WaveNeRF: Wavelet-based Generalizable Neural Radiance Fields

Muyu Xu[1]   Fangneng Zhan[2]   Jiahui Zhang[1]   Yingchen Yu[1]
Xiaoqin Zhang[3]   Christian Theobalt[2]   Ling Shao[4]   Shijian Lu[1]

[1]Nanyang Technological University   [2]Max Planck Institute for Informatics
[3]Wenzhou University   [4]UCAS-Terminus AI Lab, UCAS

Abstract

Neural Radiance Field (NeRF) has shown impressive performance in novel view synthesis via implicit scene representation. However, it usually suffers from poor scalability as requiring densely sampled images for each new scene. Several studies have attempted to mitigate this problem by integrating Multi-View Stereo (MVS) technique into NeRF while they still entail a cumbersome fine-tuning process for new scenes. Notably, the rendering quality will drop severely without this fine-tuning process and the errors mainly appear around the high-frequency features. In the light of this observation, we design WaveN-eRF, which integrates wavelet frequency decomposition into MVS and NeRF to achieve generalizable yet high-quality synthesis without any per-scene optimization. To preserve high-frequency information when generating 3D feature volumes, WaveNeRF builds Multi-View Stereo in the Wavelet domain by integrating the discrete wavelet transform into the classical cascade MVS, which disentangles high-frequency information explicitly. With that, disentangled frequency features can be injected into classic NeRF via a novel hybrid neural renderer to yield faithful high-frequency details, and an intuitive frequency-guided sampling strategy can be designed to suppress artifacts around high-frequency regions. Extensive experiments over three widely studied benchmarks show that WaveNeRF achieves superior generalizable radiance field modeling when only given three images as input.

1. Introduction

Rendering novel views from a set of posed scene images has been studied for years in the fields of computer vision and graphics. With the emergence of implicit neural representation, neural radiance field (NeRF)[25] and its variants[21, 23] have recently achieved very impressive performance in novel view synthesis with superb multi-view

---

*Shijian Lu is the corresponding author.

consistency. However, most existing works fall short in model scalability by requiring a per-scene optimization process with densely sampled multi-view images for training.

To avoid the cumbersome process of training from scratch for new scenes, a popular line of generalizable NeRF [3, 37, 32, 34, 13] introduces a pipeline that first trains a base model on the training data and then conducts fine-tuning for each new scene, which improves the scalability and shortens the per-scene training process. Their base models often extract features from the source views and then inject the features into a neural radiance field. Several previous studies [37, 32] directly use CNN networks to extract features while recent generalizable NeRF models [3, 34, 13] resort to Multi-View Stereo (MVS) technique to warp 2D source feature maps into 3D features planes, yielding better performance than merely using CNN networks. However, per-scene fine-tuning still entails a fair number of posed training images that are often challenging to collect in various real-world tasks. On the other hand, removing the per-scene fine-tuning will incur a significant performance drop with undesired artifacts and poor detail. Notably, we intriguingly observe that the rendering error mainly lies around image regions with rich high-frequency information as illustrated in Fig. 1. The phenomenon of losing high-frequency detail is largely attributed to the fact that most existing generalizable NeRFs conduct down-sampling operations at the feature extraction stage of their pipeline, i.e., the CNN networks adopted in [37, 32] or the MVS module adopted in [3, 34, 13].

In the light of the aforementioned observation, we present **Wave**lets-based **N**eural **R**adiance **F**ields (**WaveNeRF**) which incorporates explicit high-frequency information into the training process and thus obviates the per-scene fine-tuning under the generalizable and few-shot setting. Specifically, with MVS technique to construct 3D feature volumes which are converted to model NeRF in the spatial domain, we further design a Wavelet Multi-View Stereo (WMVS) to incorporate scene wavelet coefficients into the MVS to achieve frequency-domain modeling. Distinct from other frequency transformations like

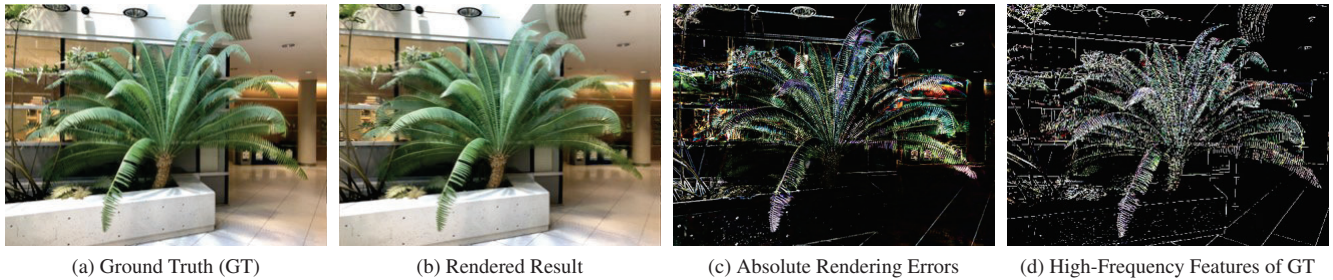| (a) Ground Truth (GT) | (b) Rendered Result | (c) Absolute Rendering Errors | (d) High-Frequency Features of GT |

Figure 1: The comparison between the absolute rendering errors (c) of GeoNeRF[13] and the high-frequency features of the ground truth (d). We can see that the errors mainly appear around the pixels with high-frequency features.

