Adaptive Wavelet-Positional Encoding for High-Frequency Information Learning in Implicit Neural Representation

Hongxu Zhao[1], Zelin Gao[1], Yue Wang[1], Rong Xiong[1], Yu Zhang[1,2] *

[1]State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China
[2]Key Laboratory of Collaborative sensing and autonomous unmanned systems of Zhejiang Province
(zhaohongxu, jamesgzl, ywang24, rxiong, zhangyu80)@zju.edu.cn

Abstract

Implicit Neural Representation (INR) has shown great potential in constructing the complex nature signal as a continuous implicit function. However, the representation results are incomplete since different components of the signal correspond to different frequencies and neural network inherently tends to low-frequency convergence. In this paper, we propose the adaptive Wavelet-Positional Encoding (WPE) to precisely represent content under different frequency distributions for coordinate-based implicit representations. The High-Frequency Perception (HFP) method is first proposed to query locations of high-frequency components from input signals, which can be indicated as local centers of WPE. Then, motivated by wavelet series regression, we present to embed these queried low-dimensional coordinate inputs into wavelet-frequency space by WPE to represent fine details of target signals. Experiments demonstrate that the proposed method can be integrated into various INR methods without modifying training frameworks while significantly improving their performance in 1D signal fitting, 2D image regression, and even 3D scene representation.

Introduction

Implicit Neural Representation (INR) has recently emerged as a potential representation paradigm in the coordinated-based complex signal modeling field, including 2D image fitting (Sitzmann, Zollhoefer, and Wetzstein 2019) and 3D scene representation (Chen and Zhang 2019; Genova et al. 2020). One of the most influential INR works is the Neural Radiance Field (NeRF) (Mildenhall et al. 2020), which employs the neural weights of MLP to learn the implicit representation function of scenes and demonstrates adaptability to various task formats (Chen et al. 2023; Zhou et al. 2023). INR methods (Zhang et al. 2020; Pumarola et al. 2021; Barron et al. 2021) can circumvent the limitations associated with classical methods like SfM (Schonberger and Frahm 2016) because of the continuous and implicit neural representations they use. However, due to the "spectral bias" problem, where neural networks inherently learn different frequency components with different convergence performances, it is difficult for these existing coordinate-based INR methods to provide a complete signal representation, especially for the complex signals with more high-frequency components (Jacot, Gabriel, and Hongler 2018). This issue limits the ability of INR methods to effectively represent complex nature signals and capture complete fine details.

To address the "spectral bias" problem, some methods embed low-dimensional coordinate inputs into higher-dimensional feature space thus make network more aware of the difference in the inputs (Mildenhall et al. 2020; Tancik et al. 2020), while other methods investigate the activation function of MLP to preserve more fine-grained gradients (Sitzmann et al. 2020; Ramasinghe and Lucey 2022).

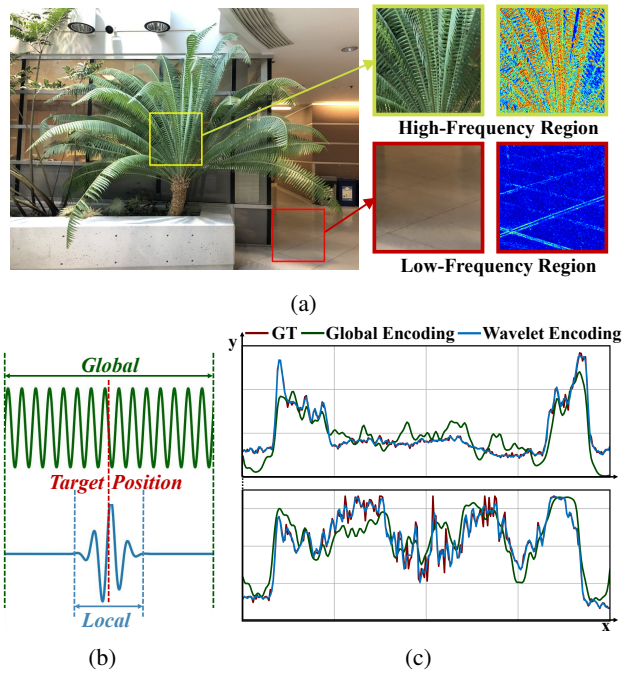

Figure 1: (a) The spatial distribution of different frequency components in the scene is non-uniform. (b) Compared to global transforms like Fourier transforms, wavelet transform contributes to local signal representations. (c) The experimental result of 1D signal regression using Fourier series embedding and Wavelet series embedding respectively.

*Corresponding author

These above methods apply global processing to the entire input without the perception of different frequency components. However, as shown in Fig. 1a, the spatial distribution of different frequency components is non-uniform in the target signal. Therefore, a local perception method should be proposed to mitigate the lack of awareness of different frequency components and improve INR's capability to represent complex signals in full frequency bands.

