INCODE: Implicit Neural Conditioning with Prior Knowledge Embeddings

Supplementary Material

Amirhossein Kazerouni [1] Reza Azad [2] Alireza Hosseini [3] Dorit Merhof [4,5] Ulas Bagci [6]
[1] School of Electrical Engineering, Iran University of Science and Technology, Iran
[2] Institute of Imaging and Computer Vision, RWTH Aachen University, Germany
[3] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Iran
[4] Faculty of Informatics and Data Science, University of Regensburg, Germany
[5] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
[6] Department of Radiology, Northwestern University, Chicago, USA
{amirhossein477, rezazad68}@gmail.com, {arhosseini77}@ut.ac.ir
{dorit.merhof}@ur.de, {ulas.bagci}@northwestern.edu
https://xmindflow.github.io/incode

A. Experimental Results

In this section, we broaden our experimental scope to encompass a more comprehensive comparison between our approach and state-of-the-art (SOTA) methods. We have demonstrated that the inherent simplicity of INCODE contributes to enhanced performance compared to its counterpart SOTA methods, specifically in terms of expressiveness and representation capacity. These findings underscore the efficacy of our approach in pushing the boundaries of INR networks and facilitating their applicability across diverse domains. We now present additional visualizations that distinctly show the advantage of our approach.

A.1. Image representation

As depicted in Figure 1 and Figure 5, it is evident that INCODE achieves superior qualitative and quantitative performance. Particularly in Figure 1, INCODE exhibits an approximate accuracy improvement of +2.98 dB and +4.59 dB compared to FFN [5] and WIRE [3], respectively. The zoomed-in image distinctly illustrates INCODE's ability to grasp intricate details of the Eiffel Tower. In contrast, ReLU+P.E. and MFN [1] yield blurry outcomes, while Gauss [2] displays slight color alteration, although it captures certain intricate features. Gauss also struggles to recognize the orange object positioned at the tower's center. Likewise, SIREN [4] fails to capture the full complexity of the tower's structure, leading to a smoothed and blurred representation.

Additionally, Figure 5 presents a challenging image with intricate patterns, posing a challenge for representation. Notably, INCODE and FFN emerge as the sole methods achieving a PSNR value over 30 dB, with INCODE exhibiting a +1.56 improvement over the second-ranking FFN. As evidenced in the zoom-in image, ReLU+P.E. expectedly yields a blurred output, given the inherent properties of its ReLU activation function. Interestingly, WIRE and Gauss encounter difficulty in precisely grasping the image's color characteristics, leading to slight color differences. While MFN effectively addresses this color challenge, it falls short in capturing the image's intricate details, particularly its edges.

Overall, our study shows that INCODE excels in image representations. It consistently outperforms other methods across various images, even with intricate patterns. This success is due to INCODE's ability to capture intricate details. While alternative methods faced challenges in representing complex patterns, colors, or high-frequency information, INCODE exhibited competence in addressing these challenges. Thus, our findings highlight INCODE as one of the optimal choices for robust and superior image representation.

A.2. Audio representation

We present audio representation visualization results along with its error maps in Figure 2. These visualizations help to understand the strength of our approach. We have provided a detailed analysis of these results in the main section of the paper to ensure a comprehensive understanding of our findings. In terms of sound playback quality, Gauss introduces a noticeable squeak-like sound that accompanies the main audio. With ReLU+P.E., noise dominance becomes more pronounced, making it difficult to discern the original sound. While employing SIREN, some moments

