i)$ denote the predicted color and depth corresponding to the ray $\mathbf{r}_i$, respectively.

To optimize the neural radiance field, we employ a combination of RGB reconstruction, depth supervision, and multi-view consistency losses. We adopt the standard volumetric rendering loss from NeRFs, comparing the rendered color $\hat{\mathbf{C}}(\mathbf{r}_i)$ against the ground truth pixel color $\mathbf{C}(\mathbf{r}_i)$ for sampled rays $\mathcal{R} = \{\mathbf{r}_i\}$. To further constrain the 3D geometry, we leverage explicit depth supervision Deng et al. (2022); Yang et al. (2024) for $\mathbf{r}_i$ against LiDAR measurements $D_{lidar}(\mathbf{r}_i)$ where available. Furthermore, while LiDAR provides direct supervision, it suffers from sparse scan patterns and cannot capture regions such as the sky, transparent surfaces (*e.g.*, windows), or distant backgrounds where depth is undefined or unprojectable. To address this without additional annotations Yang et al. (2023b) or distillation from 2D foundation models Oquab et al. (2023); Kirillov et al. (2023), we further

enforce multi-view consistency Godard et al. (2019); Cao & De Charette (2023) by minimizing the discrepancy in predicted depth distributions across different views as:

$$\mathcal{L}_{reproj} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r}_i \in \mathcal{R}} \sum_{\mathbf{x}_j \in \mathbf{r}_i} w_j |I_t(\mathbf{r}_i) - I_s(\pi_s(\mathbf{x}_j))|, \tag{5}$$

where $I_t(\mathbf{r}_i)$ denotes the color value of a pixel in a target or current image I_t corresponding to the ray $\mathbf{r}_i$. $\pi_s(\mathbf{x})$ denotes the projection matrix from 3D points to 2D pixels on a source image I_s, such as a previous I_{t-1} or future image I_{t+1}. Consequently, the sampled 3D point $\mathbf{x}_j = \mathbf{o}_i + t_j \mathbf{d}_i$ along the ray $\mathbf{r}_i$ is projected on the source image, and the corresponding pixel color $I_s(\pi_s(\mathbf{x}_j))$ is compared with $I_t(\mathbf{r}_i)$ in weighted sum $\{w_j\}$. The overall loss for pre-training consists of RGB reconstruction loss, depth supervision loss, and reprojection loss:

$$\mathcal{L}_{pretrain} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{reproj}\mathcal{L}_{reproj} \tag{6}$$

where λ_{rgb}, λ_{depth}, and λ_{reproj} are the loss scale factors for each pre-training loss. $\mathcal{L}_{rgb}$ is RGB reconstruction loss and $\mathcal{L}_{depth}$ is depth estimation loss directly supervised by LiDAR measurements.

Open-World Perception Unlike view-dependent volumetric rendering, perception tasks require comprehensive spatial coverage of the vehicle's surroundings. All we need to do with NeRP3D is scatter the points $\{\mathbf{x}\} \in \mathbb{R}^{N \times 3}$ throughout the space and reshape the resulting representations $\{\mathbf{z}\} \in \mathbb{R}^{N \times C}$ from Eq. 2 to be compatible with task-specific detection heads, for example, $\{\mathbf{z}\} \in \mathbb{R}^{(X \times Y \times Z) \times C}$ for occupancy prediction. This straightforward adaptation maintains the enhanced geometric and appearance information learned during pre-training while enabling seamless integration with established perception architectures.

4 Experiments

We demonstrate NeRP3D on the nuScenese Caesar et al. (2020) dataset against the *state-of-the-art* NeRF-based pre-training approaches as well as comparable methods. Our experiments are designed to assess both pre-trained 3D representations by scene reconstruction and the effectiveness of finetuning for downstream tasks.

4.1 Dataset

We conduct experiments using the nuScenes dataset Caesar et al. (2020), which provides 700, 150, and 150 scenes for training, validation, and testing, respectively. We follow this data split for both the pretext and downstream tasks. Each scene provides 6 RGB camera images that cover a full 360° field of view, along with a 32-beam LiDAR point cloud and 3D radar point cloud data. The key samples are annotated at 2 Hz and support multiple tasks for autonomous driving, including 3D object detection, HD map construction, and segmentation. Recently, the annotations for occupancy prediction have been made available through Occ3D Tian et al. (2023) and SurroundOcc Wei et al. (2023b), providing dense 3D semantic occupancy labels. In our experiments, we adopt the Occ3D benchmark for the occupancy prediction.

Moreover, to evaluate generalization across different data distributions and sensor configurations, we additionally utilize Argoverse 2 (AV2) Wilson et al. (2023) dataset. AV2 provides 1,000 driving sequences with a distinct sensor suite comprising seven high-resolution ring cameras (2048×1550) covering a 360° field of view and two 32-beam LiDARs. This setup introduces significant domain shifts in environmental statistics and sensor layouts compared to nuScenes Caesar et al. (2020). This distinct setup serves to assess the model's robustness to domain changes and its data efficiency under limited supervision. For our experiments, we resized the input images to 800×450 and utilized only a $1/4$ subset of the training data.

4.2 Evaluation Metrics

We evaluate performance across two pretext scene reconstruction tasks and three downstream 3D perception tasks by following standard evaluation protocols for each task to ensure comparability with existing methods.

Table 1: **Downstream detection performance**

(a) **3D object detection**

Method	Pre-train	NDS↑	mAP↑
UVTR-C	ImageNet	44.1	37.2
BEVFormerV2	ImageNet	46.7	39.6
TPVFormer†	SelfOcc	33.5	31.0
UVTR-C†	UniPAD	37.1	33.7
NeRP3D†	Ours	**39.2**	**35.8**
UVTR-C	UniPAD	45.5	41.6
NeRP3D	Ours	**47.3**	**42.8**

(b) **Occ prediction**

Method	mIoU
BEVDet	19.38
BEVFormer	26.88
TPVFormer	27.83
CTF-Occ	28.53
SelfOcc	29.65
UniPAD	34.05
NeRP3D	**35.49**

(c) **HD map construction**

Method	Pre-train	Epochs	mAP
HDMapNet	ImageNet	30	23.0
VectorMapNet	ImageNet	110	40.9
MapTR-tiny	ImageNet	24	49.9
TPVFormer	SelfOcc	24	53.9
UVTR-C	UniPAD	24	57.8
NeRP3D	Ours	24	**59.1**

Scene Reconstruction Tasks We compare scene reconstruction quality by generating rendered RGB and depth images 1:2 the size of the input images. RGB reconstructed images are evaluated for all rendered pixels by Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), following standard NeRF evaluation protocols. For depth estimation, we report relative errors (AbsRel & SqRel), root mean squared error (RMSE & RMSE log), and accuracy under threshold δ metrics up to $80m$, only for pixels where the lidar point cloud with 20 sweeps is projected.

Downstream Tasks We evaluate the performance of 3D object detection using the mean Average Precision (mAP) and nuScenes Detection Score (NDS) under the standard nuScenes evaluation protocol. The perception range for object detection is set to $[-51.2m,\ 51.2m]$ along both the X and Y axes. For vectorized HD map construction, we follow the nuScenes map annotation benchmark and report mAP under *Chamfer* distance thresholds ($\tau \in \{0.5, 1.0, 1.5\}$). The evaluation range is set to $[-15.0m,\ 15.0m]$ for the X axis and $[-30.0m,\ 30.0m]$ for the Y axis. Occupancy prediction aims to predict the semantic classes of $0.4m \times 0.4m \times 0.4m$ voxels covering $[-40m,\ 40m]$ in both the X and Y axes and $[-1.0m,\ 5.4m]$ along the Z axis. The prediction result is evaluated using mean Intersection over Union (mIoU) across 17 semantic classes.

4.3 Implementation Details

To ensure fair comparison with prior works Yang et al. (2024); Huang et al. (2024), we adopt identical pre-training architectures and detection heads. We leverage NeuS Wang et al. (2021) for radiance field pre-training, following previous studies. Furthermore, we conduct downstream tasks based on UVTR-C Li et al. (2022), MapTR Liao et al. (2023), and Occ3D (CTF-Occ) Tian et al. (2023) for 3D object detection, HD map construction, and occupancy prediction, respectively. Class-balanced sampling (CBGS) or specialized data augmentations are not applied for finetuning, and all downstream tasks are trained and evaluated using single-timestep images only, without temporal information or frame stacking.

