# Representing Sounds as Neural Amplitude Fields: A Benchmark of Coordinate-MLPs and A Fourier Kolmogorov-Arnold Framework

**Linfei Li[1], Lin Zhang[1]*, Zhong Wang[2], Fengyi Zhang[3], Zelin Li[4], Ying Shen[1]**
[1]School of Computer Science and Technology, Tongji University
[2] Department of Automation, Shanghai Jiao Tong University
[3] School of Electrical Engineering and Computer Science, The University of Queensland
[4]McCormick School of Engineering, Northwestern University
cslinfeili@tongji.edu.cn, cslinzhang@tongji.edu.cn, cszhongwang@sjtu.edu.cn,
fengyi.zhang@uq.edu.au, zelinli2025@u.northwestern.edu, yingshen@tongji.edu.cn

## Abstract

Although Coordinate-MLP-based implicit neural representations have excelled in representing radiance fields, 3D shapes, and images, their application to audio signals remains underexplored. To fill this gap, we investigate existing implicit neural representations, from which we extract 3 types of positional encoding and 16 commonly used activation functions. Through combinatorial design, we establish the first benchmark for Coordinate-MLPs in audio signal representations. Our benchmark reveals that Coordinate-MLPs require complex hyperparameter tuning and frequency-dependent initialization, limiting their robustness. To address these issues, we propose Fourier-ASR, a novel framework based on the Fourier series theorem and the Kolmogorov-Arnold representation theorem. Fourier-ASR introduces Fourier Kolmogorov-Arnold Networks (Fourier-KAN), which leverage periodicity and strong nonlinearity to represent audio signals, eliminating the need for additional positional encoding. Furthermore, a Frequency-adaptive Learning Strategy (FaLS) is proposed to enhance the convergence of Fourier-KAN by capturing high-frequency components and preventing overfitting of low-frequency signals. Extensive experiments conducted on natural speech and music datasets reveal that: (1) well-designed positional encoding and activation functions in Coordinate-MLPs can effectively improve audio representation quality; and (2) Fourier-ASR can robustly represent complex audio signals without extensive hyperparameter tuning. Looking ahead, the continuity and infinite resolution of implicit audio representations make our research highly promising for tasks such as audio compression, synthesis, and generation.

**Code and Appendix** — https://github.com/lif314/NeAF

## Introduction

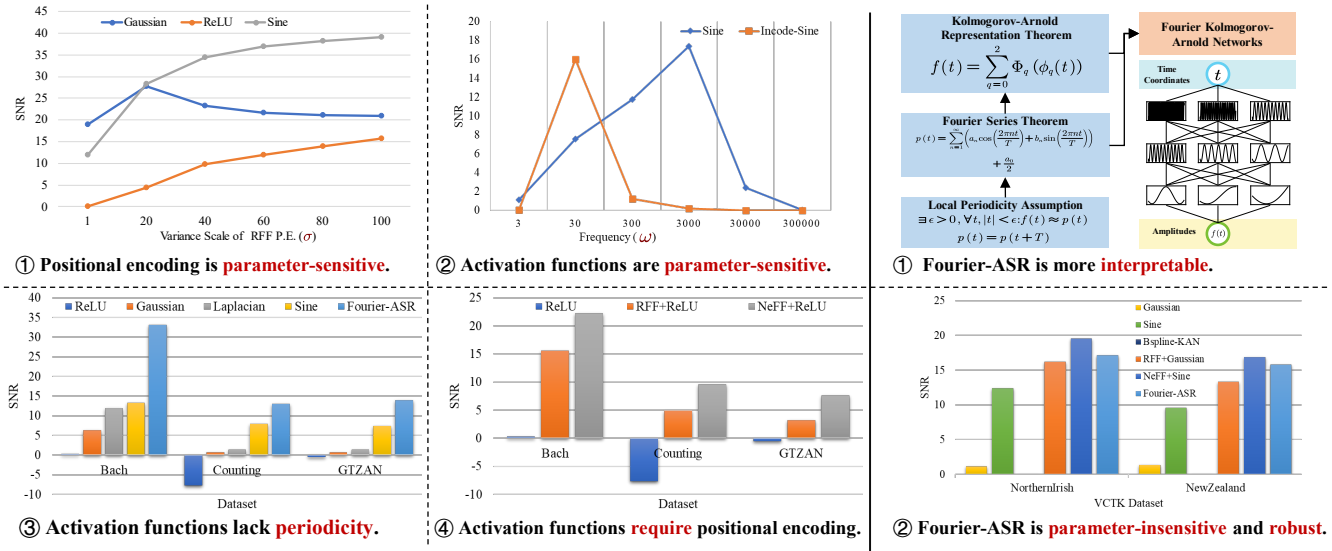
Implicit Neural Representations (INRs) provide an innovative approach to signal parameterization by representing arbitrary discrete signals as continuous functions. These functions map the domain of the signal (coordinates, e.g., timestamps in audio) to the corresponding content at those coordinates (such as the amplitude of an audio signal). Typically, these functions are approximated using neural networks, and since current neural networks are primarily constructed using multilayer perceptrons (MLPs), these types of INRs are referred to as Coordinate-MLPs.

Compared to traditional discrete signal representation schemes, INRs offer continuous implicit representation that decouples from spatial resolution and allows for infinite resolution. Therefore, the storage required for parameterized signals is independent of spatial resolution, allowing these signals to be sampled at any desired resolution. Owing to these advantages, Coordinate-MLPs have been successfully applied to various modalities of data, including neural radiance fields (Mildenhall et al. 2020), 3D occupancy grids (Mescheder et al. 2019), Signed Distance Functions (Park et al. 2019), images (Sitzmann et al. 2020), 2D computed tomography, and 3D magnetic resonance imaging (Tancik et al. 2020; Saragadam et al. 2023; Kazerouni et al. 2024).

Regarding audio signals, continuous representations offer the advantages of infinite resolution, enabling natural generation, efficient compression, and smooth processing. However, the representation of continuous audio signals using Coordinate-MLPs poses profound challenges due to the high noise, high frequency, nonlinearity, and local periodicity inherent in audio signals. According to the Weber-Fechner law, even relatively small reconstruction errors in audio signals can become perceptible due to the logarithmic nature of human auditory perception, thereby imposing high demands on the quality of audio reconstruction. Moreover, the simple combination of linear transformations and nonlinear activation functions in MLP networks makes it difficult to capture the periodicity and high-frequency components of audio signals. Through a comprehensive review, we find that till now only SIREN (Sitzmann et al. 2020) has attempted to represent audio signals using sinusoidal activation functions and provided a simple comparison with ReLU-MLPs, yet no further investigations have been conducted.

