

INSERTNeRF: INSTILLING GENERALIZABILITY INTO NeRF WITH HYPERNET MODULES

Yanqi Bao¹, Tianyu Ding², Jing Huo^{1*}, Wenbin Li¹, Yuxin Li¹, Yang Gao¹

¹State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
 yq_bao@smail.nju.edu.cn, {huojing, liwenbin, gaoy}@nju.edu.cn,
 liyuxin16@smail.nju.edu.cn

²Applied Sciences Group, Microsoft Corporation, Redmond, USA
 tianyuding@microsoft.com

ABSTRACT

Generalizing Neural Radiance Fields (NeRF) to new scenes is a significant challenge that existing approaches struggle to address without extensive modifications to vanilla NeRF framework. We introduce **InsertNeRF**, a method for **IN**stilling **g**ene**R**alizability into **NeRF**. By utilizing multiple plug-and-play HyperNet modules, InsertNeRF dynamically tailors NeRF’s weights to specific reference scenes, transforming multi-scale sampling-aware features into scene-specific representations. This novel design allows for more accurate and efficient representations of complex appearances and geometries. Experiments show that this method not only achieves superior generalization performance but also provides a flexible pathway for integration with other NeRF-like systems, even in sparse input settings. Code will be available at: <https://github.com/bbbby-99/InsertNeRF>.

1 INTRODUCTION

Novel view synthesis, a fundamental discipline in computer vision and graphics, aspires to create photorealistic images from reference inputs. Early works (Debevec et al., 1996; Lin & Shum, 2004) primarily focused on developing explicit representations, facing challenges due to the absence of 3D supervision. This issue has been alleviated by recent advancements in implicit neural representation research, which have led to improved performance. In particular, Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) has attracted significant interest. NeRF, and its derivative works, extract scene-specific implicit representations through overfitting training on posed scene images. Although NeRF uses neural scene representations effectively to yield realistic images, the scene-specific nature of these representations requires retraining when faced with novel scenarios.

An emerging topic known as Generalizable NeRF (GNeRF) has recently garnered considerable attention for this challenge. GNeRF aims to learn a scene-independent inference approach that facilitates the transition from references to target view. Current methods enhance the NeRF architecture by adding structures that aggregate reference-image features, or reference features. Examples include pixel-wise feature cost volumes (Johari et al., 2022), transformers (Wang et al., 2022; Suhail et al., 2022), and 3D visibility predictors (Liu et al., 2022). However, fitting these additions into conventional NeRF-like frameworks such as mip-NeRF (Barron et al., 2021), NeRF++ Zhang et al. (2020), and others, often proves challenging and may fail to effectively harness the guiding potential of reference features. Furthermore, the extensive use of transformers or cost volumes can be time-consuming. Thus, an intriguing question arises: Is it possible to directly **IN**still **g**ene**R**alizability into **NeRF** (**InsertNeRF**) while staying faithful to the original framework?

A straightforward way to accomplish this goal is to adaptively modify the NeRF network’s weights, or implicit representations, for different reference scenes while preserving the original framework. The concept of hypernetwork (Ha et al., 2016), which conditionally parameterizes a target network, is an effective strategy in this scenario. The features extracted from the reference scene can be used as inputs to generate scene-specific network weights. However, initial experiments in Tab. 2 indicate that constructing a hypernetwork directly based on the NeRF framework can be inadequate, and often fails to predict different attributes like emitted color and volume density. To address this, we

*Corresponding author.

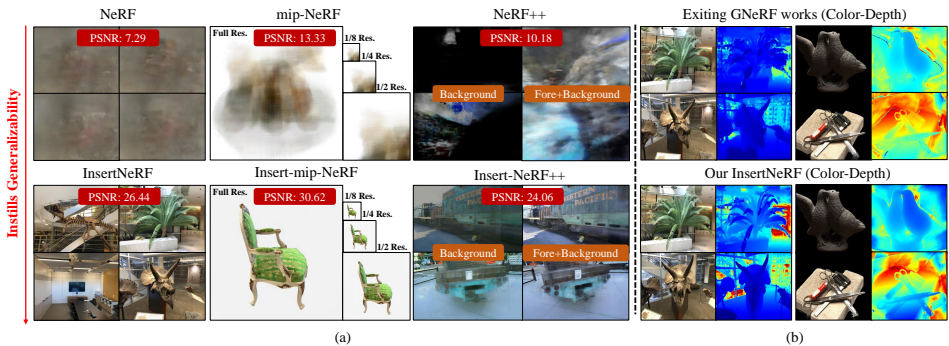


Figure 1: Overview of motivation. (a) We instill generalizability into NeRF-like systems, including vanilla NeRF, mip-NeRF, and NeRF++ frameworks, to achieve consistent performance across scenes without modifying the base framework or requiring scene-specific retraining. (b) InsertNeRF significantly improves depth estimation compared to its original counterpart.

propose to use *HyperNet modules*, which are designed to serve as easily integrable additions to existing NeRF-like frameworks. Owing to their flexibility, the resulting InsertNeRF excels at predicting the NeRF attributes by capitalizing on sampling-aware features and various module structures.

In InsertNeRF, we insert multiple HyperNet modules to instill generalizability throughout the framework’s progression. This approach allows us to fully leverage the guiding role of scene features in determining the entire network’s weights. Unlike existing works that solely utilize reference features as inputs, InsertNeRF exhibits a thorough grasp of *reference scene knowledge*. To further unlock the full potential of the HyperNet modules, it is crucial to aggregate scene features from a set of nearby reference images. To achieve this, we introduce a *multi-layer dynamic-static* aggregation strategy. Compared to existing works, it not only harnesses the inherent completion capabilities of global features, but it also implicitly models occlusion through dynamic-static weights, as demonstrated on the depth renderings shown in Fig. 1b. By feeding the aggregated scene features into the HyperNet modules, we can generate scene-related weights based on the well-understood reference scene.

In summary, we make the following specific contributions:

- We introduce InsertNeRF, a novel paradigm that inserts multiple plug-and-play HyperNet modules into the NeRF-like framework, endowing NeRF-like systems with instilled generalizability.
- We design two types of HyperNet module structures tailored to different NeRF attributes, aiming for predicting scene-specific weights derived from sampling-aware scene features. For these features, we further propose a multi-layer dynamic-static aggregation strategy, which models the views-occlusion and globally completes information based on the multi-view relationships.
- We demonstrate that InsertNeRF achieves state-of-the-art performance with extensive generalization experiments by integrating the modules into the vanilla NeRF. Furthermore, we show the significant potential of our modules in various NeRF-like systems, such as mip-NeRF (Barron et al., 2021), NeRF++ (Zhang et al., 2020), as shown in Fig. 1a, and in task with sparse inputs.

2 RELATED WORKS

2.1 GENERALIZABLE NEURAL RADIANCE FIELDS

Neural Radiance Fields (NeRF) by (Mildenhall et al., 2021) and its subsequent derivatives (Barron et al., 2022; Isaac-Medina et al., 2023; Bao et al., 2023) have gained momentum and are capable of producing realistic images. However, a significant drawback is the need to retrain them for every new scene, which is not efficient in real-world applications. Recent works by (Wang et al., 2021; 2022) introduce Generalizable Neural Radiance Fields that can represent multiple scenes, regardless of whether they are in the training set. To achieve this, many studies have focused on understanding the relationships between reference views and refining NeRF’s sampling-rendering mechanism. For instance, NeuRay (Liu et al., 2022) and GeoNeRF (Johari et al., 2022) use pre-generated depth maps or cost volumes as prior to alleviate occlusion issues. On the other hand, IBRNet (Wang et al., 2021) and GNT (Wang et al., 2022) implicitly capture these relationships through MLPs or transformers. Regarding the sampling-rendering process, most works (Xu et al., 2023; Suhail et al., 2022; Wang et al., 2021; Zhu et al., 2023) utilize the transformer-based architectures to aggregate the sampling point features and replace traditional volume rendering with a learnable technique. However, a

common limitation is that most of these methods replace NeRF’s network with transformers, making it challenging to apply to NeRF derivatives and leading to increased computational complexity. Our research aims to address this by instilling generalizability into NeRF-like systems with scene-related weights while preserving its original framework and efficiency.

2.2 HYPERNETWORKS

The hypernetwork (Ha et al., 2016; Chauhan et al., 2023), often abbreviated as hypernet, is invented to generate weights for a target neural network. Unlike traditional networks that require training from scratch, hypernets offer enhanced generalization and flexibility by adaptively parameterizing the target network (Alaluf et al., 2022; Yang et al., 2022; Li et al., 2020). Leveraging these benefits, hypernets have found applications in various domains including few-shot learning (Li et al., 2020), continual learning (Von Oswald et al., 2019), computer vision (Alaluf et al., 2022), etc. In the realm of NeRF, there have been efforts to incorporate hypernets to inform the training of the rendering process. For instance, (Chiang et al., 2022) propose to train a hypernet using style image features for style transfer, while (Zimny et al., 2022) employ encoded point-cloud features for volume rendering. On a related note, (Peng et al., 2023) utilize a dynamic MLP mapping technique to create volumetric videos and (Kania et al., 2023) use a hypernet for 3D-aware NeRF GAN. In our work, instead of using the hypernet in NeRF framework directly, we introduce a plug-and-play HyperNet module, with a focus on providing reference scene knowledge to enable generalization to new scenarios.

3 METHOD

3.1 BACKGROUND

Neural Radiance Fields. Neural radiance fields (NeRF) (Mildenhall et al., 2021) is a neural representation of scenes. It employs MLPs to map a 3D location $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{S}^2$ to an emitted color $\mathbf{c} \in [0, 1]^3$ and a volume density $\sigma \in [0, \infty)$, which can be formalized as:

$$\mathcal{F}(\mathbf{x}, \mathbf{d}; \Theta) \mapsto (\mathbf{c}, \sigma), \quad (1)$$

where \mathcal{F} is the MLPs, and Θ is the set of learnable parameters of NeRF. Note that \mathcal{F} can be further split into an appearance part \mathcal{F}_{app} and a geometry part \mathcal{F}_{geo} for the view-dependent attribute \mathbf{c} and view-invariant attribute σ , respectively (Zhang et al., 2023).