Our implementation builds upon the MMDetection3D Contributors (2020) framework, and training is conducted on 4 NVIDIA A6000 GPUs. The input image resolution varies by tasks, set to 1600×900 for object detection and 800×450 for rendering, HD map construction, and occupancy prediction. We both pre-train and fine-tune the model for 24 epochs using the AdamW optimizer, with an initial learning rate of 2e-4 and a weight decay of 0.01. The loss scale factors are set to $\lambda_{rgb} = \lambda_{depth} = \lambda_{reproj} = 10$. Unless otherwise specified, we fine-tune the models on a 1/2 subset for 12 epochs with 800×450 images in ablation studies.

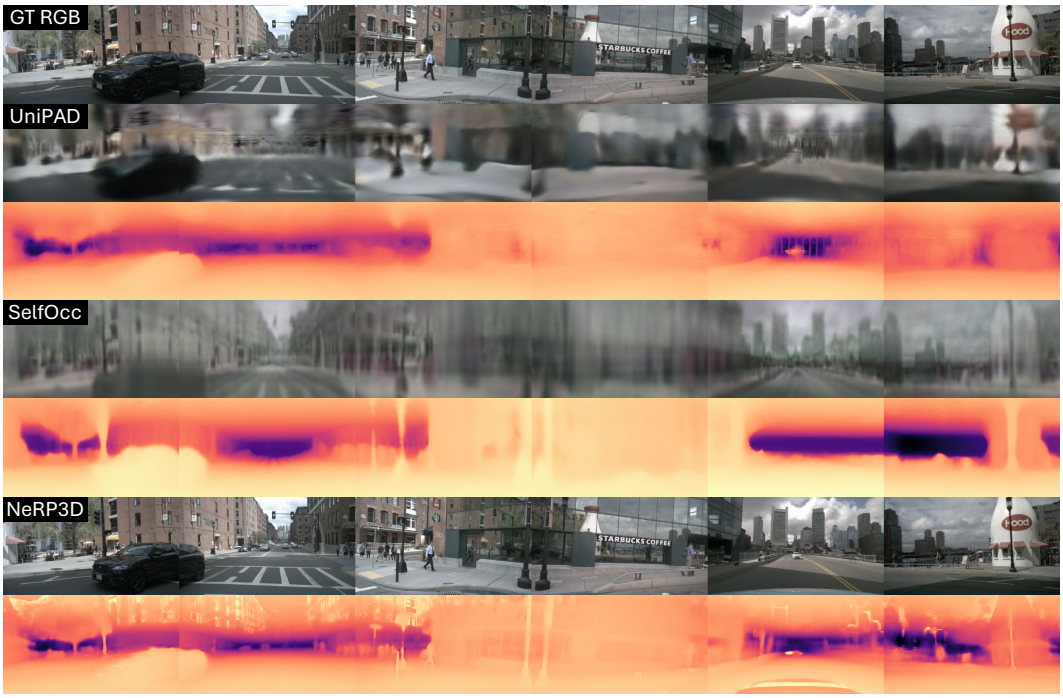

Figure 4: Qualitative comparison on rendered RGB & depth. NeRP3D outperforms *state-of-the-art* methods on both RGB and depth reconstruction. Our approach maintains high fidelity in urban scenes without any blur and pattern artifacts. For depth estimation, NeRP3D distinguishes individual people in crowded areas rather than merging them into indistinct blobs, and precisely captures thin structures such as poles that are often missed or reconstructed as thick structures by competing methods.

Table 2: **Pretext scene reconstruction performance**

(a) **Depth estimation**

Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓
SelfOcc	0.311	3.808	8.503	0.391
SelfOcc*	<u>0.215</u>	2.743	**6.706**	**0.316**
UniPAD	0.218	<u>2.512</u>	7.937	0.356
NeRP3D	**0.183**	**2.274**	<u>7.884</u>	<u>0.353</u>

(b) **RGB reconstruction**

Method	PSNR↑	SSIM↑	LPIPS↓
SelfOcc	18.82	0.536	0.657
UniPAD	21.14	0.549	0.634
NeRP3D	**33.42**	**0.969**	**0.070**

4.4 Main Results

3D Object Detection We compare NeRP3D with previous 3D object detection approaches Li et al. (2024; 2022); Liu et al. (2022a); Shu et al. (2023); Yang et al. (2023a); Yan et al. (2023) on the nuScenes *val* set. To compare with previous NeRF-based pre-training methods on detection, we follow the fine-tuning framework of UniPAD Yang et al. (2024) and also reproduce the results of both UVTR-C (UniPAD) Li et al. (2022); Yang et al. (2024) and TPVFormer (SelfOcc) Huang et al. (2023; 2024) by replacing the NeRF network for pre-training with UVTR's object detection head. † in Tab .1 (a) denotes the result evaluated on input resolutions of 800×450. Compared to the *state-of-the-art* NeRF-based self-supervision methods, our method outperforms 1.8 mAP and 2.1 NDS on 800×450 1.2 mAP and 1.8 NDS on 1600×900 over UniPAD, as shown in Tab. 1 (a). This improvement stems from NeRP3D's ability to learn fine-grained 3D representations, which enables more precise localization of bounding boxes and better separation of nearby objects, as qualitatively suggested by the detailed features in Fig. 1 and sharp reconstructions in Fig. 4.

Occupancy Prediction In Tab. 1 (b), we evaluate the performance of our method on Occ3D-nuScenes for 3D occupancy prediction. Similar to 3D object detection, we fine-tune the backbones with the same occupancy prediction head Tian et al. (2023) after pre-training. Our approach inherits NeRF's strength in modeling fine-grained representations, leading to improved mIoU and consistent

Table 3: **Zero-shot scene reconstruction performance (Argoverse 2 → nuScenes)**

Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	PSNR↑	SSIM↑	LPIPS↓
UniPAD	0.985	11.767	14.963	4.390	18.668	0.432	0.577
NeRP3D	0.626	6.251	10.728	0.921	28.238	0.905	0.111

gains over UniPAD and SelfOcc. As a result, our NeRP3D outperforms UniPAD and SelfOcc by 2.8 and 9.2 mIoUs, respectively. The continuous and high-fidelity representations learned by NeRP3D are particularly beneficial for this dense prediction task, enabling the model to accurately discern object boundaries and capture intricate geometric details often missed by other methods.

HD Map Reconstruction We evaluate the accuracy of HD map construction to assess each method's capability for understanding static driving environments, particularly in detecting road boundaries, dividers, and pedestrian crossings. To facilitate this task, we commonly utilized the detection head of MapTR Liao et al. (2023) for fair comparison. As shown in Tab. 1 (c), our method achieves improved mAP compared to both UniPAD and SelfOcc, with gains of 1.3 and 5.2 mAP, respectively. HD map reconstruction is particularly challenging as it requires a nuanced semantic understanding to differentiate map elements like pedestrian crossings that are geometrically coplanar with the drivable surface. As visually evidenced in Fig. 1, the feature representations from NeRP3D make these elements distinctly separable, which is critical for precise map construction.

RGB & Depth Reconstruction To validate the effectiveness of the pre-training, the performance of NeRP3D on the pretext tasks is also compared with the previous NeRF-based pre-training methods Yang et al. (2024); Huang et al. (2024) on the nuScenes *val* set. As shown in Tab. 2, NeRP3D achieves remarkable enhancements in both depth estimation and RGB reconstruction. More specifically, the qualitative depth maps in Fig. 4 consistently demonstrate that our method yields more accurate and detailed depth estimations, particularly in complex regions, whereas UniPAD and SelfOcc struggle to resolve fine structures and depth discontinuities. For RGB reconstruction, UniPAD generates blurry and imprecise reconstructions lacking detailed textures, while SelfOcc produces grayish images with unidentified vertical patterns. In contrast, our approach reconstructs sharper images with rich colors, closely matching the ground truth without introducing patterned signals.

