PINs: Progressive Implicit Networks for Multi-Scale Neural Representations

Zoe Landgraf [1]
Alexander Sorkine Hornung [2] **Ricardo Silveira Cabral** [2]

Abstract

Multi-layer perceptrons (MLP) have proven to be effective scene encoders when combined with higher-dimensional projections of the input, commonly referred to as *positional encoding*. However, scenes with a wide frequency spectrum remain a challenge: choosing high frequencies for positional encoding introduces noise in low structure areas, while low frequencies result in poor fitting of detailed regions. To address this, we propose a progressive positional encoding, exposing a hierarchical MLP structure to incremental sets of frequency encodings. Our model accurately reconstructs scenes with wide frequency bands and learns a scene representation at progressive level of detail *without explicit per-level supervision*. The architecture is modular: each level encodes a continuous implicit representation that can be leveraged separately for its respective resolution, meaning a smaller network for coarser reconstructions. Experiments on several 2D and 3D datasets show improvements in reconstruction accuracy, representational capacity and training speed compared to baselines.

1. Introduction

Neural Implicit Functions are gaining popularity as alternative 2D image and 3D shape representations. Using a simple MLP encoder, these networks approximate a function mapping between spatial coordinates and a quantity of interest such as colour, occupancy or SDF values. They have proven to be very effective at fitting natural images (Tancik et al., 2020; Martel et al., 2021) and 3D shapes (Mescheder et al., 2019; Park et al., 2019; Chabra et al., 2020) and have been applied to various computer vision tasks, including

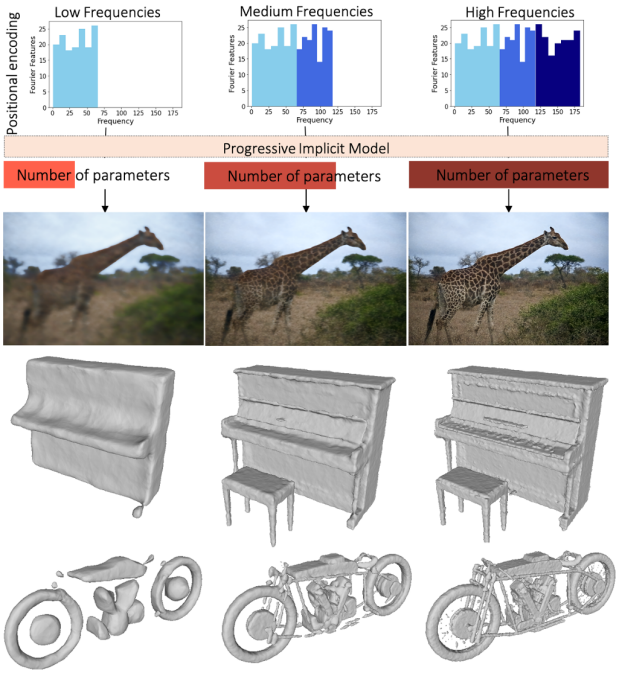

Figure 1. Method overview. The input coordinates are projected by incremental sets of frequency encodings (**top row**) and processed by a hierarchical MLP structure which produces reconstructions in 2D (**row** 2) and 3D (**row** 3 **and** 4) at progressive level of detail.

novel view synthesis (Niemeyer et al., 2020b; Mildenhall et al., 2020) and generative shape modelling (Chen & Zhang, 2019).

However, as simple MLP encoders, Neural Implicit Functions suffer from spectral bias (Rahaman et al., 2019; Basri et al., 2020), which prevents them from learning high frequency detail in signals. Projecting the input onto a manifold containing high frequency components however, reduces the spectral bias (Rahaman et al., 2019). Recent works (Mildenhall et al., 2020; Zhong et al., 2020) have demonstrated this experimentally, mapping the input through a set of sinusoidal functions (*positional encodings*). In a parallel approach, (Sitzmann et al., 2020) use periodic activation functions instead of ReLUs to enable MLPs to learn high frequency content. (Tancik et al., 2020) propose an improved positional encoding based on Fourier Features (Rahimi & Recht, 2007) and show that one can essentially *tune* the range of frequencies that can be learnt by an MLP through

[1]Department of Computing, Imperial College of London, London, UK. Research done during an internship at Meta, Zuerich [2]Meta, Zuerich, Switzerland. Correspondence to: Zoe Landgraf <zoelandgraf@fb.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Figure 2. 2D reconstruction example. **Left:** image residuals predicted by the network. **Right:** progressive reconstruction levels obtained from combining the base image c and the respective image residuals.

the frequencies in the positional encoding. However, MLPs with these encodings struggle to fit signals with a wide frequency spectrum: high frequencies in the encoding enable fitting small detail in a signal, but introduce noise in smoother regions. To improve on that, (Hertz et al., 2021) propose a training scheme based on a spatial mask, which gradually introduces frequency encodings to the network. Whilst achieving compelling results, this method requires spatial masks, which don't scale well and reduce training speed, and its training scheme for unwrapping frequencies is non-differentiable and requires manual hyperparameter tuning.