In this paper, inspired by wavelet series regression which enhances the signal representative ability for localized information as Fig. 1b, we propose an adaptive Wavelet-Positional Encoding (WPE) method to realize a complete signal representation of INR with fine details including different frequency components. Using the local nature of wavelet series, we achieve local perception and process of coordinate inputs before putting them into MLP, while retaining global Positional Encoding (PE) to capture the entirety feature. Moreover, the proposed method can realize a more complete representation of signal details and provide generalization ability to various dimensional target signals, thus exhibiting excellent adaptability to INR. Specifically, we first propose the novel High-Frequency Perception (HFP) method to decompose input signals in target space and obtain spatial locations of high-frequency components. Then we propose to embed low-dimensional coordinate inputs into wavelet-frequency space by our WPE function which is set at locations of high-frequency components obtained by HFP, capitalizing on the inherently local nature of wavelet series. Meanwhile, the retained PE method can preserve global features and low-frequency components in the entire signal space. In this way, our method can realize a complete signal representation containing different frequency information and construct complex fine details. Furthermore, since our proposed method lifts input into higher dimensions without any additional modification, it can be directly applied to various INR-based signal representation tasks with their original training frameworks. In experiments, we conduct 1D signal fitting, 2D image regression, and 3D scene representation. Experimental results demonstrate that our method can significantly improve the performance of INR especially in fine details, and even benefit explicit representations. Ablation study is carried out to further investigate the effectiveness of HFP and WPE.

The main contributions are as follows:

- We propose the High-Frequency Perception (HFP) method that enables adaptive acquisition of frequency distribution corresponding to the target signal space.
- We propose the Wavelet-Positional Encoding (WPE) method to embed low-dimensional coordinate inputs into wavelet-frequency space before putting them into MLP.
- The proposed method can be integrated into various INR-based methods to improve performance without changing the training framework.

Related Work

Implicit Neural Representation

Recent years Implicit Neural Representation (INR) methods have shown the capability of deep networks in signal modeling tasks (e.g.,2D regression and 3D reconstruction) as an implicit continuous function (Park et al. 2019; Chen and Zhang 2019). Many works are presented using network to represent 2D images (Sitzmann, Zollhoefer, and Wetzstein 2019) or 3D scenes (Genova et al. 2020). The most influential method among these works is Neural Radiance Field (NeRF) (Mildenhall et al. 2020), which utilizes MLP weights for scene representation and novel view synthesis. Different from traditional methods (Chaurasia et al. 2013) and (Germann et al. 2012) which indicate pixel color through geometric information or interpolation, NeRF and its related INR methods (Zhang et al. 2020; Pumarola et al. 2021) can represent scenes continuously and implicitly by network weights, setting off a wave of neural representation and rendering. Mip-NeRF (Barron et al. 2021) improves the sampling approach by replacing ray with frustum and realizes a high-quality multi-scale scene representation against spectral aliasing, Mip-NeRF 360 (Barrona et al. 2022) further extends the implicit representation to un-bounded scenes while maintaining reconstruction quality. However, these methods can not perfectly achieve complete reconstruction results in fine-detailed regions where the geometry or texture changes high-frequently in the target space, which is crucial in practical real-world scenarios.

High Frequency Convergence

INR methods are based on deep networks to represent and reconstruct target signals. As the theory of (Jacot, Gabriel, and Hongler 2018) and (Rahaman et al. 2019), deep networks tend to converge low-frequency components and disregard high-frequency components during training process, which is referred to as "spectral bias". To alleviate this phenomenon, many works have been presented, vanilla NeRF (Mildenhall et al. 2020) incorporates the idea of Positional Encoding (PE) from (Vaswani et al. 2017) to map input coordinates into higher dimensional space before feeding into MLP. FFN (Tancik et al. 2020) replaces the mapping transform function with Fourier functions. (Ramasinghe and Lucey 2023; Gao, Dai, and Zhang 2023) integrate PE into optimization process to adaptive adjust encoding parameters. On the other hand, SIREN (Sitzmann et al. 2020) and similar methods (Ramasinghe and Lucey 2022; Hertz et al. 2021; Chng et al. 2022) replace the activation function of network to provide higher-order gradient. WIRE (Saragadam et al. 2023) further use complex Gabor wavelet activation, which retains the advantages of periodic functions while maintaining space and frequency compactness. But as we know, all these methods are done on a global scale and process entire space indiscriminately without distinguishing different frequency components. Methods such as Instant-NGP (Müller et al. 2022) and TensoRF (Chen et al. 2022) introduce explicit representations to optimize local feature, which require modification to overall framework, thus can not be integrated into other works. We propose our WPE to enhance high-frequency selectively by using the spatial frequency distribution obtained from inputs. In this way, we can particularly enhance the mapping function in interested regions due to the local nature of wavelet transform, while can be easily integrated into other representations