Generalization To assess the robustness of our method against domain shifts and varying sensor configurations, we conducted cross-dataset transfer experiments using Argoverse 2 (AV2) Wilson et al. (2023) for pre-training and nuScenes for evaluation. AV2 possesses distinct camera geometries and environmental statistics compared to nuScenes, serving as a rigorous testbed for generalization.

We first evaluated zero-shot scene reconstruction by directly applying the AV2 Wilson et al. (2023) pre-trained weights to nuScenes Caesar et al. (2020) without any fine-tuning. As presented in Tab. 3, NeRP3D demonstrates remarkable robustness, achieving an Abs Rel of 0.626 and PSNR of 28.24, significantly outperforming UniPAD (Abs Rel 0.985, PSNR 18.67). While the view transformation method Yang et al. (2024) suffer from severe degradation due to their dependency on fixed grid priors aligned with specific sensor layouts, NeRP3D's continuous point-based architecture effectively adapts to new sensor geometries. Qualitative results in Fig. 10 of Appendix further visualize this, showing that NeRP3D preserves structural details while the view transformation method produces blurry artifacts.

Moreover, we evaluated the transferability for 3D object detection. When pre-trained on AV2 Wilson et al. (2023) and fine-tuned on nuScenes Caesar et al. (2020), NeRP3D achieved 27.46 mAP, surpassing UniPAD Yang et al. (2024) (26.29 mAP) by a significant margin. This confirms that NeRP3D learns universal geometric representations that are not overfitted to specific sensor configurations or dataset distributions but are effectively transferable across domains.

Overall, these results demonstrate that our approach effectively leverages the inherent advantages of continuous and fine-grained representations derived from NeRF. NeRP3D not only significantly benefits pretext scene reconstruction tasks and downstream detection tasks but also ensures robust generalization across different sensor configurations and data distributions for autonomous driving. More comprehensive comparison and quantitative analysis of the experimental results are provided by Tab. 4−8 in Appendix A and B.

4.5 Ablation Studies

We conduct comprehensive ablation studies to analyze different model variants and evaluate their impact. Ablation results are reported in Appendix C and summarized in the following sections.

Cross-Task Generalization We further investigate whether the learned 3D representation remains valid across different task objectives. By performing volumetric rendering using the backbone fine-tuned for occupancy prediction, we observe that NeRP3D successfully retains structural details, whereas view-transformation methods suffer from catastrophic forgetting, collapsing into mean regression. This confirms that NeRP3D learns a task-agnostic continuous representation that preserves geometric fidelity regardless of downstream optimization pressure.

Adaptability View transformation is dependent on the range and voxel size, leading to severe performance degradation if the voxel-related parameters are changed against pre-training. In contrast, NeRP3D aims for a continuous representation without voxel-related parameters, and variations only correspond to simple changes in the range of interest.

Effectivness We analyze the effectiveness of NeRP3D in reducing the reliance on annotations by comparing previous works, ranging from the full dataset to a 1/8 subset. Consequently, NeRP3D maintains strong detection performance even with significantly reduced supervision, indicating the robustness of its NeRF-based pre-training.

Multi-view Consistency LiDAR-based supervision ensures more consistent depth estimation accuracy. However, we found that the sparsity and scan patterns of LiDAR are ultimately insufficient for reconstructing dense 3D geometry. To address LiDAR's sparsity and patterns, we not only rely on LiDAR supervision but also consider multi-view consistency and our sampling strategy tailored to this approach.

Design Validation We verify the necessity of our architectural choices through comprehensive comparisons. First, applying NeRF pre-training to existing point-based detectors Liu et al. (2023a) fails to transfer knowledge due to query mismatch, confirming the importance of our unified design. Second, comparisons between SDF (NeuS Wang et al. (2021)) and density (standard NeRF Mildenhall et al. (2021)) priors validate that SDF enforces clearer object boundaries beneficial for perception. Finally, we demonstrate that deformable attention outperforms standard attention by providing a necessary locality inductive bias, ensuring that the points remain faithful to their local spatial context.

5 Conclusion

In this paper, we present NeRP3D, a novel point-based 3D architecture for scene reconstruction and downstream perception tasks for autonomous driving. Our approach addresses the fundamental misalignment between view transformation and neural radiance fields. Through its NeRF-resembled design, NeRP3D fully inherits NeRF's continuous representation capabilities, enabling the model to maintain consistent geometric and appearance information at arbitrary spatial locations across both scene reconstruction and open-world perceptions. Although NeRP3D outperforms previous approaches, it struggles with depth beyond its ROI, relying on LiDAR. Additionally, its point-based architecture incurs high computational costs from adapting NeRF's output to existing detection heads. Future enhancements include temporal RGB reconstruction for consistency, density/opacity filtering for efficiency, and Gaussian splatting for real-time performance with point queries.

References

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19697–19705, 2023.

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

Anh-Quan Cao and Raoul De Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9387–9398, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14124–14133, 2021.

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

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12882–12891, 2022.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5501–5510, 2022.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 270–279, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023.

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

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

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17615–17626, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022.

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*, volume 37, pp. 1477–1485, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. In *International Conference on Learning Representations*, 2023.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*, pp. 531–548. Springer, 2022a.

Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3262–3272, 2023a.

Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7824–7833, 2022b.

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

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. *Communications of the ACM*, 65(1):99–106, 2021.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8555–8564, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12892–12901, 2022.

Juyeb Shin, Hyeonjun Jeong, Francois Rameau, and Dongsuk Kum. Instagram: Instance-level graph modeling for vectorized hd map learning. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Changyong Shu, Jiajun Deng, Fisher Yu, and Yifan Liu. 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3580–3589, 2023.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023.

Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14895–14904, 2024.

Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8406–8415, 2023.

Letian Wang, Seung Wook Kim, Jiawei Yang, Cunjun Yu, Boris Ivanovic, Steven Waslander, Yue Wang, Sanja Fidler, Marco Pavone, and Peter Karkus. Distillnerf: Perceiving 3d scenes from single-glance images by distilling neural fields and foundation model features. *Advances in Neural Information Processing Systems*, 37:62334–62361, 2024.

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. *arXiv preprint arXiv:2106.10689*, 2021.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5610–5619, 2021.

Yi Wei, Shaohui Liu, Jie Zhou, and Jiwen Lu. Depth-guided optimization of neural radiance fields for indoor multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (9):10835–10849, 2023a.

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

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15449–15458, 2024.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Shaoqing Xu, Fang Li, Shengyin Jiang, Ziying Song, Li Liu, and Zhi-xin Yang. Gaussianpretrain: A simple unified 3d gaussian representation for visual pre-training in autonomous driving. *arXiv preprint arXiv:2411.12452*, 2024.

Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18268–18278, 2023.

Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17830–17839, 2023a.

Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15238–15250, 2024.

Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023b.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023c.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4578–4587, 2021.

Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *CoRR*, 2023.

Haiming Zhang, Wending Zhou, Yiyao Zhu, Xu Yan, Jiantao Gao, Dongfeng Bai, Yingjie Cai, Bingbing Liu, Shuguang Cui, and Zhen Li. Visionpad: A vision-centric pre-training paradigm for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17165–17175, 2025.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1851–1858, 2017.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Table 4: **3D object detection** on the nuScenes *val* set. † denotes the result evaluated on input resolutions of 800×450 using MMDetection3D Contributors (2020) by integrating UVTR detection head Li et al. (2022); Yang et al. (2024). The other results are based on 1600×900 input resolution.

Method	Pre-train	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFormer-S	ImageNet	44.8	37.5	-	-	-	-	-
UVTR-C	ImageNet	44.1	37.2	0.735	0.269	0.397	0.761	0.193
PETR	ImageNet	44.2	37.0	0.711	2.670	0.383	0.865	0.201
3DPPE	ImageNet	45.8	39.1	-	-	-	-	-
BEVFormerV2	ImageNet	46.7	39.6	0.709	0.274	0.368	0.768	0.196
CMT-C	ImageNet	46.0	40.6	-	-	-	-	-
TPVFormer†	SelfOcc	33.5	31.0	0.785	0.285	0.729	1.232	0.399
UVTR-C†	UniPAD	37.1	33.7	0.734	0.283	0.603	1.250	0.359
NeRP3D†	Ours	**39.2**	**35.8**	0.719	0.288	0.640	0.977	0.250
UVTR-C	UniPAD	45.5	41.6	0.674	0.277	0.418	0.930	0.234
UVTR-C	GaussianPretrain	47.2	41.7	0.676	0.278	0.394	0.815	0.200
NeRP3D	Ours	**47.3**	**42.8**	0.664	0.276	0.425	0.811	0.196

Table 5: **3D occupancy prediction**. We compare our method against *state-of-the-art* occupancy prediction approaches on the Occ3d-nuScenes *val* set. Results for BEVDet, BEVFormer, TPVFormer, and CTF-Occ are directly taken from Occ3d Tian et al. (2023). † denotes the result reproduced using MMDetection3D Contributors (2020) on input resolutions of 800×450. $*$ denotes that the result is directly taken from VisionPAD Zhang et al. (2025), which is pre-trained only with camera modality and evaluated on input resolutions of 1600×900.