We show that a continuous function can learn to reconstruct signals with wide frequency ranges and in a progressive fashion (see Figure 1): we partition the representation into hierarchical levels, each receiving a subset of our positional encoding, whereby lower layers process lower frequency sets compared to higher layers. Our final reconstruction is a simple composition of all intermediate levels. This results in a model that is trainable end-to-end and we find that our architecture naturally induces lower layers to learn a coarser reconstruction while higher layers focus on adding details to the scene (see Figure 2), without explicit per-level supervision. Our method provides progressive level of detail while yielding on par or superior reconstruction quality to baselines. In addition, our architecture is modular - for a coarser representation, one can drop the MLPs of higher levels and reconstruct the scene with only a small portion of the parameters. Overall our method provides:

- A multi-scale representation based on hierarchical implicit functions with progressive positional encoding, providing incremental level of detail.
- Improved reconstruction quality, in particular for scenes with a wide frequency spectrum.
- End-to-end trainable model without per-level supervision

2. Related Work

Neural Implicit Functions First introduced as novel shape representations for occupancy (Mescheder et al., 2019) and SDFs (Park et al., 2019), Neural Implicit Functions have become a popular alternative to classical 3D representations. Several works have used them for single-view shape inference (Saito et al., 2019; Liu et al., 2019), shape decomposition (Deng et al., 2020) and instance-aware SLAM systems (Li et al., 2020). They have been shown to support larger scene representations (Chabra et al., 2020; Huang et al., 2020), when composed as a voxelgrid of small implicit functions representing local detail. (Genova et al., 2020) demonstrate that compositional implicit functions also yield improved reconstruction accuracy for smaller shapes. Recent work has explored their use for image synthesis as object-specific implicit fields (Niemeyer & Geiger, 2021), as well as for dynamic scene graphs (Ost et al., 2021). Combined with differentiable volumetric rendering (Niemeyer et al., 2020a), neural implicit representations gave rise to an explosion of novel view synthesis approaches (Mildenhall et al., 2020). Other works address generalisation, conditioning the neural implicit representation on features extracted from images (Yu et al., 2021b) or on latents that encode shape priors (Schwarz et al., 2020; Jang & Agapito, 2021). A few approaches have focused on improving training and rendering speed (Reiser et al., 2021; Yu et al., 2021a) and the representation itself: to address the spectral bias of Neural Implicit Functions based on a simple MLP, periodic activation functions (Sitzmann et al., 2020) and positional encodings (Rahaman et al., 2019; Tancik et al., 2020; Hertz et al., 2021) were proposed. Similarly to (Park et al., 2019), our method is designed as a scene representation, in 2D and 3D. Instead of using a single MLP for scene representation we propose a hierarchical structure of small MLPs composing the scene at progressive level of detail. Similarly to (Hertz et al., 2021), we use Fourier Features as positional encodings, but instead of a spatial frequency mask, we condition each of the MLPs in our hierarchical structure on subsets of these encodings.

Multi-scale Neural Implicit Functions A few recent approaches propose to add hierarchical or multi-scale structure to Neural Implicit Functions; they can be loosely divided into those, that use a form of space partitioning and those, that use a frequency based approach.

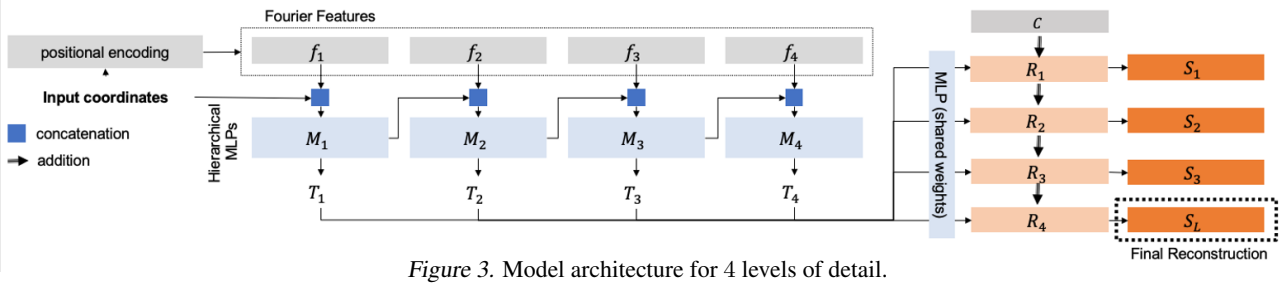

Figure 3. Model architecture for 4 levels of detail.

Of the space partitioning methods, most approaches leverage Octrees and have as primary goal a more efficient rendering or training time (Yu et al., 2021a; Martel et al., 2021; Tang et al., 2021). The method presented by (Chen et al., 2021) generates a multi-scale representation: A feature encoder generates a hierarchy of latent code grids, which represent the scene at different resolutions in a lower dimensional manifold. To reconstruct the scene at a specific resolution, the latent grid is interpolated based on continuous coordinate samples and decoded using a Neural Implicit Function. (Skorokhodov et al., 2021) use a multi-scale representation within their GAN-based image generation model, whereby the image is first generated at a coarse resolution which is then increased incrementally by subsequent layers.

The methods based on frequency decomposition add multi-scale structure in *continuous function space*: (Barron et al., 2021) introduce an integrated positional encoding to the original NeRF model (Mildenhall et al., 2020), which allows the scene to incorporate multi-scale information and has an anti-aliasing effect on the rendered views. (Wang et al., 2021b) use a composition of two SIREN models (Sitzmann et al., 2020), to represent 3D shapes as implicit displacement fields: a smooth base surface is refined with predicted displacements along the surface normal of the base surface. (Lindell et al., 2021) propose a multi-scale representation based on multiplicative filter networks (MFN) (Fathony et al., 2021), a simple linear combination of Fourier or Gabor wavelet functions applied to the input. To reconstruct at multiple resolutions, linear layers are added at different depths of the MFN, to generate intermediate outputs. These outputs, when supervised, generate reconstructions at increasing level of detail. Similarly to the aforementioned methods, we model our multi-scale representation in continuous function space. However, unlike (Barron et al., 2021), our architecture is based on a set of hierarchical MLPs, allows for per-level modularity and the representation of each reconstruction level is constrained through the frequencies in the Fourier Feature encoding of the input. Compared to (Sitzmann et al., 2020), our method is designed for multiple levels of detail and contrary to (Lindell et al., 2021), our intermediate levels compose the final reconstruction and do not require supervision during training.