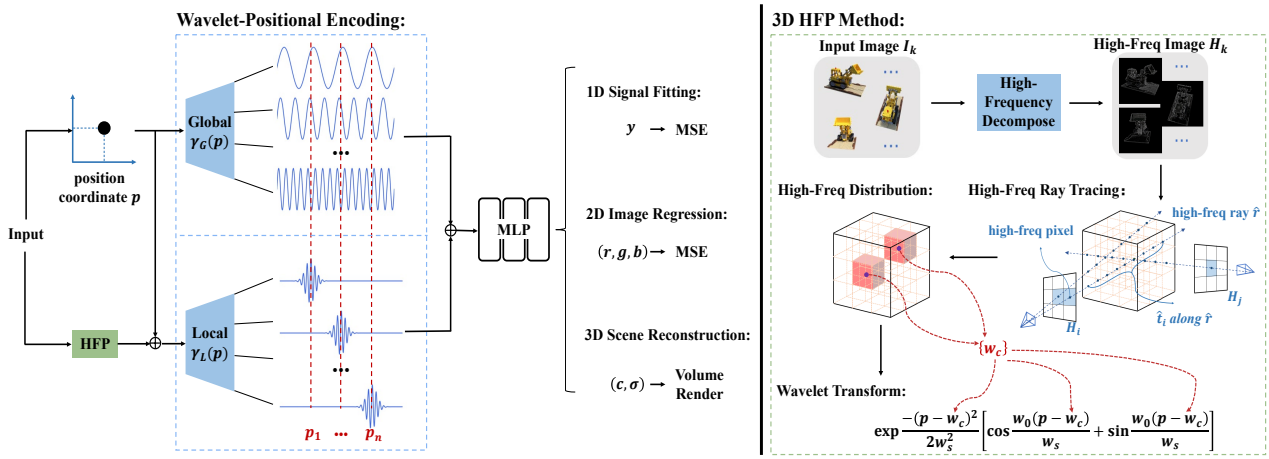

Figure 2: **Overview of Our Method.** (**Left**) Position coordinates p are fed into Wavelet-Positional Encoding (WPE) with output of High-frequency Perception(HFP) before putting into MLP. The HFP method is proposed to acquire the spatial distribution of frequency components from the input signal, and thereby determine the center of wavelet functions in encoder. The WPE encoder utilizes both global Fourier functions and local wavelet functions to embed the input coordinates into the wavelet-frequency space. Our proposed method is applicable to various implicit neural representation frameworks. (**Right**) 3D HFP method: Given a set of input RGB images I_k, we present a High-Frequency Decompose block to obtain high-frequency decomposed images H_k. By performing ray tracing method on the high-frequency ray $\hat{r}$ passing through the high-frequency pixel on H_k and counting high-frequency points $\hat{t}_i$ in each grids, the spatial distribution of high-frequency components in the target scene can be obtained by statistic method. The positions of richer high-frequency components are set as the center w_c of local wavelet transforms.

Method

The overview of our method is shown in Fig. 2. Our goal is to address the "spectral bias" problem in INR method, and achieve the higher-fidelity representation results of complex signals. Specifically, we first propose the High-Frequency Perception method to adaptive acquire the distribution of frequency components in the target space. Then we propose the Wavelet-Positional Encoding (WPE) method to process input selectively by leveraging the localized nature of the wavelet series function. We also provide a brief introduction of regression tasks utilizing INRs in the Preliminaries section.

Preliminaries

Implicit Neural Representation (INR) is commonly formulated as learning a mapping function f_θ using MLP, which takes low-dimensional coordinate p as input and outputs expected value of the function in the target space $\mathbb{R}^n$ such as RGB values in the pixel coordinate. For instance, 1D signal fitting task can be represented as $f_{\theta-1D} : x \to y$ where x denotes input position coordinate and y denotes output signal value. 2D image regression task $f_{\theta-2D} : (x, y) \to (r, g, b)$ can also be represented as INR function where (x, y) denotes input pixel coordinate and (r, g, b) denotes output pixel value. 3D scene novel view synthesis task based on NeRF can be formulated as $f_{\theta-3D} : (x, d) \to (c, \sigma)$, where x denotes spatial coordinate (x, y, z), d denotes ray direction coordinate (θ, ϕ), c and σ denote RGB color and volume density respectively.

High-Frequency Perception Method

The goal of this paper is to represent complete fine details of the target signal especially for high-frequency components, which requests the frequency distribution in target space $\mathbb{R}^n$. The principles of the High-Frequency Perception (HFP) method presented in this section can be primarily divided into two parts: firstly HFP performs high-frequency decomposition on the input signal I, then employs space partitioning and statistical methods to obtain the spatial distribution of these components in the target signal space.

For the case of input information I being image I_k, we present a High-Frequency Decompose block inspired by WaveNeRF (Xu et al. 2023), which uses a pyramidal structure of Discrete Wavelet Transform (DWT) to obtain different frequency components, but reduces the size of input image from $(h \times w)$ to $(\frac{h}{4} \times \frac{w}{4})$ due to its down-sampling operation. To preserve the size of input images, we use a 3-layer Laplacian pyramid combined with DWT to decompose the input image I_k into frequency images D^l with progressively halved resolutions as:

$$\text{DWT}\{I_k\} = \{D_k^l\}, l \in \{1, 2, 3\} \tag{1}$$

Then at each level l of the pyramid, we do subtraction $\text{sub}(\cdot)$ between D_k^{l-1} and the up-sampling image of D_k^l as $\text{up}(D_k^l)$, resulting in a series of high-frequency decomposition images H_k^l with the same resolution as D_k^{l-1}:

$$H_k^l = \begin{cases} \text{sub}\{I_k, \text{up}(D_k^l)\}, & \text{if } l = 1 \\ \text{sub}\{D_k^{l-1}, \text{up}(D_k^l)\}, & \text{if } l > 1 \end{cases} \tag{2}$$