Fourier Transform, WaveNeRF employs Wavelet Transform which is coordinate invariant and preserves the relative spatial positions of pixels. This property is particularly advantageous in the context of MVS as it allows multiple input views to be warped in the direction of a reference view to form sweeping planes in both the spatial domain and the frequency domain within the same coordinate system. Apart from MVS, this property also enables to build a frequency-based radiance field so that a designed Hybrid Neural Renderer (HNR) can leverage the information in both the spatial and frequency domains to boost the rendering quality of the appearance, especially around the high-frequency regions. In addition, WaveNeRF is also equipped with a Frequency-guided Sampling Strategy (FSS) which enables the model to focus on the regions with larger high-frequency coefficients. The rendering quality can be improved clearly with FSS by sampling denser points around object surfaces.

The contributions of this work can be summarized in three points.

- *First*, we design a WMVS module that preserves high-frequency information effectively by incorporating wavelet frequency volumes while extracting geometric scene features.

- *Second*, we design a HNR module that can merge the features from both the spatial domain and the frequency domain, yielding faithful high-frequency details in neural rendering.

- *Third*, we develop FSS that can guide the volume rendering to sample denser points around the object surfaces so that it can infer the appearance and geometry with higher quality.

2. Related Works

2.1. Multi-View Stereo

Multi-view stereo (MVS) is a method that involves using multiple images taken from various viewpoints to create a detailed 3D reconstruction of an object or scene. Over

time, various conventional methods have been proposed and tested in this field [6, 16, 7, 29, 8, 27]. More recently, deep learning techniques have been integrated into the multi-view stereo process. One such technique is MVS-Net [35], which extracts features from all input images and warps them onto a reference image to generate probabilistic planes with varying depth values. These planes are then combined to create a variance-based cost volume that accurately represents the specific scene.

Although MVS methods have demonstrated promising performance, their large memory requirements, due to the 3D volume grid and operations, severely limit the resolution of input images and subsequent development of deep learning-based MVS research. To address this issue, R-MVSNet [36] sequentially regularizes the cost volume with GRU, making MVSNet more scalable. In addition, cascade MVS models [5, 11] use a coarse-to-fine strategy to generate cost volumes of various scales and compute depth output accordingly, freeing up more memory space. MVS has been shown to be effective in inferring the geometry and occlusions of a scene [3, 13]. We follow the previous MVS techniques and further introduce wavelet transform into it to achieve a higher quality of inference.

2.2. Neural Radiance Field

3D scene reconstruction and novel view synthesis have been extensively studied for many years. Researchers have used various explicit representations of scene geometry such as 3D meshes [14, 22] and point clouds [17, 1]. However, NeRF [25] employs an implicit neural representation method that uses an MLP-based network to render novel views. NeRF has demonstrated excellent rendering performance and has been further extended to various computer vision tasks [15, 4, 26, 9, 18, 10, 28, 2, 30, 38, 19, 20]. Although all of these studies showcase the impressive strength of NeRF in specific tasks, they still follow the same training process as the original NeRF and require per-scene training to complete the corresponding task.

To address this issue, several studies in the generalization

of NeRF have shown some degree of success. Specifically, PixelNeRF [37] and IBRNet [32] both rely on the notion that aggregating multi-view features at each sampled point leads to better performance than using direct encoded RGB inputs. Another typical approach that achieves generalizable NeRF is using multi-view stereo (MVS) techniques. For instance, MVSNeRF [3], which is the first to combine MVSNet and NeRF, simply concatenates the cost volume in MVSNet with the 5D input in NeRF. More recent generalizable NeRF, PointNeRF [34] and GeoNeRF [13], both use MVS techniques to obtain a coarse 3D representation, but PointNeRF uses point cloud growing to enhance the inference ability, while GeoNeRF uses attention-based transformer modules.

Although some of the NeRF models are generalizable, they typically require a specific number of inputs, such as 10 source views in IBRNet [32]. In addition, almost all of them need per-scene optimization to achieve photorealistic outcomes. Per-scene optimization is actually an additional training process which greatly impairs the generalizability. It is worth noting that without this optimization process, the rendering quality of these existing models can drop significantly, with most errors occurring around high-frequency features. Based on this observation, we integrate wavelet frequency decomposition into NeRF to achieve generalizable yet high-quality synthesis without any per-scene optimization. We believe that this approach is much more realistic, as it mimics situations where intelligent vehicles have limited sensors and need to reconstruct 3D scenery immediately.