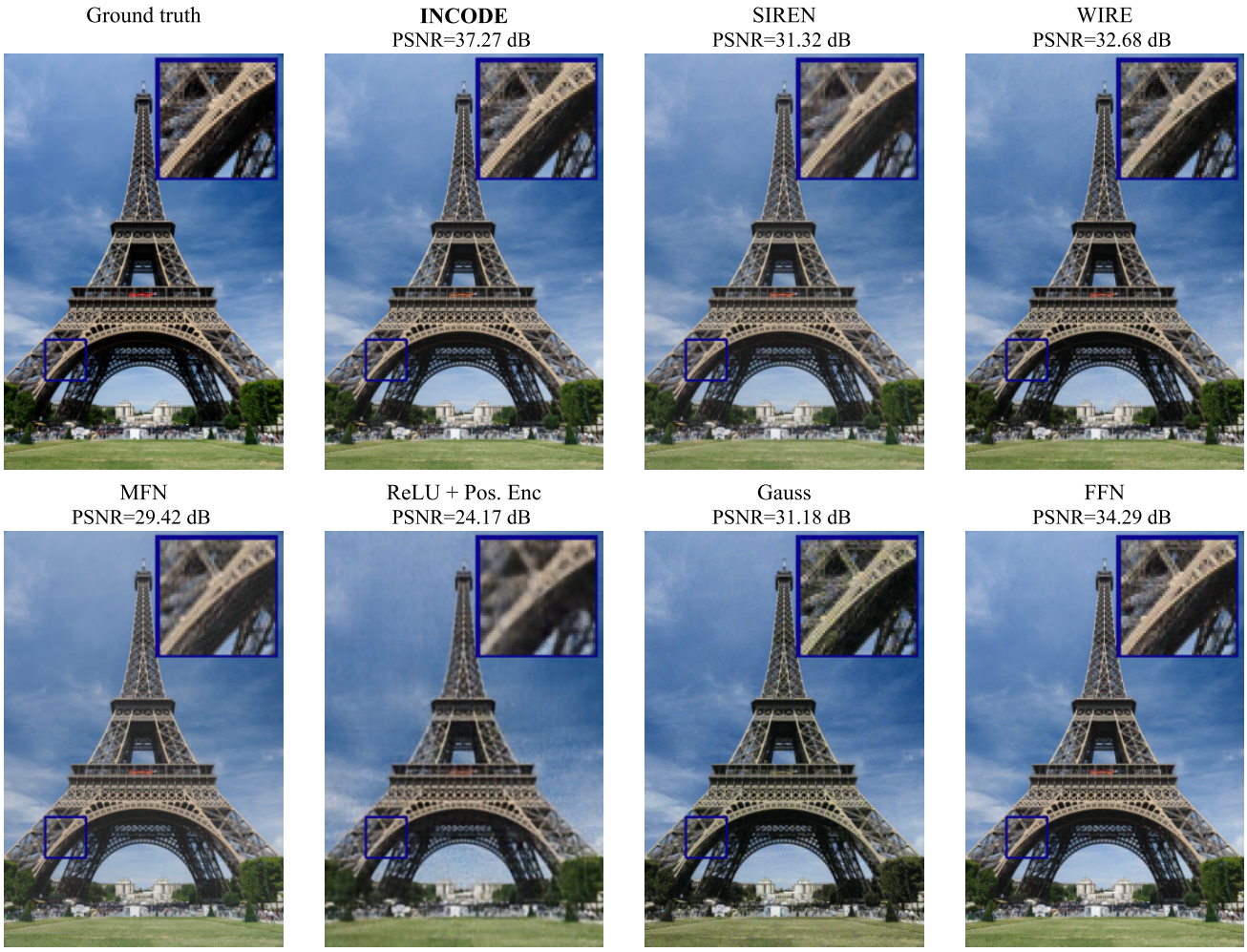

Figure 1. **Image representation:** Comparison of INCODE with SOTA methods.

are marred by bothersome noise, as indicated by the error map. However, INCODE significantly outperforms these methods by having notably less noise interference. This aspect positions INCODE as a favorable choice for encoding audio data with improved quality.

A.3. Super resolution

To illustrate the efficacy of our approach in the super-resolution task, we have included a visual comparison of 4× super-resolution in Figure 6. From a quality perspective, INCODE produces sharper results with finer details in the butterfly's wing, while the blurred outcomes of SIREN, FFN, Gauss, and ReLU+P.E. are evident, even though the quantitative values are relatively close. This visual comparison supports our quantitative findings in table 1 (see the main paper) and affirms INCODE's proficiency in super-resolution tasks, where it offers better quality when performing upsampling.