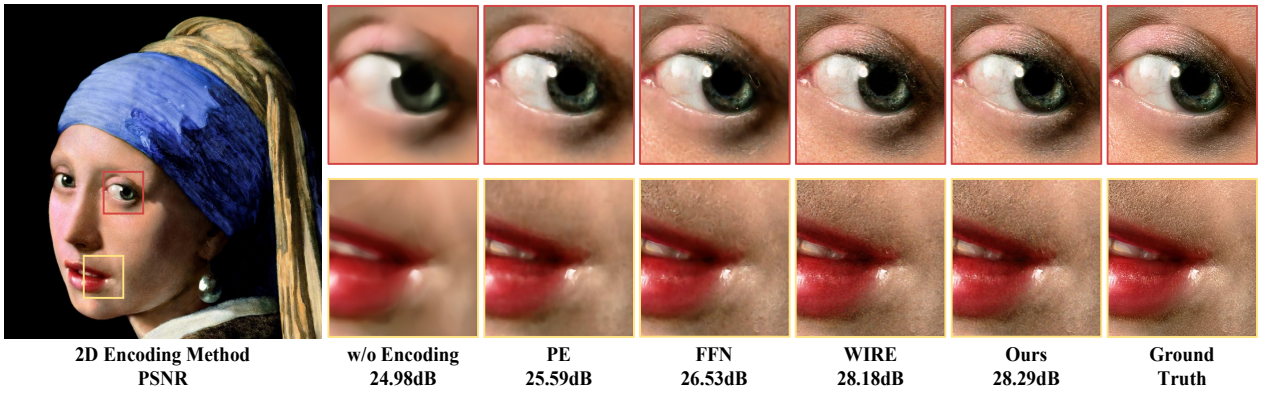

Figure 3: **2D Image Regression Experiment Results.** We compare our WPE method with global methods PE, FFN and WIRE in 2D task. Quantitative and Qualitative comparisons show the proposed method can provide the best performance and more fine details.

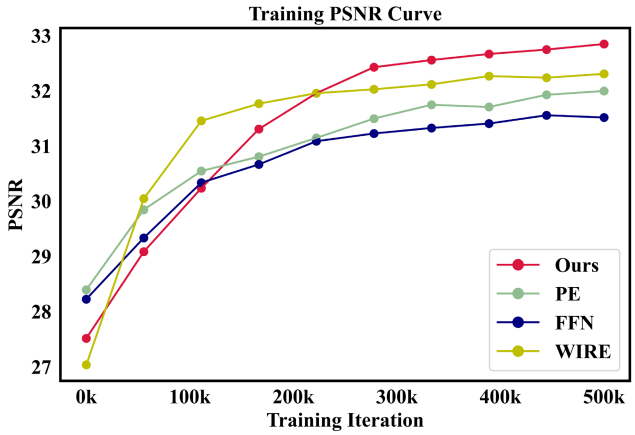

Figure 4: **Comparison of Optimization Efficiency.** We visualize the training PSNR curves including our method (red), PE (green), FFN (blue) and WIRE (yellow) in "Lego" scene of NeRF-Synthetic dataset.

We select the top-level decomposition image H_k^l, which has the same resolution ($h \times w$) with the input image I_k, as the output of the High-Frequency Decompose block, and denote H_k^l of I_k as H_k. In our 3-level pyramid, the top-level l value is 3, and all H_k^l have the same spatial resolution with input I_k due to the up-sampling we employ when solving higher-level H_k^l.

Next, we partition the target signal space $\mathbb{R}^n$ into grids based on its spatial dimension n (e.g., set grid's dimension as 2 in 2D image regression). It is worth emphasizing that, we solely take the spatial positional dimension into account. For example, the neural radiance field dimension is set as $n(x, y, z) = 3$ instead of $n(x, y, z, \theta, \phi) = 5$. Based on the obtained spatial positions of the decomposed high-frequency components, high-frequency point $\hat{t}$ statistics can be performed within each grid, thereby approximating the spatial distribution of high-frequency components in the signal space $\mathbb{R}^n$. We select M grids that have the highest number of occurrences of $\hat{t}$, and denote the central positions of these grids as $\{w_c\}$. In this way, HFP method can obtain the spatial distribution of high-frequency components in the target space:

$$\{w_c\} = \text{HFP}\{I\} \tag{3}$$

Fig. 2(right) illustrates our HFP method for 3D scene representation, where the input coordinate dimension and spatial dimension are inconsistent. We incorporate camera parameters of input images and perform high-frequency point $\hat{t}_i$ sampling on the high-frequency rays $\hat{r}$ using ray tracing way (Kajiya and Herzen 1984), which orders to obtain the high-frequency distribution.

Wavelet-Positional Encoding

In the second stage of our work, we propose a novel encoding method called Wavelet-Positional Encoding (WPE) to address the problem of "spectral bias" caused by the insufficient convergence capability of implicit neural representations to high-frequency components. Existing encoding methods such as PE (Mildenhall et al. 2020) and FFN (Tancik et al. 2020) globally process the entire spatial input, lacking targeted perception and processing of detailed high-frequency components. We combine the frequency distribution obtained in above section to selectively process complex components, which is particularly helpful in mitigating "spectral bias" issues.