3. Method

This section presents our novel wavelet-based generalizable NeRF, designed for synthesizing high-quality novel views of a scene from three-shot source views without any per-scene fine-tuning process. Inspired by the observation that the rendering errors of the previous models mainly gather around the high-frequency regions, we design a Wavelet Multi-view Stereo (WMVS) module to obtain feature volumes in both the spatial domain and the frequency domain so that the high-frequency information can be maintained and represented separately. Besides, since the renderer in prior studies is unable to directly disentangle the errors around high-frequency features, we implement a Hybrid Neural Renderer (HNR) that can adjust the rendered colors based on the high-frequency information obtained from WMVS. During this rendering process, we also notice that previous sampling strategies necessitate an additional sampling step based on the outcome of the initial sampling, or they simultaneously sample all the points at the expense of sampling quality. Therefore, to achieve higher sampling quality where more samples are around objects in the scene in a one-round sampling process, we adopt a Frequency-guided Sampling Strategy (FSS) where the coordinates of the sampled points are determined by the distribution of the features in the frequency feature volume.

The overall architecture of WaveNeRF is shown in Fig. 2. We elaborate on our designed WMVS, FSS, and HNR, in Section 3.1, 3.2, and 3.3 respectively.

3.1. Wavelet Multi-view Stereo

Since Wavelet Transform can decompose an image into components with different scales, it naturally fits with the pyramid structure of the CasMVSNet[11]. Therefore, given a set of input source views $\{I_v\}_{v=0}^{V}$ with the size of $H \times W$, we design a Wavelet Multi-View Stereo (WMVS) module to construct cascaded spatial feature volumes as well as a high-frequency feature volume following the similar way of CasMVSNet as shown in Fig. 2. We make several modifications to both the feature extraction process and the volume construction process of CasMVSNet. First, we utilize level-2 Discrete Wavelet Transform (DWT) to obtain different frequency components, where w_L represents the low-frequency component and $w_H^{(l)}$ represents the high-frequency components of level l. The low-frequency components w_L have the smallest size ($\frac{H}{4}, \frac{W}{4}$) and are directly used to generate the lowest level of semantic feature maps $f_s^{(0)}$ via a CNN-based feature extractor. For each level of high-frequency components, it is infeasible to generate spatial features by naively adding different frequency components together due to the domain gap. We thus design an Inverse Wavelet Block (IWB) that simulates the inverse discrete wavelet transform by combining frequency features of the previous level with high-frequency features of the current level via dilated deconvolution to generate latent spatial feature maps $f_L^{(l)}$. Then the latent spatial feature maps are used to generate semantic feature maps of the current level by CNN as below:

$$f_s^{(l)} = \mathbf{CNN}(f_s^{(l-1)}, \mathbf{IWB}(f_L^{(l-1)}, w_H^{(l)})), \ l \in 1, 2. \quad (1)$$

In addition, all the high-frequency features are eventually gathered to form the 2D compounded high-frequency components which are used to generate frequency feature maps f_w by a CNN-based network.

After having the spatial semantic feature maps and the wavelet feature maps, we follow the same approach as in CasMVSNet [11] to build sweep planes and spatial feature volumes $P_s^{(l)}$ at three levels. Besides, thanks to the nice property of Wavelet Transform that it does not affect the relative coordinates, we can follow the same manner to construct the high-frequency feature volume. Since high-frequency information is often sparsely distributed, it is sufficient to represent the high-frequency features in a relatively small volume. Here we choose to use the second coarsest level ($l = 1$) to balance the depth range and the
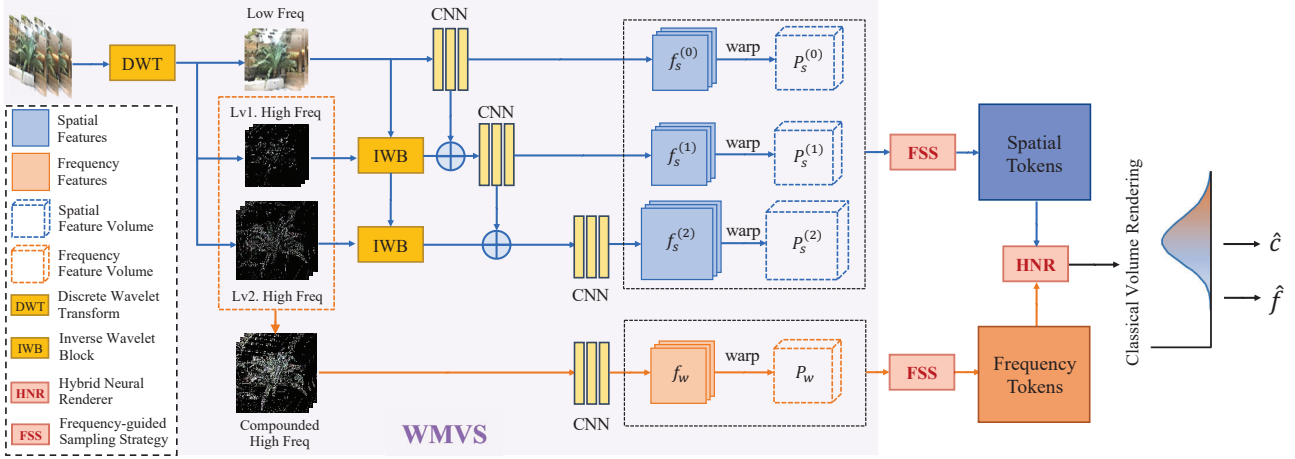
Figure 2: The overview of the proposed WaveNeRF. With sparse input views, wavelet multi-view stereo (WMVS) is designed to produce frequency feature volume f_w and multi-level spatial feature volumes $[f_s^{(0)}, f_s^{(1)}, f_s^{(2)}]$. Specifically, the input views are first divided into different frequency components with level-2 discrete wavelet transform. The spatial and frequency features are then obtained via our designed Inverse Wavelet Blocks (IWB) and CNN-based feature extractors, and warped into corresponding 3D feature volumes$[P_s^{(0)}, P_s^{(1)}, P_s^{(2)}, P_w]$. With 2D features and 3D volumes, a novel Frequency-guided Sampling Strategy (FSS) is introduced to yield more precise samples with spatial and frequency tokens. These tokens are fed into a subsequent Hybrid Neural Renderer (HNR) to infer the volume density, colors, and frequency coefficients.