Volume Rendering. Given a ray in a NeRF, $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where \mathbf{o} is the camera center and \mathbf{d} is the ray’s unit direction vector, we sample K points, $\{\mathbf{r}(t_i) | i = 1, \dots, K\}$, along the ray and predict their color values \mathbf{c}_i and volume densities σ_i . The ray’s color is then calculated by:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^K w_i \mathbf{c}_i, \quad \text{where } w_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) (1 - \exp(-\sigma_i \delta_i)), \quad (2)$$

where δ_i is the distance between adjacent samples, and w_i is considered to be the hitting probability or the weight of the i -th sampling point (Liu et al., 2022).

Generalizable NeRF. Given N reference scene views with known camera poses $\{\mathbf{I}_n, \mathbf{P}_n\}_{n=1}^N$, the goal of GNeRF is to synthesize a target novel view \mathbf{I}_T based on these reference views, even for scenes not observed in the training set, thereby achieving generalizability. Current works (Wang et al., 2021; Liu et al., 2022; Wang et al., 2022) primarily focus on aggregating features along with the ray $\mathbf{r}(t)$ from multiple reference views. The overall process can be outlined as:

$$\mathcal{F}_{\text{sample}}\left(\left\{\mathcal{F}_{\text{view}}\left(\left\{\mathbf{F}_n\left(\Pi_n(\mathbf{r}(t_i))\right)\right\}_{n=1}^N\right)\right\}_{i=1}^K\right) \mapsto (\mathbf{c}, \sigma). \quad (3)$$

Here, $\Pi_n(\mathbf{x})$ projects \mathbf{x} onto \mathbf{I}_n , and $\mathbf{F}_n(\mathbf{z})$ queries the corresponding feature vectors according to the projected points in n -th reference. $\mathcal{F}_{\text{view}}$ and $\mathcal{F}_{\text{sample}}$ specifically denote the aggregation of multi-view features and the accumulation of multiple sampling point features along the ray. These aggregations are often carried out using common techniques such as MLPs and transformers.

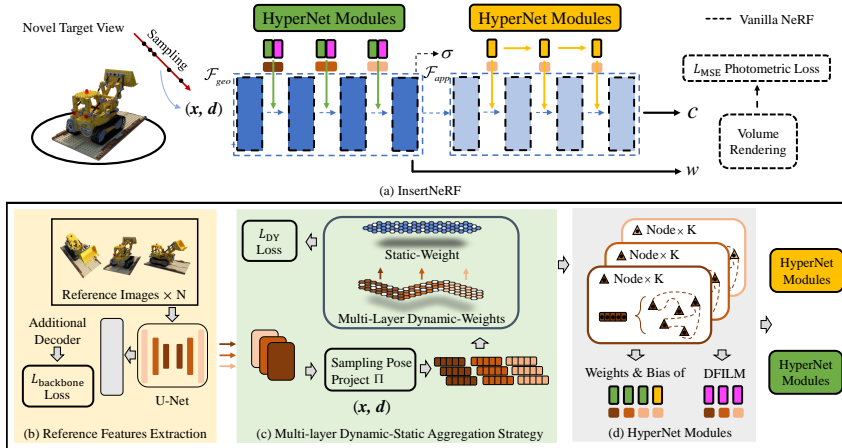


Figure 2: Overview of InsertNeRF. (a) Within the NeRF framework, two types of HyperNet modules are inserted into \mathcal{F}_{geo} and \mathcal{F}_{app} . The HyperNet modules begin by (b) extracting features among multiple (N) reference images, and (c) using a multi-layer dynamic-static aggregation strategy to aggregate the scene representations. Based on these scene representations and specially designed sampling-aware filters, (d) we develop dynamic MLPs and activation functions to guide the weights and instill generalizability into vanilla NeRF. Finally, (a) standard volume rendering is performed.

3.2 INSERTNERF

We introduce InsertNeRF, a novel paradigm that instills generalizability into the NeRF framework, as illustrated in Fig. 2. While this method can be adapted to a variety of NeRF-based systems (Sec. 4.4), we focus on its application on the vanilla NeRF in this section.

Overview. InsertNeRF achieves generalizability by inserting multiple HyperNet modules into NeRF. These modules dynamically generate scene-specific weights for NeRF that are tailored to specific reference scene, denoted by Ω_T . Based on it, by incorporating Ω_T into Eq. (1) and combining Θ as well as Ω_T , NeRF’s implicit representation (or weights) gains the generalizability across multi-scenes, which is explained in Appendix D.3. Specifically, it can be described as follows:

$$\mathcal{F}(x, d; \Theta, \Omega_T) \mapsto (c, w),$$

$$\text{where } \Omega_T = \text{HyperNet} \left(\left\{ \mathcal{F}_{\text{view}} \left(\left\{ F_n (\Pi_n(\mathbf{r}(t_i))) \right\}_{n=1}^N \right) \right\}_{i=1}^K \right). \quad (4)$$

Comparing Eq. (4) to Eq. (3), the key to InsertNeRF is the newly introduced architectures with dynamic weights Ω_T , guided by the HyperNet modules based on specific reference inputs. The process begins with reference features extraction (Sec. 3.2.1), then a multi-layer dynamic-static aggregation strategy is employed to fuse reference features from multi-views into scene features (Sec. 3.2.2). Subsequently, these aggregated scene features are used to adaptively generate NeRF’s sampling-aware weights via the HyperNet modules, which consist of sampling-aware filters, dynamic MLPs and dynamic activation functions (Sec. 3.2.3). These novel HyperNet modules are inserted before each MLP layer in the original NeRF, serving as an enhancement to the original MLP layers.

A notable aspect of InsertNeRF is its ability to directly calculate the hitting probability w_i in Eq. (2) for volume rendering, rather than simply outputting the volume density from \mathcal{F}_{geo} . This capability stems from the implicit modeling of the relationships between spatial points and the advantage of using multi-scale features. By combining \mathcal{F}_{geo} with \mathcal{F}_{app} , the entire pipeline is trained end-to-end. Our unique design not only leads to superior rendering performance in GNeRF but also offers improved computational efficiency compared to transformer-based structures (Wang et al., 2022).

3.2.1 REFERENCE FEATURES EXTRACTION

In the exploration of reference images, generalizable methods often combine U-Net (Ronneberger et al., 2015) and ResNet (He et al., 2016) to extract local dense feature maps. These have proven effective in dealing with occlusion problems (Liu et al., 2022). Yet, there is a risk that an overemphasis on local dense features might neglect global features, which are key to occlusion completion

and global inference (Iizuka et al., 2017). In our work, we take advantage of the spatial representation capabilities of multi-scale features to model complex geometry and detailed appearance. Specifically, we bring in global-local features to successively update the HyperNet module’s weights for \mathcal{F}_{geo} and dense feature for \mathcal{F}_{app} . Here, geometry requires multi-scale information to deduce occluded portions, while appearance concentrates on dense fine-grained details. This process begins with multi-scale features $\mathbf{F}_{l,n}$ from U-Net for each reference input \mathbf{I}_n , and can be expressed as:

$$\mathbf{F}_{l,n} \in \mathbb{R}^{\frac{W}{2^{l+1}} \times \frac{H}{2^{l+1}} \times C_l}, \quad l = 2, 1, 0; n = 1, \dots, N. \quad (5)$$

Here, $W \times H$ defines the image resolution, and C_l is the number of channels. During feature upsampling (as l decreases), we output each layer’s features, transitioning from global to local.

3.2.2 MULTI-LAYER DYNAMIC-STATIC AGGREGATION STRATEGY

Following the feature extraction, the next essential step is the aggregation of scene features. This is not only foundational for scene generalizability but also significantly impacts the effectiveness of the HyperNet modules. Most existing techniques focus primarily on preserving local geometry and appearance consistency, often employing visibility to model occlusions. A straightforward approach is to deduce the view-weight based on differences between reference and target views (Wang et al., 2021). We refer to it as *static weight*, denoted by $M^{ST} \in \mathbb{R}^{B \times K \times N}$, where B represents the batch size, and it assigns higher weights to closer views in a fixed manner. However, it may be unreliable as it overlooks the correlation among the features. To remedy this, we introduce a dynamic prediction of multi-layer weights based on multi-scale features, involving a blend of Maxpool-MLPs and Softmax layers, termed *dynamic weights* and denoted by $M_l^{DY} \in \mathbb{R}^{B \times K \times N}$. Our approach hence adopts a *dynamic-static* aggregation strategy for more nuanced multi-view scene feature aggregation.

Formally, given the corresponding features $\mathbf{F}_l \in \mathbb{R}^{B \times K \times N \times d_l}$ of $B \times K$ points in the space, where d_l is the latent feature dimension, we calculate the weighted means and variances as $\boldsymbol{\mu}_l = \mathbb{E}_n [\mathbf{F}_l \odot M_l^{DY}] \in \mathbb{R}^{B \times K \times d_l}$ and $\mathbf{v}_l = \mathbb{V}_n [\mathbf{F}_l \odot M_l^{DY}] \in \mathbb{R}^{B \times K \times d_l}$, respectively. After concatenating \mathbf{F}_l for each reference view with $\boldsymbol{\mu}_l$ and \mathbf{v}_l and halvely projecting its dimension, denoted as $\tilde{\mathbf{F}}_l \in \mathbb{R}^{B \times K \times N \times d_l/2}$, it is applied to the static weight to obtain $\tilde{\boldsymbol{\mu}}_l = \mathbb{E}_n [\tilde{\mathbf{F}}_l \odot M^{ST}] \in \mathbb{R}^{B \times K \times d_l/2}$ and $\tilde{\mathbf{v}}_l = \mathbb{V}_n [\tilde{\mathbf{F}}_l \odot M^{ST}] \in \mathbb{R}^{B \times K \times d_l/2}$. With $\mathbf{F}_l^{\max} \in \mathbb{R}^{B \times K \times d_l}$ representing the maximum features among all the reference views, and by concatenating $\tilde{\boldsymbol{\mu}}_l$ and $\tilde{\mathbf{v}}_l$, and adding it to \mathbf{F}_l^{\max} , we accomplish the feature aggregation phase \mathcal{F}_{view} in Eq. (4).¹

The use of global-local dynamic weights leads to a significant enhancement in edge sharpness and the thorough completion of detail in the depth rendering images, as evidenced in Fig. 1b. Note that unlike static weights, dynamic weights are guided by the relationships between multi-scale reference features and are learned with auxiliary supervision (Sec. 3.3).