The core of our proposed WPE lies in leveraging the local characteristics of wavelet series to enable targeted local wavelet-space embedding of low-dimensional input coordinates p. It allows for more complex encoding in regions with richer high-frequency details, while relatively simpler encoding in other regions. In this paper, the base of WPE we chose is the discrete morlet function:

$$f(p) = \exp\left(\frac{-(p - w_c)^2}{2w_s^2}\right)\left[\cos\frac{w_0(p - w_c)}{w_s} + \sin\frac{w_0(p - w_c)}{w_s}\right] \tag{4}$$

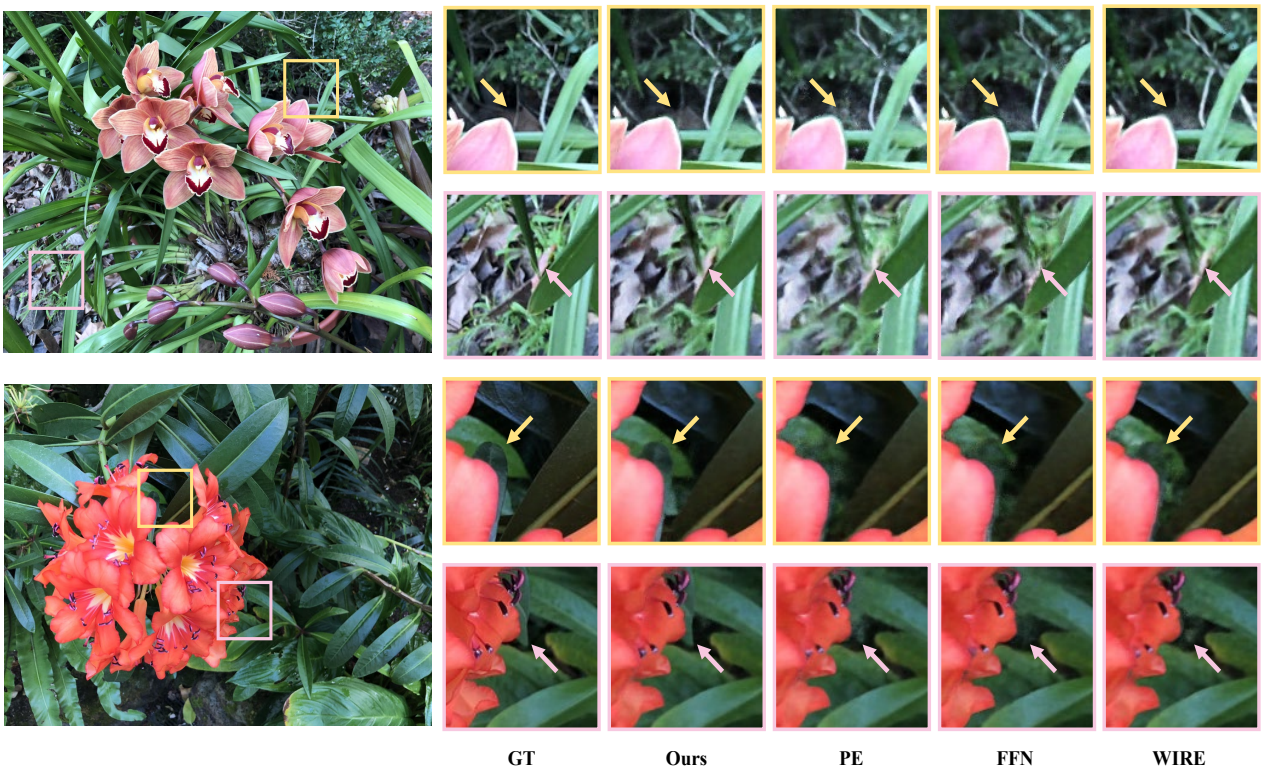

Figure 5: **Qualitative Comparison of 3D Reconstruction Novel View Synthesis on LLFF dataset.** We compare the high-frequency fine details of our Wavelet-Positional Encoding method with other global methods including PE, FFN and WIRE. All experiments are conducted on a single RTX 3090 GPU with the same experimental setup.

where p denotes input coordinate, w_0, w_s, and w_c respectively denote the frequency, scale, and center of the wavelet transform. In above section, we have obtained the central positions of M grids with rich high-frequency components. Therefore, we can directly set them as the center parameter, denoted as w_c in the wavelet function. As for w_0 and w_s, we set them as trainable parameters and incorporate them into the training process of the network, allowing for an adaptive adjustment of the frequency and scale parameters in the wavelet series. When selecting M grids, we need to configure M local wavelet functions, with each center corresponding to a w_c position. Therefore, the local encoding function $\gamma_L(p)$ for input coordinates p using wavelet series can be summarized as:

$$\gamma_L(p) = \begin{bmatrix} \exp[-(p-w_c^1)^2/2{w_s^1}^2]\cos[w_0^1(p-w_c^1)/w_s^1] \\ \exp[-(p-w_c^1)^2/2{w_s^1}^2]\sin[w_0^1(p-w_c^1)/w_s^1] \\ \cdots \\ \exp[-(p-w_c^M)^2/2{w_s^M}^2]\cos[w_0^M(p-w_c^M)/w_s^M] \\ \exp[-(p-w_c^M)^2/2{w_s^M}^2]\sin[w_0^M(p-w_c^M)/w_s^M] \end{bmatrix} \tag{5}$$