depth sampling precision and construct a wavelet frequency feature volume P_w with the size of $\frac{H}{2} \times \frac{W}{2}$. In a nutshell, given a set of input source views $\{I_v\}_{v=0}^V$, our WMVS module generates 2D feature maps $f_s^{(l)}, f_w$ and their corresponding 3D features volumes $P_s^{(l)}, P_w$ for subsequent modules as below:

$$(f_s^{(l)}, f_w, P_s^{(l)}, P_w) = \textbf{WMVS}(\{I_v\}_{v=0}^V), \ l \in 0, 1, 2. \tag{2}$$

3.2. Frequency-guided Sampling Strategy

After generating features from the WMVS module, we use the ray-casting approach to create new views. To cover the depth range, we sample N_c points uniformly along each camera ray at a novel camera pose. Many previous studies [21, 23, 37, 32] follow the classic NeRF [25], sampling N_f points based on the volume density distribution inferred by the N_c points to approximate the object surfaces. However, this coarse-to-fine sampling strategy requires training two NeRF networks at the same time. MVSNeRF [3] directly discards the fine sampling and claims that adding a fine sampling process cannot significantly improve the performance. GeoNeRF [13] first estimates a set of valid coarse sample points by checking whether the coordinates lie within the valid NDC (Normalized Device Coordinate) coordination system and then randomly samples N_f points around these valid coarse points. Although GeoNeRF simultaneously samples a mixture of $N_c + N_f$ points, it cannot ensure the sampled points are near the objects.
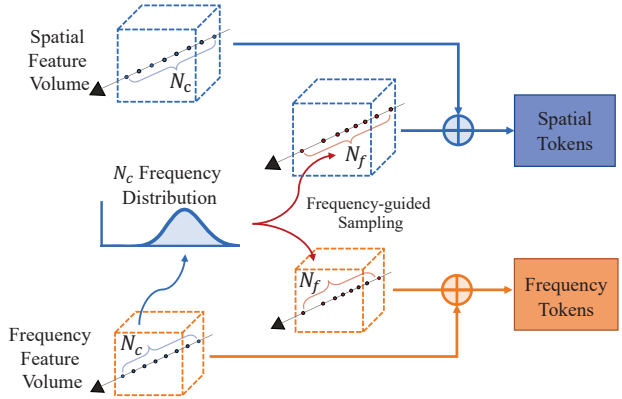

Figure 3: The illustration of our Frequency-guided Sampling Strategy (FSS). We utilize the distribution of the coarse sampling points in the frequency volume to determine the distribution of the fine sampling points. Areas having higher wavelet feature values are more likely to be sampled in the fine sampling process.

We propose a frequency-guided sampling strategy (FSS) (as shown in Fig. 3) based on the observation that high-frequency features often indicate valuable scene information. Our strategy first uses the coordinates of coarse sampling points to fetch corresponding high-frequency features from the wavelet feature volume P_w. Then, we use these frequency features to create a probability density function

p_0 along the ray, which determines the distribution of the fine sampling points. Regions with higher wavelet feature values have a higher probability of being sampled in the fine sampling process which yields better sampling quality.
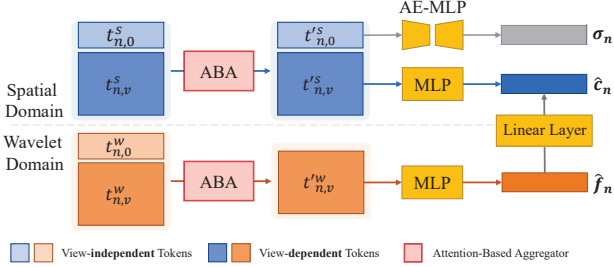

Figure 4: The overall structure of the Hybrid Neural Renderer (HNR). First, attention-based modules are employed to obtain refined tokens $\{t'^{s/w}_{n,v}\}$ for each domain. These tokens are then sent to MLP-based modules introduced in GeoNeRF [13] to generate volume density σ_n, color $\hat{c}_n$, and frequency coefficient $\hat{f}_n$ for each point x_n. The frequency coefficient $\hat{f}_n$ is further used to adjust the color after passing through the linear layers.

