SUBTRACTIVE MODULATIVE NETWORK WITH LEARNABLE PERIODIC ACTIVATIONS
SUPPLEMENTARY MATERIALS

Tiou Wang[*]*, Zhuoqian Yang*[†]*, Markus Flierl*[*]*, Mathieu Salzmann*[†]*, Sabine Süsstrunk*[†]

[*]School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden
[†]School of Computer and Communication Sciences, EPFL, Switzerland
[*]{tiou, mflierl}@kth.se, [†]first.last@epfl.ch

ABSTRACT

In this document we supply theoretical analysis, details of implementation and additional experiment results. Finally, we discuss limitations of this work.

1. THEORETICAL ANALYSIS

In this analysis we compare additive and multiplicative modulation from spectral and local perspectives. The primary advantage of multiplication lies in its intrinsic ability to generate new frequency components. When two signals, $s_1 = \sin(\omega_1 z)$ and $s_2 = \sin(\omega_2 z)$, which are representative of features in our network, are multiplicatively combined, the result is:

$$s_1 \cdot s_2 = \frac{1}{2}\left[\cos((\omega_1 - \omega_2)z) - \cos((\omega_1 + \omega_2)z)\right]. \quad (1)$$

This operation explicitly creates sum $(\omega_1 + \omega_2)$ and difference $(\omega_1 - \omega_2)$ frequencies. In contrast, an additive combination, $s_1 + s_2$, merely results in a linear superposition of the existing frequencies. This capacity for harmonic generation is fundamental for representing the intricate details of natural signals. Therefore, to progressively enrich the signal's spectrum, multiplicative modulation is the theoretically superior choice for deeper interactions within the network. Let the input to a given hidden layer be denoted as $x \in \mathbb{R}$. Define two modulation schemes as follows:

- **Additive modulation:**

$$f_{\text{add}}(x) = \sin(x + \beta), \quad \beta \in \mathbb{R}. \quad (2)$$

- **Multiplicative modulation:**

$$f_{\text{mul}}(x) = \sin(\gamma x), \quad \gamma \in \mathbb{R}. \quad (3)$$

1.1. Spectral Analysis

Consider the sine function's Fourier transform to evaluate the spectral properties of these two modulation schemes.

For additive modulation, We have:

$$f_{\text{add}}(x) = \sin(x + \beta). \quad (4)$$

Using trigonometric identities:

$$\sin(x + \beta) = \sin(x)\cos(\beta) + \cos(x)\sin(\beta). \quad (5)$$

Thus, additive modulation corresponds to a linear combination of two fixed basis functions, $\sin(x)$ and $\cos(x)$, and the frequency remains constant at 1.

For multiplicative modulation:

$$f_{\text{mul}}(x) = \sin(\gamma x). \quad (6)$$

This introduces frequency scaling by a factor γ. The function spans the set:

$$\{\sin(\gamma x) \mid \gamma \in \mathbb{R}\}, \quad (7)$$

covering an infinite frequency range as γ varies.

1.2. Local Behavior

To analyze local behavior around $x = 0$, we employ the Taylor expansions of the modulation functions.

The Taylor expansion of additive modulation around $x = 0$ is:

$$f_{\text{add}}(x) = \sin(\beta) + \cos(\beta)x - \frac{\sin(\beta)}{2}x^2 - \frac{\cos(\beta)}{6}x^3 + \mathcal{O}(x^4). \quad (8)$$

In this expansion, the frequency structure depends only on fixed trigonometric functions of β. The power series in x is fixed and not scaled.

The Taylor expansion of Multiplicative modulation is:

$$f_{\text{mul}}(x) = \gamma x - \frac{\gamma^3}{6}x^3 + \frac{\gamma^5}{120}x^5 + \mathcal{O}(x^7). \quad (9)$$

Here, the coefficients depend explicitly on powers of γ. Thus, frequency and scale control are directly introduced via γ.

From both global spectral and local Taylor expansion perspectives, multiplicative modulation enables dynamic control over frequency content and significantly enhances representational capacity. In contrast, additive modulation merely shifts phase and maintains fixed frequency content.