To fill the gap, we establish, to our knowledge, the **first** open-source benchmarking framework to fully explore the potential and limitations of Coordinate-MLPs in continuous audio signal representations. Specifically, since the performance of a Coordinate-MLP is primarily determined by the choice of the activation function and optional positional encoding, we identify 3 types of positional encoding mappings and 16 commonly used activation functions from existing

*Corresponding Author




(a) Benchmark of Coordinate-MLPs in Audio Signal Representations.

(b) Fourier-ASR.

Figure 1: Properties of Coordinate-MLPs and Fourier-ASR. Validations are in the appendix (`Appendix A`).

Coordinate-MLP methods. This results in 48 possible network configurations for audio signal representation, which we evaluate on speech and music datasets to assess their performance.

As shown in Fig. 1(a), our benchmark reveals the following findings. (1) Most activation functions, except those with strong linearity (e.g., Gaussian) or periodicity (e.g., Sine), are unable to effectively represent audio signals. (2) Although some activation functions, such as Gaussian and Sine, are proposed to overcome spectral bias and the tedious parameter tuning associated with positional encoding, positional encoding remains indispensable for representing audio signals. It efficiently maps time coordinates to high-dimensional spaces, allowing the network to capture high-frequency components in audio signals. (3) Due to the local periodicity of audio signals, periodic activation functions (e.g., Sine) significantly outperform other activation functions in representational capacity. Moreover, incorporating Fourier feature-based positional encoding can further enhance their ability. (4) While Sine-type activation functions are effective at representing audio signals due to their periodic nature, they unfortunately require hyperparameter-sensitive positional encodings and frequency-dependent initialization schemes, which negatively impact their robustness and generalization capabilities.

The aforementioned issues of Coordinate-MLPs fundamentally arise from the inadequate nonlinearity and lack of periodicity inherent in MLPs. As illustrated in Fig. 1(b), to enhance the nonlinear and periodic representational capabilities of neural networks, we propose a novel implicit audio representation framework, Fourier-ASR, based on the Fourier series theorem and the Kolmogorov-Arnold representation theorem. Firstly, we introduce a Kolmogorov-Arnold Network (Fourier-KAN) that utilizes Fourier basis functions to represent audio signals. This network implicitly decomposes any complex audio signal into a series of locally periodic Fourier series. Unlike MLPs, Fourier-KAN does not require additional positional encoding or activation functions, thereby avoiding cumbersome hyperparameter tuning. Furthermore, due to the use of Fourier basis functions, it more effectively captures the high-frequency components and local periodicity of signals. Secondly, to accelerate the convergence of Fourier-KAN, we introduce a Frequency Adaptive Learning Strategy (FaLS). FaLS employs an inverted frequency pyramid configuration to capture signals at various frequencies and utilizes a frequency-adaptive weight initialization scheme based on forward propagation theory to mitigate issues of gradient explosion or vanishing, thereby expediting convergence. Experimental results demonstrate that Fourier-ASR not only offers enhanced interpretability but is also robust to hyperparameter variations, effectively representing complex audio signals.

In summary, our contributions are summarized as follows:

- We introduce the **first benchmark** for Coordinate-MLPs in audio representation, incorporating 3 types of positional encodings and 16 commonly used activation functions. Our benchmark provides an in-depth analysis of the impact of positional encoding and activation functions on the representation of continuous audio signals.
- To avoid spectral bias from positional encoding and complex parameter tuning of activation functions, we propose a novel audio signal representation framework, **Fourier-ASR**, based on the Fourier series theorem and the Kolmogorov-Arnold theorem. Fourier-ASR includes Fourier Kolmogorov-Arnold Networks (Fourier-KAN) and a Frequency-adaptive Learning Strategy (FaLS). Due to the periodicity and strong nonlinearity of Fourier basis functions, Fourier-ASR effectively represents audio signals and provides enhanced interpretability.

- As shown in Fig. 1, extensive experiments conducted on speech and music datasets reveal that (1) careful tuning of positional encoding and activation function parameters can significantly enhance the representational capacity of Coordinate-MLPs for audio signals; and (2) Fourier-ASR can robustly represent audio signals without requiring cumbersome parameter tuning.

## Related Work

**Coordinate-MLPs.** The usage of Coordinate-MLPs differs significantly from that of traditional MLPs in two main aspects: (a) traditional MLPs typically operate on high-dimensional inputs, such as images, sounds, or 3D shapes; (b) traditional MLPs are primarily employed as classification heads, where the decision boundaries need not be smooth. In contrast, Coordinate-MLPs encode signals into weights, where the input is low-dimensional coordinates and the output must maintain smoothness. The success of NeRF (Mildenhall et al. 2020) demonstrates that Coordinate-MLPs, when trained with a limited number of perspective images, can reconstruct photometric projections from any angle and at any resolution. This breakthrough spurs the application of Coordinate-MLPs in numerous fields, including radiance field reconstruction (Barron et al. 2022; Chen et al. 2022; Müller et al. 2022), 3D shape representation (Wang et al. 2021; Yu et al. 2022; Yariv et al. 2023), 2D image regression (Tancik et al. 2020; Saragadam et al. 2023; Lindell et al. 2022; Ramasinghe and Lucey 2022), audio signal regression (Sitzmann et al. 2020; Kazerouni et al. 2024), and inverse rendering problems in 2D CT and 3D MRI (Tancik et al. 2020).

**Positional Encoding.** Positional encoding facilitates the learning of high-frequency representations in radiance fields, images, and 3D shapes. NeRF (Mildenhall et al. 2020) improves the ability of ReLU-MLPs to capture high-frequency signals by mapping the input coordinates to a high-dimensional Fourier space. Building upon NeRF, FFN (Tancik et al. 2020) incorporates Gaussian noise to improve the robustness of ReLU-MLPs. Although positional encodings enable MLPs to represent high-frequency components, selecting the appropriate frequency scale is crucial and often involves cumbersome parameter tuning. Specifically, when the signal bandwidth is excessively increased, Coordinate-MLPs tend to produce noisy signal interpolations (Ramasinghe, MacDonald, and Lucey 2022; Hertz et al. 2021).