3.3. Hybrid Neural Renderer

Since we have feature volumes $P_s^{(l)}, P_w$ in both the spatial domain and the frequency domain and the coordinates of the sampled points from FSS, we can fetch the features from the feature volumes and represent them as sets of tokens. For a point x_n in both domains, we generate a view-independent (i.e., global) token $t_{n,0}$ and V view-dependent tokens $t_{n,v}$. We define t^s and t^w as tokens in the spatial domain and tokens in the wavelet frequency domain, respectively. For a sample n, $t_{n,0}^{s/w}$ could be considered as a global understanding of the scene at point x_n, while $t_{n,v}^{s/w}$ represents the view-dependent understanding of the scene. We then implement a Hybrid Neural Renderer (HNR) which integrates these tokens to estimate both the colors and the frequency coefficients of the rays. The overall structure of the HNR is shown in Fig. 4.

We first adopt an Attention-Based Aggregator(ABA) in GeoNeRF [13] to refine the feature tokens. The refined view-independent tokens are used to estimate the volume density while the refined view-dependent tokens are utilized to predict the colors and frequency coefficients. Since the global information of wavelet high-frequency is often sparse and we demand local high-frequency enhancement, we only reserve the view-independent tokens in the spatial domain for the subsequent volume density estimation. Hence, the output of ABA only contains one set of view-independent tokens $\{t'_{n,0}\}_{n=1}^N$ which have access to all necessary data to learn the geometry of the scene and estimate

volume densities. These view-independent tokens are then regularized using an auto-encoder-style MLP network (AE-MLP) [13]. The AE-MLP network learns the global geometry along the ray using convolutional layers and predicts more coherent volume densities σ_n. Notably, only the tokens in the frequency domain $\{t'^w_{n,v}\}_{v=1}^V$ are used to predict the frequency coefficients $\hat{f}_n$ while the color prediction utilizes all the view-dependent tokens. The prediction of color and frequency coefficients for each point relies on a weighted sum of the source view samples. The weight of each view, denoted as $w_{n,v}^{s/w}$, is determined using a MLP-based module. To obtain the color and wavelet samples for each point x_n, we project them onto the source images and the source wavelet frequency maps, resulting in the samples $c_{n,v}$ and $f_{n,v}$, respectively. We first estimate the wavelet coefficients via this weighted sum process. These wavelet coefficients form another set of weights by two linear layers which are further used to adjust the color prediction based on the weighted sum of the color samples as:

$$\hat{f}_n = \sum_{v=1}^V w_{n,v}^w f_{n,v} , \tag{3}$$

$$\hat{c}_n = (\sum_{v=1}^V w_{n,v}^s c_{n,v}) * (\mathbf{LT}(\hat{f}_n) + 1). \tag{4}$$

We argue that this design can increase the significance of the color samples around the surfaces of the objects and can reconstruct more details of the objects in the novel view.

Once we have the prediction of the volume densities, colors, and frequency coefficients, the color and the wavelet coefficient of the camera ray at a novel pose can be estimated via the classic volume rendering technique in NeRF [25]. Besides the color and the wavelet coefficient, we also predict the depth value of each ray for the depth supervision (see supplementary materials for more details). The volume rendering can be represented as:

$$\{\hat{c}, \hat{f}, \hat{d}\} = \sum_{n=1}^N \exp(- \sum_{k=1}^{n-1} \sigma_k)(1 - \exp(-\sigma_n))\{\hat{c}_n, \hat{f}_n, z_n\}, \tag{5}$$

where z_n is the depth of point x_n with respect to the novel pose.

3.4. Loss Function

Based on previous studies, we adopt the same primary color loss $\mathcal{L}_c$ and depth loss $\mathcal{L}_D$ as GeoNeRF [13]. For more details about these losses, please refer to the supplementary materials.

In addition to these losses, we introduce two frequency losses on the predicted wavelet coefficients to supervise the training in the frequency domain. The base frequency loss

Method	DTU [12]			NeRF Synthetic [25]			LLFF [24]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
PixelNeRF [37]	19.31	0.789	0.382	7.390	0.658	0.411	11.24	0.486	0.671
MVSNeRF [3]	20.68	0.875	0.243	16.70	0.845	0.278	20.07	0.726	0.318
PointNeRF [34]	23.89	0.874	0.203	22.73	0.887	0.193	N/A	N/A	N/A
GeoNeRF [13]	27.67	0.920	0.117	24.80	0.891	0.182	23.22	0.757	0.248
GeoNeRF*	29.02	0.940	0.0864	25.83	0.907	0.137	24.31	0.793	0.213
WaveNeRF	29.55	0.948	0.0749	26.12	0.918	0.113	24.28	0.794	0.212

