Multiplicative Fourier Level of Detail

Yishun Dou[2] Zhong Zheng[2] Qiaoqiao Jin[1] Bingbing Ni[1,2*]
[1]Shanghai Jiao Tong University, Shanghai 200240, China [2] Huawei
yishun.dou@gmail.com nibingbing@sjtu.edu.cn

Abstract

We develop a simple yet surprisingly effective implicit representing scheme called Multiplicative Fourier Level of Detail (MFLOD) *motivated by the recent success of multiplicative filter network. Built on multi-resolution feature grid/volume (*e.g.*, the sparse voxel octree), each level's feature is first modulated by a sinusoidal function and then element-wisely multiplied by a linear transformation of previous layer's representation in a layer-to-layer recursive manner, yielding the scale-aggregated encodings for a subsequent simple linear forward to get final output. In contrast to previous hybrid representations relying on interleaved multilevel fusion and nonlinear activation-based decoding, MFLOD could be elegantly characterized as a linear combination of sine basis functions with varying amplitude, frequency, and phase upon the learned multilevel features, thus offering great feasibility in Fourier analysis. Comprehensive experimental results on implicit neural representation learning tasks including image fitting, 3D shape representation, and neural radiance fields well demonstrate the superior quality and generalizability achieved by the proposed MFLOD scheme.*

1. Introduction

Classical geometric modeling techniques in computer graphics represent signals by storing discrete samples in array- or grid-based formats. It is nontrivial to adapt them to learning-based framework due to the lack of differentiability. Recently, neural implicit functions have emerged as an attractive alternative, which parameterize the continuous mapping between low dimensional coordinates and image/object domain signals using neural network, for example, as the representation of 3D shapes [5, 23, 29, 42] and radiance fields [16, 25].

Prior works commonly use a large multi-layer perceptron (MLP) to parameterize the learning function. To circumvent the well-known low frequency spectral bias of neural networks [31], frequency encoding [15, 25, 33, 34, 40] is usually adopted to map input coordinates to a higher dimensional space. A useful property of these pure MLP methods is that the entire function can be theoretically characterized as a linear combination of Fourier bases [6, 49], facilitating the design of neural representation from the spectral point of view, such as explicitly manipulating the bandwidth [15], representing overall signal and details separately [46], balancing representation generalization and spectrum coverage [40]. However, solely relying on the Fourier bases to express local high-frequency details is inefficient since these bases normally have (infinite) global support [49], resulting in the requirement of employing over-sized MLP to accommodate a large set of basis coefficients.

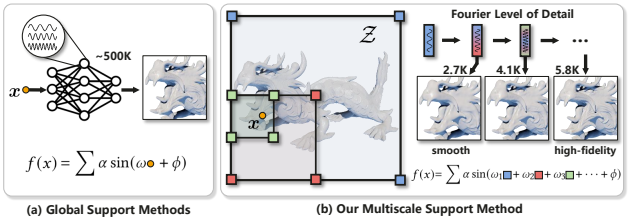

Figure 1. **Implicit Neural Representations from Spectral Perspective.** (a) Most implicit neural representations can be categorized into global support-based method. A large network is required to enrich the basis coefficients for representing high-frequency details. (b) Our method can also be well-characterized like the global method, enabling explicit bandwidth control for each LOD. This allows for representing high-fidelity details at fine levels and smooth overall shape at coarse levels.

Most recently, hybrid representations [27, 38, 39] emerge for their efficiency and high-fidelity. They employ a multilevel feature grid/volume to capture local details, and thus allow the use of a much smaller MLP as decoder. Yet, the interleaved multilevel fusion and nonlinear activation-based decoding make the entire function hard to characterize and not amenable to Fourier analysis like pure MLP methods. Specifically, little is known of how multilevel features are combined to get the final output.

We address the above limitations with our proposed implicit feature representation framework named *Multiplica-*

*Corresponding author: Bingbing Ni.

tive Fourier Level of Detail (MFLOD). Within a multi-resolution feature grid framework, MFLOD inherits the efficiency merits of hybrid methods, and in the meantime, as multilevel local features are modulated with a rich set of frequency basis functions, the resulting representation is feasible in Fourier analysis. More concretely, each feature point within a multi-resolution feature grid/volume (*e.g.*, the sparse voxel octree) is interpolated according to its spatial distances to the griding points, modulated by a sinusoidal filter, and then element-wisely multiplied by a linear transformation of previous layer's representation in a layer-to-layer recursive manner. After that, a simple linear forward is sufficient to decode these multilevel encodings to the final implicit function value. To enable meaningful levels of detail, we explicitly manipulate the spectral bandwidth of each level. This allows for representing high-fidelity details at fine levels and smooth overall shape at coarse levels, as shown in Fig. 1.

In addition to introducing MFLOD, we conduct in-depth theoretical study from the spectral and neural tangent kernel (NTK) [10] approximating perspectives, showing that the proposed method has better spectrum coverage and generalization. Comprehensive experimental results on implicit neural representation learning tasks including image fitting, 3D shape representation, and neural radiance fields well demonstrate the superior quality and generalizability achieved by the proposed MFLOD scheme.

2. Related Work

Our work is most related to previous research on implicit neural representation, random Fourier features, and level of detail.

Implicit Neural Representation. Representing signals as a continuous function parameterized by neural network is gaining popularity. The networks can be optimized as either signed distance functions (SDFs) [8, 11, 14, 24, 29, 42, 50] or occupancy functions [5, 23]. Using differentiable rendering [28, 41], it can also be trained using multiview 2D images, showing promising results in 3D shape reconstruction [12,28,44,45] and novel view synthesis [3,7,13,16,20, 25, 36, 47, 48]. Most of these approaches use MLP with ReLU activation. To learn the high-frequency variation, motivated by the success of Fourier transform in machine learning, some approaches have suggested integrating sinusoidal mapping into the networks [6,15,25,34,40]. Surprisingly, both of these models are under a unified formulation and thus have similar expressive power.