**Activation Functions.** The nonlinear representation capability of Coordinate-MLPs primarily arises from activation functions. In the field of INRs, various activation functions have been employed to approximate different types of signals. ReLU is frequently employed as the activation function in NeRF-related studies due to its simplicity and effective initialization scheme (Mildenhall et al. 2020; Barron et al. 2022; Yu et al. 2021; Chen et al. 2022). However, ReLU struggles to capture high-frequency information in radiance fields, necessitating additional positional encoding. To avoid the cumbersome parameter tuning associated with positional encoding, GARF (Chng et al. 2022) uses the Gaussian activation function, which can effectively capture high-frequency information but fails to capture periodic signals and tends to overfit both noise and signal equally. To address these issues, WIRE (Saragadam et al. 2023) utilizes the complex Gabor wavelet activation function to improve the robustness. SIREN (Sitzmann et al. 2020) employs the sine activation function to capture signal periodicity, though it is sensitive to initialization schemes, limiting its generalization to audio reconstruction tasks. Building on SIREN, INCODE (Kazerouni et al. 2024) makes the parameters of the sine activation functions learnable, thereby reducing the parameter sensitivity to some extent, but it still relies on the frequency-aware initialization scheme.

## Method

### Problem Formulation

As illustrated in Fig. 2(a), natural audio signals are continuous functions of time, representing the variation in amplitude of sound signals over time. To convert this continuous signal into a digital format for storage and processing, the signal is discretely sampled, resulting in a discrete signal $a(t)$ with respect to the time coordinate $t$. However, in fields such as audio super-resolution, synthesis, and compression, researchers aim to leverage implicit neural representation techniques to preserve the continuity and differentiability of the signal as much as possible. Specifically, by receiving a discrete time coordinate $t$, a neural network regresses the amplitude $f(t)$ corresponding to $t$, thereby encoding the audio signal within the network weights. We refer to this representation as the **Neural Amplitude Fields (NeAF)**. Optimization is performed by fitting $f(t)$ to the sampled waveform $a(t)$ using an MSE loss function,

$$\mathcal{L} = \int \left\| \amalg_a(f(t)) - a(t) \right\|^2 dt, \tag{1}$$

where $\amalg_a$ samples $f(t)$ at the waveform measurement locations. Given that NeAF is independent of spatial resolution, audio can be processed at any desired resolution.

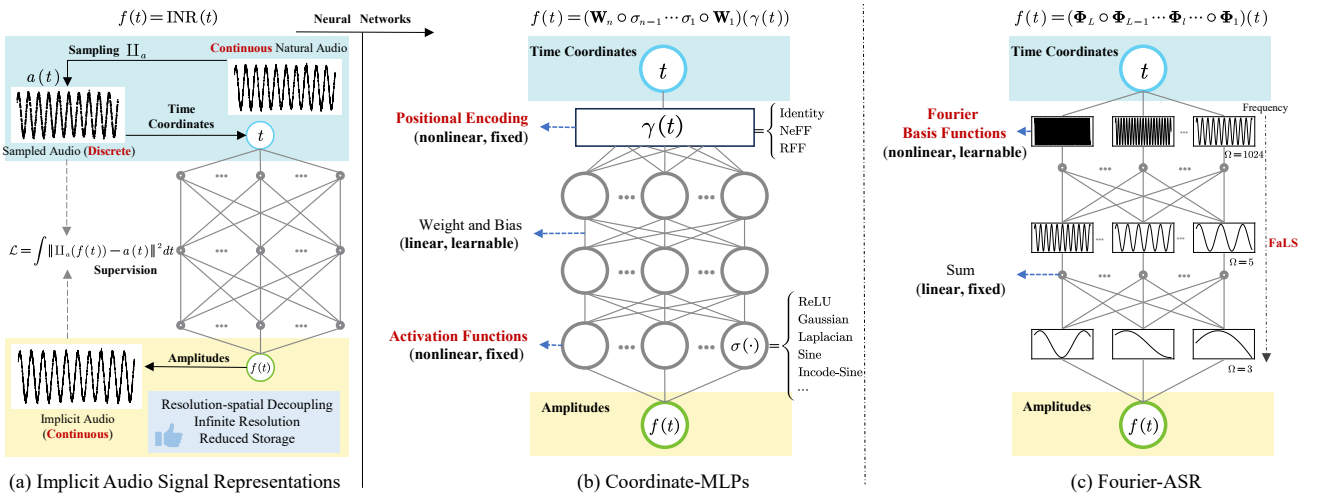
### Benchmark of Coordinate-MLP-based NeAF

As depicted in Fig. 2(b), to represent arbitrary complex audio signals, a $k$-layer Coordinate-MLP $f : \mathbb{R} \to \mathbb{R}$ is employed, which takes the time coordinate $t \in \mathbb{R}$ as input and outputs the amplitude $f(t) \in \mathbb{R}$. Thus, $f(t)$ can be defined through the following recursive relations,

$$\begin{aligned} \mathbf{z}^{(1)} &= \gamma(t) \\ \mathbf{z}^{(i+1)} &= \sigma\left(\mathbf{W}^{(i)}\mathbf{z}^{(i)} + \mathbf{b}^{(i)}\right), i = 1, \ldots, k-1 \\ f(t) &= \mathbf{W}^{(k)}\mathbf{z}^{(k)} + \mathbf{b}^{(k)}, \end{aligned} \tag{2}$$

where $\gamma(\cdot)$ denotes an optional positional encoding function that maps the input coordinate $t$ to a higher-dimensional space, $\sigma(\cdot)$ represents the element-wise applied nonlinear activation function, $\mathbf{W}^{(i)} \in \mathbb{R}^{d_{i+1} \times d_i}$ and $\mathbf{b}^{(i)} \in \mathbb{R}^{d_{i+1}}$ denote the weights and biases of the $i$-th layer, respectively, while $\mathbf{z}^{(i)} \in \mathbb{R}^{d_i}$ represents the hidden units of the $i$-th layer.