Table 1: Quantitative comparison of our proposed WaveNeRF with existing generalizable NeRF models in terms of PSNR↑, SSIM↑, and LPIPS↓ metrics. The results in red are the best, the results in orange are the second best, and the third best ones are in yellow. '*' denotes that the model (i.e., GeoNeRF) was trained based on a pre-trained Cascade MVSNet checkpoint, while our model is trained from scratch. When we train the GeoNeRF model from scratch using their training scripts, its performance degrades to the values shown in the row of GeoNeRF.

function is similar to the color loss function and calculates the mean squared error between the predicted wavelet coefficients and the ground truth pixel wavelet coefficients as below:

$$\mathcal{L}_{f_b} = \frac{1}{|R|} \sum_{r \in R} ||\hat{f}(r) - f_{gt}(r)||^2, \quad (6)$$

where R is the set of rays in each training batch and f_{gt} is the ground truth frequency coefficients.

To improve learning around high-frequency features, we have also designed a Weighted Frequency Loss (WFL), which is a modified color loss. This loss amplifies the error around the high-frequency features based on the value of the wavelet coefficients in that region. It can be represented as:

$$\mathcal{L}_{f_w} = \frac{1}{|R|} \sum_{r \in R} f_{gt}(r) ||\hat{c}(r) - c_{gt}(r)||^2. \quad (7)$$

Finally, by combining all the losses mentioned above, the complete loss function of our model is represented as:

$$\mathcal{L} = \mathcal{L}_c + 0.1\mathcal{L}_{f_b} + 0.5\mathcal{L}_{f_w} + 0.1\mathcal{L}_D. \quad (8)$$

4. Experiment

Dataset. We have trained our generalizable network using the DTU dataset [12], IBRNet dataset [32], and a real forward-facing dataset from LLFF [24]. For the partition of DTU dataset, we follow the approach of PixelNeRF [37], resulting in 88 training scenes and 16 testing scenes while maintaining an image resolution of 600×800 as in GeoNeRF [13]. For depth supervision, we only use ground truth depths from MVSNet [35] for DTU dataset. For samples from the forward-facing LLFF dataset and IBRNet dataset, we use self-supervised depth supervision. Specifically, we used 35 scenes from LLFF and 67 scenes from IBRNet as in GeoNeRF.

To evaluate our model, we test it on three datasets: DTU test data, Synthetic NeRF data [25], and LLFF Forward-Facing data. DTU dataset contains 16 test scenes and the other two datasets both have 8 test scenes. We followed the same evaluation protocols as NeRF [25] for the synthetic dataset and LLFF dataset, and the same protocol in MVSNeRF [3] for the DTU dataset.

Implementation details. To fit the pyramid structure, we adopt a two-scale (J=2) wavelet transform for the WMVS module. Increasing the number of scales does not improve the rendering quality significantly, but it largely increases the difficulty of implementation due to the complicated padding operations. In contrast to the three different granularities ($D_s = [8, 32, 48]$) for the spatial sweep planes in WMVS, we uniformly sample 32 frequency sweep planes ($D_w = 32$) from near to far because high-frequency features are usually sparsely distributed. We set the number of points in our sampling strategy to be $N_c = 96$ and $N_f = 32$ on a ray for all scenes, and set the number of input views to be $V = 3$ for both the training and evaluation process. For more implementation details, please refer to the supplementary.