Since our method perceives frequency distribution in spatial dimension, the local encoding function $\gamma_L(p)$ also embeds only spatial coordinates when p contains non-spatial coordinates. Taking NeRF framework as example, where signal space dimension $n = 3$ but input coordinates (x, y, z, θ, ϕ) dimension is 5, we only embed the spatial coordinates (x, y, z) using $\gamma_L(p)$ and do not embed the ray direction coordinates (θ, ϕ) locally.

Due to the localized nature of wavelet encoding function and significant variations of frequency information distribution across different signals, we retain the global method of PE (Mildenhall et al. 2020) to achieve a comprehensive embedding of the entire input signal, which ensures the convergence of information from all frequency bands. Setting the embedding dimension for global encoding from R to R^{2L}, the global encoding function $\gamma_G(p)$ for input coordinates p is:

$$\gamma_G(p) = \begin{bmatrix} \sin(2^0\pi p) \\ \cos(2^0\pi p) \\ \cdots \\ \sin(2^{L-1}\pi p) \\ \cos(2^{L-1}\pi p) \end{bmatrix} \tag{6}$$

For the low-dimensional coordinate input p, WPE achieves local and global embedding through $\gamma_L(p)$ and $\gamma_G(p)$ respectively, embedding the input to a high-dimensional wavelet-frequency space before putting into MLP:

$$\text{WPE}(p) = \begin{bmatrix} \gamma_L(p) \\ \\ \gamma_G(p) \end{bmatrix} \tag{7}$$

Scene	View synthesis quality											
	PSNR ↑				SSIM ↑				LPIPS ↓			
	PE	FFN	WIRE	Ours	PE	FFN	WIRE	Ours	PE	FFN	WIRE	Ours
Fern	24.51	24.69	25.08	**25.40**	0.80	0.79	0.81	**0.83**	0.129	0.147	0.125	**0.121**
Flower	26.90	26.77	27.49	**27.83**	0.82	0.81	0.85	**0.85**	0.108	0.106	0.101	**0.094**
Fortress	30.81	30.07	30.41	**30.86**	0.87	0.85	0.86	**0.87**	0.052	0.066	0.057	**0.051**
Horns	25.35	**25.73**	25.38	25.49	0.79	**0.82**	0.77	0.81	0.190	**0.172**	0.197	0.177
Leaves	22.17	21.51	22.80	**23.25**	0.75	0.72	0.75	**0.78**	0.139	0.160	0.132	**0.123**
Orchids	21.64	21.63	23.18	**23.39**	0.71	0.69	0.76	**0.77**	0.158	0.174	0.133	**0.129**
Room	36.87	34.99	35.68	**37.23**	0.96	0.94	0.94	**0.97**	0.034	0.043	0.039	**0.032**
T-rex	28.39	28.99	29.41	**29.71**	0.89	0.91	0.92	**0.92**	0.067	**0.047**	0.054	0.052
Mean	27.08	26.80	27.43	**27.89**	0.82	0.81	0.83	**0.85**	0.110	0.114	0.105	**0.097**

Table 1: **Quantitative Comparison on LLFF Dataset.** "↑"means more is better, and "↓"means less is better. We select PSNR, SSIM, and LPIPS as evaluation metrics. The best results are in bold.

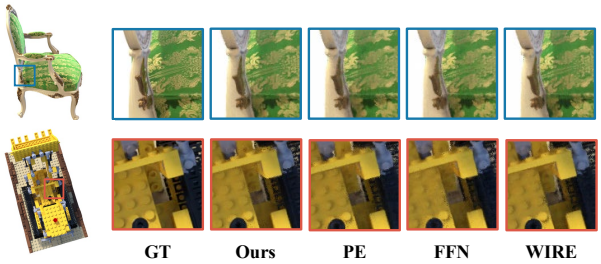

Figure 6: **Qualitative Comparison of 3D Reconstruction Detail on NeRF-Synthetic Dataset.** The representation and reconstruction details of "Lego" and "Chair" scene using PE, FFN, WIRE and Ours.

Experiments

We conduct experiments on three implicit neural representation tasks: 1D signal fitting $f_{\theta-1D} : x \rightarrow y$, 2D image regression $f_{\theta-2D} : (x, y) \rightarrow (r, g, b)$, and 3D scene reconstruction $f_{\theta-3D} : (x, d) \rightarrow (c, \sigma)$ to demonstrate the effectiveness of our method in various INR framework. They all under the category of regression tasks utilize coordinate-based MLPs for implicit neural representation. The comparison and ablation study experiment results show the effectiveness of our method. The more detail experimental setup and results are shown in supplementary materials.