3.2.3 HYPERNET MODULES

We now turn our attention to the HyperNet modules, the core element of InsertNeRF, integrated within both \mathcal{F}_{geo} and \mathcal{F}_{app} . These modules are composed of three basic components: sampling-aware filters, dynamic MLPs (D-MLP), and dynamic activation functions.

Sampling-aware Filter. Unlike traditional hypernetworks, where reference features are generally stable, those based on pose-related epipolar geometric constraints in GNeRF are noisy. This noise complicates their direct use for weights generation. To address this challenge, we introduce a sampling-aware filter that seeks to implicitly find correlations between inter-samples and reduce noise within the reference features through graph reasoning. Specifically, following the aggregation phase \mathcal{F}_{view} , each aggregated point-feature is regarded as a node within a graph structure. The relationships between these K points are then modeled using graph convolutions, formulated as:

$$\mathbf{H}_l = (\mathbf{I} - \mathbf{A}_l) \mathbf{F}_{view} \mathbf{W}_l^a, \quad (6)$$

where $\mathbf{F}_{view} \in \mathbb{R}^{B \times K \times d_l}$ denotes the aggregated K point-features after \mathcal{F}_{view} , and \mathbf{A}_l and \mathbf{W}_l^a represent the $K \times K$ node adjacency matrix and the learnable state update function, respectively. \mathbf{I} here denotes the identity matrix. This specific graph structure helps filter out noise by state-updating,

¹The advantages of this strategy in comparison with Wang et al. (2021) are discussed in the Appendix D.1.

enabling the network to concentrate on key features more effectively. Additionally, for intricate tiny structures, we adopt an approach inspired by Chen et al. (2019), where linear layers across different dimensions are utilized instead of standard matrix multiplications within the graph convolutions.

Dynamic MLP. Using the filtered features \mathbf{H}_l , the HyperNet module is designed to generate corresponding $\text{Weight}_{\mathbf{H}_l}$ and $\text{Bias}_{\mathbf{H}_l}$ within specific MLPs. This instills scene-awareness into vanilla NeRF, ensuring compatibility with $\mathbf{F}_{\text{input}}$, the output of the previous layer in the original NeRF framework. To enhance efficiency, these MLPs are integrated within the sampling-aware filter.

Dynamic Activation Function. Activation functions plays an essential role in the NeRF framework (Sitzmann et al., 2020). Traditional options, such as the ReLU function, may struggle with detail rendering and hinder the performance of D-MLPs due to their static nature. To address this, we introduce a dynamic activation function. This function adaptively activates features in accordance with the unique characteristics of a given scene. Inspired by Perez et al. (2018), we propose the Dynamic Feature-wise Linear Modulation (DFiLM), in which the frequencies ($\text{Freq}_{\mathbf{H}_l}$) and phase-shifts ($\text{Shift}_{\mathbf{H}_l}$) are dynamically determined from \mathbf{H}_l , allowing for more responsive activation.

The entire MLP-Block, including both the D-MLP and the activation function, can be expressed as:

$$\mathbf{F}_{\text{output}} = \text{Shift}_{\mathbf{H}_l}(\text{Freq}_{\mathbf{H}_l}(\text{Weight}_{\mathbf{H}_l} \times \mathbf{F}_{\text{input}} + \text{Bias}_{\mathbf{H}_l})), \quad (7)$$

To insert the HyperNet modules into the NeRF framework, $\mathbf{F}_{\text{output}}$ is subsequently fed into an original NeRF’s MLP layer for the final result. This yields superior performance, as validated through experimental results. We remark that the parameters are not shared among the HyperNet modules. Moreover, their compact structures ensure that the impact on rendering efficiency is negligible.

HyperNet Modules in \mathcal{F}_{geo} and \mathcal{F}_{app} . In vanilla NeRF, \mathcal{F}_{geo} and \mathcal{F}_{app} serve distinct purposes but employ similar MLP structures, albeit with varying complexities. \mathcal{F}_{geo} focuses on geometric properties, whereas \mathcal{F}_{app} encodes view-dependent features using a smooth BRDF prior for surface reflectance. This smoothness can be facilitated by progressively exploiting guided scene features, along with a reduction in both MLP parameters and activation functions for variable d (Zhang et al., 2020). Recognizing this need, we propose a modified HyperNet module architecture specifically for \mathcal{F}_{app} . Our design employs a progressive guidance mechanism within \mathcal{F}_{app} , incorporating multiple parallel dynamic branches into the NeRF framework. The weights of the D-MLP in each branch are progressively generated from the preceding branch, enabling the capture of reference features at different levels for complex appearance modeling. Finally, the results of all branches are summed and used as input to the original MLP for predicting the RGB value. In accordance with our analysis, the DFiLM is not used in \mathcal{F}_{app} , setting it apart from other elements in the architecture. [Appendix C.1]

3.3 LOSS FUNCTIONS

The InsertNeRF pipeline is trained end-to-end utilizing three carefully designed loss functions.

Photometric loss. First, we employ the photometric loss in NeRF (Mildenhall et al., 2021), i.e., the Mean Square Error (MSE) between the rendered and true pixel colors:

$$\mathcal{L}_{\text{MSE}} = \sum_{\mathbf{r} \in \mathcal{R}} \left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2, \quad (8)$$

where \mathcal{R} is the set of rays in a batch, and $C(\mathbf{r})$ is the ground-truth RGB color for ray $\mathbf{r} \in \mathcal{R}$.

Backbone loss. During end-to-end training, optimizing the feature extraction without additional guidance poses considerable challenges. To address this, we draw inspiration from auto-encoding (Kingma & Welling, 2013). By adding an additional upsampling layer and a small decoder (used exclusively for loss computation), we seek to reconstruct reference images from encoded features. The original images serve as supervision, and we refer to this particular loss term as $\mathcal{L}_{\text{backbone}}$.

Dynamic weights loss. Initiating the learning of dynamic weights from scratch introduces difficulties in understanding the connections among multi-scale features. To tackle this issue, we introduce an auxiliary supervision to encompass global-local information. Specifically, we let $C_n^{\text{ref}}(\mathbf{r}) \in \mathbb{R}^{B \times K \times N \times 3}$ represent the ground-truth RGB values in corresponding reference images for K points in ray \mathbf{r} within a batch \mathcal{R} . We compute $\mathbf{c}'_i = \sum_{n,l,\mathbf{r} \in \mathcal{R}} C_n^{\text{ref}}(\mathbf{r}) \odot M_l^{\text{DY}}$, the weighted sum of these RGB values by dynamic weights. Utilizing \mathbf{c}'_i , $\hat{C}'(\mathbf{r})$ is subsequently calculated according to Eq. (2), and supervised by the true color $C(\mathbf{r})$. We designate this loss term as \mathcal{L}_{DY} .

Table 1: Comparisons of InsertNeRF against SOTA methods with Setting I.

Methods	NeRF Synthetic			LLFF			DTU		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
PixelNeRF (CVPR2021)	22.65	0.808	0.202	18.66	0.588	0.463	19.40	0.463	0.447
MVSNeRF (ICCV2021)	25.15	0.853	0.159	21.18	0.691	0.301	23.83	0.723	0.286
IBRNet (CVPR2021)	26.73	0.908	0.101	25.17	0.813	0.200	25.76	0.861	0.173
ContraNeRF (CVPR2023)	-	-	-	25.44	<u>0.842</u>	<u>0.178</u>	27.69	0.904	<u>0.129</u>
GeoNeRF [†] (CVPR2022)	28.33	<u>0.938</u>	<u>0.087</u>	25.44	0.839	0.180	-	-	-
WaveNeRF [†] (ICCV2023)	26.12	0.918	0.113	24.28	0.794	0.212	-	-	-
NeuRay(CVPR2022)	<u>28.92</u>	0.920	0.096	<u>25.85</u>	0.832	0.190	<u>28.30</u>	<u>0.907</u>	0.130
InsertNeRF (Ours)	30.35	0.938	0.065	26.44	0.844	0.169	29.75	0.925	0.077

Table 2: Comparisons and ablations with Setting II.

Methods	NeRF Synthetic			LLFF		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
GNT (ICLR2023)	27.29	0.937	0.056	25.59	0.858	0.128
Baseline (NeRF)	7.29	0.512	0.690	11.46	0.328	0.582
NeRF with HyperNetwork	25.86	0.902	0.081	24.25	0.793	0.177
InsertNeRF w/o MLDS	25.12	0.896	0.098	24.41	0.814	0.156
InsertNeRF (Ours)	27.57	<u>0.936</u>	0.056	25.68	0.861	0.126

Table 3: Results with sparse inputs.

Methods	3-view		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
DietNeRF (ICCV 2021)	14.94	0.370	0.496
RegNeRF (CVPR 2022)	19.08	0.587	0.336
GeCoNeRF (ICML 2023)	18.77	0.596	0.338
FreeNeRF (CVPR 2023)	19.63	0.612	0.308
InsertNeRF (w/o retrain)	<u>19.41</u>	0.618	<u>0.330</u>

We formulate our final loss function as

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_1 \mathcal{L}_{\text{backbone}} + \lambda_2 \mathcal{L}_{\text{DY}} \quad (9)$$

where λ_1 and λ_2 are hyperparameters controlling the relative importance of these terms.

4 EXPERIMENTS

We conduct comparative experiments with state-of-the-art (SOTA) methods across different settings on mainstream datasets. Additionally, we validate the effectiveness of the proposed paradigm in the context of derivative NeRF-like systems generalizations and tasks involving sparse inputs.

4.1 EXPERIMENTAL PROTOCOL AND SETTINGS

Following IBRNet (Wang et al., 2021), GNeRF exploits the target-reference pairs sampling strategy during both the training and inference phases. Here, reference views are selected from a set of nearby views surrounding the target view. Specifically, N reference views are chosen from a pool of $P \times N$ ($P \geq 1$) neighboring views of target, ensuring that the target view is excluded from the reference views. During the evaluation phase, we conduct evaluations using three metrics: PSNR, SSIM, and LPIPS, on well-established datasets such as NeRF Synthetic, LLFF, and DTU. More training and inference details are provided in the Appendix A and Algorithm in the Appendix C.2.