Feature grid/volume [17, 30, 35] is another effective choice to represent high-frequency local details, which discretizes the spatial space into a multi-resolution regular grid and stores local features in grid points, managed in memory as octree [39] or hashtable [27]. Given a query point, grid sampling is conducted at each scale. The interpolated features from different scales are fused together and then decoded by a small MLP with nonlinear activation. Compared with pure MLP methods, this hybrid representation seems more difficult to characterize. In this work, we step further in theoretical interpretability based on local features' modulation with Fourier basis functions.

Random Fourier Features (RFF). A seminal work by Rahimi & Recht [32] shows that projecting the inputs into random Fourier bases vastly improves the expressiveness of models. Many subsequent machine learning algorithms apply RFF to improve the performance in many domain areas [1, 9, 37, 43]. Specifically, RFF in deep implicit functions [6, 25, 40] acts as an encoding to improve the high-frequency representing capability. Instead of applying RFF to raw coordinate inputs, we first map the coordinates to multilevel learnable embeddings which are then transformed into Fourier space, enabling explicit bandwidth control for each level.

Level of Detail (LOD). Level of Detail [18] in computer graphics is used to mitigate flickering and accelerate rendering by reducing model complexity. The creation of 3D shapes LOD usually depends on mesh decimation, which has difficulty in blending between LODs, while SDF methods such as NGLOD [39] can reduce blending flickering. Our MFLOD represents 3D shapes by SDF, and thus inherits this property. Also, the bandlimited behavior of each LOD leads to smoother results at coarse levels.

3. Prerequisite: Implicit Neural Representation from Frequency Perspective

The goal of an implicit neural representation is to encode a continuous target signal using a neural network $f : \mathbb{R}^n \rightarrow \mathbb{R}^c$, by representing the mapping between input coordinates $\boldsymbol{x} \in \mathbb{R}^n$, *e.g.*, positions, and signal values $\bar{\boldsymbol{y}} \in \mathbb{R}^c$, *e.g.*, signed distances.

Classical neural network architectures are known for their strong spectral bias towards lower frequencies [31], which prevents them from being used in implicit representation tasks. A series of recent studies circumvent the spectral bias of neural network by mapping the coordinate to a high-dimensional Fourier space. Main solutions such as Position Encoding [25], Fourier Feature Network (FFN) [40], and SIREN [34] can be decomposed into a Fourier mapping $\gamma(\boldsymbol{x}) = \sin(\omega \boldsymbol{x} + \phi)$ followed by a multi-layer perceptron, where the filter $\omega \in \mathbb{R}^{d \times n}$ (d is the mapping dimension) can modulate the spectral bias of neural network, with larger scale ω biasing these networks towards higher frequencies.

Additionally, Yüce *et al.* [49] derive a unified formulation by observing that all analytic activation functions, *e.g.*, ReLU and sinusoidal, can be approximated using polynomials with a naïve Taylor expansion [22]. Thus, the entire implicit function can be expressed as a linear combination

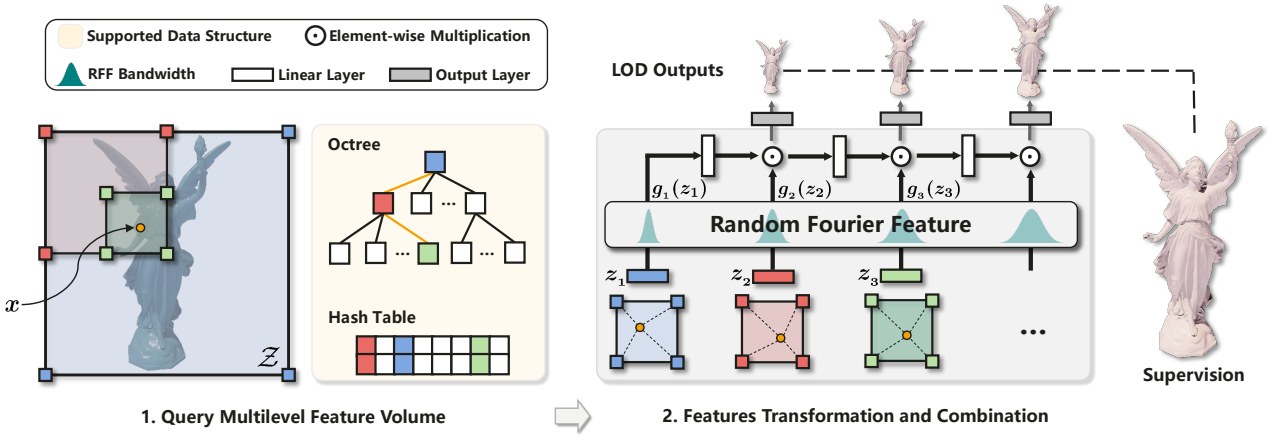

Figure 2. **Illustration of the MFLOD in 2D.** Given a query point $\boldsymbol{x}$, we traverse the multilevel feature-volume to find all voxels containing $\boldsymbol{x}$. Per-voxel vector $\boldsymbol{z}_\ell$ is obtained by interpolating the corner features of the voxel at $\boldsymbol{x}$, which is then transformed to Fourier space by applying a sinusoidal filter with explicit controlled bandwidth. Then, the linear transformation (no nonlinear activation) and multiplication are applied in a layer-to-layer recursive manner. We introduce simple linear forward to each intermediate layer to obtain LOD outputs.

of certain integer harmonics of the initial Fourier mapping $\gamma(\boldsymbol{x})$:

$$\boldsymbol{y} = \sum_{j=0}^{N_{\text{sine}}} \bar{\alpha}^j \sin(\bar{\omega}^j \boldsymbol{x} + \bar{\phi}^j), \tag{1}$$

where the support frequencies $\bar{\omega}^j$ are *completely* determined by the $\gamma(\boldsymbol{x})$, and the support number N_{sine} grows exponentially with the network depth.

Limitation. Although the improvement from the Fourier mapping is remarkable, we still need to tune the frequencies of the initial mapping $\gamma(\boldsymbol{x})$ to balance the recovery of high-frequency details and the suppression of aliasing artifact caused by the initial Fourier mapping with high-frequency [49]. We hold that the expression limitation stems from the fact that the Fourier bases have global support, which results in high-frequency components with large weight in Eq. (1) when fitting the spikes in target signals, thereby increasing the chances of yielding aliased reconstruction in other spatial locations due to the periodicity of Fourier bases.