Method	mIoU	car	bus	bicycle	barrier	vegetation	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	others
BEVDet	19.4	34.5	32.3	0.2	30.3	15.1	13.0	10.3	10.4	6.3	8.9	23.7	52.7	24.6	26.1	22.3	15.0	4.4
BEVFormer	26.9	42.4	40.4	17.9	37.8	17.7	7.4	23.9	21.8	21.0	22.4	30.7	55.4	28.4	36.0	28.1	20.0	5.9
TPVFormer	27.8	45.9	40.8	13.7	38.9	16.8	17.2	20.0	18.9	14.3	26.7	34.2	55.7	35.5	37.6	30.7	19.4	7.2
CTF-Occ	28.5	42.2	38.3	20.6	39.3	18.0	16.9	24.5	22.7	21.1	23.0	31.1	53.3	33.8	38.0	33.2	20.8	8.1
SelfOcc†	29.7	43.8	40.0	10.0	36.3	30.6	13.7	11.8	16.5	15.7	23.2	29.3	79.1	37.3	47.7	28.0	34.8	6.2
UniPAD†	34.1	45.8	42.3	13.0	39.7	38.1	19.4	14.3	20.0	17.7	27.4	33.1	80.0	38.7	49.4	50.6	42.8	6.5
VisionPAD*	35.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NeRP3D†	**35.5**	49.4	43.9	15.0	41.0	38.8	19.2	20.0	23.6	16.5	27.9	36.7	81.0	37.4	49.8	53.6	43.9	5.5

A Downstream Detection Tasks

A detailed analysis of NeRP3D's performance is provided on three downstream perception tasks: 3D object detection, 3D occupancy prediction, and HD map construction. We expand upon the results presented in Sec. 4.4 and Tab. 1, with a focus on comprehensive comparisons against state-of-the-art methods, including those leveraging 3DGS (3D Gaussian Splatting)-based pre-training.

As shown in Tab. 4, NeRP3D achieves state-of-the-art performance in 3D object detection among NeRF-based pre-training methods, with an NDS of 47.3 and an mAP of 42.8. This represents a significant improvement over UniPAD, with gains of 1.8 NDS and 1.2 mAP when both are fine-tuned on the UVTR-C detector. Crucially, NeRP3D also outperforms GaussianPretrain Xu et al. (2024), which still relies on a view transformation backbone. In comparison, NeRP3D achieves a higher NDS (47.3 vs. 47.2) and a more substantial lead in mAP (42.8 vs. 41.7). The enhanced performance is attributed to NeRP3D's fine-grained 3D representation, which provides the necessary detail to identify far or occluded targets and resolve individuals within dense crowds, as shown in Fig. 8

For 3D occupancy prediction, NeRP3D's ability to model continuous geometry and appearance translates into superior performance. As demonstrated in Tab. 5, our method achieves an mIoU of 35.5, surpassing both UniPAD (34.1 mIoU) and SelfOcc (29.7 mIoU) by a significant margin. We further compare NeRP3D with VisionPAD Zhang et al. (2025), a vision-centric pre-training also based on 3D Gaussians. Even though VisionPAD is pre-trained only with camera modality, but

Table 6: **HD map construction** on the nuScenes *val* set. "C" and "L" denote camera and LiDAR modalities, respectively. Results for HDMapNet and VectorMapNet are directly taken from MapTRLiao et al. (2023).

Method	Modality	Pre-train	Epochs	mAP	AP_{ped}	$AP_{divider}$	$AP_{boundary}$
HDMapNet	C	ImageNet	30	23.0	14.4	21.7	33.0
HDMapNet	L	ImageNet	30	24.1	10.4	24.1	37.9
HDMapNet	C & L	ImageNet	30	31.0	16.3	29.6	46.7
VectorMapNet	C	ImageNet	110	40.9	36.1	47.3	39.3
VectorMapNet	L	ImageNet	110	34.0	25.7	37.6	38.6
VectorMapNet	C & L	ImageNet	110	45.2	37.6	50.5	47.5
MapTR-tiny	C	ImageNet	24	49.9	52.0	45.3	52.4
TPVFormer	C	SelfOcc	24	53.9	47.8	55.6	58.3
UVTR-C	C	UniPAD	24	57.8	**54.8**	58.5	61.5
NeRP3D	C	Ours	24	**59.1**	52.9	**62.2**	**62.2**

Table 7: **Depth estimation** on nuScenes *val* set. We conduct evaluation at a downsampled resolution of 114×64 for EmerNeRF Yang et al. (2023b) and DistillNeRF Wang et al. (2024) and 400×225 for others. † denotes per-scene optimization, not feedforward model. $*$ denotes only *depth-optimized* variant of SelfOcc Huang et al. (2024). The results of EmerNeRF and DistillNeRF are taken from the paper of DistillNeRF.

Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
EmerNeRF†	0.073	0.346	2.696	0.159	0.942	0.975	0.986
DistillNeRF	0.248	3.090	6.096	0.312	0.704	<u>0.885</u>	<u>0.947</u>
DistillNeRF-D	0.233	2.890	<u>5.890</u>	<u>0.296</u>	0.703	0.881	0.945
DistillNeRF-DV	0.223	**1.776**	**5.461**	**0.293**	<u>0.763</u>	**0.903**	**0.961**
SelfOcc	0.311	3.808	8.503	0.391	0.641	0.803	0.888
SelfOcc*	0.215	2.743	6.706	0.316	0.753	0.875	0.932
UniPAD	<u>0.218</u>	2.512	7.937	0.356	<u>0.763</u>	0.869	0.921
NeRP3D	**0.183**	<u>2.274</u>	7.884	0.353	**0.799**	0.883	0.926

evaluated on the higher resolution 1600×900, NeRP3D achieves a competitive overall mIoU (35.5 vs. 35.4). A class-level breakdown reveals that NeRP3D shows notable improvements in thin and small categories, as shown in Fig. 9, such as bicycle (15.0 vs. 13.0), motorcycle (20.0 vs. 14.3), and pedestrian (23.6 vs. 20.0).

The comprehensive results for downstream perception tasks indicate that our NeRP3D, which avoids the conflicting priors between the pre-training method and 3D backbone, enables the learning of continuous and fine-grained 3D representations that directly benefit downstream detection tasks.

B Pretext Scene Reconstruction Tasks

The overall performance of RGB reconstruction and depth estimation is compared with previous NeRF-based pre-training methods Yang et al. (2024); Huang et al. (2024) and comparable methods on the nuScenes *val* set, as shown in Tab. 7 and 8. Specifically, EmerNeRF Yang et al. (2023b) is a *per-scene* optimization model, and the variants of DistillNeRF Wang et al. (2024) are *without* distillation, *with* depth distillation (noted as "D"), and *with* virtual camera distillation (noted as "V").

The depth estimation results in Tab. 7 demonstrate clear benefits from our NeRF-inherited representation learning. SelfOcc* shows competitive depth estimation, but this variant does not support

Method	PSNR ↑	SSIM ↑	LPIPS ↓
EmerNeRF	30.88	0.879	-
DistillNeRF-D	30.11	0.917	-
SelfOcc	18.82	0.536	0.657
UniPAD	21.14	0.549	0.634
NeRP3D	**33.42**	**0.969**	**0.070**

Table 8: **RGB reconstruction** on nuScenes *val* set at a resolution of 228×114 for EmerNeRF Yang et al. (2023b) and DistillNeRF Wang et al. (2024) and 400×225 for others. The results of EmerNeRF and DistillNeRF are taken from the paper of DistillNeRF.