Following the architecture of Coordinate-MLPs, we conduct an extensive review of implicit neural representations




(a) Implicit Audio Signal Representations (b) Coordinate-MLPs (c) Fourier-ASR

Figure 2: (a) The problem definition of implicit audio representations; (b) The audio representation framework based on Coordinate-MLPs; (c) Fourier-ASR, a novel audio signal representation framework based on Fourier-KAN.

and identify 3 types of positional encoding and 16 potential activation functions. Specifically, as shown in Table 1, the three positional encoding schemes are identity mapping (`Identity`), NeRF Fourier features (`NeFF`) (Mildenhall et al. 2020), and random Fourier features (`RFF`) (Tancik et al. 2020). The primary activation functions include ReLU, Gaussian (Chng et al. 2022), Laplacian (Ramasinghe and Lucey 2022), Sine (Sitzmann et al. 2020), Incode-Sine (Kazerouni et al. 2024), and Gabor-Wavelet (Saragadam et al. 2023), among others (details are provided in Table 2).

| PE ($\mathcal{P}$) | $\gamma \in \mathcal{P}$ | Parameter |
|---|---|---|
| `Identity` | $\gamma(t) = t$ | - |
| `NeFF` | $\gamma(t) = [\cos(2^L \pi t), \sin(2^L \pi t)]^{\mathrm{T}}$ | $[L]$ |
| `RFF` | $\gamma(t) = [\cos(2\pi b_L t), \sin(2\pi b_L t)]^{\mathrm{T}}$, $b_L \sim \mathcal{N}(0, \sigma^2)$ | $[\sigma, L]$ |
| **Activations ($\mathcal{A}$)** | **$\sigma \in \mathcal{A}$** | |
| `ReLU` | $\sigma(x) = \max(0, x)$ | - |
| `Gaussian` | $\sigma(x) = e^{\frac{-x^2}{2a^2}}$ | $[a]$ |
| `Laplacian` | $\sigma(x) = e^{\frac{-\lvert x \rvert}{a}}$ | $[a]$ |
| `Sine` | $\sigma(x) = \sin(\omega x)$ | $[\omega]$ |
| `Incode-Sine` | $\sigma(x) = a \sin(b\omega x + c) + d$ | $a, b, c, d, [\omega]$ |
| ... | ... | ... |

Table 1: The nonlinear mappings in Coordinate-MLPs. Note that $a$ denotes a learnable parameter, while $[a]$ denotes a hyperparameter.

It is noteworthy that Gaussian, Sine, and Incode-Sine activation functions are proposed to eliminate the dependence on positional encoding in radiance fields and image representations. However, high-dimensional positional encoding mappings may be beneficial for learning high-frequency features in audio signals and their structural variations at different time scales. Therefore, in our benchmark, we apply positional encoding mappings to all three activation functions.

Based on Table 1, the Coordinate-MLPs used for benchmarking audio signal representations can be expressed as follows,

$$f(t) = (\mathbf{W}_n \circ \sigma_{n-1} \cdots \sigma_1 \circ \mathbf{W}_1)(\gamma(t)), \quad \gamma(\cdot) \in \mathcal{P}, \sigma_i(\cdot) \in \mathcal{A}, \tag{3}$$

where $t$ denotes the input time coordinate normalized to the interval $[0, 1]$, $\mathcal{P}$ represents the set of positional encodings, and $\mathcal{A}$ denotes the set of activation functions.

## Fourier-ASR

Our benchmark indicates that only through carefully designed positional encoding and activation functions can some Coordinate-MLPs effectively represent audio signals. However, their flexibility and generality are reduced due to complex parameter tuning and high sensitivity to initialization. To address this issue, as shown in Fig. 2(c), drawing from the Fourier series theorem and the Kolmogorov-Arnold representation theorem, we introduce a novel framework for audio signal representation, Fourier-ASR, which incorporates Fourier Kolmogorov-Arnold Networks (Fourier-KAN) and a Frequency-adaptive Learning Strategy (FaLS).

**Fourier Kolmogorov-Arnold Networks (Fourier-KAN).** Unlike Coordinate-MLPs based on the Universal Approximation Theorem (Hornik, Stinchcombe, and White 1989), which use combinations of linear transformations and nonlinearities, Fourier-ASR follows the Kolmogorov-Arnold Representation Theorem (Kolmogorov 1956; Arnold 1957) to represent any continuous function as a finite composition of single-variable functions and addition. For a continuous signal $f(t)$, this simplifies to,

$$f(t) = \sum_{q=0}^{2} \Phi_q \left( \phi_q(t) \right), \tag{4}$$

where $\Phi_q : \mathbb{R} \to \mathbb{R}$ and $\phi_q : [0, 1] \to \mathbb{R}$ denote the outer and inner functions, respectively. To enhance the capacity and

learnability of this representation, we employ the KAN (Liu et al. 2024) approach to extend the network to an arbitrary number of layers.

**Assumption 1** ***Local Periodicity Assumption.*** *For a complex non-stationary signal $f(t)$, there exists a sufficiently small time interval $\epsilon > 0$ where $f(t)$ can be approximated by a periodic function $p(t)$:*

$$\exists \epsilon > 0, \quad \forall t, \quad |t| < \epsilon : \quad f(t) \approx p(t)$$

*where $p(t)$ is a periodic function with period $T$, satisfying $p(t + T) = p(t)$.*

**Theorem 1** ***Fourier Series Theorem.*** *Any periodic signal $p(t)$ with period $T$ can be represented as an infinite series of sine and cosine functions,*

$$p(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n \cos\left(\frac{2\pi n t}{T}\right) + b_n \sin\left(\frac{2\pi n t}{T}\right) \right)$$

*where $a_0$, $a_n$, and $b_n$ are the Fourier coefficients.*

Specifically, consider a neural network with a shape of $[n_0, n_1, \cdots, n_L]$, where $n_l$ denotes the number of neurons on the $l$-th layer of the computational graph. For the $i$-th node on the $l$-th layer, denoted by $(l, i)$, the activation value of this neuron is $t_{l,i}$. Between the $l$-th and $(l+1)$-th layers, there are $n_l \times n_{l+1}$ non-linear basis functions. Based on Assumption 1 and Theorem 1, any audio signal within short time intervals can be approximated as combinations of cosine and sine functions. Therefore, unlike using B-Splines in KAN (Liu et al. 2024), we employ Fourier basis functions as the non-linear units connecting neurons $(l, i)$ and $(l+1, j)$,

$$\begin{aligned} &\phi_{l,j,i}(t_{l,i}) = a_{l,i} \cos(\omega t_{l,i}) + b_{l,i} \sin(\omega t_{l,i}) + c_{l,i}, \\ &l = 0, \cdots, L-1, i = 1, \cdots, n_l, j = 1, \cdots, n_{l+1}, \end{aligned} \tag{5}$$

where $a_{l,i}, b_{l,i}$ are learnable Fourier coefficients, $c_{l,i}$ is a learnable bias term, and $\omega$ is a frequency hyperparameter. Then, the activation value $t_{l+1,j}$ of the $(l+1, j)$ neuron is simply the sum of all incoming post-activations,

$$t_{l+1,j} = \sum_{i=1}^{n_l} \phi_{l,j,i}\left(t_{l,i}\right), j = 1, \cdots, n_{l+1}. \tag{6}$$