Another line of solutions turns to multilevel feature grid for detailed representation. However, these methods rely on interleaved multilevel fusion and nonlinear activation-based decoding, which make the entire implicit function more difficult to characterize than that based on a pure MLP.

4. Methodology

Motivated by the above limitations, we aim to design a well-characterized representation which can reconstruct high-fidelity details and scale to different LOD, based on the idea of multilevel Fourier modulation at the learnable local feature space (*i.e.*, in contrast to coordinate values). Figure 2 shows a visual overview of our method. We present our method in Sec. 4.1 and then provide a frequency spectrum analysis in Sec. 4.2, followed by implementation details in Sec. 4.3.

4.1. Multiplicative Fourier Level of Detail

Feature Volume. MFLOD builds on a multilevel feature grid/volume [27, 38, 39], which contains a collection of learnable features that are organized in tree-structured regular grid, such as the sparse voxel octree.

We denote the tree-structured feature-volume as $\mathcal{Z}$. Each voxel V in $\mathcal{Z}$ holds an m dimensional learnable feature vector $\boldsymbol{z}_V^{(j)} \in \mathcal{Z}$ at each of its eight corners (indexed by j), which are shared if neighbor voxels exist. Each level $\ell \in \mathbb{N}$ of the feature-volume defines a LOD for the geometry. As the tree depth L in the multi-resolution hierarchy increases, the surface is represented with finer discretization, allowing reconstruction quality to scale with memory usage.

Given a query point $\boldsymbol{x} \in \mathbb{R}^3$ and desired LOD level L, we traverse the tree up to level L to find all voxels $V_{1:L} = \{V_1, \ldots, V_L\}$ containing $\boldsymbol{x}$. For each level $\ell \in \{1, \ldots, L\}$, we compute per-voxel shape vector $\boldsymbol{z}_\ell = \psi(\boldsymbol{x}; \ell, \mathcal{Z})$ by trilinearly interpolating the corner features of the voxels at $\boldsymbol{x}$.

Transformation and Combination. After traversing the tree and extracting multi-resolution features $\boldsymbol{z}_{1:L}$, typical methods would aggregate these features using summation [39] or concatenation [27], where the multilevel fusion is then decoded by an MLP with non-linear activation.

In order to promote the amenability of neural LOD representation to Fourier analysis, in this work, we replace the previous paradigm of ***aggregate-decode*** with ***transform-combine***. Namely, multilevel features $\boldsymbol{z}_{1:L}$ are *transformed* into a Fourier feature space and then *linearly combined*

from coarse to fine, allowing their analysis much more feasible than that for *aggregate-decode* paradigm.

First, we apply layer normalization [2] right after the grid interpolation, followed by a sinusoidal transform with learnable filters, both of which are performed independently at each level:

$$\begin{aligned} \hat{\boldsymbol{z}}_\ell &= \text{LayerNorm}(\boldsymbol{z}_\ell), \\ g_\ell(\boldsymbol{z}_\ell) &= \sin(\omega_\ell \hat{\boldsymbol{z}}_\ell + \phi_\ell), \end{aligned} \tag{2}$$

with filter parameters $\theta_\ell^{\text{filter}} = \{\omega_\ell \in \mathbb{R}^{d\times m}, \phi_\ell \in \mathbb{R}^d\}$, where m and d are grid feature dimension and mapping dimension, respectively. LayerNorm is introduced before sinusoidal filter for better spectral manipulation. We term such a representation the Fourier LOD, as the above sinusoidal filter corresponds naturally to Random Fourier Features [32] representation.

After that, we employ multiplicative network [6] to achieve linear combination of Fourier LOD, where the network is composed of elementwise multiplication and fully connected layer without non-linear activation. The network's configuration is therefore determined, with depth equal to LOD maximum level L and hidden dimension matching the mapping dimension d. We refer to the intermediate activation as $\boldsymbol{t}_\ell \in \mathbb{R}^d$, and we allow intermediate outputs of the network $\boldsymbol{y}_\ell \in \mathbb{R}^c$ at the ℓth layer, defined as follows (see also Fig. 2):

$$\begin{aligned} \boldsymbol{t}_1 &= g_1(\boldsymbol{z}_1), \\ \boldsymbol{t}_\ell &= g_\ell(\boldsymbol{z}_\ell) \circ (\boldsymbol{W}_\ell \boldsymbol{t}_{\ell-1} + \boldsymbol{b}_\ell), \quad 1 < \ell \le L \\ \boldsymbol{y}_\ell &= \boldsymbol{W}_\ell^{out} \boldsymbol{t}_\ell + \boldsymbol{b}_\ell^{out}, \end{aligned} \tag{3}$$

where $\circ$ denotes elementwise multiplication. The parameters of the network are $\theta_\ell^{\text{net}} = \{\boldsymbol{W}_\ell \in \mathbb{R}^{d\times d}, \boldsymbol{b}_\ell \in \mathbb{R}^d, \boldsymbol{W}_\ell^{\text{out}} \in \mathbb{R}^{c\times d}, \boldsymbol{b}_\ell^{\text{out}} \in \mathbb{R}^c\}$.

A useful and compelling property of this formulation is that the network output can be expressed equivalently as a sum of sines, each of which consists of a linear combination of linear modulated grid features $\boldsymbol{z}_{1:L}$. Ultimately, the output of MFLOD at level ℓ can be characterized as (see supplemental §1.1):

$$\boldsymbol{y}_\ell = \sum_{j=0}^{N_{\text{sine}}^\ell - 1} \bar{\alpha}^j \sin(\underline{\bar{\omega}_1^j \boldsymbol{z}_1 + \bar{\omega}_2^j \boldsymbol{z}_2 + \cdots + \bar{\omega}_\ell^j \boldsymbol{z}_\ell} + \bar{\phi}^j), \tag{4}$$

where the coefficients $\bar{\alpha}^j$, $\bar{\omega}^j$ and $\bar{\phi}^j$ are determined by the parameters of filters and network. Moreover, linear layer $\boldsymbol{W}_\ell \boldsymbol{t}_{\ell-1}$ increases the number of sine terms in $\boldsymbol{t}_{\ell-1}$ exponentially and elementwise multiplication $\circ$ results it to double, along with the additional terms contributed by bias $\boldsymbol{b}_\ell$, the number of terms in the sum at LOD level ℓ is given as $N_{\text{sine}}^\ell = \sum_{i=0}^{\ell} 2^i d^{i+1}$. The key element of the proof for Eq. (4) is the trigonometric identity:

$$\begin{aligned} &\sin(\tau_1 z_1 + \varphi_1) \circ \sin(\tau_2 z_2 + \varphi_2) \\ =&\frac{1}{2}[\sin(\underline{\tau_1 z_1 + \tau_2 z_2} + \varphi_1 + \varphi_2 - \frac{\pi}{2}) \\ &+ \sin(\underline{\tau_1 z_1 - \tau_2 z_2} + \varphi_1 - \varphi_2 + \frac{\pi}{2})]. \end{aligned} \tag{5}$$