A.4. Computed Tomography (CT) reconstruction

Under-measurement in CT samples results from a range of factors that reduce the accuracy of the imaging process. Artifacts, stemming from issues like patient movement during scanning, metallic objects causing beam distortion, and equipment calibration problems, contribute to discrepancies. INRs address these concerns and solve this inverse problem by leveraging their inductive bias. We investigate the impact of varying the number of measurements (ranging from 50 to 400, with increments of 50) as shown in Figure 3. Notably, SIREN, WIRE, and ReLU+P.E. yield consistent results across all measurements. Particularly, WIRE excels in CT reconstruction with 50 measurements; however, increasing the data information in such models doesn't enhance their performance, indicating saturation. In contrast, INCODE exhibits considerable improvement as measurements increase from 100 to 400, showcasing the effectiveness of incorporating deep prior information. Notably, INCODE with 150 measurements outperforms all nonlin-

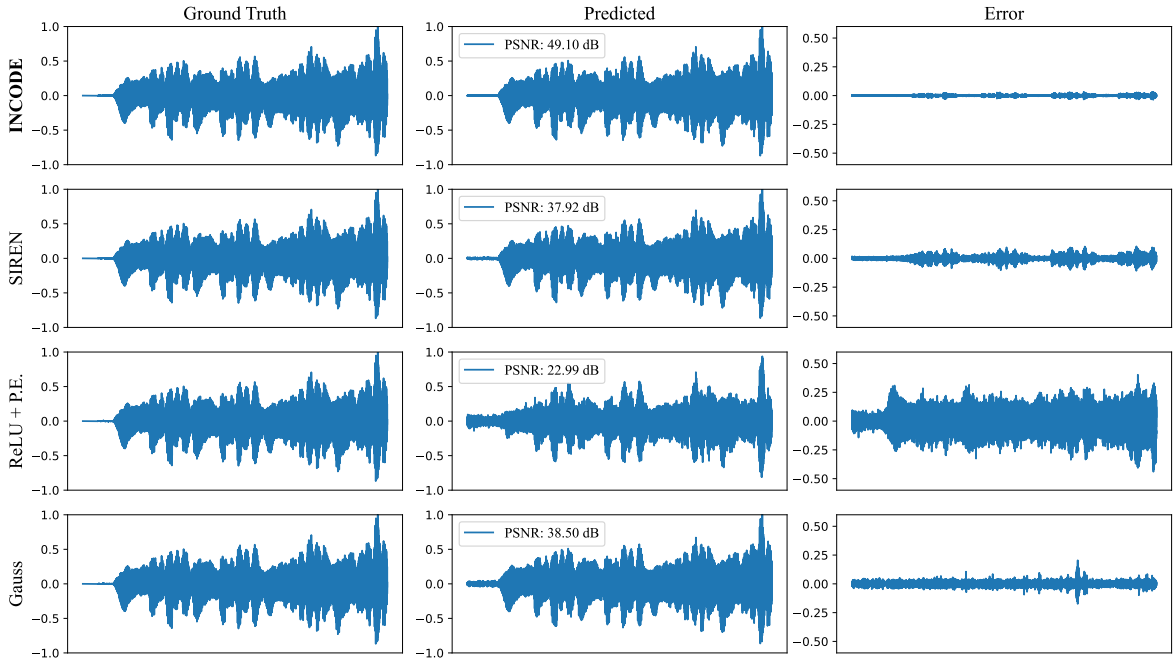

Figure 2. **Audio representation:** We compare INCODE with SOTA methods for audio representation. In the third column, we display the reconstruction error.