Table 9: **Multi-resolution reconstruction analysis.** We evaluate reconstruction quality across varying image scales (from $1/16$ to $1/4$) to isolate the impact of discretization levels on representational fidelity.

Method	Scale	PSNR↑	SSIM↑	LPIPS↓
UniPAD Yang et al. (2024)	1/16	23.55	0.796	0.250
	1/8	22.49	0.664	0.442
	1/4	21.14	0.549	0.634
NeRP3D	1/16	26.00	0.862	0.116
	1/8	29.40	0.919	0.090
	1/4	33.42	0.969	0.070

Table 10: **Evaluation of Cross-Task Generalization and Structural Retention.** (a) Per-pixel evaluation: Standard metrics measuring absolute reconstruction errors, which are sensitive to scale shifts. (b) Structural evaluation: Scale-invariant and perceptual metrics assessing geometric fidelity and structural integrity, independent of feature scale variations induced during fine-tuning.

(a) **Per-pixel evaluation**

Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	PSNR↑	SSIM↑
UniPAD Yang et al. (2024)	0.477	6.914	15.104	1.056	11.623	0.283
NeRP3D	2.192	12.372	19.459	1.185	9.308	0.135

(b) **Structural and scale-invariant evaluation**

Method	SI RMSE↓	Grad Loss↓	GMSD↓	LPIPS↓	PSNR-HM↑	SSIM-HM↑
UniPAD Yang et al. (2024)	0.859	90.164	0.306	0.863	12.319	**0.300**
NeRP3D	**0.643**	**83.739**	**0.289**	**0.671**	**12.839**	0.285

RGB reconstruction. On the other hand, the variant of SelfOcc that supports both RGB and depth reconstruction exhibits comparatively lower accuracy. Compared to UniPAD, our method achieves better performance across multiple metrics, such as AbsRel (0.183 vs. 0.218), SqRel (2.274 vs. 2.512), and RMSE (7.884 vs. 7.937). Moreover, accuracy within specific depth thresholds (δ metrics) further underscores the robustness of our model in reconstructing precise depth values. When compared with DistillNeRF, which is specifically designed for scene reconstruction, our NeRP3D achieves competitive depth estimation accuracy despite not relying on dense depth maps obtained from per-scene optimization Yang et al. (2023b) or distillation from 2D foundation models Radford et al. (2021); Oquab et al. (2023).

For RGB reconstruction, NeRP3D significantly outperforms previous approaches, as shown in Tab. 8. Compared to previous feedforward methods and EmerNeRF, PSNR and SSIM are improved by 33.42 and 0.969, respectively. Our method also notably reduces LPIPS, reflecting more perceptually accurate reconstructions over UniPAD and SelfOcc by 0.070.

To quantitatively verify the conflicting prior between discrete view transformation and continuous neural rendering representations, we evaluated reconstruction performance across varying resolutions as shown in Tab. 9. The view transformation method (UniPAD Yang et al. (2024)) degrades at higher resolutions, confirming that discrete voxel grids act as a "low-pass filter". As a result, UniPAD masks errors at coarse scales but fails to capture high-frequency details due to the rigid bottleneck. In contrast, NeRP3D demonstrates superior representational fidelity with a widening performance gap at finer scales. This quantitatively proves that our continuous architecture resolves the structural mismatch, successfully modeling fine-grained geometry that fixed grids cannot capture.

C Ablation Studies

C.1 Cross Task Generalization

We investigate whether the learned 3D representations remain valid across different task objectives, specifically evaluating the "Radiance Modeling Ability" of the backbone after fine-tuning for occupancy prediction. In this experiment, we utilize the backbone encoder fine-tuned for the downstream task while keeping the pre-trained rendering heads (RGB and SDF decoder) frozen.

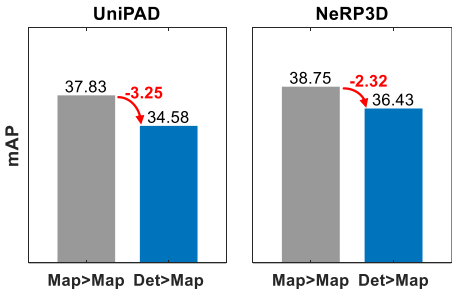

Figure 5: Comparison of performance variation with changes in detection range between the pre-training and fine-tuning phases.

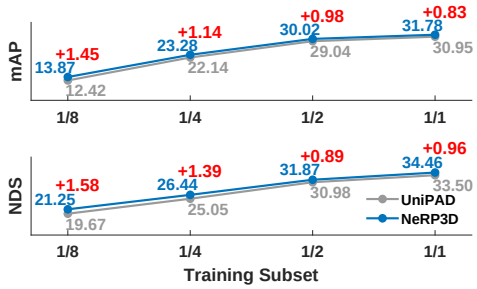

Figure 6: Comparison of pre-training effectiveness: Impact of 3D backbone and pre-training network alignment on performance retention across varying annotated training data sizes.

As shown in Fig. 11, there is a stark contrast in the retained representations; the view transformation method like UniPAD Yang et al. (2024) suffers from catastrophic forgetting, producing blurry outputs that indicate a loss of fine-grained 3D information and a collapse into mean regression. In contrast, NeRP3D successfully retains structural understanding, with key elements remaining clearly distinguishable. Quantitative results in Tab. 10 further support this. While standard per-pixel metrics are sensitive to feature scale shifts induced during fine-tuning, often favoring the mean regression of UniPAD, NeRP3D significantly outperforms the view transformation method in scale-invariant (SI-RMSE) and perceptual (LPIPS, GMSD) metrics. This confirms that, unlike view transformation methods that overfit to specific tasks and collapse into mean regression, NeRP3D learns a robust and continuous representation that maintains geometric fidelity across diverse downstream objectives.

C.2 Adaptability

We evaluate the adaptability of NeRP3D compared to the previous NeRF-based pre-training method when transferring from one detection range for pre-training to another for fine-tuning. We pre-train UniPAD Yang et al. (2024) and our NeRP3D on the detection range optimized for 3D object detection with the full training set and subsequently fine-tune for HD map construction on a 1/2 training set. Detailed detection range for 3D object detection and HD map construction is described in Sec. 4.2.

In Fig. 5, "Map>Map" denotes that the detection range remains the same for HD map construction in both phases, while "Det>Map" indicates a change in detection range from 3D object detection during pre-training to HD map construction during fine-tuning. As a result, while view transformation-based approaches suffer substantial performance drops due to the fundamental modification of volumetric features (the size of a tensor and voxels) when changing detection range with voxel size, NeRP3D maintains consistent representation quality across different spatial configurations. This is because NeRP3D's point-based architecture only requires adjusting the coordinates of sampled points without altering the underlying representation itself. The continuous nature of our NeRF-resembled architecture highlights a key advantage of NeRP3D, namely the ability to generalize across tasks with different spatial requirements without compromising the quality of learned representations, further demonstrating the benefits of our unified point-based approach over discretized view transformation approaches.

C.3 Effectiveness

We investigate the effectiveness of pre-training knowledge transfer in terms of the alignment of the 3D backbone and pre-training network by evaluating its performance when fine-tuned with varying amounts of annotated data. We compare the performance between UniPAD and NeRP3D when fine-tuned on 1/8, 1/4, 1/2, and the full training set.