Each additional multiplicative layer creates both a sum and difference combination of multilevel features in the sinusoidal terms. When the ℓth level Fourier features $g_\ell(\boldsymbol{z}_\ell)$ is fed into the network through multiplication, $\boldsymbol{z}_\ell$ would be linearly combined with previous coarser features $\boldsymbol{z}_{1:\ell-1}$ within all sinusoidal terms.

Note that the sinusoidal transform is conducted upon learnable feature instead of raw coordinate values [6,15,33]. In other words, the coordinate $\boldsymbol{x}$ is enriched with multilevel features, where the fine levels endow the entire function the capability to reconstruct local high-frequency components without introducing aliasing artifact [49] that stems from the deficient global support.

4.2. Spectrum Analysis

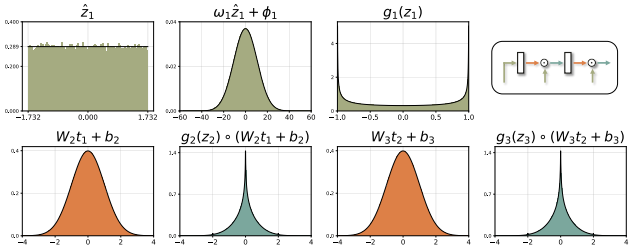

Figure 3. **Distribution of activations at initialization.** The proposed initialization scheme maintains a standard normal distribution after sinusoidal transform and each linear layer, and activations closely match the analytical derivations (black lines).

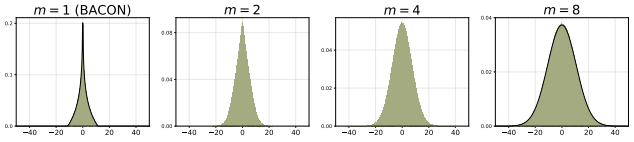

Figure 4. **Distribution of $\omega_\ell \hat{\boldsymbol{z}}_\ell + \phi_\ell$ at initialization.** When grid feature dimension m=1, the distribution is similar with that shown in BACON [15]. As m increases, it approximates a Gaussian distribution that has a wider spectrum coverage in theory.

Bandlimited Initialization Scheme. We derive a principled initialization scheme to provide a reasonable scaling of activations throughout the multiplicative layer, sinusoidal filter, and feature-volume. Activations of intermediate layers are illustrated in Fig. 3. More importantly, the behavior of incorporating finer levels gradually gives us the chance to explicitly manipulate the bandwidth of each level for smoother reconstruction at coarse levels (see Fig. 7).

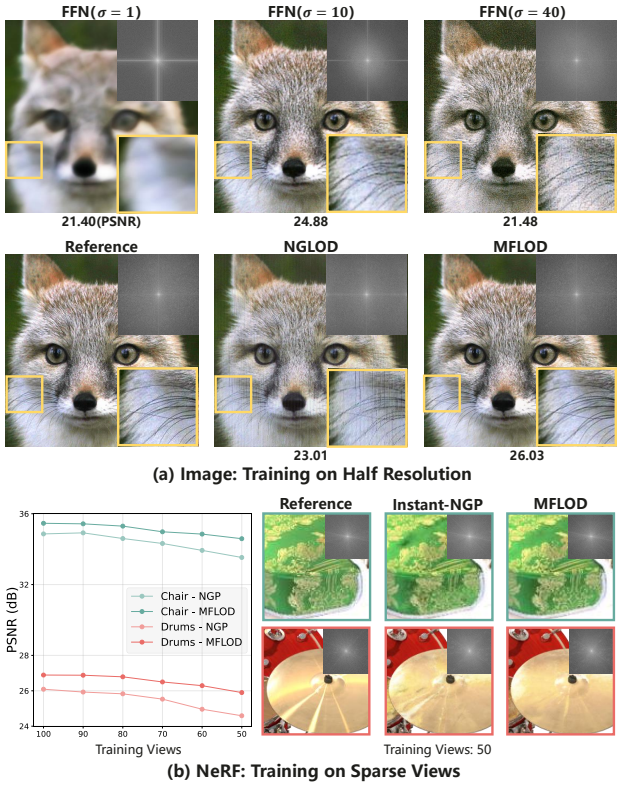

Figure 5. **Comparison of Generalization Ability.** (a) Models are trained on half resolution of a 512×512 image and required for reconstruction at full resolution. FFN [40] with high-frequency initial mapping and NGLOD show noisy interpolation. While MFLOD achieves smoother interpolation and better spectrum coverage. (b) We gradually reduce the NeRF training views to demonstrate the interpolating capability in novel view synthesis. The PSNR curves show that our approach is more insensitive to the reduction of training views.