Dataset. We select the restored picture of "Girl With a Pearl Earring" in resolution 13240 × 15500 as dataset for 1D signal fitting and 2D image regression. For the former, we sample random row pixel values in the image as target signals and divide them into training and test sets in an alternating 1:1 pattern. For the latter, we divide the whole image into training and testing sets using a pixel-wise alternating pattern. 3D scene representation and reconstruction task is the primary part of our experimental study, we mainly carry out our experiments on the following two public datasets: NeRF-Synthetic dataset (Mildenhall et al. 2020) and Real-World LLFF dataset (Mildenhall et al. 2019). We use NeRF-Synthetic dataset in resolution 800 × 800, and down-scale LLFF dataset to resolution 1008 × 756.

1D and 2D Comparison

1D Signal Fitting. As shown in Fig. 1c, the qualitative experiment is conducted to illustrate that the localization property of wavelet encoding can enhance the convergence of neural networks towards high-frequency information. We employ 30 dimension Fourier function as global encoding, and add 30 uniformly distributed wavelet functions with Fourier function as our WPE, to embed the input coordinate x before putting into training. The experiment result demonstrates that local wavelet encoding can significantly enhance the convergence capability of coordinate networks. Note that the HFP module is not used here, which verify that using wavelets embedding alone can enhance the signal fitting performance.

2D Image Regression. Fig. 3 shows the qualitative and quantitative results of 2D image regression, and the local enlargements are shown to compare the fine details. We train an MLP to regress from a 2D input pixel coordinate to the corresponding RGB value of an image. The High-Frequency Perception method in 2D task follows the same principles as 3D HFP described in above section. It involves performing high-frequency decomposition on image and dividing the 2D plane into grids. Adding two-dimensional wavelets to the plane based on the distribution of high-frequency information on a 2D spatial grid. We compared our method with several global methods: PE, FFN and WIRE. Experimental results prove that the proposed method can significantly outperform global methods and preserve more details of 2D image.

3D Comparison

We evaluate our method in 3D NeRF task and compare it with existing methods that aim to solve the "spectral bias" problem, including global encoding methods PE, FFN and periodic activation method WIRE. We quantitatively compare the models in terms of PSNR, SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018), thus demonstrate the superiority of our WPE method over previous methods. The LPIPS results vary when different network models are used, and we employ the "alex" network for computation.

Method	View Synthesis Quality		
	PSNR ↑	SSIM ↑	LPIPS ↓
PE	28.07	0.91	0.079
FFN	27.93	0.91	0.081
WIRE	28.49	0.93	0.071
Ours	**28.81**	**0.94**	**0.067**

Table 2: **Quantitative Results on the NeRF-Synthetic Dataset.** Specific experimental results are shown in the Supplementary Material.

Configures			View Synthesis Quality		
PE	WE	HFP	PSNR ↑	SSIM ↑	LPIPS ↓
✓			27.08	0.82	0.110
	✓	✓	27.31	0.83	0.102
✓	✓		27.14	0.82	0.113
✓	✓	✓	**27.89**	**0.85**	**0.097**

Table 3: **Quantitative Results of Ablation Study.** We conduct the ablation study on LLFF dataset in 3D novel view synthesis task. Where PE denotes global Positional Encoding, WE denotes Wavelet Encoding, HFP denotes High-Frequency Perception method. The best results are in bold.

Notably, to achieve a fair comparison, we set all methods under the same setting and conduct experiments instead of directly quoting the results from original papers.

Fig. 4 shows the optimization efficiency in "Lego" scene of NeRF-Synthetic dataset. In the early stage of training, global methods like WIRE converge faster. However, as the training progresses, our WPE method eventually surpasses others and achieves the highest PSNR value. We attribute it to the fact that local wavelet embedding primarily handles detailed information, which does not have a significant impact on metrics in the early-training stage. And introducing higher-dimensional embedding also makes it more challenging for MLP to converge at first. But in later stage of training, the network benefits from the complete reconstruction of better details, resulting in higher PSNR values.

Fig. 5 presents the qualitative results and enlarged details of the synthetic novel views on Real-World LLFF dataset. And Fig. 6 showcases the reconstruction detail in Nerf-Synthetic Dataset. Both trained for same iterations, our proposed WPE method can reach the best implicit neural representation and reconstruction results than global processing methods PE, FFN and WIRE. More importantly, the proposed method is capable of achieving better reconstruction details in regions with rich high-frequency information, which enhances the representation capability of implicit neural radiance fields for capturing fine details.

Table. 1 shows the quantitative results of views synthesis quality on Real-World LLFF dataset. It shows that our proposed WPE method achieved the best performance in terms of the evaluation metrics. Table. 2 presents the quantitative results on NeRF-Synthetic dataset, we compute the average experimental results for each sequence and compare them with other global methods, and specific metrics of each case are shown in the supplement material. The quantitative results demonstrate that by incorporating our method, the performance of the 3D coordinate-based implicit scene representation and novel view synthesis can be improved. Note that in scenes like "Horns", where high-frequency information is lacking, our method does not make significant improvement. This is because global methods are already sufficient for simple scenes, while it demonstrates the effective improvement of our method in fine details representation.