3. Preliminaries

Neural Implicit Functions encode signals in continuous space using a neural network parametrisation. Often referred to as coordinate-based networks, they are defined as a function mapping $f : \mathbf{x} \rightarrow V$, with $\mathbf{x} \in \mathbb{R}^{1,2,3}$; V is a quantity of interest such as colour or occupancy. In 3D, the surface is modelled as the levelset of a continuous function $f : f(\mathbf{x}) = 0$. By design, the representation predicts a single point in the signal's domain at a time and multiple values can be obtained by querying the representation at the corresponding set of coordinates $\{\mathbf{x}_1 \dots \mathbf{x}_N\}$.

Positional Encoding was first introduced in Natural Language Processing (Vaswani et al., 2017) to inject information about the relative position of word tokens into the Transformer Model. In the context of Neural Implicit Functions, positional encoding refers to the projection of the input (spatial coordinates) to a higher dimensional space $P : \mathbb{R}^{1,2,3} \rightarrow \mathbb{R}^N$ (Mildenhall et al., 2020). Several function mappings have been proposed, including simple sinusoidal mappings to fit neural radiance fields (Mildenhall et al., 2020) and reconstructing 3D protein structures (Zhong et al., 2020), non-axis aligned Fourier basis functions: 'Fourier Features' (Tancik et al., 2020) and a positional encoding based on multi-scale B-Splines (Wang et al., 2021a). **Fourier Features** have become a popular framework for fast kernel method approximations (Li et al., 2019). First proposed by (Rahimi & Recht, 2007), the approximation is based on Bochner's theorem (Bochner, 1932), which shows that any continuous shift-invariant kernel is the Fourier Transform of a positive bounded measure (*spectral measure*). The approach proposed by (Rahimi & Recht, 2007) approximates such a continuous kernel by its Monte-Carlo estimate using samples from this spectral measure: *Fourier Features* (Li et al., 2019). (Tancik et al., 2020) relate this method to the frequency mapping applied to inputs of Neural Implicit Functions and propose Fourier Features as an improved form of positional encoding: a set of Fourier bases $\mathbf{F} : \{cos(\omega_i x + b_i) \dots cos(\omega_n x + b_n\}$ where ω and b are sampled from a parametric distribution (e.g $\mathcal{N}(\mu = 0, \sigma)$). As a positional encoding, Fourier Features map input coordinates to a higher dimensional manifold as $P : \mathbf{x} \in \mathbb{R}^{1,2,3} \rightarrow \mathbf{F(x)} \in \mathbb{R}^{\mathrm{n}}$. We choose a Fourier Feature mapping as positional encoding for our progressive

implicit representation and refer to it in the rest of this paper as Fourier Features (FF).

(Tancik et al., 2020) show that the frequencies of the input mapping function directly affect the spectral falloff of the Neural Tangent Kernel: positional encoding influences how fast, and if, specific frequencies of a signal can be learnt. They show through parameter search that sampling the frequencies of their proposed Fourier Feature encoding at $\sigma = 10$ generates the best reconstruction of natural images; however, no connection is made to the frequency composition of the target signal. We perform a similar parameter search for σ (see Section 5.4) and provide a short analysis relating σ to the frequency composition of the target signal in Appendix B.

4. Method

4.1. Progressive Fourier Feature encoding

We formulate the task of reconstructing a scene S as a composition of a base component c and a set of residuals $R_{1...L}$:

$$S(\mathbf{x}) = c + \sum_{l=1}^{L} \omega_l R_l(f_l(\mathbf{x})), \tag{1}$$

where $\mathbf{x}$ is a coordinate vector and each residual R_l is parameterised by an MLP. The contribution of R_l to the final representation is controlled by w_l, and c is a constant, set according to the reconstruction task domain. R_l takes as input a subset $f_l(\mathbf{x})$ of Fourier Feature mapping $\mathbf{F}$ with a set F of frequencies sampled from a Gaussian distribution $F \sim \mathcal{N}(0, \sigma)$. The sampled frequencies are sorted by increasing value and divided into L subsets $\{f_{1...L}\}$. By definition, each subset f_l now contains frequencies $\{k_{1...N}\}$ and for any two adjacent subsets, f_l and f_{l-1}, it holds that $k_{l,1} > k_{l-1,N}$, whereby $k_{l,1}$ and $k_{l-1,N}$ are the first and last frequency of the subsets l and $l-1$ respectively. Intermediate levels of reconstruction S_l can be obtained by composing part of the residuals: $S_{l=2}(\mathbf{x}) = \sum_{l=0}^{l=2} \omega_l R_l(f_l(\mathbf{x}))$. Intuitively, the residual R_l learnt by one specific level is guided by the frequency encodings it is exposed to. Given the structure of our progressive Fourier Feature encoding $\{f_{1...L}\}$, the first residuals are encouraged to focus on low frequency content while later residuals focus on high frequency content in the scene.