We initialize the feature-volume $\mathcal{Z}$ entries using the uniform distribution $\mathcal{U}(-10^{-4}, 10^{-4})$ to provide a small amount of randomness. After normalization, $\hat{\boldsymbol{z}}$ distributes as $\mathcal{U}(-\sqrt{3}, \sqrt{3})$ at initialization. Since sinusoidal transform is performed independently at each level and corresponding to random Fourier features, each LOD explicitly controls a certain bandwidth. We initialize the sinusoidal filter parameter for each level separately: $\omega_\ell \sim \mathcal{U}(-B_\ell, B_\ell)$ and $\phi_\ell \sim \mathcal{U}(-\pi, \pi)$, where B_ℓ is a hyperparameter controlling the bandwidth. Note that the distribution of $\omega_\ell \hat{\boldsymbol{z}}_\ell + \phi_\ell$ depends on the grid feature dimension m, as shown in Fig. 4. When there is only one dimension ($\omega_\ell \in \mathbb{R}^{d \times 1}$), the distribution degenerates to (see supplemental §1.2):

$$\omega_\ell \hat{\boldsymbol{z}}_\ell + \phi_\ell \;\sim\; \frac{1}{2\sqrt{3}B_\ell} \log\left(\frac{B_\ell}{\min(|x/\sqrt{3}|, B_\ell)}\right), \qquad (6)$$

which is similar to that shown in BACON [15]. When $m > 1$, the probability density is the sum of independent random variables sampled from the above distribution. As m increases, it will approach the Gaussian distribution according to the central limit theorem. After applying sinusoidal function, $g_\ell(\boldsymbol{z}_\ell)$ is approximately arcsine distributed with variance 0.5. Now, we set $\mathbf{W}_\ell \sim \mathcal{U}(-\sqrt{6/d}, \sqrt{6/d})$ following BACON. Then we have that $\mathbf{W}_\ell t_{\ell-1} + \mathbf{b}_\ell$ converges to the standard normal distribution with increasing d. Finally, the elementwise multiplication is the product of arcsine distributed and standard normal random variables which again has a variance of 0.5. Applying the next linear layer results in another standard normal distribution, as well as after all subsequent linear layers.

Spectrum Coverage and Inductive Bias. In the context of implicit neural representation, the inductive bias is mainly related to the spectrum coverage [49]. For the pure MLP methods, the set of frequencies that define the initial mapping $\gamma(\boldsymbol{x})$ completely determines the frequency support of the entire function in Eq. (1), it is consequently fundamental to guarantee these supports cover the spectrum of target signal. However, it has been proven that the low frequency initial mapping $\gamma(\boldsymbol{x})$ cannot cover the spectrum and thus exhibits underfit, while the high-frequency often leads to overfitting and noisy interpolation [40] (see Fig. 5 (a)). Specifically, for instance, matching the spikes in target signals introduces high-frequency components with large weight in Eq. (1), thereby increasing the chances of yielding aliased reconstruction in other spatial locations due to the periodicity of Fourier bases [49].

As for the feature-volume methods, there is little theoretical understanding from spectrum perspective because those existing are not amenable to Fourier analysis. Intuitively, an inherent flaw of these methods is that the fine levels trained on local signals tend to overfit. Experimentally, we find that underfit is rare for the models with fine-grained feature-volume discretization, but overfit often occurs, as illustrated in Fig. 5. Feature-volume methods with *aggregate-decode* mechanism (*e.g.* NGLOD [39] and NGP [27]) show severe artifacts in interpolating new pixels and novel views, exposing that they may fail to achieve harmonious collaboration across multilevel. Instead, our proposed paradigm of *transform-combine* gives a more effective way to coordinate the features across multilevel and, as a consequence, shows more convincing interpolation results.

NTK for Generalization Analysis. To further understand why MFLOD has better generalization, we use neural tangent kernel (NTK) [10] to describe the implicit networks. In the context of approximating deep network with kernel regression, implicit neural representations can be interpreted as signal dictionaries whose atoms are the eigenfunctions of their NTK at initialization [49]. Under this view, the study of the inductive bias (trend to underfit or overfit) is equivalent to the study of the capabilities of its NTK dictionary. Eigenfunctions of the empirical NTK are demonstrated in Fig. 6. As σ values of FFNs increase, the eigen-

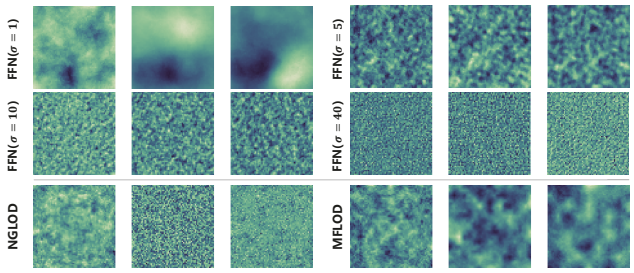

Figure 6. **First Eigenfunctions of the Empirical NTK [10]** of different representations at initialization. Eigenfunctions of FFN become more intricate as σ increases, corresponding to more overfitting tendencies (Fig. 5). NGLOD [39] and MFLOD share the same dense octree feature-volume. Evaluated at initialized $\boldsymbol{z}_{1:L}$, MFLOD has less tendencies to overfit than NGLOD.

functions demonstrate more intricate patterns, meaning that FFNs gradually tend to overfit, which is consistent with experimental phenomenon in Fig. 5 (a).

Analyzing hybrid neural representation is slightly different from pure MLP methods. To the best of our knowledge, we are the first to explain the hybrid neural representation using kernel regression theory. Specifically, we use a large grid feature dimension to separate raw coordinates in feature space $\boldsymbol{z}_{1:L}$ at initialization as much as possible. Then the empirical NTK for two coordinates $\boldsymbol{x}^1, \boldsymbol{x}^2$ corresponds to the matrix product between the Jacobian of the network evaluated at $\boldsymbol{z}^1_{1:L}$ and $\boldsymbol{z}^2_{1:L}$ (see supplemental §2). Following the conclusions of [49], the eigenfunctions indicate that MFLOD may more easily learn representation than NGLOD [39] because the eigenfunctions of MFLOD do not exhibits highly high-frequency patterns that are non-compatible with natural signals.

4.3. Implementation

Multiscale Supervision. Recall that MFLOD gradually incorporates finer level feature as the network depth ℓ increases. Thus, it is straightforward to train MFLOD to fit a signal at multiple scales simultaneously, by introducing output layers at intermediate stages throughout the network and supervising these outputs. Compared with the *aggregate-decode* methods (*e.g.* NGLOD [39]) that require training individual MLP decoder for each of the L levels, as a side benefit, our method is more succinct and natural in generating LOD outputs. Moreover, because the outputs are bandlimited, MFLOD can be trained in a semi-supervised fashion where the supervisory signal does not need to match the desired bandwidth of the output.