4.1. Experiment Results

We evaluate our model and compared it with existing generalizable NeRF models, including PixelNeRF [37], MVSNeRF [3], PointNeRF [34], and GeoNeRF [13]. We quantitatively compare the models in terms of PSNR, SSIM [33], and LPIPS [39] as shown in Table 1, which demonstrates the superiority of our WaveNeRF model over previous generalizable models. Notably, for a fair comparison, we evaluate all methods under the same setting with only three input views, and do not quote the results reported in original papers. Specifically, MVSNeRF [3] has a nearest-view evaluation mode that uses three nearest source views for novel views, which actually imports more than three input views. We thus adopt its fixed-views evalua-
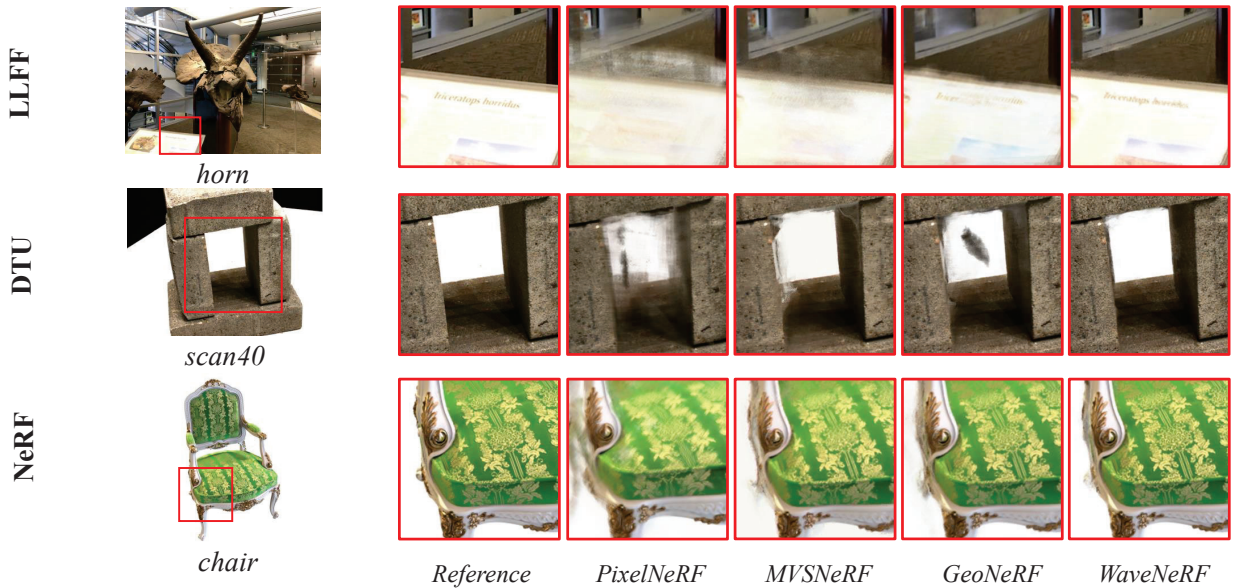
Figure 5: The qualitative results of our WaveNeRF and the comparison with PixelNeRF [37], MVSNet [3], and GeoNeRF [13]. We show the scenes from LLFF dataset [24] (*horn*), DTU dataset [12] (*scan40*), and NeRF synthetic dataset [25] (*chair*). Our WaveNeRF model can preserve more details than the previous generalizable NeRF.

Experiments	DTU [12]			NeRF Synthetic [25]			LLFF [24]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline	27.67	0.920	0.117	24.80	0.891	0.182	23.22	0.757	0.248
+ WMVS	27.97	0.922	0.113	24.63	0.887	0.183	23.23	0.762	0.244
+ WMVS + FSS	28.90	0.942	0.084	25.63	0.912	0.119	23.99	0.782	0.227
+ WMVS + FSS + HNR	29.16	0.942	0.083	25.89	0.916	0.118	24.02	0.795	0.206
+ WMVS + FSS + HNR + WFL	**29.55**	**0.948**	**0.075**	**26.12**	**0.918**	**0.113**	**24.28**	**0.794**	**0.212**

Table 2: The quantitative results of the Ablation studies in terms of PSNR↑, SSIM↑, and LPIPS↓ metrics. The experiments are carried out on the DTU dataset, the NeRF Synthetic dataset, and the LLFF dataset. Please refer to Section 4.2 for the details of the design of our ablation studies

tion mode that has three fixed source views. Additionally, the pretrained checkpoints provided by GeoNeRF [13] are based on the pretrained weights from CasMVSNet [11], while our model is trained end-to-end. We thus train a GeoNeRF from scratch using their scripts and evaluate both the end-to-end version and the complete version. The results show that our model can outperform GeoNeRF even if it is trained based on the pretrained weights from CasMVS-Net.

In addition to quantitative comparisons, we also provide qualitative comparisons of our model with existing methods on different datasets in Fig. 5. Our WaveNeRF model produces images that better preserve the details of the scene and contain fewer artifacts.

4.2. Ablation Study

We conducted several ablation studies to validate the effectiveness of our designed modules on three evaluation datasets (DTU dataset [12], NeRF synthetic dataset [25], and LLFF dataset [24]). The evaluation of WaveNeRF includes the following variants: 1) the baseline model without any of our novel modules, 2) the baseline model + our WMVS module, 3) the baseline model + our WMVS module + our FSS sampling strategy, 4) the baseline model + all three of our proposed modules but without the WFL loss $\mathcal{L}_{f_w}$, and 5) The complete version of our WaveNeRF model. Table 2 shows the quantitative results of the ablation study, indicating the effectiveness of our proposed modules.

4.3. Evaluation of High-frequency Components

To assess how well our model renders high-frequency features in images, we rely on a metric called HFIV [31]. This metric measures the proportion of high-frequency components (HF_c) in an image, which is indicative of its high-frequency quality. To facilitate comparisons across our test data, we modify HFIV to calculate the difference between the HF_c of the ground truth and the HF_c of the rendered results. The smaller this difference, the better the performance of the model.

We compare HFIV of our WaveNeRF, GeoNeRF [13], and MVSNeRF [3] on the same three datasets as the previous experiments. The quantitative (see Table 3) results indicate that our WaveNeRF model can reconstruct better high-frequency details than the previous generalizable NeRFs.