As shown in Fig. 6, NeRP3D demonstrates robustness to reduced annotation quantities, with less performance degradation compared to UniPAD as the training set size decreases. This enhanced data efficiency can be attributed to the rich geometric and appearance information captured during pre-training, which provides a strong foundation for downstream tasks even with limited supervision. The alignment of the 3D backbone and the principle of NeRF-based pre-training enhances the

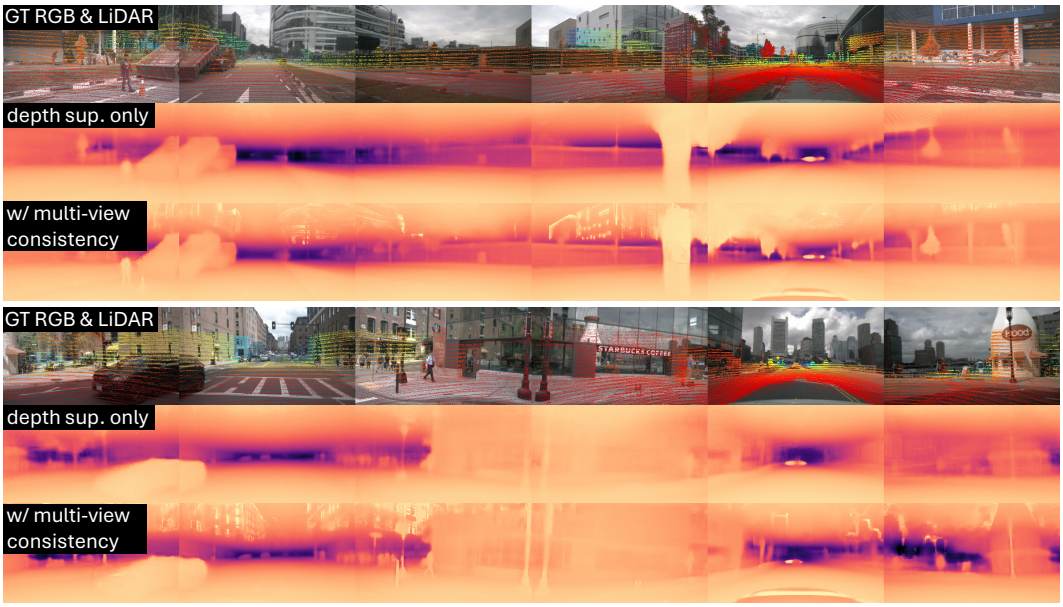

Figure 7: Qualitative comparison of depth estimation results. While LiDAR-based depth supervision alone shows limited improvement, incorporating multi-view consistency significantly enhances fine-grained and spatial accuracy, enabling plausible predictions of geometric structures even beyond the detection range.

Table 11: Ablation study on depth estimation performance with and without multi-view consistency. Sparse LiDAR scans define the *ground truth* of depth in this experiment.

Multi-view Consistency	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
✗	0.202	2.264	7.716	0.348	0.764	0.874	0.926
✓	0.183	2.274	7.884	0.353	0.799	0.883	0.926

effectiveness of knowledge transfer from pre-training to fine-tuning, allowing the model to generalize better from fewer examples in autonomous driving perception tasks.

C.4 Depth Supervision & Multi-view Consistency

We compare two approaches for depth pre-training: relying solely on LiDAR point cloud ground truth and incorporating multi-view consistency during training. When supervision is limited to LiDAR point clouds, depth estimation is accurate within the regions covered by the sensor. However, it cannot provide meaningful predictions in areas lacking LiDAR point cloud returns. In contrast, multi-view consistency enables the model to leverage geometric cues from overlapping camera views, but it is not as accurate as LiDAR point cloud supervision.

Qualitatively, the addition of multi-view consistency provides fine-grained depth quality, allowing the model to infer plausible geometric structures in regions where LiDAR supervision is unavailable or out of range, as shown in Fig. 7. However, since depth evaluation metrics are restricted to areas with sparse LiDAR point cloud ground truth, these improvements are not fully reflected in quantitative results. In fact, as shown in Tab. 11, a model explicitly trained to optimize these evaluation metrics may achieve slightly better numerical scores on some metrics by focusing exclusively on accurate prediction at sparse LiDAR points, while potentially sacrificing overall geometric coherence and depth consistency in regions without ground truth supervision.

Furthermore, Tab. 12 demonstrates how depth supervision during pre-training impacts downstream 3D object detection. The experiment is conducted on input resolutions of 800×450 with full data. Pre-trained with only cameras using multi-view consistency, our NeRP3D model establishes a strong baseline, achieving 38.6 NDS and 34.5 mAP, which already outperforms the LiDAR-assisted UniPAD model. Moreover, incorporating LiDAR-based depth supervision during pre-training further enhances

Table 12: Ablation study on 3D object detection performance with and without depth supervision from LiDAR. "C" and "L" under Pre-train Modality denote camera for multi-view consistency and LiDAR for depth supervision, respectively.

Method	Pre-train	Pre-train Modality	NDS↑	mAP↑
UVTR-C Li et al. (2022)	UniPAD Yang et al. (2024)	C & L	37.1	33.7
NeRP3D	Ours	C	<u>38.6</u>	<u>34.5</u>
NeRP3D	Ours	C & L	**39.2**	**35.8**

this performance, boosting performance to 39.2 NDS and 35.8 mAP. This result demonstrates both the inherent effectiveness of the NeRP3D architecture and the significant, additive benefit of using explicit geometric priors from LiDAR to improve detection accuracy.

C.5 Design Validation

All experiments to validate our design choice were conducted on a 1/4 subset with 704×256 images.

Consistency of 3D Point Priors. To verify the importance of our unified architecture rather than a point-based architecture, we applied rendering pre-training strategy to PETRv2 Liu et al. (2023a), a point-based architecture. While PETRv2 learned 3D representations from pre-training, the performance failed to transfer to detection (approx. 0.0 mAP). This failure stems from a fundamental disruption in the consistency of 3D point priors. In pre-training, queries represent specific spatial locations to encode geometry and radiance. However, PETR's fine-tuning forces a drastic shift where queries act as object instances, invalidating the learned spatial priors. In contrast, NeRP3D maintains consistent spatial point queries across tasks, ensuring effective knowledge transfer (20.70 mAP).

SDF Prior. We evaluated the impact of the geometric prior by replacing SDF (NeuS Wang et al. (2021)) with standard density (NeRF Mildenhall et al. (2021)). The SDF-based model achieved 20.70 mAP, outperforming the density-based variant (19.35 mAP). Since SDF enforces a hard surface constraint and creates sharp and unambiguous object boundaries. This structural clarity is critical for localizing and distinguishing precise objects in perception tasks, validating our design choice of using NeuS over standard NeRF.

Deformable vs. Standard Attention. We validated the effectiveness of deformable attention against standard global attention. Deformable attention achieved 20.70 mAP, significantly surpassing standard attention (18.38 mAP). Since each 3D point corresponds to a specific physical location, attending to the global context (standard attention) introduces irrelevant noise. Deformable attention enforces a necessary locality inductive bias by restricting the receptive field to the projected neighborhood. This proves that focusing on relevant local features is essential for accurate continuous representation.

D Computation Analysis

We evaluated the practicality and scalability by varying sampling densities. Tab. 13 demonstrates that NeRP3D operates at a computational level comparable to UniPAD Yang et al. (2024) while delivering enhanced performance. Crucially, detection accuracy scales linearly with sampling density. This scalability allows the model to be tuned for performance or efficiency, depending on the resource budget.

Table 13: **Computation analysis**

Method	GFLOPS	FPS	mAP
UniPAD	1250.10	5.59	19.12
NeRP3D	1903.77	4.47	20.70
	1621.35	4.91	19.69
	1492.60	5.25	19.20
	1315.03	5.54	18.89

E The Use of Large Language Models

During the preparation of this paper, we utilized publicly available large language models (LLMs) only to aid in polishing the writing. The model's role was strictly limited to improving grammar, refining sentence structure, and enhancing the overall clarity and readability of the text. All scientific contributions, including the core ideas, experimental design, and analysis of results, are exclusively our work. We carefully reviewed and edited all model-generated suggestions and retain full responsibility for the final content of this paper.