In our experiments, we follow two GNeRF settings of existing methods:

Setting I. Following NeuRay (Liu et al., 2022), we use three types of training datasets for training GNeRF, including three forward-facing datasets, the synthetic Google Scanned Object dataset and the DTU dataset. Note that we only select training scenes in the DTU dataset, excluding the four evaluation scenes. Following their setting in the experiments, we set $N = 8$.

Setting II. Following GNT (Wang et al., 2022), we train GNeRF using three forward-facing datasets and the Google Scanned Object dataset. Unlike Setting I, the DTU dataset is not used for either training or evaluation. In addition, we set $N = 10$ in this setting.

4.2 COMPARATIVE EXPERIMENTS

We evaluate InsertNeRF for its generalization based on the vanilla NeRF framework, comparing its performance with SOTA methods under two GNeRF settings. Through extensive quantitative and qualitative experiments, we explore the advantages of our approach, even with fewer references.

Quantitative comparisons. We present quantitative comparisons with SOTA methods under Setting I and Setting II, as reported in Tab. 1 and Tab. 2. For Setting I, the quantitative comparisons

[†]GeoNeRF (Johari et al., 2022) and WaveNeRF (Xu et al., 2023) are trained on original rectified images and evaluated on the distinct scenes with us in the DTU dataset.

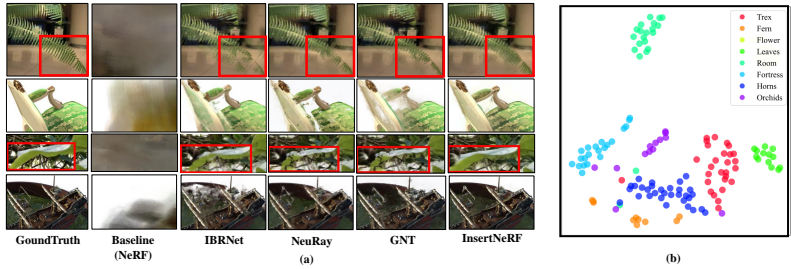


Figure 3: (a) Qualitative comparisons of InsertNeRF against SOTA methods. (b) A t-SNE plot of the scene-specific representations from our HyperNet modules. More analysis in the Appendix D.2

Table 4: HyperNet modules ablations.

Methods	PSNR \uparrow	LLFF SSIM \uparrow	LPIPS \downarrow
w/o D-MLP	23.33	0.774	0.198
w/o Sampling Filter	24.67	0.815	0.158
w/o DFILM	25.04	0.832	0.152
w/o original MLP	25.44	0.848	0.131
InsertNeRF (Ours)	25.68	0.861	0.126

Table 5: MLDS aggregation strategy ablations.

Static-Weight	Dynamic-Weight	Auxiliary-Supervision	Multi-Layers	Single-Layer	LLFF PSNR \uparrow	LLFF SSIM \uparrow	LPIPS \downarrow
✓			✓		24.88	0.827	0.154
✓	✓		✓		25.55	0.851	0.128
	✓	✓	✓		25.53	0.850	0.131
✓	✓	✓		✓	25.15	0.838	0.139
✓	✓	✓	✓		25.68	0.861	0.126

in Tab. 1 display our model’s competitive results in evaluation datasets, with significant improvements in PSNR, SSIM and LPIPS in comparison to existing SOTA methods. Specifically, PSNR and LPIPS exhibit substantial enhancements by $\sim 1.16\text{dB} \uparrow$ and $\sim 23.6\% \downarrow$ respectively. For Setting II, InsertNeRF consistently outperforms the SOTA method (Wang et al., 2022), as substantiated by the results in Tab. 2. We observe that these improvements become even more pronounced with fewer reference images, alongside higher efficiency, as demonstrated in subsequent sections.

Qualitative comparisons. Fig. 1 and Fig. 3 (a) show the qualitative performances of our method against baseline and SOTA methods. InsertNeRF achieves improved geometric fidelity and clear edges, attributable to the completion capability of global features and the modeling of sample spatial relationships from graph structures. For more analysis and results, please refer to the Appendix E.5.

4.3 ABLATION STUDIES

In Tab. 2, we analyze the core components of our method. The findings underscore the vital role of HyperNet modules in enhancing rendering performance, as further evidenced in Fig. 3 (b) that they instill scene-specific capabilities into NeRF’s representation. Additionally, the multi-layer dynamic-static aggregation strategy proves to be essential. By integrating both modules, our novel paradigm instills generalizability into the NeRF framework, leading to a performance boost of approximately two to three times compared to the baseline model, i.e., vanilla NeRF. Additionally, we explore the underlying mechanisms driving the effectiveness of these components. More experiments, including single-scene setting, fine-tuning and ablation studies about \mathcal{F}_{app} in the Appendix E

HyperNet modules. Tab. 4 demonstrates that both the sampling-aware filters and dynamic activation functions are vital in the HyperNet modules, with the sampling-aware filters having a more substantial impact. This could be due to the need to consider relationships between sampled points in the rendering process, which implicitly models occlusions, as noted in Liu et al. (2022). Solely using dynamic activation functions without D-MLP leads to a marked decline in performance, highlighting the essential role of MLPs in neural representation. Furthermore, using only the HyperNet modules and omitting the original NeRF’s MLP layers results in inferior performance, reducing training stability.

Multi-layer dynamic-static aggregation strategy. In Tab. 5, ablation studies reveal the significance of dynamic-static weights and multi-layer features. Using only dynamic weights appears more effective than static weight, likely because they are adaptively generated to suit different scene features. The auxiliary supervision for dynamic weights and multi-layer global-local features also play essential roles in aggregating multi-view features, underlining their importance in this strategy.

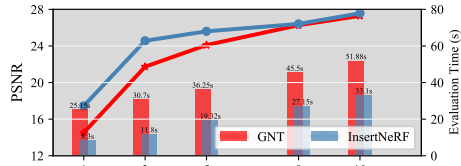


Figure 4: Performance and efficiency under different input-number N on NeRF Synthetic.

Table 6: Quantitative results of InsertNeRF and Insert-mip-NeRF on multi-scale NeRF Synthetic.

Methods	PSNR \uparrow				SSIM \uparrow				LPIPS \downarrow			
	Full Res.	1/2 Res.	1/4 Res.	1/8 Res.	Full Res.	1/2 Res.	1/4 Res.	1/8 Res.	Full Res.	1/2 Res.	1/4 Res.	1/8 Res.
mip-NeRF	12.94	13.03	13.18	13.33	0.700	0.636	0.563	0.469	0.424	0.460	0.470	0.530
InsertNeRF	27.60	28.58	29.45	29.85	0.926	0.943	0.960	0.972	0.066	0.054	0.045	0.036
Insert-mip-NeRF	28.15	29.17	30.22	30.62	0.935	0.951	0.966	0.977	0.056	0.045	0.037	0.029

Input number (N) and efficiency. Since feature extraction is time-consuming, reducing the number of reference images substantially improves the training and inference efficiency of the network. Fig. 4 illustrates the performance of InsertNeRF as the number of reference images (N) varies for training on NeRF Synthetic. In comparison to GNT (Wang et al., 2022), InsertNeRF consistently demonstrates superior rendering performance and inference efficiency. This success can be attributed to our novel generalization paradigm and the compact structures of the HyperNet modules.

4.4 INSERT-NeRF-LIKE FRAMEWORKS

Thanks to the plug-and-play advantage of the HyperNet modules, we extend the study of generalization to derived domains of NeRF, such as mip-NeRF (Barron et al., 2021) and NeRF++ (Zhang et al., 2020), areas that have rarely been discussed before. More details are provided in the Appendix A.

Insert-mip-NeRF. Mip-NeRF is a multi-scale NeRF-like model used to address the inherent aliasing of NeRF, a significant challenge for GNeRF. Unlike Huang et al. (2023), we explore how to instill generalizability into mip-NeRF, following its original setup. We report the qualitative and quantitative performance of mip-NeRF, InsertNeRF, and Insert-mip-NeRF on multi-scale NeRF Synthetic in a cross-scene generalization setting (see Tab. 6, Fig. 1 and Fig. 5). One can observe that incorporating the HyperNet modules not only enhances generalization for mip-NeRF but also addresses the inherent aliasing of InsertNeRF and improves the performance in the task of multi-scale rendering.

Insert-NeRF++. NeRF++, an unbounded NeRF. Fig. 1 and Fig. 13 depicts qualitative and quantitative rendering results of Insert-NeRF++. It is evident that our approach has successfully instilled generalizability into the NeRF++ framework, doubling its PSNR compared to the original.

Sparse Inputs. Training NeRF with sparse inputs has become a notable focus recently (Niemeyer et al., 2022; Yang et al., 2023). Unlike our nearby reference views setting (Sec. 4.1), this task often involves training from a limited number of fixed viewpoints to represent the entire scene. Under this setting, we relax constraints on selecting nearby viewpoints and uniformly select fixed sparse seen viewpoints to infer on arbitrary unseen viewpoints. Unlike existing works, our method trains on extensive auxiliary datasets, allowing us to represent the entire evaluation scene from sparse inputs without retraining (see Tab. 3). To ensure fairness, all scenes in evaluation are excluded in the training phase. In conclusion, InsertNeRF offers a novel insight that employs pre-training on auxiliary datasets to enhance representation capabilities with sparse inputs. We believe that, through fine-tuning on the evaluation scene and incorporating existing technologies like geometry and color regularization, our paradigm will achieve even better performance under sparse inputs.

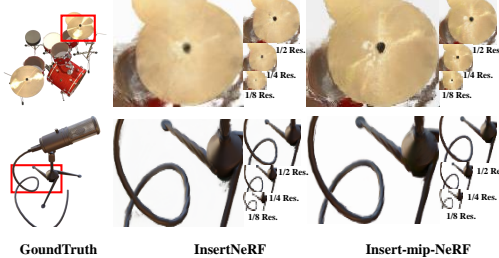


Figure 5: Qualitative results of Insert-mip-NeRF. Please refer to the Appendix E.5 for more results.