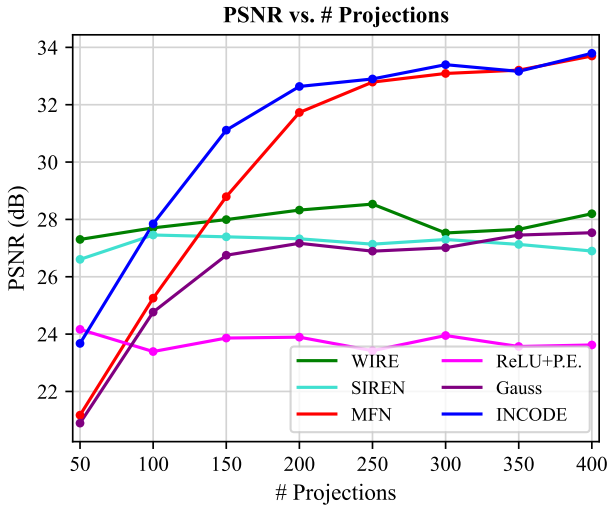

Figure 3. INCODE vs. SOTAs in CT reconstruction across different numbers of projections.

earities in the full range of projection numbers, except for MFN, which closely competes after reaching 200 projections and performs the second best. These findings acknowledge the robustness and power of INCODE in addressing under-measurement challenges within CT reconstruction.

A.5. Inpainting

Image inpainting poses a formidable challenge as models are tasked with predicting entire pixel values based on only a fraction of trained pixel data. The high capacity of INR provides the opportunity to accomplish this inverse problem challenge. The strong prior ingrained within the space of INR functions paves the way for applications like inpainting from limited observations, where it uses the learned representation of the trained model to predict inpainting missing values. Our approach involves randomly sampling 20% of the pixels and then employing the model's learned representation to predict the missing pixels. The comparison result is shown in Figure 4. As observed in other tasks, INCODE's power in capturing intricate features, particularly edges, stands out compared to other methods that tend to yield blurred outcomes. While a modest +0.38 dB improvement is noted compared to SIREN, the visual presentation demonstrates that SIREN, much like ReLU+P.E., struggles to comprehensively capture high-frequency details.

A.6. Neural radiance fields

In our approach, we followed a strategy akin to [3], making use of the publicly available torch-ngp package [6, 7] to train the NeRF model. Our NeRF architecture encompasses two main networks: one for predicting sigma (σ) and the other for determining color (RGB). These networks are constructed as 4-layer MLPs, each with 182 hidden features.

Additionally, we introduced two harmonizer networks, one for the sigma network and another for the color network. These harmonizers employ 4-layer MLPs, featuring 32, 16, 8, and 4 nodes, with each layer followed by Lay-

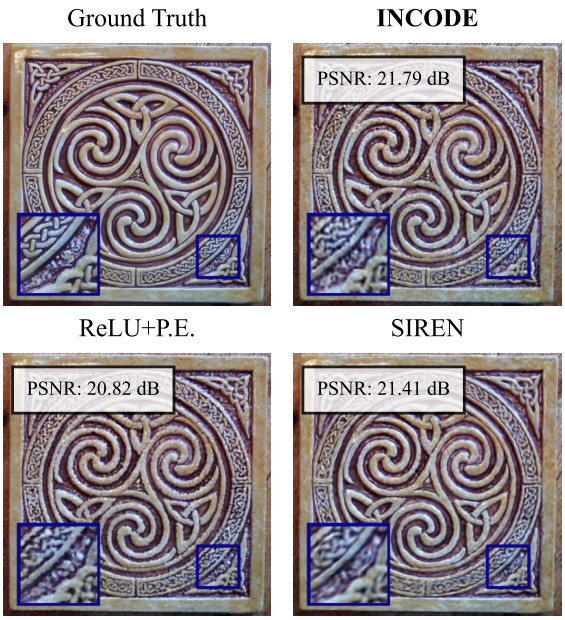

Figure 4. **Image inpainting:** Comparison of INCODE with SOTA methods.

erNorm and the SiLU activation function. They receive a latent code and condition their corresponding composer networks, which are initialized similarly to the denoising task.

To generate the latent code, we utilized a truncated ResNet34 model at its fifth layer, followed by adaptive average pooling. During training, a single random image from the training dataset was used, and for testing and validation, we again employed one random training image. The color MLP took positional coordinates (x, y, z) and direction parameters (θ, ϕ) as inputs, while the sigma MLP solely required positional information.