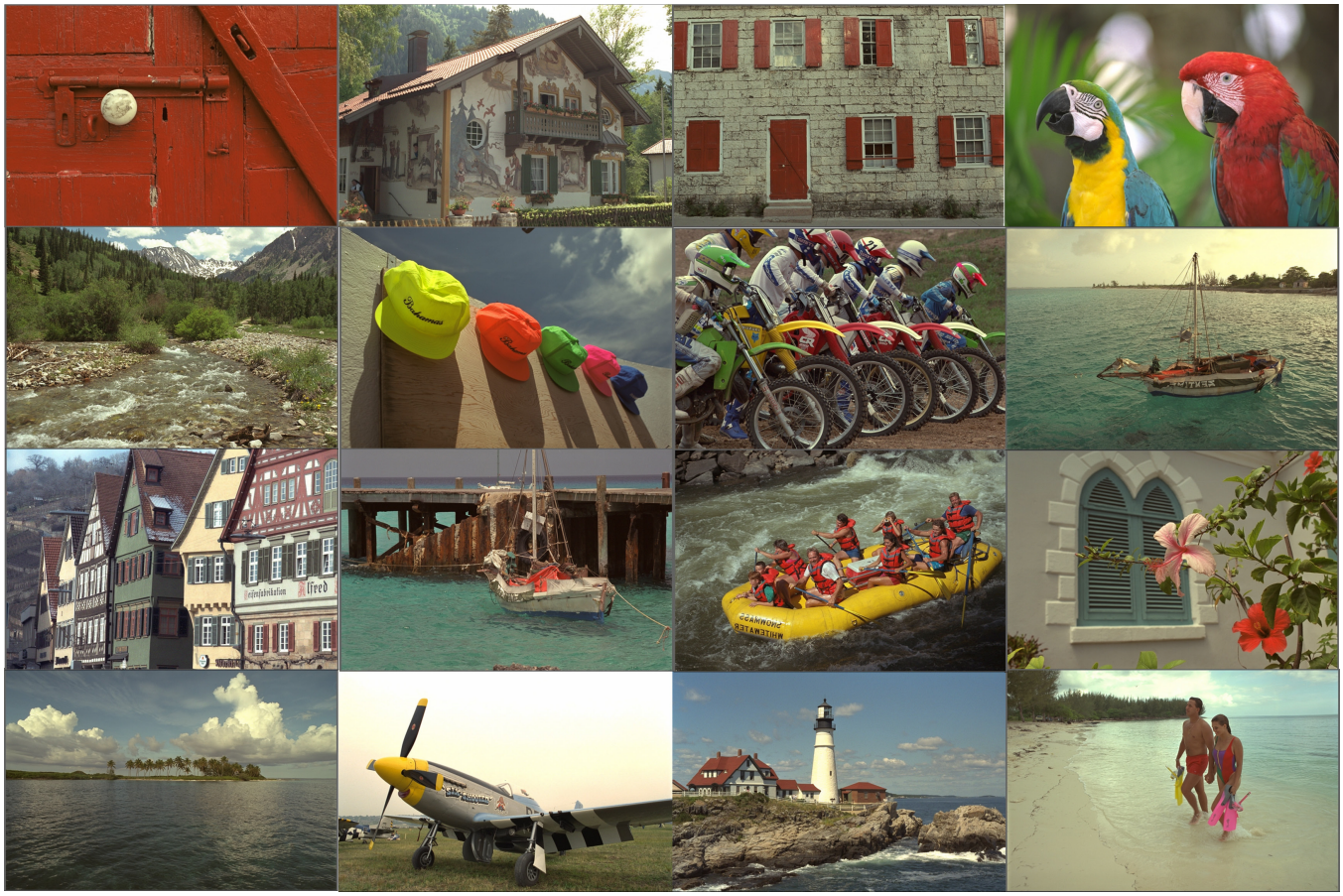

Fig. 1: Qualitative reconstruction results on Kodak dataset.

2. ADDITIONAL RESULTS

2.1. Ablation Study on the Self-Mask Module

We commence our ablation study by quantifying the impact of the final non-linear stage in our network: the Self-Mask module. This parameter-free squaring operation is hypothesized to refine the feature representation before the final linear projection. To validate this hypothesis, we compare the performance of our full SMN against an ablated variant, denoted "SMN w/o Self-Mask", where this final squaring operation is omitted. We conduct this comparison across a range of model capacities by varying the dimensionality of the hidden features. Table 1 presents the average PSNR on the Kodak dataset for models with hidden feature dimensions of 256, 300, and 312.