5 CONCLUSION

We present InsertNeRF, a novel paradigm that instills generalizability into NeRF systems. Unlike popular transformer-based structures, our HyperNet modules are efficiently incorporated into the original NeRF-like framework, leveraging reference scene features to generate scene-specific network weights. To achieve this, we design a multi-layer dynamic-static feature aggregation strategy for extracting scene features from reference images and employ sampling-aware filters to explore relationships between sample points. Experiments on well-established datasets show that InsertNeRF and other Insert-NeRF-like frameworks can render high-quality images across different scenes without retraining. This offers insights for future works on: (i) generalization tasks for additional NeRF-like systems such as mip-NeRF 360; and (ii) sparse inputs tasks based on auxiliary datasets.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62276128, Grant 62192783, and Grant 62106100; in part by the Jiangsu Natural Science Foundation under Grant BK20221441; in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2023QNRC001; in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

- Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18511–18521, 2022.
- Yanqi Bao, Yuxin Li, Jing Huo, Tianyu Ding, Xinyue Liang, Wenbin Li, and Yang Gao. Where and how: Mitigating confusion in neural radiance fields from sparse inputs. *arXiv preprint arXiv:2308.02908*, 2023.
- Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.
- 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.
- SC Chan, Heung-Yeung Shum, and King-To Ng. Image-based rendering and synthesis. *IEEE Signal Processing Magazine*, 24(6):22–33, 2007.
- Vinod Kumar Chauhan, Jiandong Zhou, Ping Lu, Soheila Molaei, and David A Clifton. A brief review of hypernetworks in deep learning. *arXiv preprint arXiv:2306.06955*, 2023.
- 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.
- Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pp. 333–350. Springer, 2022.
- Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 433–442, 2019.
- Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1475–1484, 2022.
- Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 11–20, 1996.
- Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, and Qi Shan. Equivariant neural rendering. In *International Conference on Machine Learning*, pp. 2761–2770. PMLR, 2020.
- Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4857–4866, 2020.
- David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Xin Huang, Qi Zhang, Ying Feng, Xiaoyu Li, Xuan Wang, and Qing Wang. Local implicit ray function for generalizable radiance field representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 97–107, 2023.
- Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.
- Brian KS Isaac-Medina, Chris G Willcocks, and Toby P Breckon. Exact-nerf: An exploration of a precise volumetric parameterization for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 66–75, 2023.
- Nishant Jain, Suryansh Kumar, and Luc Van Gool. Enhanced stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13208–13217, 2023.
- Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6001–6010, 2020.
- Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18365–18375, 2022.
- Adam Kania, Artur Kasymov, Maciej Zieba, and Przemysław Spurek. Hypernerf-gan: Hypernetwork approach to 3d nerf gan. *arXiv preprint arXiv:2301.11631*, 2023.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.
- Johannes Kopf, Fabian Langguth, Daniel Scharstein, Richard Szeliski, and Michael Goesele. Image-based rendering in the gradient domain. *ACM Transactions on Graphics (TOG)*, 32(6):1–9, 2013.
- Jonáš Kulháněk, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pp. 198–216. Springer, 2022.
- Yawei Li, Shuhang Gu, Kai Zhang, Luc Van Gool, and Radu Timofte. Dhpr: Differentiable meta pruning via hypernetworks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 608–624. Springer, 2020.
- Zhouchen Lin and Heung-Yeung Shum. A geometric analysis of light field rendering. *International Journal of Computer Vision*, 58:121–138, 2004.
- Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems*, 32, 2019.
- 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, 2022.
- Zhen Liu, Yao Feng, Michael J Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. *arXiv preprint arXiv:2303.08133*, 2023.
- Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

- 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.
- Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5480–5490, 2022.
- Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Representing volumetric videos as dynamic mlp maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4252–4262, 2023.
- Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 523–540. Springer, 2020.
- Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- Gernot Riegler and Vladlen Koltun. Free view synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pp. 623–640. Springer, 2020.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.
- Sudipta Sinha, Drew Steedly, and Rick Szeliski. Piecewise planar stereo for image-based rendering. In *2009 International Conference on Computer Vision*, pp. 1881–1888, 2009.
- Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.
- Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.
- Przemysław Spurek, Sebastian Winczowski, Jacek Tabor, Maciej Zamorski, Maciej Zieba, and Tomasz Trzeciński. Hypernetwork approach to generating point clouds. *arXiv preprint arXiv:2003.00802*, 2020a.
- Przemysław Spurek, Maciej Zieba, Jacek Tabor, and Tomasz Trzeciński. Hyperflow: Representing 3d objects as surfaces. *arXiv preprint arXiv:2006.08710*, 2020b.
- Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*, pp. 156–174. Springer, 2022.
- Johannes Von Oswald, Christian Henning, Benjamin F Grewe, and João Sacramento. Continual learning with hypernetworks. *arXiv preprint arXiv:1906.00695*, 2019.
- Chu Wang, Babak Samari, and Kaleem Siddiqi. Local spectral graph convolution for point set feature learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–66, 2018.
- Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, Zhangyang Wang, et al. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022.

- Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2021.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- Muyu Xu, Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Xiaoqin Zhang, Christian Theobalt, Ling Shao, and Shijian Lu. Wavenerf: Wavelet-based generalizable neural radiance fields. *arXiv preprint arXiv:2308.04826*, 2023.
- Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8254–8263, 2023.
- Lingfeng Yang, Xiang Li, Renjie Song, Borui Zhao, Juntian Tao, Shihao Zhou, Jiajun Liang, and Jian Yang. Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10945–10954, 2022.
- Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.
- 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.
- Maciej Zamorski, Maciej Zieba, Piotr Klukowski, Rafał Nowak, Karol Kurach, Wojciech Stokowiec, and Tomasz Trzcinski. Adversarial autoencoders for compact representations of 3d point clouds. *Computer Vision and Image Understanding*, 193:102921, 2020.
- Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.
- Zicheng Zhang, Yinglu Liu, Congying Han, Yingwei Pan, Tiande Guo, and Ting Yao. Transforming radiance field with lipschitz network for photorealistic 3d scene stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20712–20721, 2023.
- Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5826–5835, 2021.
- Haidong Zhu, Tianyu Ding, Tianyi Chen, Ilya Zharkov, Ram Nevatia, and Luming Liang. Caesarnerf: Calibrated semantic representation for few-shot generalizable neural rendering. *arXiv preprint arXiv:2311.15510*, 2023.
- Dominik Zimny, T Trzcinski, and Przemysław Spurek. Points2nerf: Generating neural radiance fields from 3d point cloud. *arXiv preprint arXiv:2206.01290*, 2022.
- Dominik Zimny, Jacek Tabor, Maciej Zieba, and Przemysław Spurek. Multiplanenerf: Neural radiance field with non-trainable representation. *arXiv preprint arXiv:2305.10579*, 2023.

A EXPERIMENTAL DETAILS

During the training process, the feature extraction and InsertNeRF-systems are trained with different learning rates in an end-to-end manner, and we apply the Adam to optimize the entire network with an exponentially decaying learning rate. In our experiments, λ_1 is set as 0.1 and λ_2 is set as 1. During the cross-scene evaluation phase, we individually test the average PSNR, SSIM, and LPIPS metrics of all testing-view renderings for each scene and report the average values across all scenes, as shown in Tab. 1. For the sake of fair experimental comparison and efficiency with (Wang et al., 2022), we set the rendering stride size to 2 across all experiments. Empirical evidence demonstrates that this choice influences the evaluation of metrics, as shown in Tab. 8. Our model is implemented using PyTorch 1.11.0 and all experiments are conducted on Nvidia RTX 3090 GPUs with CUDA 11.4.

Metrics. We calculate the Peak Signal to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) (Wang et al., 2004) to evaluate rendering quality for target novel viewpoints. Additionally, the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) is adopted as a perceptual metric. For all our experiments, we report the average of the PSNR, SSIM and LPIPS under different testing views in multiple scenes to verify the generalization. It is noteworthy that, similar to the majority of GNeRF research (Liu et al., 2022; Wang et al., 2022), our focus lies on foreground-centric metrics for both the NeRF Synthetic and DTU datasets during the evaluation process.

A.1 EVALUATION DATASETS.

NeRF Synthetic. The dataset consists of 8 synthetic objects with viewpoints uniformly sampled on the upper hemisphere. Each scene comprises 200 test images, wherein a sampling strategy with an interval of 8 images is employed during the evaluation phase.

Local Light Field Fusion (LLFF). The dataset consists of 8 complex real-world scenes. Each scene includes real images, 1/8 of which are used as the evaluation dataset.

DTU Dataset. The dataset consists of 128 object-scenes, which is a classic dataset of MVS. In the experiments, we select 4 scenes (birds, tools, bricks and snowman) for evaluation (Liu et al., 2022).

A.2 EXPERIMENTAL DETAILS FOR INSERTNERF

For all our experiments, we maintain the experimental protocols of NeuRay (Liu et al., 2022) and GNT (Wang et al., 2022). In Setting I, we also employ depth maps (Liu et al., 2022) as priors to assist the Hypernet modules to generate adaptive weights. Following (Liu et al., 2022), we randomly sample 2048 rays from Target-Reference pairs, and it trains for a total of 600,000 steps. In Setting II (Wang et al., 2022), we randomly sample 512 rays from Target-Reference pairs, and it trains for a total of 400,000 steps without any priors. In order to enhance training and inference efficiency, we sample $K = 64$ points along each ray and simplify the volume rendering process in our paradigm.

A.3 EXPERIMENTAL DETAILS FOR INSERT-MIP-NeRF

Insert-mip-NeRF substitutes point-samplings with a sequence of conical-frustums and introduces the integrated positional encoding into the InsertNeRF framework. In contrast to InsertNeRF, we sample $K = 65$ positions for integrated positional encoding on conical-frustums. Throughout all experiments, we employ the multi-scale variants of NeRF Synthetic, such as Full, 1/2, 1/4 and 1/8 Resolutions, to simulate multi-resolution scenes for training and evaluation as (Barron et al., 2021). The remaining experimental configurations follow those of mipNeRF and our Setting II.

A.4 EXPERIMENTAL DETAILS FOR INSERT-NeRF++

Insert-NeRF++ divides the InsertNeRF’s scene representations into foreground and background components and combines both for finally rendering (Zhang et al., 2020). In Insert-NeRF++, we avoid using inverted sphere parametrization and chose standard parameterization across foreground-background spatial ranges, which possesses the capability to offer precise spatial positions for projection process $\Pi_n(\mathbf{x})$. Following (Zhang et al., 2020), we employ the Tanks and Temples dataset (Knapsitsch et al., 2017), a real-world unbound dataset captured with hand-held cameras, for training and evaluation under cross-scenes setting. During the evaluation process, we separately

report the renderings for the foreground and background as (Zhang et al., 2020). The remaining experimental configurations also follow those of vanilla NeRF++ and our Setting II.