Inspired by the fact that power spectral densities of natural signals decay exponentially (i.e., larger proportions of the signal come from lower frequency ranges) (Ricker, 2003), we reduce the weight of higher levels compared to lower levels for the final reconstruction. Empirically, we set w to a decreasing geometric progression such that $w_l = \frac{1}{l+2}$ and show that this leads to better results than equal weightings of R_l in our ablation studies.

In 1D, our model maps x to y coordinates: $S : x \rightarrow y$ and we set $c = 0$. In 2D, coordinates are mapped to pixel values: $S : x, y \rightarrow P(x, y)$ and we set c to the image mean $\overline{m}$ (detailed explanation provided in section 5.2). In 3D, our model defines a levelset as $V(\tau) = \{x : S(x) = \tau\}$ where V is the volume containing the 3D shape whose surface lies at $V(0)$, x is a coordinate in V and S maps coordinates to SDF values τ as $S : \mathbb{R}^3 \rightarrow \mathbb{R}$. Similarly to (Chen et al., 2021), our residuals R_l are defined as $R_l = S_l - S_{l-1}$. However, in our formulation, we set S_0 to be a constant $c = 0$.

4.2. Architecture

Our architecture is composed of multiple stacked small MLPs $M_{1...L}$, whereby each MLP receives as input a set of FF mappings $f_l(\mathbf{x})$ and the output of the previous level (level-conditioning). The first level receives as input $f_1(\mathbf{x})$ and the raw coordinates $\mathbf{x}$. The output of each level is a feature tensor T_l:

$$\begin{aligned} M_l : f_l(\mathbf{x}), T_{l-1} \rightarrow T_l, \quad l > 0 \\ M_l : f_l(\mathbf{x}), \mathbf{x} \rightarrow T_l, \quad l = 0 \end{aligned} \tag{2}$$

Each feature tensor T_l is then mapped to a residual R_l by another MLP which shares weights across each level and acts as a feature to domain mapping. Finally, the network outputs are composed into the final reconstruction according to Equation 1. Using an intermediate feature representation T_l at each level allows for a more expressive feature representation to be passed to the next level during level-conditioning. We experimented with per-level domain mapping but found no explicit benefit over using a single MLP with shared weights. While the number of layers L as well as the number and range of frequencies F are selected before training, we only apply a loss on the final reconstruction S, leaving it up to the network how to decompose the scene representation into progressive levels.

Intuitively, our architecture motivates a decomposition into progressive levels of detail by restricting the frequency range each level-specific MLP has access to. We show in our ablation studies that the conditioning of M_l on the previous level M_{l-1} yields higher reconstruction accuracy. **Per-level modularity** The architecture design allows for each level of reconstruction to be used independently at test time. For a reconstruction at an intermediate level of detail $S_{l=2}$, MLPs $M_{3...L}$ can be dropped and the reconstruction will be computed as $S_{l=2}(\mathbf{x}) = \sum_{l=0}^{l=2} \omega_l R_l(f_l(\mathbf{x}))$. An overview of our method and the architecture are provided in Figures 1 and 3.

4.3. Loss

To train our model, we apply a reconstruction loss L_r at the final reconstruction S, as well as a regularisation loss L_{reg} which encourages intermediate levels of detail S_l to

Figure 4. Qualitative results on a 2D image regression task. We compare our model against baselines on images from the COCO 2017 validation set (Lin et al., 2014)

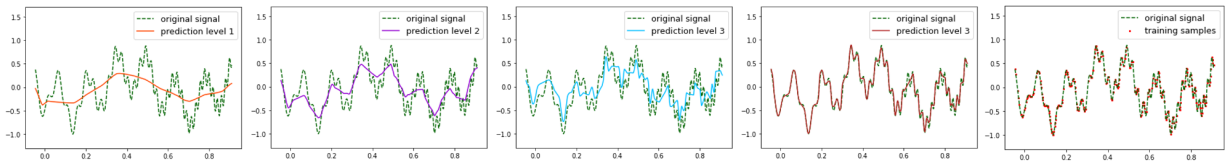

Figure 5. Fitting 1D periodic signals. **Left** to **right**: reconstruction levels 1-4 and original signal with training samples.

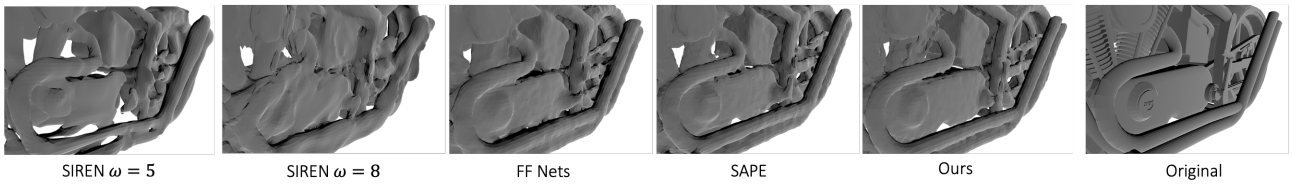

Figure 6. Qualitative results for a 3D regression task. We regress a TSDF and compare against SIREN models with two different σ and set σ to 2 for our model and the baselines using Fourier Features.

be close to the ground truth scene S_g:

$$L_{reg} = \sum_{l=1}^{L} L_r(S_l(\mathbf{x}), S_g). \tag{3}$$

Our final loss is defined as $L_r + \omega L_{reg}$ and we find that a value of $\omega = 0.01$ works well for our experiments. We experiment with simple $L1$ and $L2$ norms for L_r, as well as a perceptual loss based on VGG features. Perceptual losses based on deep network features have to be regularised using the $L1$ or $L2$ norm (Johnson et al., 2016) and we find that although the perceptual loss yields slightly more pleasing visual results for 2D regression tasks, the difference is not significant enough to motivate the introduction of additional hyperparameters. For our final experiments we therefore use the $L2$ norm.