For the $l$-th Fourier KAN layer, by rewriting Eq. 6 under the matrix form, we can have,

$$\mathbf{t}_{l+1} = \underbrace{\begin{pmatrix} \phi_{l,1,1}(\cdot) & \cdots & \phi_{l,1,n_l}(\cdot) \\ \phi_{l,2,1}(\cdot) & \cdots & \phi_{l,2,n_l}(\cdot) \\ \vdots & \vdots & \vdots \\ \phi_{l,n_{l+1},1}(\cdot) & \cdots & \phi_{l,n_{l+1},n_l}(\cdot) \end{pmatrix}}_{\mathbf{\Phi}_l} \mathbf{t}_l.$$

where $\mathbf{\Phi}_l$ is the transition matrix between the Fourier layers.

In summary, Fourier-ASR employs Fourier-KAN to derive continuous representations from discrete audio signals as,

$$f(t) = (\mathbf{\Phi}_L \circ \mathbf{\Phi}_{L-1} \cdots \mathbf{\Phi}_l \cdots \circ \mathbf{\Phi}_1)(t). \tag{7}$$

Compared to the Coordinate-MLPs (Eq. 3), our Fourier-KAN leverages Fourier basis functions to achieve not only enhanced nonlinear representation capabilities but also the ability to capture local periodicity in audio signals.

**Frequency-adaptive Learning Strategy (FaLS).** Due to the varying frequency distributions of audio signals across different time scales, a fixed frequency hyperparameter ($\omega$ in Eq. 5) can lead Fourier-KAN to predominantly learn specific frequency components, thereby hindering convergence. To address this issue, we propose a Frequency-adaptive Learning Strategy (FaLS). Specifically, we assign basis functions with varying frequency thresholds to different Fourier-KAN layers. Then, a Fourier-KAN can be represented as,

$$\begin{aligned} \mathbf{z}^{(0)} &= t \\ \mathbf{z}^{(l+1)} &= \sum_{\omega=1}^{\Omega_l} \left[ \mathbf{a}^{(l,\omega)} \cos(\omega \mathbf{z}^{(l)}) + \mathbf{b}^{(l,\omega)} \sin(\omega \mathbf{z}^{(l)}) \right] + \mathbf{c}^{(l)} \\ f(t) &= \mathbf{z}^{(n_L)} \left( \cdots \left( \mathbf{z}^{(l)} \left( \cdots (\mathbf{z}^{(0)}) \right) \right) \right), l = n_0, \ldots, n_L, \end{aligned}$$

where $\mathbf{a}^{(l,\omega)}$ and $\mathbf{b}^{(l,\omega)} \in \mathbb{R}^{d_{l+1} \times d_l}$ denote the Fourier coefficient weights for the $l$-th layer at frequency $\omega$, $\mathbf{c}^{(l)} \in \mathbb{R}^{d_{l+1}}$ is the bias term of the $l$-th layer, $z^{(l)} \in \mathbb{R}^{d_l}$ denotes the hidden units of the $l$-th layer, and $\Omega_l$ is a hyperparameter indicating the maximum frequency threshold for the $l$-th layer.

***Parameter initialization.*** Following the principles of Xavier (Glorot and Bengio 2010) and Kaiming's work (He et al. 2015), we derive the initialization scheme for the Fourier-KAN. Specifically, in the forward propagation process at layer $l$ of the Fourier-KAN, the symmetry of the Fourier basis functions ensures that the expected values of both the input and the output are zeros, i.e., $E[\mathbf{z}^{(l)}] = E[\mathbf{z}^{(l+1)}] = 0$. According to Kaiming initialization (He et al. 2015), we make the following assumptions: (1) the expected values of the Fourier parameters $\mathbf{a}^{(l,\omega)}$ and $\mathbf{b}^{(l,\omega)}$ are both zeros, and the bias term $\mathbf{c}^{l}$ is omitted; (2) the variances of the input $\mathbf{z}^{(l)}$ and the output $\mathbf{z}^{(l+1)}$ are both ones. Thereby, we can determine the variance of the output at layer $l$ as,

$$\begin{aligned} Var[\mathbf{z}^{(l+1)}] = \sum_{\omega=1}^{\Omega_l} &(\cos^2(\omega \mathbf{z}^{(l)}) Var[\mathbf{a}^{(l,\omega)}] \\ &+ \sin^2(\omega \mathbf{z}^{(l)}) Var[\mathbf{b}^{(l,\omega)}]). \end{aligned}$$

Assuming that the variances of the Fourier coefficients are equal, we can have,

$$Var[\mathbf{a}^{(l)}] = Var[\mathbf{b}^{(l)}] = \frac{1}{\Omega_l}. \tag{8}$$

Thus, each independent Fourier coefficient $a_i^{(l)}$ (and $b_i^{(l)}$) is initialized using the following normal distribution,

$$a_i^{(l)}, b_i^{(l)} \sim \mathcal{N}(0, \frac{1}{\Omega_l d_{in}^{(l)}}), \tag{9}$$

where $d_{in}^{(l)}$ denotes the dimensionality of the input to layer $l$.