For our experimental results, depicted in Figure 7, we utilized a Lego dataset comprising 100 training images, each downsampled by 1/2 to 400×400 dimensions, for training the NeRF. Subsequently, we evaluated the model's performance on an additional 200 images. Training of the NeRF models was conducted on an A-100 GPU with 20 GB of memory. Throughout training, we used learning rates of 3×10^{-4} for INCODE, 3×10^{-4} for SIREN, 6×10^{-4} for WIRE, 3×10^{-3} for Gauss, and 1×10^{-2} for ReLU+P.E. The learning rate is decreased to $0.1\times$ initial value over a total of 3000 training epochs to achieve their optimal outputs. Additionally, we set omega (ω_0) to 40 for INCODE, SIREN, and WIRE, and sigma (s_0) to 40 for WIRE and Gauss. Apart from ReLU, we did not use positional encoding for other nonlinearities to highlight their individual capabilities.

As shown in Figure 7, our approach achieves a +0.16 dB improvement over SIREN and a +0.79 improvement compared to WIRE. Qualitative results also demonstrate a superior performance of INCODE compared to SOTA models. Notably, INCODE excels in capturing fine-grained details and information. For instance, it effectively captures intricate features such as the middle black connector in the loader, while SIREN failed to learn. Also, INCODE outperforms other methods like WIRE, ReLU+P.E., and Gauss, which exhibit blurred and smooth results in comparison.

B. Experimental Analysis

B.1. Convergence rate comparison

We analyze the convergence rate of INCODE in comparison to other methods across three distinct representation tasks: image, occupancy volume, and audio, as depicted in Figure 8. The data used for each task corresponds to the respective domain in the main paper. Remarkably, INCODE consistently showcases accelerated convergence compared to SOTA architectures. This expedited convergence is most pronounced in the audio domain, where a substantial gap between SIREN and INCODE is evident. Leveraging its robust approximation capacity, INCODE achieves fast convergence with high fidelity, rendering it an apt choice for representing different signals.

B.2. Impact of depth and width of the network

The analysis of the network's depth and width are presented in Figure 9, which sheds light on the impact of architectural parameters in shaping the performance of INCODE. By systematically varying the number of hidden layers and their width, we gain insights into the trade-off between model complexity and approximation accuracy.

In the left figure, we vary the network's depth from 2-layer MLP to 6-layer, while keeping the width constant at 256. Notably, INCODE exhibits competitive performance compared to other methods in lower layers. However, as the network deepens, INCODE distinctly outperforms FFN, demonstrating its capacity to effectively capture more intricate information with increasing model depth. Shifting to the right figure, we explore the effect of hidden features by adjusting the network's width from 64 to 320, in increments of 64, while maintaining a 5-layer MLP. The trend depicted in the plot accentuates INCODE's remarkable performance, showcasing a steep ascent. Throughout the spectrum of hidden feature counts, INCODE consistently outperforms other SOTA methods. This observation highlights INCODE's proficiency in capturing broader patterns as the width of the network expands, underlining its versatility and ability to adapt to varying levels of complexity.

C. Experimental details

In all experiments, we employed a 5-layer MLP with 256 hidden features for all architectures. However, for WIRE, we followed their recommended structures as outlined in