Note that our regularisation loss encourages intermediate levels of detail to be similar to the final reconstruction, by adding a small fraction of the difference between every intermediate level S_l and the ground truth to the loss. However, by adding this regularisation, no guidance is given to the network for how to structure the residuals themselves, nor, how close each intermediate level S_l should be to the final reconstruction. In fact, it is possible to train without the regularisation loss, which we demonstrate in our ablation studies.

4.4. Training and implementation details

We use the Adam optimiser (Kingma & Ba, 2015) and a standard learning rate of $1e{-}3$ for all experiments. Unless mentioned otherwise, we train on a uniformly sampled subset of 50% of the image pixels for all 2D regression tasks. For 3D shape regression, we train on an average of $456k$ SDF samples per shape. For the presented experiments, unless specified, we train with 3 levels of detail, a hidden layer size of 256 per level and $\sigma = 15$. We extend the baseline architectures to have the same number and size of hidden layers.

5. Experiments

We evaluate our method with several 1D, 2D and 3D regression tasks as well as in terms of representational capacity and training time. We also provide a set of ablation studies and an analysis on the relation between positional encoding frequencies and signal frequencies. **Baselines** We compare our method against 3 different baselines: 1) A simple MLP with Fourier Feature encoding (FF Net) (Tancik et al., 2020) 2) SIREN, an MLP with sinusodial activations (Sitzmann et al., 2020) 3) SAPE, an MLP with spatially-adaptive progressive encoding with a mask resolution of 64 (Hertz et al., 2021) (the mask resolution is set to 64 to have similar con-

Model	ImageNet		
	MSE ↓	PSNR ↑	VGG f ↓
FF Nets	0.0048 ± 8.89e-5	74.5 ± 26.41	2.06 ± 0.56
SIREN	0.015 ± 8.4e-4	72.1 ± 49.69	2.36 ± 0.73
SAPE	0.0030 ± 8.6e-6	75.7 ± 26.26	**1.85±0.58**
Ours (nC)	0.0027±1.3e-5	**76.5±24.4**	1.92±0.59
Ours (eq)	0.0028±9.4e−6	75.9±24.82	1.91±0.54
Ours	**0.0027±1.6e-6**	75.9±23.49	1.89±0.49
	DIV2K		
	MSE ↓	PSNR ↑	VGG f ↓
FF Nets	3.3e-3±3.4e−5	75.6±19.7	2.2±0.5
SIREN	3.1e-3±7.6e−6	74.5±13.3	2.4±0.4
SAPE	2.0e-3±2.2e−6	76.5±13.9	2.1±0.5
Ours (nC)	2.07e-3±2.87e-6	76.5±17.4	2.2±0.5
Ours (eq)	1.91e-3±2.46e-6	**76.9±18.9**	**2.0±0.5**
Ours	**1.88e-3±2.26e-6**	76.8±13.6	2.1±0.4

Table 1. Evaluation on DIV2K and ImageNet (100 imgs)

vergence time to PINs). For SIREN, we set their frequency hyperparameter w to 30 as suggested by (Sitzmann et al., 2020). We observe that for regressing 3D shapes, $w = 30$ leads to divergence and choose lower values between 5 and 10. We also observe SIREN to be unstable if optimising for too many iterations and reduce the number of iterations to achieve the best possible reconstruction.

5.1. 1D signal regression

In a first instance, we qualitatively evaluate our model on periodic signals composed of multiple sinusoids. As can be seen from the example in Figure 5, 1D signals are fit at progressive level of detail by our model. Examples of the target signals and more results are provided in Appendix C.

5.2. 2D image regression

We evaluate our method on natural image reconstruction tasks with a subset of the ImageNet test dataset (Deng et al., 2009), high resolution images from the DIV2K validation dataset (Agustsson, 2017) and qualitatively, with images from the 2017 COCO validation set (Lin et al., 2014). We evaluate in terms of Mean-Squared-Error (MSE), Peak Signal to Noise Ratio (PSNR) as well as a perceptual loss based on the L2 norm between VGG features (Johnson et al., 2016). Our quantitative results (see Table 1) show that we outperform SIREN and the FF Net baseline in all metrics; SAPE outperforms our method on the perceptual loss for the ImageNet dataset. Qualitative results can be found in Figures 1, 4, 2, 11 and Appendix C. Our method reconstructs images with less noise, particularly for scenes with wide frequency bands. Thanks to the continuous representation of every level of detail, we are able to handle very fine detail against plain backgrounds (see Figure 4). In contrast, for such scenes, (Hertz et al., 2021) would require a very high mask resolution to achieve noiseless reconstruction: despite locally adapted frequency unwrapping, if the area of one mask pixel includes both fine detail and smooth background, noise will be introduced. This effect is visible in Figure 4. Although having a high enough mask resolution is feasible for small images, it becomes intractable for large images (see results of section 5.7). **Sample sparsity** We qualitatively evaluate how our method reconstructs natural images for different pixel sample densities. As can be seen from Figure 11, our model outperforms baselines for low sample densities and can achieve reasonable reconstruction accuracy even when only training on 2% of the image pixels.