***Inverted pyramid frequency setting.*** Given the depth $L$ and width of a Fourier-KAN, the hyperparameters

| Activation $\sigma(\cdot)$ | Equation | Parameter | PE $\gamma(\cdot)$ | Bach (7s) | | Counting (7s) | | Blues (30s) | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | SNR ↑ | LSD ↓ | SNR ↑ | LSD ↓ | SNR ↑ | LSD ↓ | SNR ↑ | LSD ↓ |
| PReLU | $\begin{cases} x, & \text{if } x > 0 \\ ax, & \text{otherwise} \end{cases}$ | $[a]$ | Identity | 0.00 | 4.724 | 0.00 | 4.630 | 0.00 | 7.031 | 0.00 | 5.462 |
| | | | RFF | 13.42 | 1.010 | 3.38 | 1.437 | 2.50 | 2.035 | 6.43 | 1.494 |
| | | | NeFF | 17.50 | 1.133 | 7.88 | 1.575 | 5.20 | 1.539 | 10.19 | 1.416 |
| ReLU | $\max(0, x)$ | | Identity | 0.00 | 4.623 | -7.66 | 4.546 | 0.00 | 6.774 | -2.55 | 5.314 |
| | | | RFF | 15.62 | 0.978 | 4.93 | **1.400** | 3.23 | 1.862 | 7.93 | 1.413 |
| | | | NeFF | 22.29 | 1.129 | 9.57 | 1.538 | 7.64 | 1.324 | 13.17 | 1.330 |
| Gaussian | $e^{\frac{-x^2}{2a^2}}$ | $[a]$ | Identity | 6.35 | 1.130 | 0.74 | 2.165 | 0.68 | 3.059 | 2.59 | 2.118 |
| | | | RFF | 20.85 | 2.046 | 12.14 | 3.195 | 11.80 | 1.346 | 14.93 | 2.196 |
| | | | NeFF | 19.68 | 2.127 | 9.20 | 3.438 | 7.74 | 1.597 | 12.21 | 2.387 |
| Laplacian | $e^{\frac{-\lvert x\rvert}{a}}$ | $[a]$ | Identity | 12.04 | 0.932 | 1.34 | **1.561** | 1.37 | 2.434 | 4.92 | 1.642 |
| | | | RFF | 15.57 | 2.386 | 10.97 | 2.632 | 14.74 | **1.112** | 13.76 | 2.043 |
| | | | NeFF | 15.26 | 2.434 | 8.67 | 3.191 | 8.16 | 1.526 | 10.70 | 2.384 |
| Sine | $\sin(\omega x)$ | $[\omega]$ | Identity | 13.36 | 0.838 | 7.96 | 1.660 | **7.47** | **1.722** | **9.59** | **1.407** |
| | | | RFF | **39.02** | **0.582** | **13.06** | 1.412 | **16.57** | 1.156 | **22.88** | **1.050** |
| | | | NeFF | **42.39** | **0.537** | **33.58** | **0.914** | **22.02** | **0.696** | **32.66** | **0.716** |
| Incode-Sine | $a\sin(b\omega x + c) + d$ | $[\omega], a, b, c, d$ | Identity | **15.98** | **0.778** | **8.16** | 1.611 | 0.01 | 3.865 | 8.05 | 2.085 |
| | | | RFF | 38.10 | 0.595 | 12.86 | 1.559 | 15.13 | 1.241 | 22.03 | 1.132 |
| | | | NeFF | 41.40 | 0.556 | 32.24 | 1.038 | 21.33 | 0.763 | 31.99 | 0.786 |

Table 2: Benchmark leaderboard of Coordinate-MLPs. For different positional encodings (Identity, RFF, NeFF), the best results are bold for first and underlined for second. Note that "$a$" denotes a learnable parameter, while "$[a]$" denotes a hyperparameter. The benchmarking results for the remaining 10 activation functions are provided in the appendix (Appendix D).

$[\Omega_0, \cdots, \Omega_l, \Omega_L]$ dictate the number of Fourier basis functions and the tendency to learn frequency components in each layer. With the same network capacity, a larger $\Omega$ enhances the frequency resolution, improving the network's ability to capture audio signal periodicity and fluctuations. Similar to the role of positional encoding in Coordinate-MLPs, an inverted pyramid frequency setting is beneficial for the Fourier-KAN in capturing high-frequency information, thereby accelerating convergence. For instance, a 3-layer Fourier-KAN with $\Omega$ set to $[64, 5, 3]$ outperforms $[8, 8, 8]$, which may lead to convergence issues.

# Experiments

## Experimental Setup

**Datasets.** **GTZAN** music dataset (Tzanetakis and Cook 2002) includes 1000 thirty-second music clips across ten genres. **CSTR VCTK** speech corpus (Yamagishi, Veaux, and MacDonald 2019) consists of voice recordings from 110 speakers with diverse accents, each speaking approximately 400 sentences. For the benchmark, we used two 7-second clips provided by SIREN (Sitzmann et al. 2020) ("Bach" and "Counting") and a 30-second clip from GTZAN dataset ("Blues"). To comprehensively evaluate the performance of effective methods, we selected ten audio clips of different genres from the GTZAN dataset and ten audio clips with various accents from the CSTR VCTK dataset.

**Networks.** We ensured that the network parameters were comparable, ranging between 250K and 270K. For the Coordinate-MLPs, each network has a depth of 6 and a width of 256. In contrast, the Fourier-ASR network has a depth of 6 and a width of 64, with the maximum frequency thresholds set to 1024, 5, and 3 for the input layer, hidden layers, and the output layer, respectively.

**Evaluation Metrics.** Signal-to-Noise Ratio (SNR) (Roux et al. 2019) and Log-Spectral Distance (LSD) (Gray and Markel 1976) were utilized to assess the temporal and spectral errors in the reconstructed audios, respectively. Since LSD provides an indirect measure for frequency domain evaluation, we primarily focus on the SNR metric.

## Benchmark Leaderboard

We begin by examining the impact of nonlinear mappings, which are commonly presumed but have not yet been analyzed in the context of implicit audio representation. In line with Eq. 3, Table 2 presents the evaluation results of audio signal representation using 16 different activation functions and 3 types of positional encoding (Identity, NeFF, and RFF). Based on this comprehensive benchmarking, the following conclusions can be drawn:

- Most activation functions (Sigmoid, ReLU, Tanh, etc.), aside from those with strong nonlinearity (Gaussian-type) and periodicity (Sine-type), fail to capture the high-frequency and local periodicity of audio signals.
- Positional encodings significantly enhance the ability of Coordinate-MLPs to represent audio signals due to their high-dimensional mapping capabilities, which improve the model's ability to capture high-frequency information. This enhancement is particularly notable for Gaussian (11.02dB ↑ in SNR) and Sine (18.96dB ↑ in SNR) activation functions.
- In the context of positional encoding, the introduction of random Gaussian noise by RFF makes it more suited to Gaussian-type activation functions ($\approx$ 3dB ↑ in SNR).