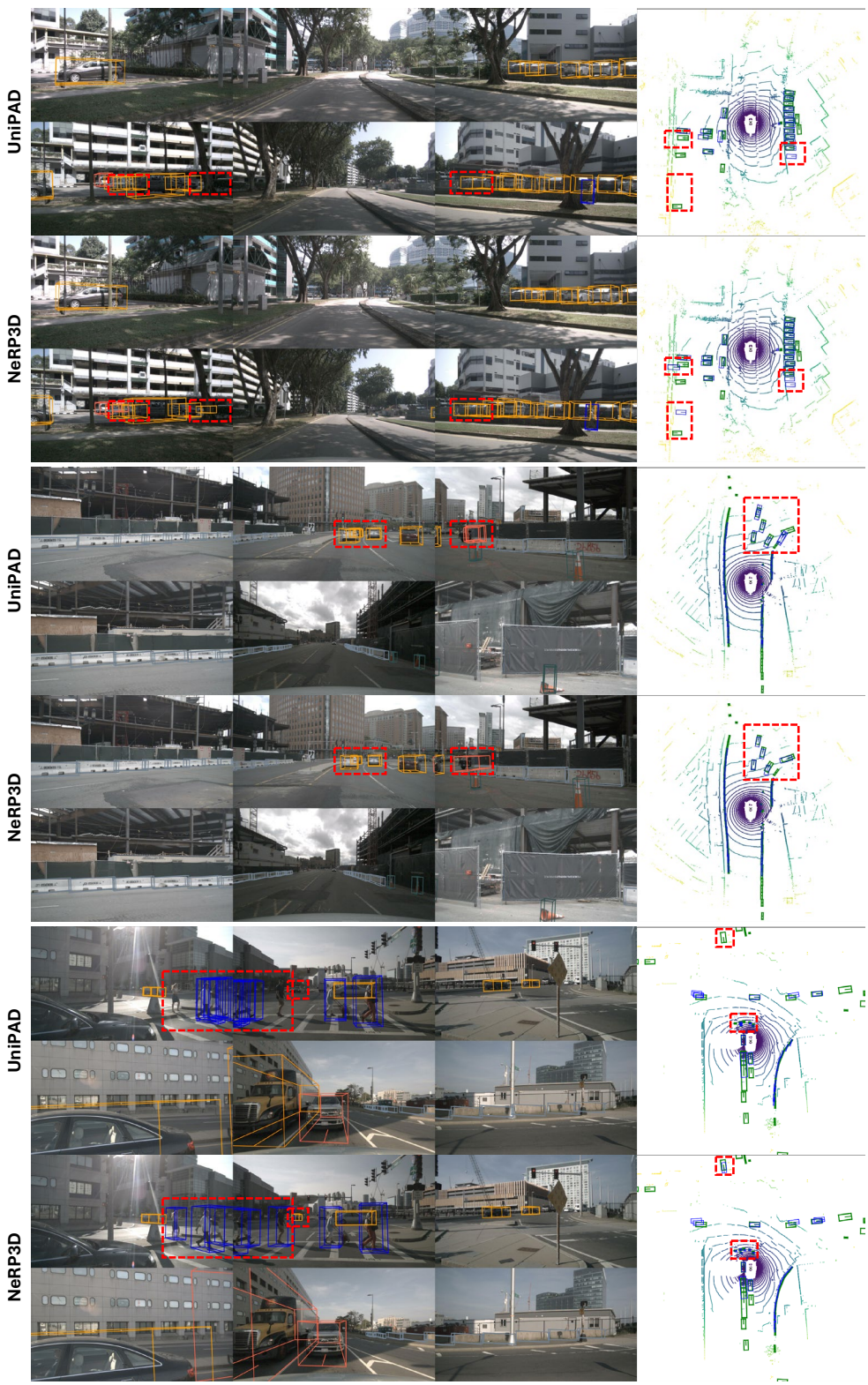

Figure 8: Qualitative comparison of 3D object detection results. NeRP3D consistently generates more accurate and reliable 3D bounding boxes. It demonstrates key advantages such as successfully detecting partially occluded objects in dense crowds (top row), reducing false positives for cleaner predictions (middle row), and more accurately localizing the position of small objects like pedestrians (bottom row).

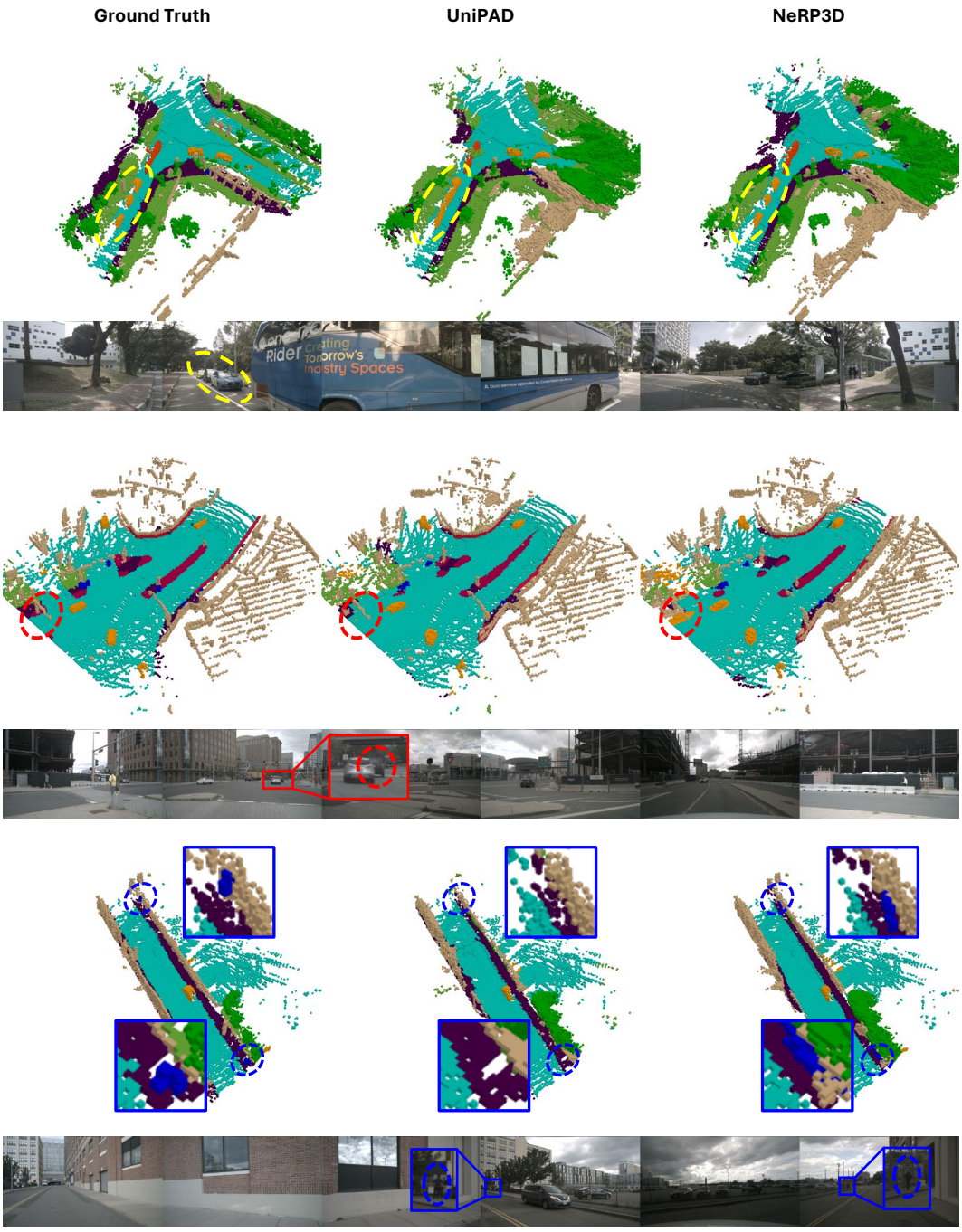

Figure 9: Qualitative comparison of occupancy prediction results. NeRP3D produces more detailed and complete occupancy predictions compared to UniPAD. NeRP3D excels at distinguishing individual objects that are close together, as shown by its clear separation of the vehicles (top row, yellow). Furthermore, it demonstrates a superior ability to detect objects that are entirely missed in the ground truth annotation, likely due to occlusion (middle row, red). The robust perception ability of NeRP3D also extends to resolving smaller, distant objects, such as pedestrians (bottom row, blue), contributing to more accurate and reliable scene understanding.