B RELATED WORKS

B.1 IMAGE-BASED RENDERING

Image-based rendering (IBR) (Chan et al., 2007) is a classic technique that aims to generate novel views within a specific scene by warping and integrating pixel information from reference-view images. To ensure spatial consistency, most existing works simplify this challenge by resorting to estimated explicit geometry or depth maps (Riegler & Koltun, 2020; Jain et al., 2023). In order to obtain proxy geometry, Structure-from-Motion (SfM) (Sinha et al., 2009) and MultiView Stereo (MVS) (Yao et al., 2018) have recently attracted the attention of researchers in the field of IBR. However, hints from explicit geometry without 3D supervision are unstable in most scenarios. Recently light field rendering (Lin & Shum, 2004) has become one of the alternatives to explicit representations, which considers the lighting and reflection properties of the scene to ensure visually plausible rendering. Moreover, some works (Kopf et al., 2013; Chen et al., 2021) focus on aggregating information from multiple reference views, which exploits the relationships between references and also implicitly solves the occlusion. As opposed to explicit geometry, these methods rely on constructing implicit representations to enable reasoning about novel views (Liu et al., 2019). Different from the above explicit or implicit scene-customized representations for IBR, our method can be used in a large number of scenes simultaneously without retraining.

B.2 NEURAL SCENE REPRESENTATION

Representing the geometry and appearance of a scenes with neural networks has been considered an alternative to 3D scene representations in recent years (Mescheder et al., 2019; Peng et al., 2020). Existing works demonstrate the potential of Multi-Layer Perceptrons (MLPs) in implicit representations, which activate spatial features by continuous functions (Genova et al., 2020; Jiang et al., 2020). Neural radiance fields (NeRF) (Mildenhall et al., 2021) apply such functions for coordinate-based representations, which use high-dimensional interpolation to produce photorealistic renderings of target views. On its basis, mip-NeRF (Barron et al., 2021) replaces rays with casting cones during volume rendering, changing the input of NeRF from points to cone frustums, and introduces an integrated positional encoding for multi-resolutions images. Subsequently, NeRF++ (Zhang et al., 2020) and mip-NeRF 360 (Barron et al., 2022) further improve NeRF and mip-NeRF to adapt to distant targets under unbounded scenes. To further enhance representation efficiency, (Chen et al., 2022) and (Zimny et al., 2023) proposed efficient representation methods to replace a large number of neural network parameters. Although derivative works has been a surge, similar to most IBR works, NeRF must also be trained for each novel scene, which is time-consuming in practice.

B.3 GENERATIVE MODELS

With the development of generative models, 3D generative models have been widely discussed, enabling the direct construction of 3D representations such as point clouds (Zamorski et al., 2020), surfaces (Spurek et al., 2020b), voxels (Zhou et al., 2021) and NeRF (Poole et al., 2022). A significant amount of works have leveraged techniques from image generative models and applied them to 3D generation, including GAN (Kania et al., 2023) and diffusion model (Liu et al., 2023). In this work, we focus on some 3D generative models with HyperNetwork. (Spurek et al., 2020a) is an early work that builds variable size representations of point clouds with HyperNetwork. Then, HyperFlow (Spurek et al., 2020b) uses a hypernetwork to model 3D objects as families of surfaces and Points2NeRF (Zimny et al., 2022) utilizes a HyperNetwork to generate NeRF from a 3D point cloud. Additionally, in recent years, there have been some NeRF works that focus on this technique, they directly incorporate HyperNetwork into NeRF, as described in Section 2.2. However, in the generalizable NeRF task, such idea is suboptimal, overlooking the characteristic of different attributes, such as volume density and color. Furthermore, they struggle to capture the relationship between the inputs (reference images) in the target’s sampling process. Therefore, this paper proposes two types of HyperNet module structures for \mathcal{F}_{geo} and \mathcal{F}_{app} and Sampling-aware Filter separately to mitigate the aforementioned two issues.

C IMPLEMENTATION DETAILS

C.1 HYPERNET MODULE ARCHITECTURE

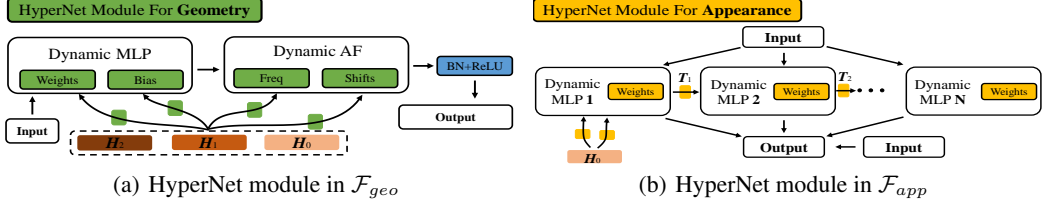


Figure 6: Detailed HyperNet module of \mathcal{F}_{geo} and \mathcal{F}_{app}

Diverging from the conventional HyperNetworks (Ha et al., 2016), the direct prediction of distinct attributes, such as emitted color and volume density, within the NeRF framework often proves sub-optimal. Drawing an analogy to the vanilla NeRF (Mildenhall et al., 2021), where different network depths were assigned for \mathcal{F}_{geo} and \mathcal{F}_{app} , it is essential for InsertNeRF to discuss distinct HyperNet module structures for them. \mathcal{F}_{geo} plays a pivotal role in the NeRF’s geometric representations, which necessitates accurate inference and completion of the relationship between references based on global-local features. Based on this analysis, multi-scale features \mathbf{H}_l are systematically introduced into the \mathcal{F}_{geo} ’s HyperNet modules, as shown in Fig. 6(a). After Eq. (7), \mathbf{F}_{output} is subsequently fed into an original NeRF’s MLP layer. Note that we also incorporate an additional ReLU and BatchNorm after MLP to ensure training stability for HyperNet modules.

$$\mathbf{F}_{final} = \text{ReLU}(\text{BatchNorm}(\text{MLP}_l(\mathbf{F}_{output}))), \quad (10)$$

Note that in \mathcal{F}_{geo} , as the network’s depth increases, we leverage denser features for guiding weights prediction, i.e., global to dense. Additionally, when the number of MLP layers surpasses that of the feature layers, \mathbf{H}_0 will be recurrently utilized in the remaining MLP layers.

Unlike \mathcal{F}_{geo} , \mathcal{F}_{app} exhibits a heightened focus on dense features (Wang et al., 2022) and the smooth BRDF prior for surface reflectance (Zhang et al., 2020). As shown in Fig. 6(b), \mathcal{F}_{app} ’s HyperNet modules employ a parallel progressive generation paradigm and residual connection that respond to the desired smoothness. Specifically, given dense feature \mathbf{H}_0 ,

$$\tilde{\mathbf{F}}_{final} = \text{MLP} \left(\sum_{z=0}^Z (\text{Weight}_{T_z} \times \mathbf{F}_{input}) + \mathbf{F}_{input} \right), \quad \begin{cases} T_z = \text{Weight}_{T_{z-1}} & z \geq 1 \\ T_z = \mathbf{H}_0 & z = 0 \end{cases}, \quad (11)$$

where Z represents the number of parallel branches. Note that DFILM and dynamic bias are not utilized in \mathcal{F}_{app} for the smooth BRDF prior.

C.2 PSEUDOCODE

In contrast to the scene-customized vanilla NeRF and its derivative works, GNeRF primarily concentrates on cross-scene rendering tasks without any retraining. As shown in Fig. 1, our HyperNet modules possess the capacity to instill generalizability into NeRF-like systems. Here, we delve further into the training processes of InsertNeRF-like systems. In Algorithm 1, due to $\Omega_T^{\text{TrScene}}$ ’s adaptive response to the stochastic sampling of scenes by $\mathcal{D}_{\text{TrScene}}$ (that represents data of training scenes), InsertNeRF-like systems acquires inherent generalizability, where NeRF-like systems encompass diverse frameworks, including but not limited to mipNeRF, NeRF++, NeRF-, and others. In the evaluation phase, for any given P_T , we sample neighboring views $\{\mathbf{I}_n, \mathbf{P}_n\}_{n=1}^N$ from $\mathcal{D}_{\text{TeScene}}$, rendering for \mathbf{I}_T with the pretrained $\Theta_{\text{NeRF-like Systems}}$ and Θ_{HyperNet} , as shown in Algorithm 2.

C.3 GRAPH REASONING

Graph-based methods have been the focus of extensive research recently and shown to be an efficient way of relation reasoning (Wang et al., 2018). Following the spatial properties, we conceptualize

Algorithm 1: Training for InsertNeRF-like Systems**Data:** Training Datasets $\mathcal{D}_{\text{Train}}$ **Result:** $\Theta_{\text{NeRF-like Systems}}, \Theta_{\text{HyperNet}}$

```

1 while  $t \leq it$  do
2   Sample:  $r(t_i) \leftarrow \left\{ \{I_T, P_T\}, \{I_n, P_n\}_{n=1}^N \right\} \leftarrow \mathcal{D}_{\text{TrScene}} \leftarrow \mathcal{D}_{\text{Train}}$ 
3    $\Omega_T^{\text{TrScene}} \leftarrow \text{HyperNet} \left( \left\{ \mathcal{F}_{\text{view}} \left( \{F_n(\Pi_n(r(t_i)))\}_{n=1}^N \right) \right\}_{i=1}^K \right)$ 
4   InsertNeRF-like Systems  $\leftarrow$  NeRF-like Systems  $(r(t_i), \Omega_T^{\text{TrScene}})$ 
5    $\Theta_{\text{NeRF-like Systems}}, \Theta_{\text{HyperNet}} \leftarrow \mathcal{L}$ 
6    $t \leftarrow t + 1$ ;
7 end