Learning the base component c For 1D and 3D regression tasks, the scene representation is w.r.t. 0: A 1D sinusoid oscillates around 0 and the SDF representation of a shape consists in knowing the distance from the surface defined at 0. For an image this is not the case, as negative RGB values have no meaning; instead, we set c to the image mean. We validate this choice with following experiment: For a 2D regression task, we set c to a trainable parameter in the network and optimise it together with the network weights. When initialised at different values, c always converges to the image mean (see Appendix A for details).

5.3. 3D shape regression

We evaluate our method on a set of 3D regression tasks. We compare against all baselines on 3D models from 3D Warehouse (3DW) categories *lamp, car, chair, sofa, motorbike, bed and camera*; quantitative results can be found in table 2, qualitative results in Figures 1, 6 and 7. We observe SIREN to easily diverge on some shapes, as already noted by others (Wang et al., 2021b). This leads to a significantly higher reconstruction error for some of the 3D shapes. Similarly to our results in 2D, our model produces smoother reconstructions without losing detail. This is particularly visible in the example shown in Figure 6.

5.4. Sampling Fourier Features at different σ

We evaluate how our model performs for different ranges of sampled frequencies on 2D and 3D regression tasks (Figure 8). We evaluate for 10 natural images from DIV2K and 10 shapes of different categories taken from 3D Warehouse. We compare to SAPE and FF Net which also use a Fourier Feature encoding. We find that for values below 30 our model overall outperforms both SAPE and FF Net. For values above 30, our model is on par with SAPE without the need to mask out frequencies for smooth regions. Thanks to our progressive architecture we achieve crisper representations for low encoding frequencies and less noise is introduced in smooth regions, even when high encoding frequencies are present in the encoding (see Appendix D for qualitative examples). Compared to the best performance for FF Net

Model	lamp (ChD ↓)	car (ChD ↓)	chair (ChD ↓)	sofa (ChD ↓)	motorbike (ChD ↓)	bed (ChD ↓)	camera (ChD ↓)
FF Nets	2.5±3.4e−3	2.1±4.8e−4	0.92±1.5e−4	1.5±5.6e−4	**1.0±1.5e-5**	2.83±5.2e−3	3.46±1.26e−2
SIREN	25.4±4.2e−3	2.2±5.2e−4	28.4±6.7e−3	1.6±5.5e−4	1.7±6.4e−4	3.22±6.9e−3	2.31±2.68e−3
SAPE	6.7±9.2e−2	2.2±6.6e−4	1.50±1.1e−3	6.6±0.42	2.8±3.9e-4	4.58±2.4e−2	4.38±2.7e−2
Ours	**1.5±1.0e-4**	**2.0±5.1e-4**	**0.87±1.7e-4**	**1.48±5.4e-4**	1.1±6.7e-5	**2.79±4.7e-3**	**2.05±2.7e-3**

Table 2. Evaluation on 3D models from 3D Warehouse (3DW) in terms of the bi-directional Chamfer Distance (mm)

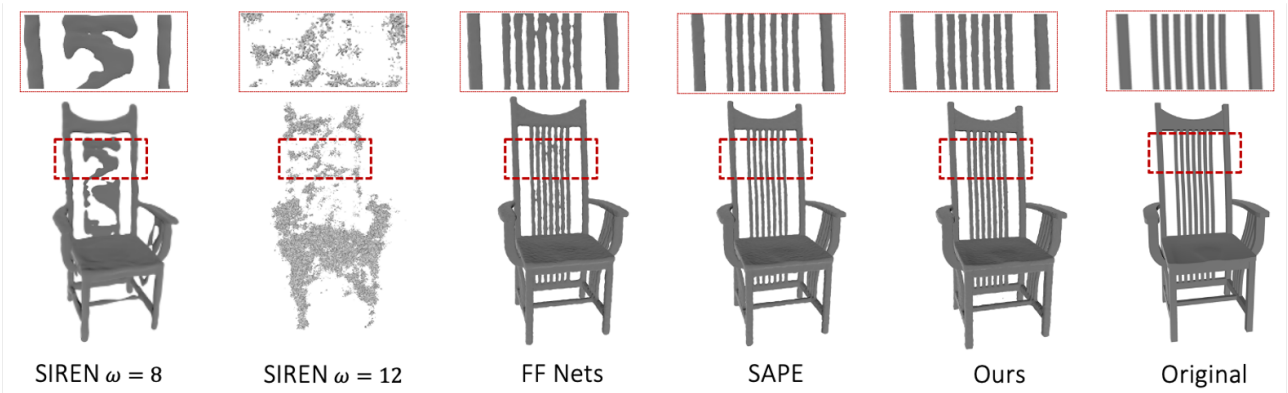

Figure 7. Qualitative results for a 3D regression task. We regress a TSDF and compare against SIREN models with two different ω and set σ to 2 for our model and the other baselines using Fourier Features.

found at $\sigma = 10$ (Tancik et al., 2020), we find that for our model, a value of $\sigma = 15$ yields the best PSNR on natural images. We attribute this to the progressive nature of our architecture that allows for higher frequencies to be used in the positional encoding without introducing noise in low structure areas. For 3D shapes we observe that for high σ values FF Net diverges resulting in a strong drop in reconstruction accuracy. This is not the case for SAPE and our method. Although the drop in reconstruction accuracy is not significant across different values of σ, higher frequencies visibly introduce noise in the reconstruction. This can be seen in the qualitative examples presented in Appendix D. We find $\sigma = 3$ to be the best value for the tested 3D shapes.