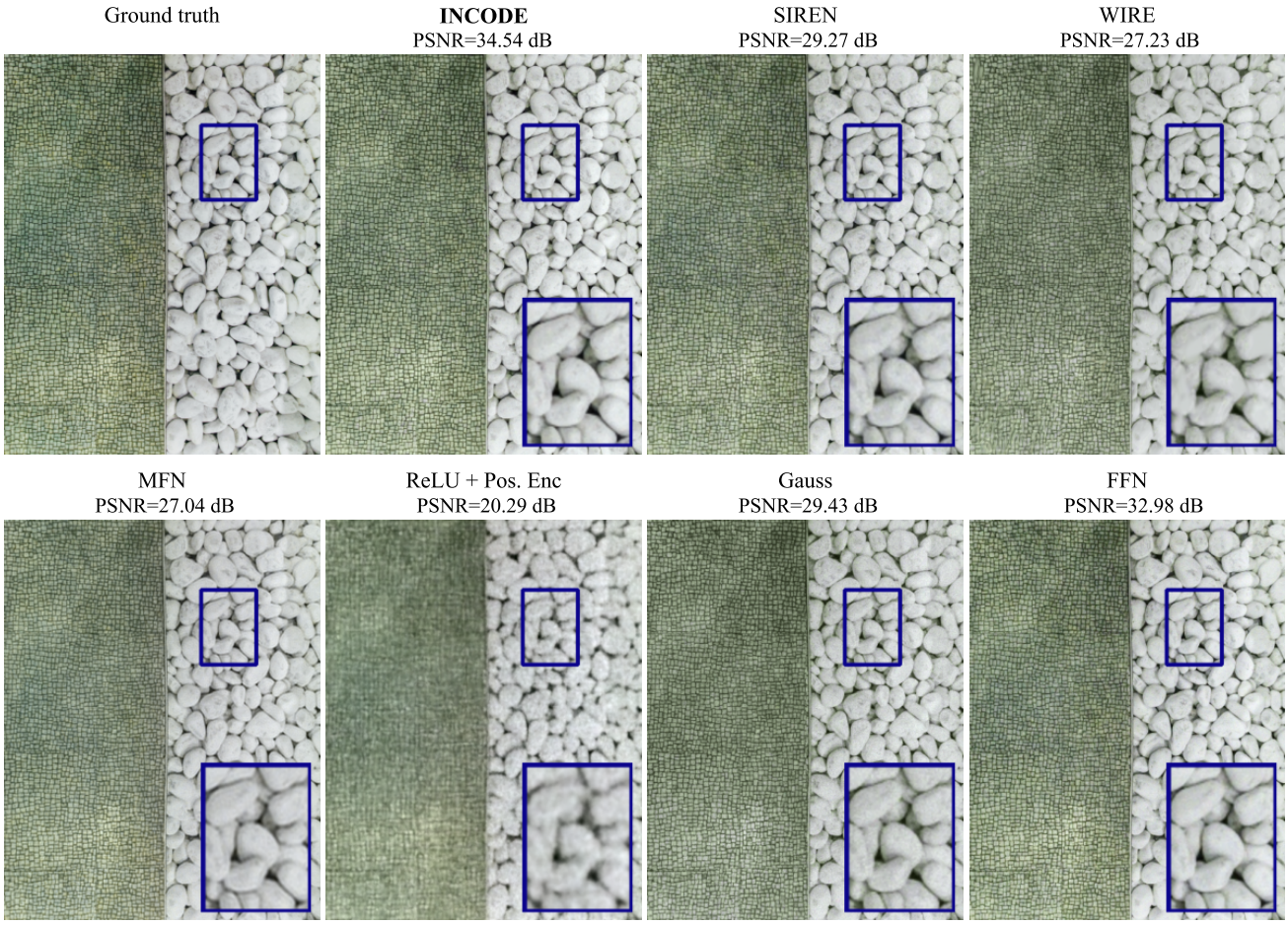

Figure 5. **Image representation:** Comparison of INCODE with SOTA methods.