Model Architecture. MFLOD is designed to be compatible with a variety of feature-volume data structures, such as dense/sparse octree [39] and multi-resolution hashtable [27]. In all experiments, we set grid feature dimensions m=8, and Fourier space dimensions d=32. Since the computational cost is related to the level of feature-volume, we restrict the max level L=6. As a result, the number of parameters introduced by *transform-combine* is 7.0K at most, in addition to a negligible simple linear output layer for each level, the total number of parameters has the same order of magnitude as the small MLP that is used in NGLOD [39] and NGP [27]. In order to adapt the hashtable with NGP's default setting (*i.e.* L=16, m=2) to MFLOD, we simply concatenate every four levels together, resulting in a proxy feature-volume with L=4, m=8.

Non-spatial Input Dimensions. The feature-volume $\mathcal{Z}$ has a relatively lower dimension. All our experiments operate either in 2D or 3D. However, it is frequently useful to input auxiliary dimensions to the neural network, such as the view direction in neural radiance fields (NeRF) [25]. In such case, we can enlarge the per-level's output dimension and view MFLOD as an *encoding*. We show more details in NeRF experiment.

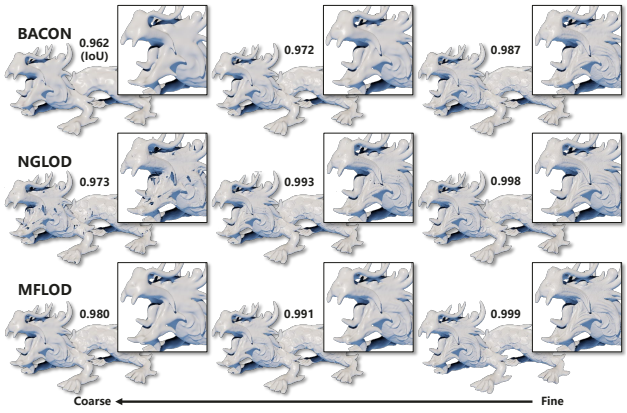

Figure 7. **Comparison of Levels of Detail.** BACON [15] struggles in recovering high-frequency details even at the finest level, while NGLOD [39] fails to generate smooth overall shape at the coarse level. In contrast, MFLOD can reconstruct both high-fidelity details and smooth overall shape, suggesting a large benefit to our novel paradigm and bandwidth control for each LOD.

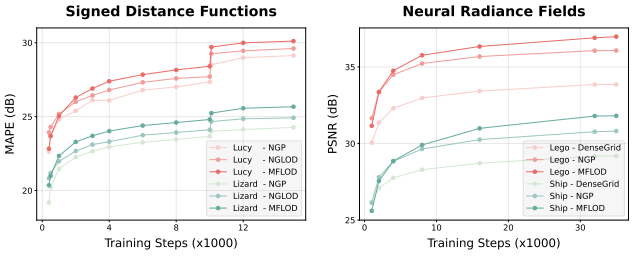

Figure 8. **Comparison of Convergence.** Test errors over training step on SDFs and NeRFs in terms of mean absolute percentage error (MAPE) and peak signal to noise ratio (PSNR), respectively. MFLOD exhibits comparable or faster convergence rates than other hybrid approaches.

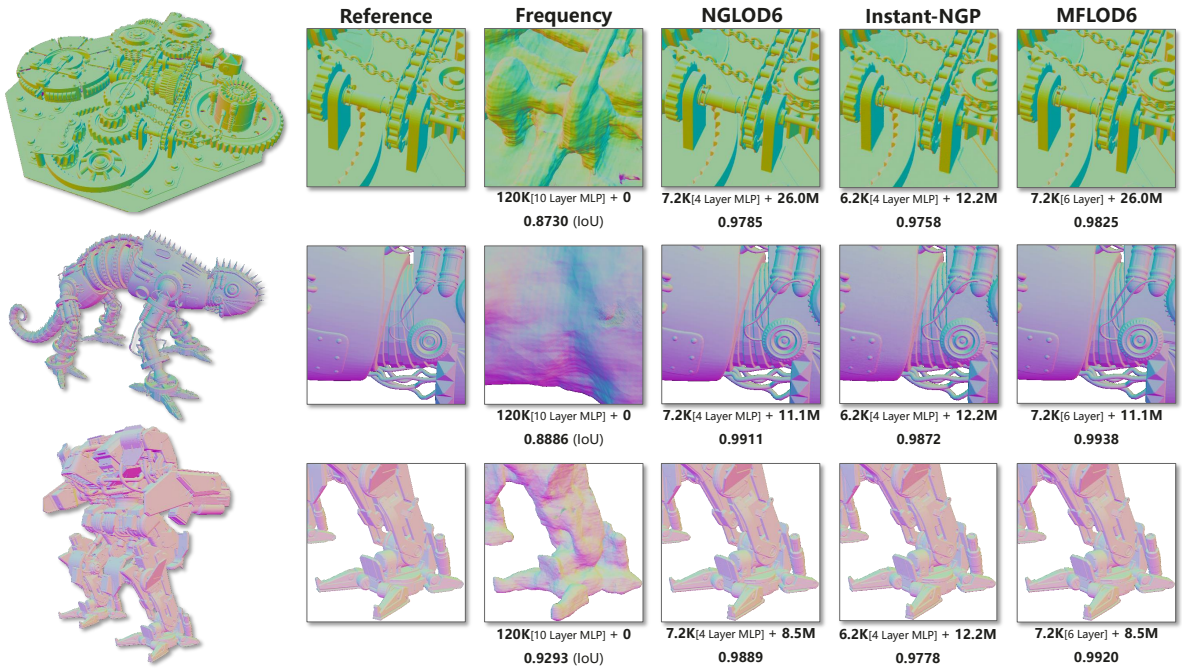

Figure 9. **Qualitative Comparison on Mesh Reconstruction.** All methods are trained for 15000 steps. We render surface normals to highlight geometric details. MFLOD achieves the highest intersection over union (IoU). Although both NGLOD [39] and our MFLOD show high-fidelity visual quality (almost saturated), MFLOD exhibits much smoother visual effect at the coarse level (Fig. 7).