```

Algorithm 2: Testing for InsertNeRF-like Systems**Input :** $\Theta_{\text{NeRF-like Systems}}, \Theta_{\text{HyperNet}}, \mathcal{D}_{\text{Test}}$, and Random P_T in Testing Scenes.**Output:** I_T

```

1 Sample:  $r(t_i) \leftarrow \left\{ \left\{ \{I_n, P_n\}_{n=1}^N \leftarrow \mathcal{D}_{\text{TeScene}} \leftarrow \mathcal{D}_{\text{Test}} \right\}, \{P_T\} \right\}$ 
2    $\Omega_T^{\text{TeScene}} \leftarrow \text{HyperNet} \left( \left\{ \mathcal{F}_{\text{view}} \left( \{F_n(\Pi_n(r(t_i)))\}_{n=1}^N \right) \right\}_{i=1}^K \right)$ 
3    $I_T \leftarrow$  InsertNeRF-like Systems  $(r(t_i), \Omega_T^{\text{TeScene}})$ 

```

all sampled points along a ray in a fully connected graph to find correlations between inter-samples and further update node features. Specifically, as shown in Eq. (6), InsertNeRF initially predicts a learnable adjacency matrix A_l to parameterize the edge weights between nodes, which models the relationships between sampled points. Subsequently, W_l^a is employed to update node states, mitigating the noise from epipolar geometric constrains. Furthermore, the identity matrix I is introduced to guide the learning process to pay more attention to the intrinsic characteristics of node features. An naive approach is to calculate the adjacency matrix based on the similarity between node features or Euclidean distance, and update node states, similar to existing graph convolution works. However, this is computationally expensive, especially for a large number of MLP blocks in InsertNeRF. Inspired by Chen et al. (2019), A_l and W_l^a are replaced by two separate linear layers operating in different dimensions, while the identity matrix is represented as a residual connection,

$$H_l = \text{Linear} \left(\text{Linear} (F_{\text{view}})^T \right)^T + F_{\text{view}}. \quad (12)$$

D DISCUSSION

D.1 DIFFERENT FROM IBRNET ABOUT DYNAMIC-STATIC

For *inputs*, we employ the global-dense features as our multi-layer inputs, compared to IBRNet’s single-layer dense feature, it not only retains the detailed information from dense features but also utilizes global features to predict occluded regions, as demonstrated on the depth renderings shown

Table 7: Ablation studys for Multi-layer Dynamic-Static Aggregation Strategy with IBRNet (Wang et al., 2021)

Methods	NeRF Synthetic			LLFF		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
InsertNeRF w/o Max	27.44	0.930	0.061	25.52	0.854	0.131
IBR-InsertNeRF	25.71	0.909	0.085	25.00	0.836	0.140
Multi-layer IBR-InsertNeRF	26.89	0.915	0.074	25.31	0.845	0.131
InsertNeRF (OUR)	27.57	0.936	0.056	25.68	0.861	0.126

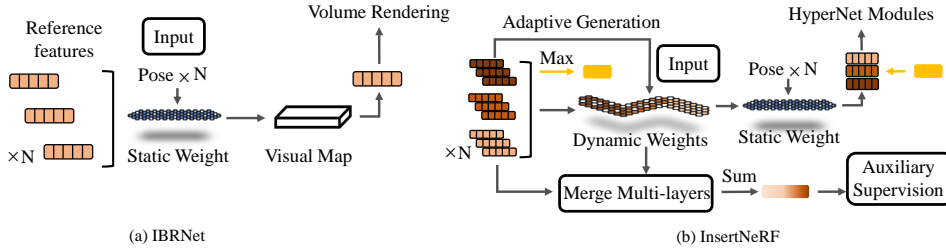


Figure 7: The structures of multi-views feature aggregation parts in IBRNet and InsertNeRF

Table 8: Ablation studies for rendering resolutions.

Methods	LLFF			DTU		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
InsertNeRF(Original Res)	26.22	0.838	0.184	29.88	0.929	0.096
InsertNeRF(1/2 Res)	26.44	0.844	0.169	29.75	0.925	0.077

in Fig. 1b, where reports the rendering results of IBRNet (top) and InsertNeRF (bottom) in terms of color-depth. It is evident that InsertNeRF produces sharper edges and achieves more accurate depth predictions for background regions, even when occluded in the reference views.

For *architecture*, IBRNet generates the visual maps using features based on the M^{ST} , while our InsertNeRF directly predicts M_l^{DY} using multi-layer features before M^{ST} . Intuitively, our strategy prevents excessive reliance on the M^{ST} and contributes to the adaptive inference of relationships between reference-target images. To provide additional validation, we conduct ablation experiments. Concretely, we replace the Multi-layer Dynamic-Static with the aggregation strategy from IBRNet and introduce identical multi-layer inputs into IBR-InsertNeRF for fairness. As shown in Tab. 7, despite its simplicity, our approach yields significant improvements under GNeRF settings. In addition, we also observe that the Multi-layer inputs still contributes to a notable enhancement in IBR-InsertNeRF, which is consistent with the findings in Tab. 5.

For *supervision*, in contrast to the direct generation of visual maps without any supervision, we introduce auxiliary supervision to guide M_l^{DY} in fully encoding global-dense features. As shown in Tab. 5, the significance of our auxiliary supervision cannot be disregarded. In summary, as shown in Fig. 7, compared to IBRNet, the Multi-layer Dynamic-Static Aggregation Strategy focuses on predicting dynamic weights and combines them with static weight based on the multi-layer inputs and the auxiliary supervision to aggregate multiple reference features.

D.2 SCENE REPRESENTATION ANALYSIS

The HyperNet modules instill generalizability into NeRF by generating scene-specific weights in the original framework. To verify this, we visualize the intermediate representations of InsertNeRF through a t-SNE plot. As shown in Fig. 3(b), it is noteworthy that the reduced-dimensional features exhibit scene clustering characteristics in LLFF evaluation data, which may be attributed to the dynamic MLPs and activation functions in our HyperNet modules. NeRF Synthetic also exhibits a similar trend. In Fig. 8, although the scenes still possess clustering characteristics, they appear relatively more dispersed, which may be attributed to significant disparities between evaluation viewpoints in the NeRF Synthetic.

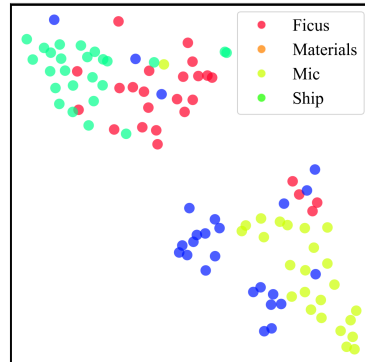


Figure 8: A t-SNE plot of the scene-specific representations in NeRF Synthetic’s scenes.

Table 10: Comparisons of HyperNet modules against SOTA methods on ShapeNet.

Methods	Chairs 1-view		Chairs 2-views		Cars 1-view		Cars 2-views	
	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow
ENR (Dupont et al., 2020)	22.83	-	-	-	22.26	-	-	-
SRN (Sitzmann et al., 2019)	22.89	0.89	24.48	0.92	22.25	0.89	24.84	0.92
ViewFormer (ECCV 2022)	14.74	0.79	17.20	0.84	19.03	0.83	20.09	0.85
pixelNeRF (CVPR 2021)	23.72	0.91	26.20	0.94	23.17	0.90	25.66	0.94
pixelNeRF+HyperNet modules	24.51	0.92	26.71	0.94	24.18	0.91	26.05	0.95

D.3 WHY INSERTNERF DESIGNS COULD HELP NERF GENERALIZATION

Vanilla NeRF can be considered as an implicit representation used to depict a scene through the parameters of a neural network, i.e. Θ , as described in Eq. (1)

A natural idea is how to alter Θ for different scenes s , so that Θ_s possesses the ability to represent this new scene. By sampling different $s \in \mathcal{S}$ and generating different Θ_s , this can be considered as endowing vanilla NeRF representation with generalizability in multi-scenes. However, unlike explicit 3D representations such as voxels, meshes, and point clouds, constructing an implicit representation Θ_s directly for a given s is challenging.

Therefore, in this paper, we introduce the HyperNet modules, which is invented to generate weights for a target neural network, to address this issue. Through two types of the HyperNet modules we propose, scene-customization weights (parameters) Ω_T^s in the NeRF framework are generated in a given s . Here, we predict Ω_T^s by combining the feature extraction from reference images and the multi-Layer dynamic-static aggregation strategy. Finally, by combining Θ and new weights Ω_T^s within the NeRF framework, we obtain Θ_s that can adapt to different scenes s , as described in Sec. 3.2

E ADDITIONAL RESULTS AND ANALYSIS

E.1 ADDITIONAL ABLATION STUDIES

Rendering Resolutions: During the evaluation phase, prior works employed different rendering resolutions, which has some impact on the metric. To investigate this issue, we evaluate the rendering performance at different resolutions without altering the training settings. In Tab. 8, reducing the rendering resolution not only improved rendering efficiency but also demonstrated performance enhancements in LLFF. However, in the DTU dataset, a contrasting trend is evident, which may be attributed to its emphasis on foreground rendering.

Parallel Branches Z : We further analyze the influence of the number of parallel branches Z on network rendering performance, as shown in Tab. 9. When $Z = 1$, i.e., \mathcal{F}_{app} is replaced by \mathcal{F}_{geo} with the same input, a significant performance drop occurs. This might be attributed to the implicit modeling of BRDF prior by \mathcal{F}_{app} . Furthermore, with an increase in the number of branches, \mathcal{F}_{app} endows the NeRF framework with enhanced fine-detail generalizability.

Table 9: Ablation studies for \mathcal{F}_{app} .

Z	DTU		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
$Z = 1$ (\mathcal{F}_{geo})	29.03	0.918	0.086
$Z = 2$	29.75	0.925	0.077
$Z = 3$	29.83	0.925	0.075

E.2 RESULTS FOR SHAPENET

In this section, we explore the performance of the our InsertNeRF in ShapeNet under Chairs and Cars scenes. Due to the primary emphasis of InsertNeRF on multi-view settings I and II, the validation for multi-layer dynamic-static aggregation strategy under few-views settings is unnecessary. Therefore, we integrate the HyperNet modules into the original pixelNeRF Yu et al. (2021), altering its training inputs accordingly. As shown in the Tab. 10, our modules exhibit significant improve-

Table 11: The performance in different scenes and the results after Fine-Tuning in NeRF Synthetic.