| GTZAN Dataset | | Metrics | blu. | cla. | cou. | dis. | hip. | jaz. | met. | pop. | reg. | roc. | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | Gaussian (MLP) | SNR ↑ | 0.68 | 0.25 | 2.64 | 2.40 | 1.02 | 0.15 | 1.66 | 4.61 | 1.33 | 1.06 | 1.58 |
| | | LSD ↓ | 3.059 | 3.175 | 3.028 | 3.379 | 3.761 | 3.494 | 3.634 | 3.030 | 3.047 | 3.219 | 3.383 |
| | Sine (MLP) | SNR ↑ | 7.47 | 2.86 | 5.92 | 7.34 | 3.04 | 6.08 | 4.32 | 8.776 | 4.87 | 6.84 | 5.76 |
| | | LSD ↓ | 1.722 | 1.755 | 2.338 | 2.204 | 2.754 | 1.771 | 2.830 | 2.595 | 1.969 | 1.900 | 2.184 |
| | B-Spline (KAN) | SNR ↑ | 0.00 | 0.01 | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 | 2.07 | 0.00 | 0.01 | 0.23 |
| | | LSD ↓ | 4.643 | 4.278 | 6.685 | 5.799 | 4.359 | 4.899 | 6.000 | 3.864 | 7.533 | 4.899 | 5.30 |
| Ours | RFF+Gaussian (MLP) | SNR ↑ | 11.80 | 10.76 | 11.98 | 12.00 | 11.30 | 12.75 | 11.25 | 12.07 | 11.57 | 11.84 | 11.73 |
| | | LSD ↓ | 1.346 | 1.721 | 1.474 | 1.299 | 1.148 | 1.936 | 1.731 | 1.439 | 1.362 | 1.481 | 1.494 |
| | NeFF+Sine (MLP) | SNR ↑ | **22.02** | **25.95** | **16.35** | **17.70** | **13.92** | **19.22** | **13.05** | **15.27** | **17.79** | **19.16** | **18.04** |
| | | LSD ↓ | **0.696** | **0.585** | **1.064** | **1.036** | **0.741** | **0.983** | **0.902** | **1.245** | **0.883** | **0.714** | **0.885** |
| | Fourier-ASR (KAN) | SNR ↑ | 13.80 | 15.05 | 12.54 | 12.87 | 12.22 | 13.67 | 12.21 | 13.27 | 12.42 | 12.65 | 12.76 |
| | | LSD ↓ | 1.245 | 0.913 | 1.249 | 1.158 | 1.059 | 1.244 | 1.203 | 1.302 | 1.110 | 1.399 | 1.110 |
| **CSTR VCTK Dataset** | | Metrics | p225 | p234 | p238 | p245 | p248 | p253 | p335 | p345 | p363 | p374 | Avg. |
| Baselines | Gaussian (MLP) | SNR ↑ | 1.88 | 2.09 | 1.16 | 4.06 | 0.23 | 2.39 | 1.37 | 3.06 | 5.56 | 2.25 | 2.41 |
| | | LSD ↓ | 2.126 | 2.034 | 2.557 | 1.831 | 2.884 | 2.065 | 2.277 | 1.896 | 1.791 | 1.827 | 2.129 |
| | Sine (MLP) | SNR ↑ | 14.86 | 10.88 | 12.38 | 14.41 | 10.32 | 13.85 | 9.61 | 15.89 | 12.78 | 12.53 | 12.75 |
| | | LSD ↓ | 1.743 | 1.588 | 1.748 | 1.665 | 1.630 | 1.672 | 1.619 | 1.556 | 1.716 | 1.500 | 1.644 |
| | B-Spline (KAN) | SNR ↑ | 0.01 | 0.02 | 0.01 | 0.01 | 0.00 | 0.01 | 0.02 | 0.02 | 0.05 | 0.11 | 0.03 |
| | | LSD ↓ | 3.312 | 3.113 | 3.317 | 3.151 | 3.506 | 3.160 | 3.000 | 2.957 | 2.705 | 2.631 | 3.085 |
| Ours | RFF+Gaussian (MLP) | SNR ↑ | 11.67 | 12.93 | 16.19 | 11.99 | 15.52 | 12.21 | 13.32 | 15.95 | 12.79 | 12.28 | 12.81 |
| | | LSD ↓ | 2.401 | 2.128 | 1.789 | 2.218 | 2.183 | 2.258 | 2.076 | 1.704 | 2.012 | 2.059 | 1.983 |
| | NeFF+Sine (MLP) | SNR ↑ | **25.20** | **31.63** | **19.56** | **32.03** | **27.00** | **27.11** | **16.87** | **28.38** | **29.25** | **30.83** | **26.79** |
| | | LSD ↓ | **1.015** | **0.734** | **1.235** | **0.866** | **1.134** | **0.917** | **1.207** | **1.032** | **0.753** | **0.877** | **0.977** |
| | Fourier-ASR (KAN) | SNR ↑ | 18.30 | 20.68 | 17.12 | 18.26 | 21.34 | 17.40 | 15.79 | 17.34 | 17.86 | 20.20 | 18.43 |
| | | LSD ↓ | 1.495 | 1.228 | 1.615 | 1.310 | 1.464 | 1.397 | 1.456 | 1.417 | 1.321 | 1.267 | 1.397 |

Table 3: Evaluation of Fourier-ASR and new nonlinear mapping designs on GTZAN and CSTR VCTK dataset.

Conversely, `NeFF` employs Fourier mappings, which are more compatible with Sine-type activation functions ($\approx$ 9dB ↑ in SNR).

## Evaluation of Fourier-ASR and New Designs

Based on the benchmark leaderboard presented in Table 2, we selected effective nonlinear mappings for comparison with Fourier-ASR on the GTZAN (Tzanetakis and Cook 2002) and CSTR VCTK (Yamagishi, Veaux, and MacDonald 2019) datasets. It is noteworthy that although Gaussian (Ramasinghe and Lucey 2022) and Sine (Sitzmann et al. 2020) activation functions were introduced to mitigate the complex parameter adjustments and pectral bias associated with positional encoding, we found that positional encoding remains essential due to the high-frequency nature and local periodicity of audio signals. Consequently, we designed **new nonlinear mappings**, namely `RFF+Gaussian` and `NeFF+Sine`, to address these challenges.