The results presented in Table 1 provide compelling evidence for the efficacy of the Self-Mask module. Across all tested model capacities, the inclusion of the final squaring operation yields a substantial and consistent performance improvement of over **1 dB** in PSNR. For our main model with 256 hidden dimensions, the gain is a significant 1.15 dB.

Table 1: Ablation study on the Self-Mask module across different model capacities (hidden feature dimensions). Performance is measured by average PSNR (dB) on the Kodak dataset.

Hidden Dim	w/o Self-Mask	Full Model	Δ PSNR
256	40.25	**41.40**	+1.15
300	42.24	**43.33**	+1.09
312	42.67	**43.68**	+1.01

2.2. Validity Analysis in Super-Resolution

To validate our model's capacity to represent an image as a continuous function queryable at arbitrary resolutions, we performed a single-image super-resolution task. The model was trained to fit an image at its native resolution (e.g., 768×512 pixels) and then queried on a denser coordinate grid to produce a super-resolved output without retraining. Our qualitative evaluation, which avoids the methodological bias of quantitative metrics like PSNR towards specific interpolation algorithms, confirms the model's success. Visual results in Figure 2 show both a high-fidelity reconstruction at the origi-

(a) Ground Truth (768×512) (b) Reconstruction (768×512) (c) Super-Resolution (1152×768)

Fig. 2: Demonstration of the super-resolution capability of our learned continuous representation. (a) The original high-resolution ground truth. (b) Our model's high-fidelity reconstruction at the native training resolution. (c) The super-resolved output generated by querying the same model on a denser coordinate grid.

nal resolution and a super-resolved image with markedly improved sharpness over simple upsampling. By synthesizing plausible high-frequency details rather than merely interpolating between pixels, the model demonstrates it has learned a meaningful underlying continuous function of the image content, thereby validating a core premise of implicit neural representations.

2.3. Qualitative Results on Kodak dataset

We provide additional qualitative reconstruction results from the Kodak benchmark dataset in Figure 1.

2.4. Additional Results on 3D

We evluate across all eight scenes of the standard NeRF synthetic dataset. This section presents a detailed, per-scene breakdown of our model's reconstruction fidelity. The quantitative results, measured in PSNR, are detailed in Table 2. All experiments were conducted at a 400×400 resolution. The qualitative results are presented in Figure 3.

Table 2: Per-scene performance of our proposed **PE+SMN** model on the eight scenes of the NeRF synthetic dataset. The final row reports the average PSNR across all scenes, demonstrating strong and consistent performance.

Scene	PSNR (dB) ↑
Lego	36.07
Chair	33.64
Drums	27.09
Ficus	31.51
Hotdog	39.04
Materials	32.56
Mic	32.79
Ship	31.15
Average	**32.98**

The results in Table 2 demonstrate the strong and consistent performance of our SMN architecture across a diverse set of 3D scenes. The model achieves an impressive average PSNR of **32.28 dB**. It delivers exceptional fidelity on scenes with complex geometric details and non-Lambertian surfaces, such as Hotdog (38.31 dB) and Chair (34.45 dB).

This variance in performance across different scenes is expected and provides valuable insights. The model excels on scenes with well-defined objects and complex material properties, while performance is more challenged by scenes containing extremely fine, thin structures and complex occlusions, such as Drums and Ship. Nevertheless, the consistently high performance, especially when compared to the baseline results presented in Table 2, affirms that the architectural principles of our SMN are not tailored to a single type of structure but are broadly effective. These results collectively validate the robustness and general applicability of our modulative filtering approach for high-fidelity neural scene representation.

3. IMPLEMENTATION DETAILS

For training, we employed the Adam optimizer [1] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was initialized to 2×10^{-2} and dynamically managed by a `ReduceLROnPlateau` scheduler, which reduced the rate by a factor of 0.5 whenever the validation loss plateaued for 100 iterations, down to a minimum value of 1×10^{-5}. The network was trained for up to 5,000 iterations, with the entire set of pixel coordinates constituting a single batch in each iteration.