(a) PSNR										
Method	FT	Lego	Chair	Drums	Ficus	Hotdog	Materials	Mic	Ship	Avg.
InsertNeRF		30.00	31.97	25.59	26.08	35.04	28.91	35.13	30.08	30.35
MVSNeRF	✓	-	-	-	-	-	-	-	-	27.21
IBRNet	✓	-	-	-	-	-	-	-	-	30.05
GNT	✓	31.38	33.70	26.98	29.55	36.95	29.11	33.35	30.54	31.45
NeuRay	✓	-	-	-	-	-	-	-	-	32.35
InsertNeRF	✓	31.93	34.41	27.76	29.08	37.25	31.46	36.43	31.98	32.54

(b) SSIM										
Method	FT	Lego	Chair	Drums	Ficus	Hotdog	Materials	Mic	Ship	Avg.
InsertNeRF		0.939	0.969	0.910	0.915	0.969	0.922	0.972	0.888	0.936
MVSNeRF	✓	-	-	-	-	-	-	-	-	0.888
IBRNet	✓	-	-	-	-	-	-	-	-	0.935
NeuRay	✓	-	-	-	-	-	-	-	-	0.960
InsertNeRF	✓	0.962	0.981	0.937	0.959	0.988	0.957	0.978	0.931	0.962

(c) LPIPS										
Method	FT	Lego	Chair	Drums	Ficus	Hotdog	Materials	Mic	Ship	Avg.
InsertNeRF		0.056	0.033	0.079	0.085	0.037	0.076	0.029	0.129	0.066
MVSNeRF	✓	-	-	-	-	-	-	-	-	0.162
IBRNet	✓	-	-	-	-	-	-	-	-	0.066
NeuRay	✓	-	-	-	-	-	-	-	-	0.048
InsertNeRF	✓	0.044	0.024	0.065	0.054	0.027	0.048	0.021	0.102	0.048

ments compared to pixelNeRF Yu et al. (2021), especially in the 1-view setting. It’s also evident that compared to NeRF Synthetic, LLFF, and DTU, InsertNeRF shows less improvement on ShapeNet. This might be due to the relatively simplistic appearance and geometry of ShapeNet-Scenes, and our work primarily focuses on multi-view settings as mentioned in (Kulhánek et al., 2022).

E.3 RESULTS FROM FINE-TUNING

We also explore the rendering performance of InsertNeRF after fine-tuning in various scenes. In contrast to the fine-tuning methodology adopted by (Liu et al., 2022), we fine-tune directly on the pre-trained model. Tab. 11 presents the performance of InsertNeRF across different scenes and the results after fine-tuning in NeRF Synthetic.

E.4 SINGLE SCENE RESULTS

Existing works (Wang et al., 2022) also tend to focus on single-scene rendering within the framework of GNeRF. We conduct a quantitative comparison with existing works in the single-scene setting and achieve satisfactory performance, as shown in Tab. 12.

E.5 MORE QUALITATIVE RESULTS

We present additional qualitative results to further analyze the superiority of InsertNeRF. i). Fig. 9 and Fig. 10 report qualitative results in LLFF and NeRF Synthetic. ii). Fig. 11 showcase more Color-Depth results in LLFF and DTU datasets. iii). We also qualitatively analyze the generalizability of the InsertNeRF-systems including Insert-mip-NeRF Fig. 12, and Insert-NeRF++ Fig. 13. Note that the presence of color distortion in the Insert-NeRF++’s foreground rendering is observed, yet it does not impact the combined results, possibly attributable to the replaced sampling process.

Table 12: Comparison of InsertNeRF for single scene rendering on the NeRF Synthetic.

(a) PSNR								
Method	Room	Fern	Leaves	Fortress	Orchids	Flower	T-Rex	Horns
LLFF	24.54	28.72	21.13	21.79	18.52	20.72	27.48	23.22
NeRF	<u>32.70</u>	<u>25.17</u>	20.92	31.16	20.36	<u>27.40</u>	26.80	27.45
GNT	32.96	24.31	<u>22.57</u>	32.28	<u>20.67</u>	<u>27.32</u>	28.15	29.62
InsertNeRF	32.55	24.88	22.59	<u>31.82</u>	21.18	28.39	<u>27.49</u>	<u>29.37</u>

(b) SSIM								
Method	Room	Fern	Leaves	Fortress	Orchids	Flower	T-Rex	Horns
LLFF	0.932	0.753	0.697	0.872	0.588	0.844	0.857	0.840
NeRF	0.948	0.792	0.690	0.881	0.641	0.827	0.880	0.828
GNT	0.963	0.846	<u>0.852</u>	0.934	<u>0.752</u>	<u>0.893</u>	0.936	0.935
InsertNeRF	<u>0.961</u>	0.846	0.853	<u>0.925</u>	0.756	0.904	<u>0.928</u>	<u>0.932</u>

(c) SSIM								
Method	Room	Fern	Leaves	Fortress	Orchids	Flower	T-Rex	Horns
LLFF	0.155	0.247	0.216	0.173	0.313	0.174	0.222	0.193
NeRF	0.178	0.280	0.316	0.171	0.321	0.219	0.249	0.268
GNT	0.060	0.116	0.109	0.061	<u>0.153</u>	<u>0.092</u>	0.080	<u>0.076</u>
InsertNeRF	<u>0.063</u>	<u>0.121</u>	0.109	<u>0.062</u>	0.152	0.070	<u>0.085</u>	0.074

F LIMITATIONS

In the majority of scenarios, a higher number of sample points along the rays often leads to improved rendering performance. In essence, thanks to the transformer architecture, existing works (Wang et al., 2022; 2021) can be trained on a limited number of sample points (64 training samples) and evaluated on the more sample points (192 evaluation samples), resulting in elevated training efficiency and improved rendering performance. However, in the InsertNeRF-system, it is essential to maintain consistency in the number of sample points between the training and evaluation processes. In order to ensure fairness in comparative experiments and strike the trade-off between training efficiency and rendering performance, we set $K = 64$ both during training and evaluation, which imposes certain limitations on the rendering performance. Naturally, as we increase the number of sample points for training, the rendering performance will further improve.

G FUTURE WORK

We aspire to construct an all-encompassing InsertNeRF framework, endowing generalizability into various NeRF-derived works, such as TensorRF, NeRF-, NeuS, and so forth. This can facilitate existing or future NeRF research to transcend the constraints of scene-customization.

In addition, we have provided a pre-trained model to address NeRF under sparse inputs in Sec. 4.4. While such models have demonstrated satisfactory performance without any retraining Tab. 3, we still plan to design a fine-tuning approach for sparse inputs to further enhance rendering quality.



Figure 9: Qualitative comparisons of InertNeRF against SOTA methods under LLFF scenes.

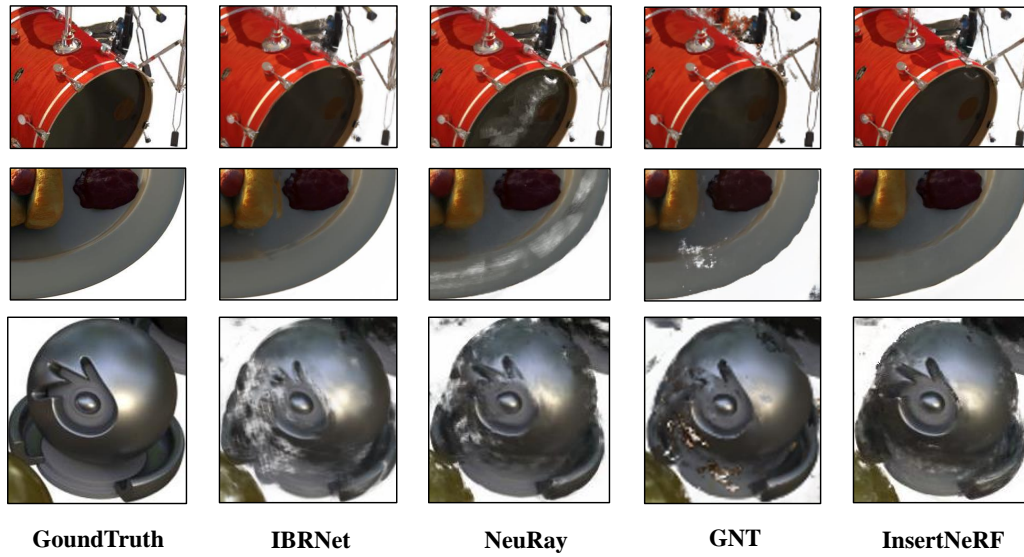


Figure 10: Qualitative comparisons of InsertNeRF against SOTA methods under NeRF Synthetic scenes.

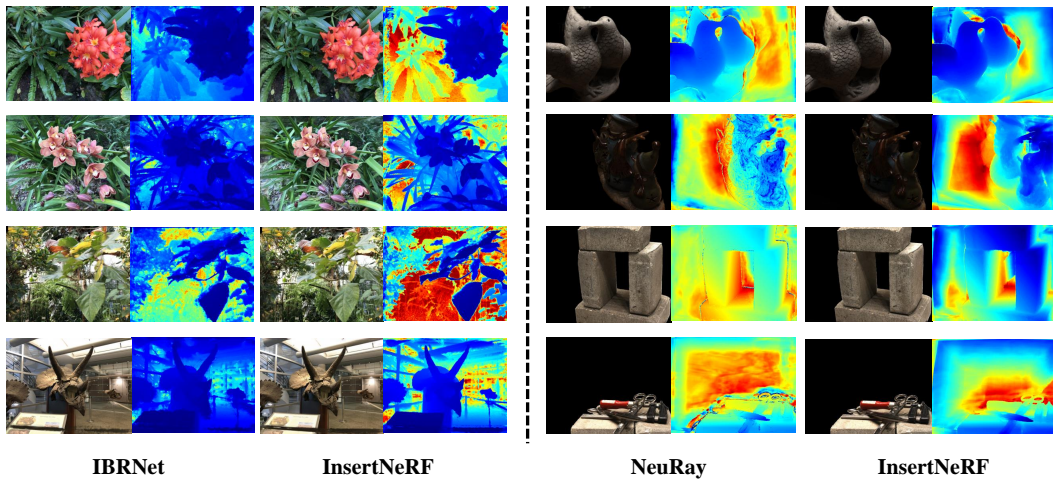


Figure 11: Color-Depth results of InsertNeRF with existing works under LLFF and DTU's scenes.