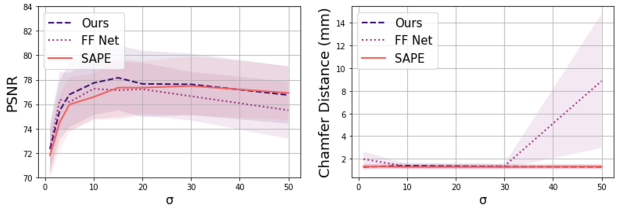

Figure 8. Sensitivity to σ (standard deviation) when sampling Fourier Features for 2D regression tasks (**left**) and 3D regression tasks (**right**).

5.5. Encoding feature density

We study how the encoding feature density (number of sampled frequencies) affects the reconstruction quality of images and 3D shapes. For a subset of 10 natural images from the DIV2K dataset and 10 shapes (of different classes) from 3D Warehouse, we plot the PSNR against the number of frequencies in the encoding (see Figure 10). We set $\sigma = 15$ and $\sigma = 3$ for natural images and 3D shapes respectively. We find that the reconstruction quality saturates when using more than 100 frequencies in the encoding for natural images and at about 30 frequencies for 3D shapes.

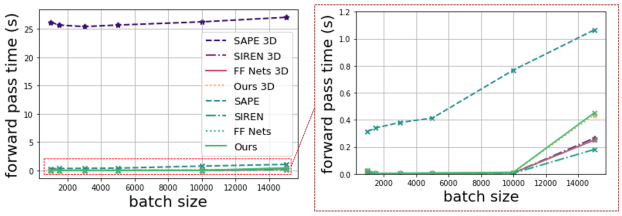

Figure 9. Forward pass time for 2D and 3D regression tasks at different batch sizes.

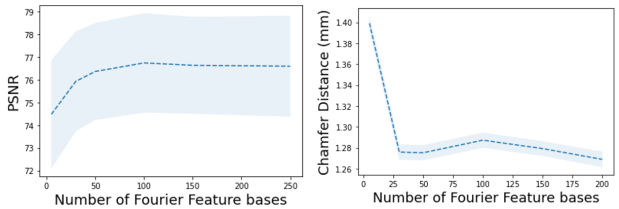

Figure 10. Sensitivity to frequency density of the positional encoding for 2D regression tasks (**left**) and 3D regression tasks (**right**).

5.6. Encoding Frequencies & Signal Frequencies

The presented experiments as well as previous approaches (Tancik et al., 2020) provide empirical results for the best values of σ for sampling the frequencies of Fourier Features. We believe a more principled approach will select encoding frequencies based on the frequency composition of the scene itself and we provide a few initial experiments in Appendix B to pave the road for future work.

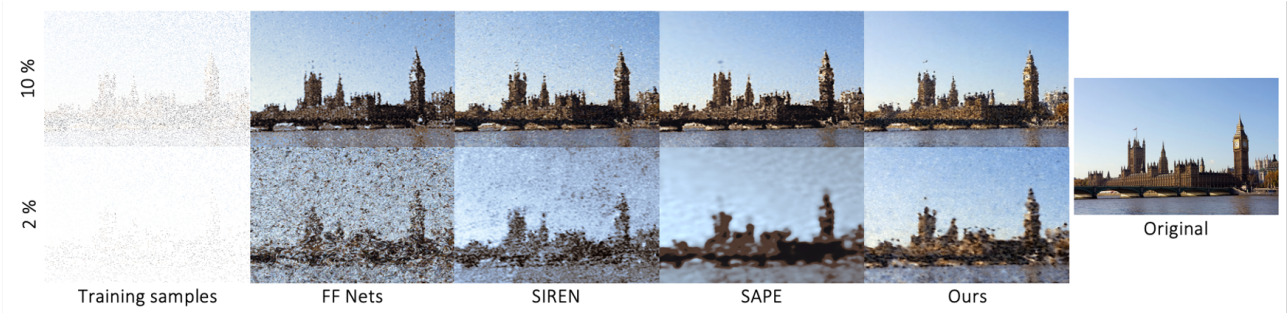

Figure 11. Fitting natural images with sparse sample sets of 10% (**top row**) and 2% (**bottom row**) of all image pixels.

5.7. Model size and training speed

We evaluate our architecture in terms of its representational capacity and training speed on 2D and 3D regression tasks. We fit a subset of the DIV2K dataset for different sizes of our model (gradually increasing both network depth and width). Our method outperforms SIREN for all tested sizes and FF Net for sizes above $100k$ parameters. In particular, note how all baselines experience a performance drop as the number of parameters increases. We attribute this to overfitting due to overparameterisation. For SIREN this could also be due to instability at greater depth which we noted leads to divergence. Our model's accuracy steadily increases for the tested model sizes, suggesting that the architecture is less prone to overfitting. SAPE outperforms us in representational capacity at lower sizes, however, it is significantly slower in training speed (see Figure 9). SAPE's forward pass is slowed down by regular mask interpolation and update steps. This is particularly noticeable for 3D regression tasks. At the cost of a small increase in network size, our architecture can achieve the same reconstruction quality as SAPE, while training much faster and being more robust to overfitting.

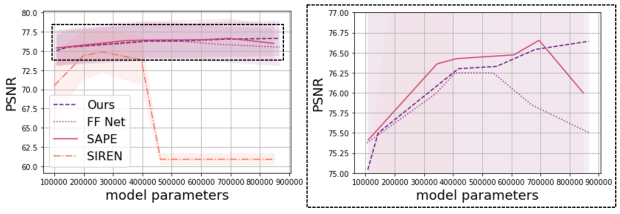

Figure 12. Reconstruction accuracy wrt. model size, evaluated on a subset of the DIV2K dataset.