4. LIMITATIONS

A key limitation of this work is the performance of our proposed Learnable Sine Layer as a standalone input encoder for high-dimensional tasks like NeRF. While highly effective for 2D image representation, it struggled to converge when

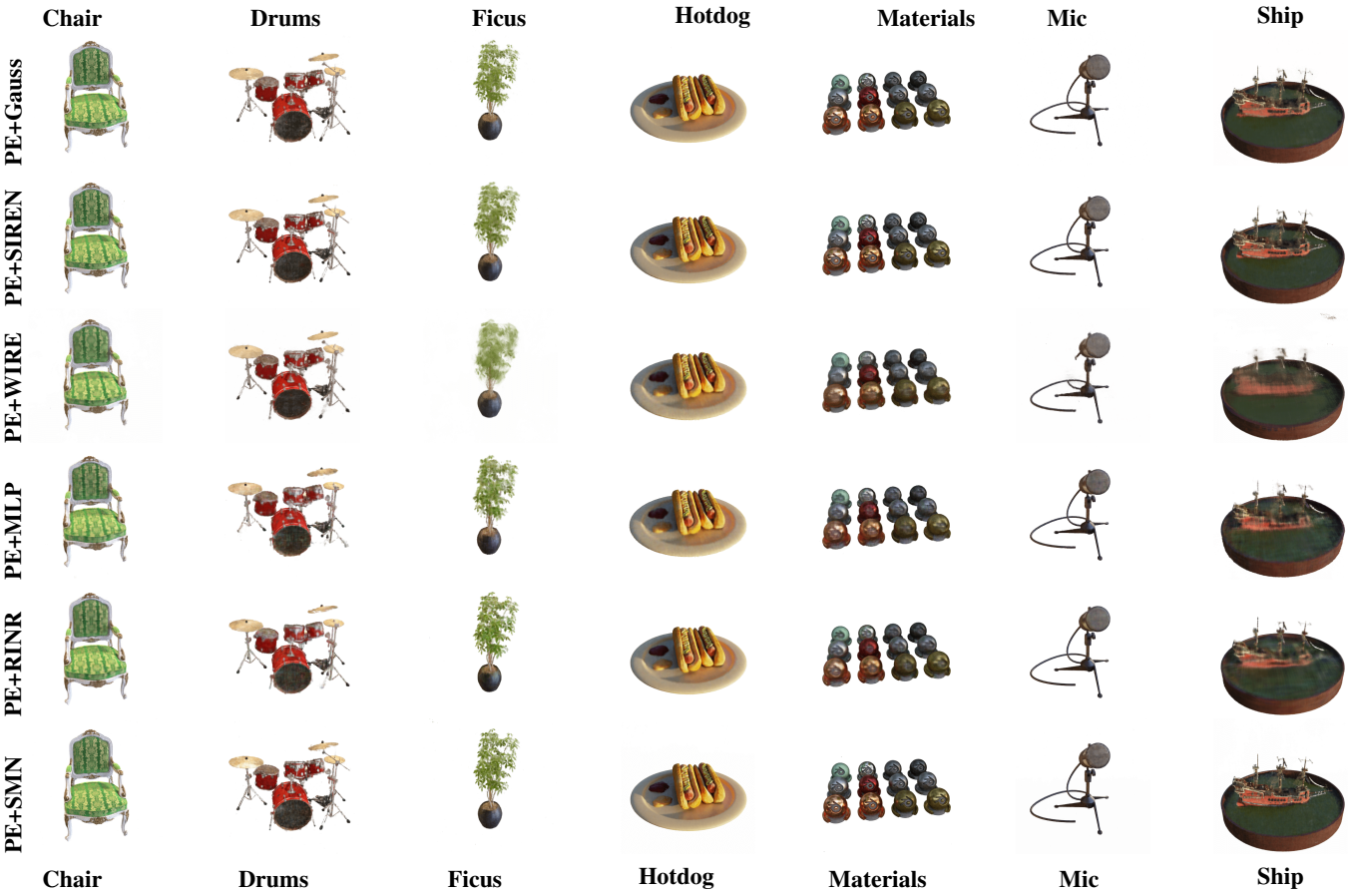

Fig. 3: Qualitative comparison on NeRF Synthetic scenes.

applied to the 6D NeRF input without the aid of a standard Positional Encoding. This suggests that the optimization landscape for discovering an effective high-dimensional frequency basis from scratch is exceptionally challenging. The strong inductive bias from the fixed, exponentially-spaced frequencies in standard Positional Encoding appears crucial for stabilizing training in such complex scenarios, a mechanism our current design does not yet sufficiently replicate.

5. REFERENCES

[1] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.