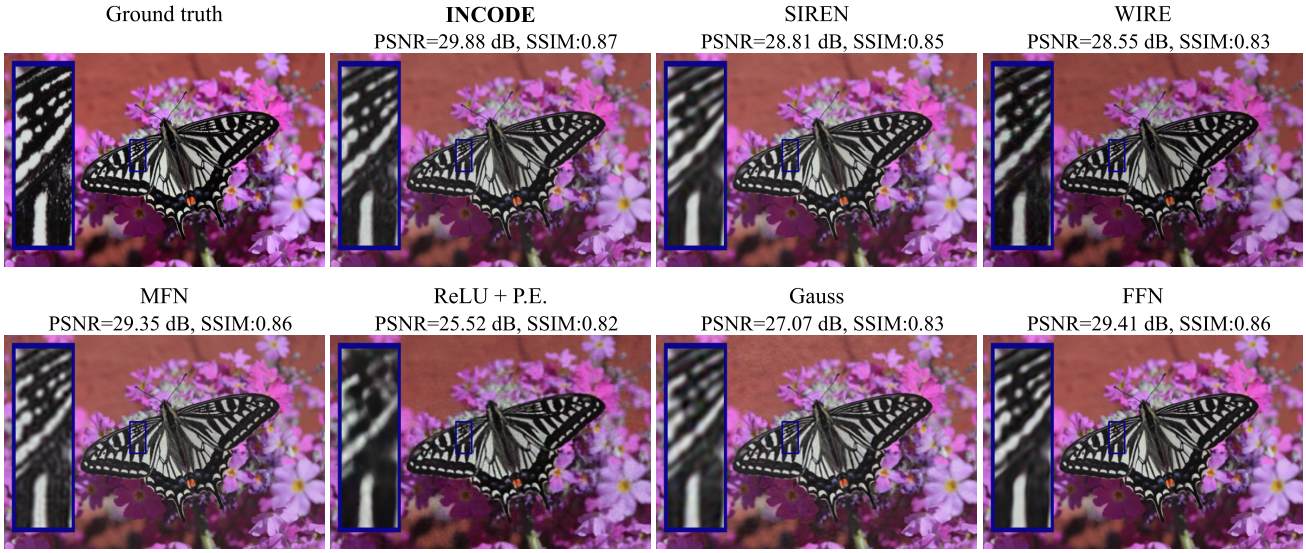

Figure 6. **Super Resolution.** Results of a 4× single image super-resolution using various approaches.

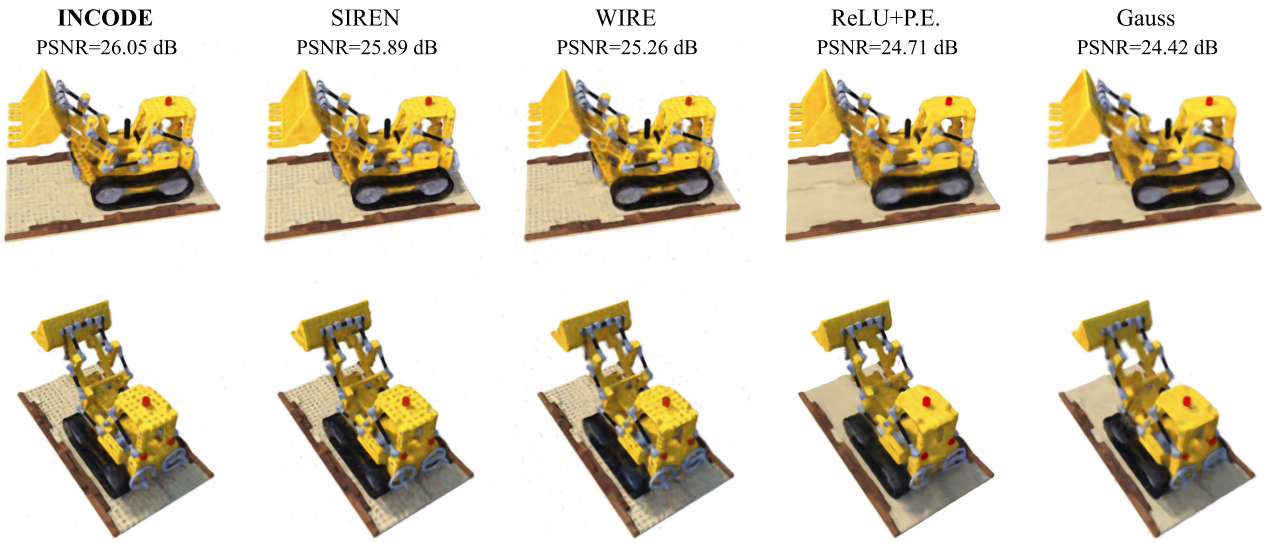

Figure 7. **Neural Radiance Fields:** The figure presented above illustrates rendered images generated by a neural radiance field using different methods. Notably, INCODE consistently outperforms all other methods in terms of visual reconstruction quality, highlighting its robust feature representation.