Method	DTU	NeRF Synthetic	LLFF
MVSNeRF [3]	0.129	0.1910	0.241
GeoNeRF [13]	0.103	0.0455	0.128
WaveNeRF	**0.0521**	**0.0362**	**0.115**

Table 3: Quantitative comparisons of rendered high-frequency components among MVSNeRF [3], GeoNeRF [13], and our WaveNeRF. The metric used here is HFIV↓ which can measure the difference between two images on the high-frequency bands.

4.4. Evaluation of the Frequency-Guided Sampling

In the classic NeRF [25], the fine-sampled N_f points are selected based on a normalized weight distribution obtained by estimating the volume density of coarse-sampled points, which allows to sample dense points around the region with visible content. To simplify this coarse-to-fine process, GeoNeRF [13] randomly samples fine points around the valid coarse points to calculate the color and the volume density of all points simultaneously. However, this randomly-sampling strategy cannot ensure that the fine-sampled points exist around the surfaces of the objects, which motivates our frequency-guided sampling strategy. In this section, we evaluate the sampling quality of our frequency-guided strategy by comparing the distribution of the volume density of the sampled points from WaveneRF, GeoNeRF [13], and MVSNeRF [3]. As shown in Fig. 6, we can observe that our WaveNeRF model can sample more points with high volume density values, which means our FSS strategy effectively guides the model to have more samples around the surfaces of the objects.

5. Limitation

Our model is designed to be trained and evaluated using three-shot source views (V=3) on a single GPU with
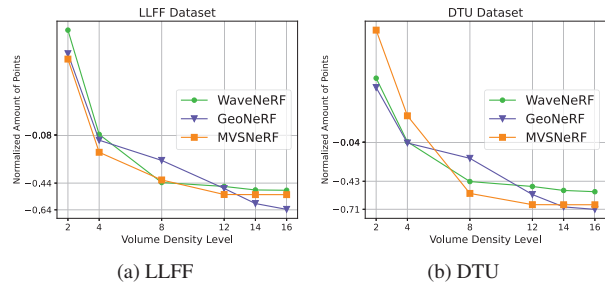

(a) LLFF   (b) DTU

Figure 6: Comparisons of the distribution of volume density of our WaveNeRF, GeoNeRF [13], and MVSNeRF [3] on LLFF dataset [24] and DTU dataset [12]. The horizontal axis represents the level of volume density where large levels indicate a higher possibility of being around objects. The vertical axis means the number of sampled points whose values are standardized to a standard normal distribution for better visualization.

16 GB memory. For cases with more input views, larger memory is required or the batch size should be decreased to accommodate the additional inputs. It is worth noting that our WMVS module is based on the MVS technique, which means that artifacts may appear if stereo reconstruction fails. The artifacts can manifest as noise in textureless regions or as view-dependent noisy floating-point clusters.

6. Conclusion

In this paper, we present a new generalizable NeRF model that is capable of generating high-quality novel view images under the few-shot setting, without requiring per-scene optimization. Our proposed model constructs MVS volumes and NeRF in the wavelet frequency domain where the explicit frequency information can be incorporated to boost the rendering quality. Additionally, we utilize frequency features to guide the sampling in NeRF, yielding densely sampled points around objects. We demonstrate that our model outperforms existing models on three datasets: the DTU dataset [12], the NeRF synthetic dataset [25], and the LLFF real forward-facing dataset [24], each with fixed-three input source views.

7. Acknowledgements

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.

[2] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.

[3] 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*, pages 14124–14133, 2021.

[4] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Feng Ying, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild, 2021.

[5] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

[6] Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *Proceedings of International Conference on Computer Vision (ICCV)*, volume 2, 1999.

[7] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[8] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[9] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021.

[10] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021.

[11] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[12] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

[13] 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*, pages 18365–18375, 2022.

[14] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

[15] Shuja Khalid and Frank Rudzicz. wildnerf: Complete view synthesis of in-the-wild dynamic scenes captured using sparse monocular data. *arXiv preprint arXiv:2209.10399*, 2022.

[16] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002.

[17] Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021.

[18] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[19] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8338–8348, 2023.

[20] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. 3d open-vocabulary segmentation with foundation models. *arXiv preprint arXiv:2305.14093*, 2023.

[21] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.

[22] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[23] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.

[24] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[25] 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.

[26] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[27] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for

unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016.

[28] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

[29] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006.

[30] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021.

[31] Li Chien Tan, Haniza Yazid, and Yen Fook Chong. Image quality assessment (iqa) using high-frequency and image variance (hfiv) for colour image. In *Journal of Physics: Conference Series*, volume 1372, page 012034. IOP Publishing, 2019.

[32] 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*, pages 4690–4699, 2021.

[33] 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.

[34] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022.

[35] 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)*, pages 767–783, 2018.

[36] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019.

[37] 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*, pages 4578–4587, 2021.

[38] Jiahui Zhang, Fangneng Zhan, Rongliang Wu, Yingchen Yu, Wenqing Zhang, Bai Song, Xiaoqin Zhang, and Shijian Lu. Vmrf: View matching neural radiance fields. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6579–6587, 2022.

[39] 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*, pages 586–595, 2018.