5. Experiments

To demonstrate/highlight the versatility and high quality of MFLOD, we extensively experiment: 1) 3D shape representation to verify its ability in high-fidelity detail reconstruction at fine levels and overall smoothness achieved at coarse levels, 2) image fitting and 3) novel view synthesis to show its generalizing ability.

We implement MFLOD in CUDA and integrate it with *instant-ngp* [26, 27] framework for meaningful comparison. All our experiments are conducted on an RTX 3090 GPU.

5.1. 3D Shape Representation

To evaluate the quality of 3D shape representation, we fit our network on ShapeNet [4], Stanford 3D scanning repository[1], and selected models from TurboSquid[2]. We compare MFLOD to Frequency Encoding [25], SIREN [34], BACON [15], NGLOD [39], and Instant-NGP [27]. Since the ground-truth geometry is accessible, we use a sparse octree tailored to the reference shape as the feature-volume, and thus the difference with NGLOD is how to decode the multilevel features to final outputs.

Table 1 shows comparative results on ShapeNet and Stanford3D, in terms of commonly used Chamfer distance and intersection over union (IoU) over uniformly sampled points, where the Chamfer distances are evaluated on 10^5 uniformly sampled points on the surface of reconstruction. For ShapeNet, we sample 100, 50, and 50 shapes respectively from *chair*, *car*, and *airplane* categories. BACON needs to tune the per-shape's predefined frequency to achieve optimal performance, thus we only report the results on Standford3D using their released configuration. Our method exhibits similar numeric quality with NGLOD at finer level, but better results at coarser level. Figure 7 gives a visual comparison at different levels of detail. Compared with NGLOD, both BACON and our method can generate smoother surface at coarse levels, while BACON cannot achieve high fidelity even at the finest level. We also attempt to recover high-frequency details for BACON by increasing the predefined frequency (from 384) to 512 and 640, but find unexpected aliasing artifacts occurred, reflecting the limitation of global supports in representing high-frequency details.

To further demonstrate the effectiveness of MFLOD, we qualitatively compare the reconstruction quality on TurboSquid which contains much more intricate geometry details, as shown in Fig. 9.

5.2. Image Fitting

Learning the 2D coordinate to RGB color mapping has become a popular benchmark for testing a model's ability to represent high-frequency detail [19, 34, 40]. We train models on half resolution and evaluate them on full resolution to show the generalization on unseen pixels during training.

[1] http://graphics.stanford.edu/data/3Dscanrep/
[2] https://www.turbosquid.com/

		ShapeNet-200		Standford3D	
	# Inference Param.	IoU↑	Chamfer↓	IoU↑	Chamfer↓
Frequency	125K	89.7	0.038	0.958	0.0419
SIREN	125K	91.1	0.299	0.982	0.0057
BACON	531K	-	-	0.979	0.0054
Instant-NGP ($T = 2^{19}$)	6.2K	93.2	0.0094	0.995	0.0020
NGLOD / LOD 2	7.2K	89.1	0.0413	0.974	0.0069
MFLOD / LOD 2	**2.7K**	90.9	0.0231	0.984	0.0055
NGLOD / LOD 3	7.2K	92.5	0.0099	0.989	0.0043
MFLOD / LOD 3	4.1K	92.6	0.0106	0.990	0.0045
NGLOD / LOD 4	7.2K	93.8	0.0097	0.996	0.0026
MFLOD / LOD 4	4.4K	93.9	0.0072	0.995	0.0023
NGLOD / LOD 5	7.2K	93.9	0.0072	0.998	0.0022
MFLOD / LOD 5	5.8K	**94.3**	**0.0063**	**0.999**	**0.0021**

Table 1. **Per-shape Mesh Reconstruction.** MFLOD achieves the highest quality. For coarse level such as LOD2, it outperforms NGLOD significantly even with a smaller number of parameters. (Chamfer distance multiplied by 10^3).

	Mic	Ficus	Chair	Hotdog	Materials	Drums	Ship	Lego	avg.
NeRF (∼hours)	32.91	30.13	33.00	36.18	29.62	25.01	28.65	32.54	31.005
NSVF (∼hours)	34.27	31.23	33.19	37.14	32.68	25.18	27.93	32.29	31.739
BACON (∼hours)	28.45	23.75	30.73	31.94	24.30	24.18	25.67	30.42	27.430
MIPNeRF (∼hours)	38.04	33.19	37.14	39.31	32.56	27.02	33.08	35.74	34.510
Instant-NGP (5 min)	36.22	33.51	35.00	37.40	29.78	26.02	31.10	36.39	33.176
MFLOD-A (5 min)	37.33	33.48	35.42	37.51	31.40	26.42	31.80	36.44	33.727
MFLOD-B (∼7 min)	37.44	33.53	35.46	37.51	31.67	26.89	31.81	36.98	33.910

Table 2. **Novel View Synthesis on NeRF Synthetic Dataset**. MFLOD-A and -B represent training NeRF wo/w multilevel supervision. MFLOD-B is trained for the same iteration as -A, showing consistent improvement at the cost of about two more minutes training overhead.

For the image in Fig. 5, FFN [40] achieves a PSNR of 24.88 dB with a Fourier mapping scale σ=10. While increasing σ can improve the high-frequency expressions in theory, overfitting often occurs in practice and results in noisy interpolation. For hybrid methods, NGLOD [39] and MFLOD share the same dense 2D feature-grid (L=4, m=8) with the finest grid resolution equal to half of the training resolution. MFLOD achieves 26.03 dB PSNR while NGLOD exhibits overfitting. We attribute this to the MLP cannot achieve as good collaboration across multilevel features as ours.

5.3. Neural Radiance Fields

In NeRF setting, a volumetric shape is represented in terms of a spatial (3D) density function and a spatiodirectional (5D) emission function. To adapt our architecture to NeRF, we change the output $\boldsymbol{y}_L$ to 16 values, the first of which we treat as log-space density. The view direction is projected onto the first 16 coefficients of the spherical harmonics basis, which is a natural frequency encoding over unit vectors. It is then concatenated with $\boldsymbol{y}_L$, followed by a 1-hidden-layer color MLP with ReLU activation (64 neurons wide). Note that the ReLU can be effectively approximated using Chebyshev polynomial [21], thus the final functions of MFLOD on NeRF can still be characterized as linear combination of Fourier bases.