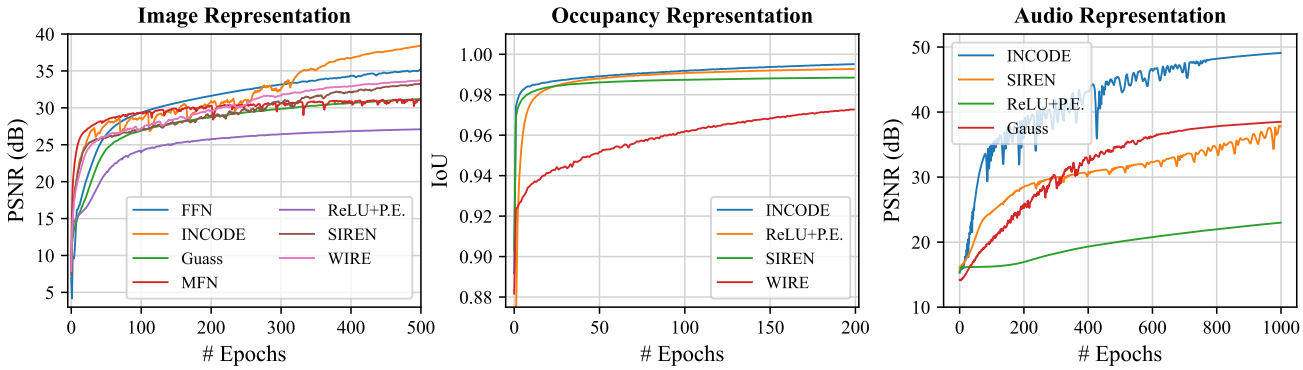

Figure 8. **Convergence rates in different representations:** Explore the convergence rates of Image, Occupancy volume, and Audio representations.

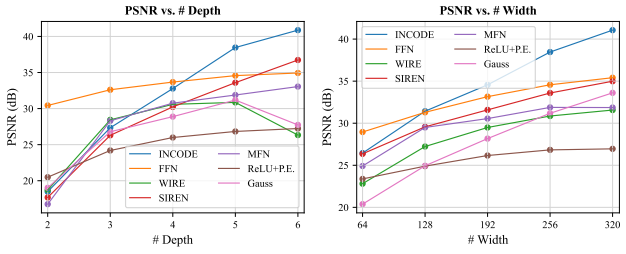

Figure 9. **Impact of network depth and width:** Explore the influence of network depth and width on performance.

their paper to achieve optimal performance. Specifically, for image-based tasks, we used a 4-layer MLP with $s_0 = 30$ and $\omega_0 = 20$, featuring 300 hidden features. For the occupancy task, we utilized a 4-layer MLP with 256 hidden features, alongside $s_0 = 40$ and $\omega_0 = 10$. In the case of CT reconstruction, we employed a 5-layer MLP with 256 hidden features and set $s_0 = 10$ and $\omega_0 = 10$. Lastly, for the denoising task, we opted for the same architecture as the image representation and for $s_0 = 4$ and $\omega_0 = 4$. In FFN, a mapping input size of 256 is utilized, for instance, to map image coordinates from 2 to 512, and the parameter $\mathcal{B}$, a random Gaussian matrix, is scaled by a factor of 10. We configured the value of s_0 for the Gauss model as follows: $s_0 = 30$ for image representation, $s_0 = 100$ for audio representation, and $s_0 = 10$ for the inverse problem tasks. In addition, we utilized the same initial parameters as described for INCODE in the case of SIREN.

References

[1] Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. Multiplicative filter networks. In *International Conference on Learning Representations*, 2021. 1

[2] Sameera Ramasinghe and Simon Lucey. Beyond periodicity: Towards a unifying framework for activations in coordinate-mlps. In *European Conference on Computer Vision*, pages 142–158. Springer, 2022. 1

[3] Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18507–18516, 2023. 1, 3

[4] 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

[5] 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

[6] Jiaxiang Tang. Torch-ngp: A pytorch implementation of instant-ngp, 2022. 3

[7] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. *Advances in Neural Information Processing Systems*, 35:14798–14809, 2022. 3