As shown in Table 3, the designs `RFF+Gaussian` and `NeFF+Sine` significantly enhance the ability of Coordinate-MLPs to represent audio signals. On the GTZAN dataset, these methods improve the SNR by 10.15dB ↑ and 12.28dB ↑, respectively. On the CSTR VCTK dataset, the SNR improvements are 10.40dB ↑ and 14.04dB ↑, respectively. Due to the periodic nature of Fourier basis functions and the Frequency-adaptive Learning Strategy (FaLS), our proposed Fourier-ASR(KAN) significantly outperforms Sine(MLP) ($\approx$ 6dB ↑) and B-Spline(KAN) ($\approx$ 18dB ↑). However, because existing optimization strategies are not perfectly adapted to KAN networks (Liu et al. 2024), Fourier-ASR(KAN) performs slightly worse than the locally periodic NeFF+Sine(MLP). Nonetheless, Fourier-ASR(KAN) does not require positional encoding, thereby avoiding the need for cumbersome hyperparameter tuning.

## Conclusion and Future Work

We proposed the first open-source benchmark for evaluating implicit neural audio signal representations based on Coordinate-MLPs, addressing a critical gap in standardized performance assessment. We demonstrated the effectiveness of combining positional encoding and nonlinear mapping designs of activation functions in the field of continuous audio representations. Additionally, we introduced a novel audio signal representation framework, Fourier-ASR, which integrates the Fourier series theorem and the Kolmogorov-Arnold representation theorem, offering enhanced interpretability and more stable representational capacity. Our work not only guides the selection of components for Coordinate-MLP-based audio signal representations but also advances the development of audio representation applications. Due to the superior characteristics of implicit neural representations, such as continuous differentiability and decoupling from spatial resolution, our work can be effectively applied to downstream tasks such as audio super-resolution, denoising, compression, and generation.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272343; in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23; and in part by the Fundamental Research Funds for the Central Universities.

## References

Arnold, V. I. 1957. On Functions of Three Variables. *Proceedings of the USSR Academy of Sciences*, 114: 679–681.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 5460–5469. New Orleans, LA, USA: IEEE.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022. TensoRF: Tensorial Radiance Fields. In *Proceedings of the European Conference on Computer Vision*, 333–350. Berlin, Heidelberg: Springer-Verlag.

Chng, S.-F.; Ramasinghe, S.; Sherrah, J.; and Lucey, S. 2022. Gaussian Activated Neural Radiance Fields for High Fidelity Reconstruction and Pose Estimation. In *Proceedings of the European Conference on Computer Vision*, 264–280. Berlin, Heidelberg: Springer-Verlag.

Glorot, X.; and Bengio, Y. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, 249–256. Chia Laguna Resort, Sardinia, Italy: PMLR.

Gray, A.; and Markel, J. 1976. Distance Measures for Speech Processing. *IEEE Transactions on Speech and Audio Processing*, 24(5): 380–391.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1026–1034. Santiago, Chile: IEEE.

Hertz, A.; Perel, O.; Giryes, R.; Sorkine-Hornung, O.; and Cohen-Or, D. 2021. SAPE: Spatially-Adaptive Progressive Encoding for Neural Optimization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34, 8820–8832. Red Hook, NY, USA: Curran Associates Inc.

Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5): 359–366.

Kazerouni, A.; Azad, R.; Hosseini, A.; Merhof, D.; and Bagci, U. 2024. INCODE: Implicit Neural Conditioning with Prior Knowledge Embeddings. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 1287–1296. United States: Institute of Electrical and Electronics Engineers Inc.

Kolmogorov, A. N. 1956. On the Representation of Continuous Functions of Several Variables by Superpositions of Continuous Functions of a Smaller Number of Variables. *Proceedings of the USSR Academy of Sciences*, 108: 179–182.

Lindell, D. B.; Van Veen, D.; Park, J. J.; and Wetzstein, G. 2022. Bacon: Band-limited Coordinate Networks for Multiscale Scene Representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 16231–16241. New Orleans, LA, USA: IEEE.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024. KAN: Kolmogorov-Arnold Networks. arXiv:2404.19756.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 4455–4465. Long Beach, CA, USA: IEEE.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision*, 405–421. Cham: Springer International Publishing.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 165–174. Long Beach, CA, USA: IEEE.

Ramasinghe, S.; and Lucey, S. 2022. Beyond Periodicity: Towards a Unifying Framework for Activations in Coordinate-MLPs. In *Proceedings of the European Conference on Computer Vision*, 142–158. Cham: Springer Nature Switzerland.

Ramasinghe, S.; MacDonald, L.; and Lucey, S. 2022. On Regularizing Coordinate-MLPs. arXiv:2202.00790.

Roux, J. L.; Wisdom, S.; Erdogan, H.; and Hershey, J. R. 2019. SDR – Half-baked or Well Done? In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 626–630. Brighton, UK: IEEE.

Saragadam, V.; LeJeune, D.; Tan, J.; Balakrishnan, G.; Veeraraghavan, A.; and Baraniuk, R. G. 2023. WIRE: Wavelet Implicit Neural Representations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 18507–18516. Vancouver, BC, Canada: IEEE.

Sitzmann, V.; Martel, J. N. P.; Bergman, A. W.; Lindell, D. B.; and Wetzstein, G. 2020. Implicit Neural Representations with Periodic Activation Functions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, 7462–7473. Red Hook, NY, USA: Curran Associates Inc.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, 7537–7547. Red Hook, NY, USA: Curran Associates, Inc.

Tzanetakis, G.; and Cook, P. 2002. Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing*, 10(5): 293–302.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34, 27171–27183. Red Hook, NY, USA: Curran Associates Inc.

Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). https://doi.org/10.7488/ds/2645.

Yariv, L.; Hedman, P.; Reiser, C.; Verbin, D.; Srinivasan, P. P.; Szeliski, R.; Barron, J. T.; and Mildenhall, B. 2023. BakedSDF: Meshing Neural SDFs for Real-Time View Synthesis. In *ACM SIGGRAPH 2023 Conference Proceedings*. New York, NY, USA: Association for Computing Machinery.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 4576–4585. Nashville, TN, USA: IEEE.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, volume 35, 25018–25032. Red Hook, NY, USA: Curran Associates Inc.