Ablation Study

We conduct our ablation study experiments to validate the effectiveness of our proposed method on the Real-World LLFF dataset. The evaluation includes the following variants: (1) the baseline model with only global Positional Encoding method; (2) the baseline model with Wavelet Encoding and High-Frequency Perception method; (3) the baseline model with global Positional Encoding and uniformly distributed Wavelet Encoding without High-Frequency Perception method; (4) the baseline model with our proposed Wavelet-Positional Encoding and High-Frequency Perception method; To ensure fairness in the comparison, we added 50 wavelets in (3) and (4), 70 wavelets in (2) to keep same embedding dimension. Table. 3 shows the quantitative results of the ablation study, indicating the effectiveness of our proposed WPE method. As Table. 3, WE with HFP can represent signal well even without global PE encoding, though it requires higher wavelet dimensions and relatively slower convergence. This is because the use of wavelets exclusively is more capable of representing local high-frequency components, but rises the difficulty of convergence for low-frequency global components, which is why we retain PE in our methods. Another thing worth noting is that simply adding wavelet transforms uniformly can not significantly improve the performance either, this is because, for 3D NeRF MLP models, the dimension of PE is already sufficiently high. Merely increasing the dimension of the encoding function can't further enhance performance. This also explains the observation that methods such as FFN exhibit better performance improvements in 2D tasks compared to 3D tasks. We also conduct experiments on different representation frameworks to demonstrate the effectiveness of our method, details are in the Supplementary Material.

Conclusion

In this paper, we address the "spectral bias" difficulty in Implicit Neural Representations, and effectively improve the detail quality of signals learned by coordinate-based MLP. Specifically, we propose the novel High-Frequency Perception method to obtain the frequency distribution of target signal space, and then propose a novel embedding method called Wavelet-Positional Encoding to simultaneously perform global and local embedding on input coordinates by incorporating frequency distribution. Our method is simple to implement without modifying existing frameworks, which means it can be easily integrated into various INR tasks. Experiment results demonstrate the effectiveness of our method in INR models, showcasing its significant complete representation capacity especially for details.

Acknowledgements

This research was supported by National Key R&D Program of China under Grant 2023YFB4704404, in part by NSFC 62088101 Autonomous Intelligent Unmanned Systems, and in part by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F030001.

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Peter Hedman, R. M.-B.; and Srinivasan, P. P. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Barrona, 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Chaurasia, G.; Duchene, S.; Sorkine-Hornung, O.; and Drettakis, G. 2013. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics*, 32(3): 1–12.

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.

Chen, J.; Ji, B.; Zhang, Z.; Chu, T.; Zuo, Z.; Zhao, L.; Xing, W.; and Lu, D. 2023. TeSTNeRF: Text-Driven 3D Style Transfer via Cross-Modal Learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 5788–5796.

Chen, Z.; and Zhang, H. 2019. Learning Implicit Fields for Generative Shape Modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5939–5948.

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.

Gao, Z.; Dai, W.; and Zhang, Y. 2023. Adaptive Positional Encoding for Bundle-Adjusting Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3284–3294.

Genova, K.; Cole, F.; Sud, A.; Sarna, A.; and Funkhouser, T. 2020. Local Deep Implicit Functions for 3D Shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4857–4866.

Germann, M.; Popa, T.; Keiser, R.; Ziegler, R.; and Gross, M. 2012. Novel-View Synthesis of Outdoor Sport Events Using an Adaptive View-Dependent Geometry. *Computer graphics forum*, 31(2): 325–333.

Hertz, A.; Perel, O.; Giryes, R.; Sorkine-Hornung, O.; and Cohen-Or, D. 2021. SAPE: Spatially-adaptive progressive encoding for neural optimization. *Advances in Neural Information Processing Systems*, 34: 8820–8832.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in Neural Information Processing Systems*, 31: 8571–8580.

Kajiya, J. T.; and Herzen, B. P. V. 1984. Ray tracing volume densities. *ACM SIGGRAPH Computer Graphics*, 18(3): 165–174.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, 38(4): 1–14.

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.

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

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 the IEEE Conference on Computer Vision and Pattern Recognition*, 165–174.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.

Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.; Bengio, Y.; and Courville, A. 2019. On the Spectral Bias of Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 5301–5310.

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.

Ramasinghe, S.; and Lucey, S. 2023. A Learnable Radial Basis Positional Embedding for Coordinate-MLPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(2), 2137–2145.

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

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-From-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit Neural Representations with Periodic Activation Functions. *Advances in Neural Information Processing Systems*, 33: 7462–7473.

Sitzmann, V.; Zollhoefer, M.; and Wetzstein, G. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. *Advances in Neural Information Processing Systems*, 32.

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. *Advances in Neural Information Processing Systems*, 33: 7537–7547.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Łukasz Kaiser; and Polosukhin, I. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xu, M.; Zhan, F.; Zhang, J.; Yu, Y.; Zhang, X.; Theobalt, C.; Shao, L.; and Lu, S. 2023. WaveNeRF: Wavelet-based Generalizable Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18195–18204.

Zhang, K.; Riegler, G.; Snavely, N.; and Koltun, V. 2020. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv preprint arXiv:2010.07492*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhou, X.; He, Y.; Yu, F. R.; Li, J.; and Li, Y. 2023. RePaint-NeRF: NeRF Editting via Semantic Masks and Diffusion Models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 1813–1821.