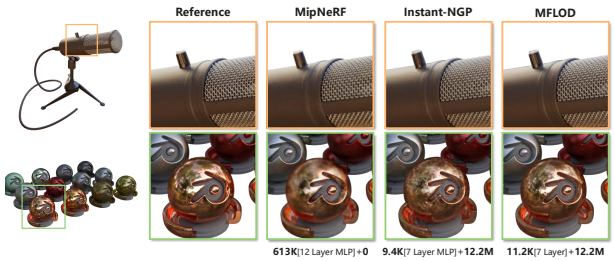

Figure 10. **Qualitative Comparison on Neural Radiance Fields**. MFLOD(-B) captures higher frequency details than global methods (such as MipNeRF) and the hybrid method baseline of NGP. The intermediate density output layer and color MLP of MFLOD are discarded after training, and the number of inference parameter is similar with NGP.

The underlying data structure used for NeRF experiments is multi-resolution hashtable with default setting in NGP (L=16, m=2), which is then adapted to a proxy (L=4, m=8) by simply concatenating every four levels together. We train the model by backpropagating through a differentiable ray marcher driven by 2D RGB images from known camera poses. Since MFLOD is designed for LOD, we also train NeRF with intermediate supervision. Table 2 shows the comparison with NeRF [25], NSVF [16], MIPNeRF [3], BACON [15], and NGP [27], in terms of peak signal to noise ratio (PSNR). When the training time is limited to 5 min, MFLOD outperforms NGP on most scenes, especially those contain high-frequency details, such as Materials. Even compared with offline methods that require training for hours, our PSNR is still competitive.

To further understand the effectiveness of MFLOD on NeRF, we compare convergence rate and qualitative results in Fig. 8 and Fig. 10.

6. Conclusion

In this work, we take steps towards making hybrid neural representations interpretable. Our approach can be elegantly characterized as a linear combination of Fourier basis function upon the learned multilevel features, enabling analysis and bandwidth manipulation at each level of detail. This allows for representing high-fidelity details at fine levels and smoother overall signal at coarse levels. Moreover, we conduct in-depth theoretical study from the spectral and NTK perspective, as well as the experiments in image interpolation and novel view synthesis, demonstrating that MFLOD has better generalizing ability than other methods.

7. Acknowledgement

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

[1] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International conference on machine learning*, pages 253–262. PMLR, 2017. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] 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*, pages 5855–5864, 2021. 2, 8

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 7

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1, 2

[6] Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J. Zico Kolter. Multiplicative filter networks. In *ICLR*. OpenReview.net, 2021. 1, 2, 4

[7] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2

[8] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3789–3799. PMLR, 2020. 2

[9] Zhen Hu, Ming Lin, and Changshui Zhang. Dependent online kernel learning with constant number of random fourier features. *IEEE transactions on neural networks and learning systems*, 26(10):2464–2476, 2015. 2

[10] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 2, 5, 6

[11] 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*, pages 6001–6010, 2020. 2

[12] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1251–1261, 2020. 2

[13] Animesh Karnewar, Tobias Ritschel, Oliver Wang, and Niloy Mitra. Relu fields: The little non-linearity that could. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[14] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4297, 2021. 2

[15] David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16252–16262, 2022. 1, 2, 4, 5, 6, 7, 8

[16] 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. 1, 2, 8

[17] Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, 2019. 2

[18] David Luebke, Martin Reddy, Jonathan D Cohen, Amitabh Varshney, Benjamin Watson, and Robert Huebner. *Level of detail for 3D graphics*. Morgan Kaufmann, 2003. 2

[19] Julien N. P. Martel, David B. Lindell, Connor Z. Lin, Eric R. Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: adaptive coordinate networks for neural scene representation. *ACM Trans. Graph.*, 40(4):58:1–58:13, 2021. 7

[20] 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. 2

[21] Christian HX Ali Mehmeti-Göpel, David Hartmann, and Michael Wand. Ringing relus: Harmonic distortion analysis of nonlinear feedforward networks. In *International Conference on Learning Representations*, 2020. 8

[22] Christian H. X. Ali Mehmeti-Göpel, David Hartmann, and Michael Wand. Ringing relus: Harmonic distortion analysis of nonlinear feedforward networks. In *ICLR*. OpenReview.net, 2021. 2

[23] 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*, pages 4460–4470, 2019. 1, 2

[24] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2019. 2

[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. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 6, 7, 8

[26] Thomas Müller. tiny-cuda-nn, 4 2021. 7

[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 1, 2, 3, 5, 6, 7, 8

[28] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[29] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1, 2

[30] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 2

[31] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 1, 2

[32] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 2, 4

[33] Shayan Shekarforoush, David B Lindell, David J Fleet, and Marcus A Brubaker. Residual multiplicative filter networks for multiscale reconstruction. *arXiv preprint arXiv:2206.00746*, 2022. 1, 4

[34] 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. 1, 2, 7

[35] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2

[36] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[37] Yitong Sun, Anna Gilbert, and Ambuj Tewari. But how does it work in theory? linear svm with random features. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[38] Towaki Takikawa, Alex Evans, Jonathan Tremblay, Thomas Müller, Morgan McGuire, Alec Jacobson, and Sanja Fidler. Variable bitrate neural fields. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 1, 3

[39] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 1, 2, 3, 5, 6, 7, 8

[40] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 1, 2, 5, 7, 8

[41] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 2

[42] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[43] Hui Xue, Zheng-Fan Wu, and Wei-Xiang Sun. Deep spectral kernel learning. In *IJCAI*, pages 4019–4025, 2019. 2

[44] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[45] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2

[46] Wang Yifan, Lukas Rahmann, and Olga Sorkine-Hornung. Geometry-consistent neural shape representation with implicit displacement fields. *arXiv preprint arXiv:2106.05187*, 2021. 1

[47] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[48] 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. 2

[49] Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19228–19238, 2022. 1, 2, 3, 4, 5, 6

[50] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1439, 2021. 2