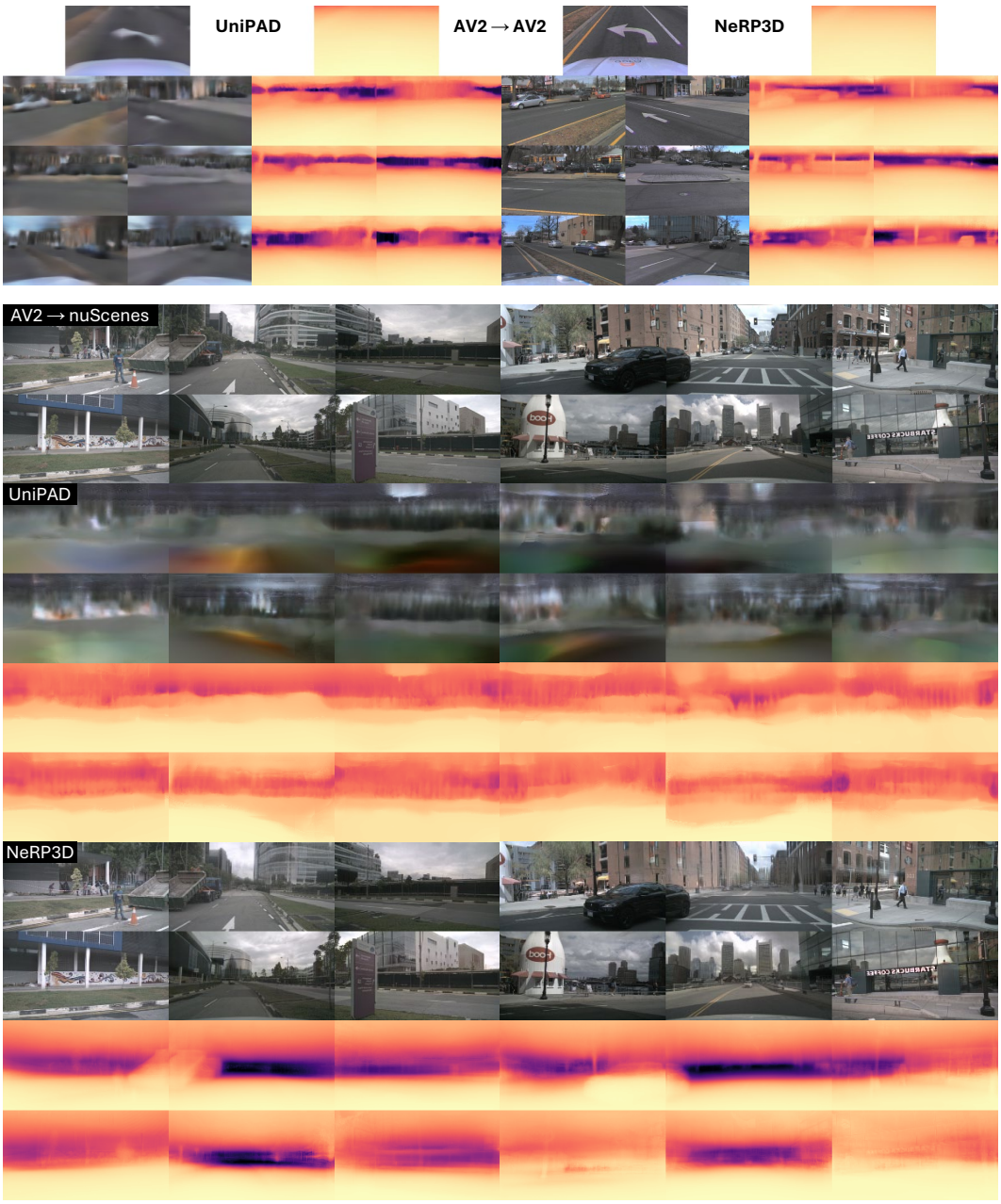

Figure 10: Qualitative comparison of rendering results on the Argoverse 2 dataset (Pre-training phase). Models were pre-trained on Argoverse 2 (AV2) and evaluated without any fine-tuning on the target domain (nuScenes). (Top: AV2 → AV2) In-domain reconstruction results. Both models demonstrate that pre-training on AV2 was successful, reconstructing scene details within the source domain. (Bottom: AV2 → nuScenes) Zero-shot transfer results to nuScenes. When applying AV2-trained weights directly to the distinct camera geometry of nuScenes, UniPAD fails to render meaningful structure (blurry artifacts), revealing the vulnerability of fixed voxel grids to sensor layout changes. In contrast, NeRP3D maintains high-fidelity rendering, demonstrating that its point-based architecture is sensor-agnostic and robust to extrinsic shifts.

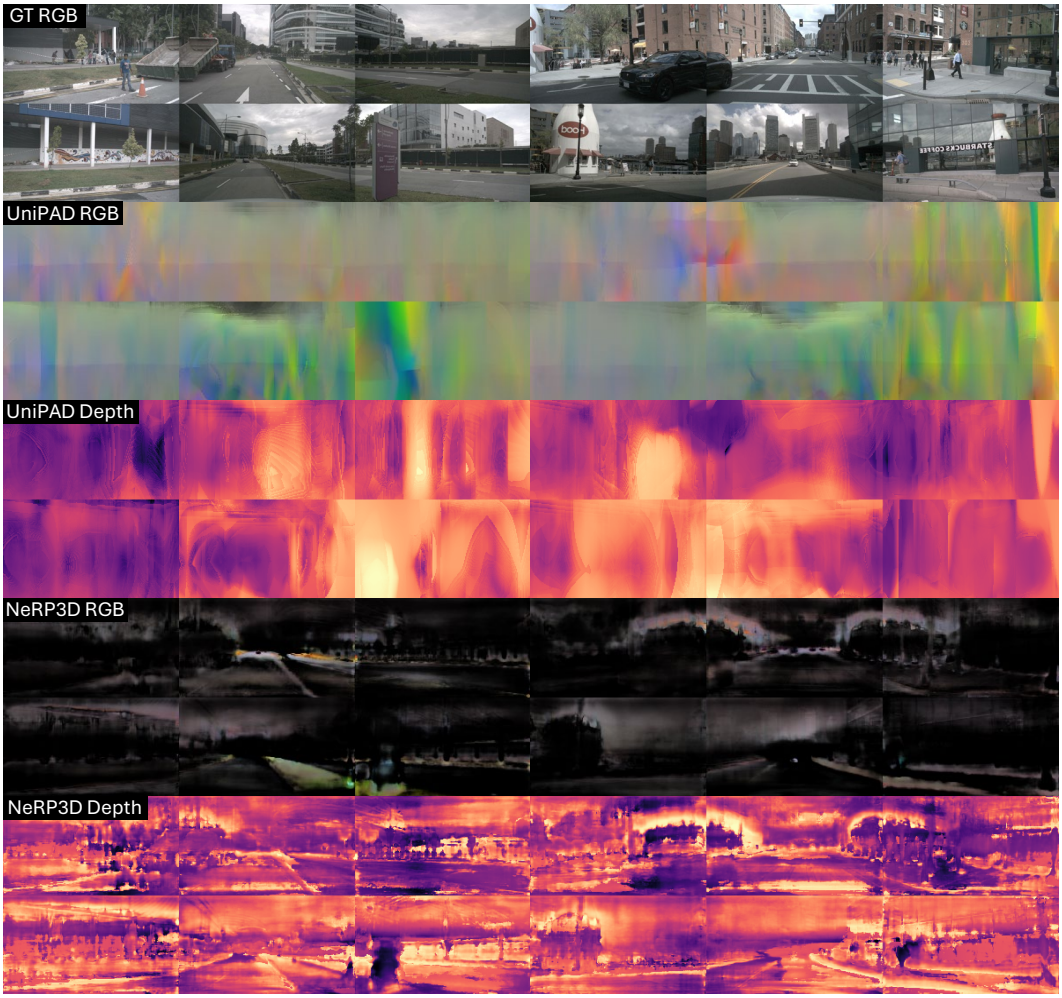

Figure 11: Qualitative evaluation of "Radiance Modeling Ability" after fine-tuning for occupancy prediction. We visualize rendering results using backbones fine-tuned for the occupancy task, with pre-trained decoders frozen. (Row 2-3) UniPAD suffers from catastrophic forgetting, producing "blurry gray" outputs. The model loses 3D structural information and resorts to mean regression to minimize loss. (Row 4-5) NeRP3D successfully retains structural understanding despite the task shift. Key elements like vehicles, road boundaries, and building structures remain clearly distinguishable in both RGB and Depth renderings. This proves that NeRP3D learns a task-agnostic continuous representation that remains valid across different downstream objectives.