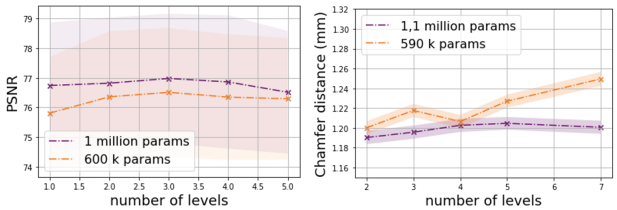

Figure 13. Reconstruction accuracy for different numbers of levels, keeping overall network size fixed. **Left**: 2D regression task. **Right**: 3D regression task.

5.8. Ablation studies

5.8.1. Level composition and level conditioning

We compare our model against an ablated version with equal weights (*Ours (eq)*) in the level composition and note that while using decreasing weights leads to a better MSE score, no improvement can be seen for perceptual metrics; in fact, equal weighting leads to a slightly higher mean for PSNR and perceptual loss (see Table 1). However, removing the level-conditioning of higher layers on lower layers (see Equation 2) (*Ours (nC)*), leads to a more noticeable drop in performance, in particular for the high resolution images of the DIV2K dataset.

5.8.2. Level decomposition

We evaluate how level decomposition affects our model's performance. Keeping the number of parameters fixed, we evaluate how the reconstruction accuracy varies with different numbers of levels. We evaluate on 10 natural images from the DIV2K dataset and 10 3D models of different categories from 3D Warehouse for model sizes of roughly $1m$ and $600k$ parameters. (see Figure 13). We find that overall, within the range of levels tested, reconstruction accuracy is not strongly affected. However, we observe the following trends: for fitting natural images, 3 levels of detail yield the best result for both tested model sizes. The lower reconstruction accuracy for more levels of detail is most likely due to the fact that each level has fewer parameters and is therefore not able to fit each scene residual well. Fewer levels of detail on the other hand may lack the representational advantage introduced by our architecture design. For fitting 3D shapes we find that for both networks, 2 levels of detail yields the best reconstruction. The reconstruction accuracy degrades faster for increasing levels for the smaller network. We attribute this to the fact that fewer parameters are available for each level which eventually leads to underfitting.

5.8.3. Frequency order

We validate our choice of increasing FF frequencies with the following experiments: 1) To ensure that the progressive nature of our results is not simply a result of the hierarchical

MLP structure, we train an ablated model with randomly sampled FF frequencies across all levels. 2) To validate the importance of sorting encoding frequencies in increasing order for obtaining a progressive scene representation, we train an ablated version with decreasing frequencies in the positional encoding (highest frequencies in the first level and lowest frequencies in the last level). As can be seen from an example in Figure 15, without sorted, increasing frequencies in the positional encoding, the progressive nature of the reconstruction is lost and the reconstruction is noisy in smooth image regions. Results in Table 3 show that using increasing frequencies yields overall better reconstruction accuracy for a set of COCO images. Interestingly, we find that decreasing frequencies produces better results than a random order. This suggests that the structure introduced by having a specific range of FF frequencies per MLP is beneficial for reconstruction quality.

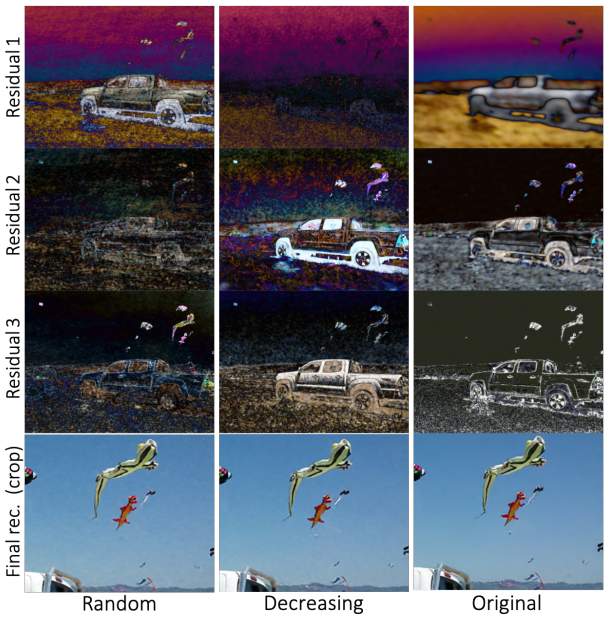

Figure 14. Fitting an image with (**left**) randomly sampled FF frequencies, (**center**) decreasing FF frequencies and (**right**) FF frequencies in the increasing order used in PINs.

5.8.4. Removing per-level regularisation

Our proposed loss function includes a per-level regularisation term, that encourages every level to be similar to the final reconstruction. When training without this regularisation we observe a slight drop in convergence speed, but find that the representation learnt by individual levels remains comparable (see an example in Figure 15).

6. Limitations and Future Work

We demonstrate the advantages of our proposed representation, however, several questions remain open and we would like to suggest directions for future work: 1) We select both

Figure 15. Fitting an image with 3 levels of detail, with (**left**) and without (**right**) per-level regularisation. **Top row**: reconstruction. **Bottom row**: residuals.

Loss	Model ablation (different FF frequency order)		
	Random	Decreasing	Original (PINs)
PSNR ↑	71.87 ± 4.6	71.96 ± 4.16	$\mathbf{72.24 \pm 6.22}$
VGG f ↓	2.24 ± 0.16	2.20 ± 0.17	$\mathbf{2.09 \pm 0.17}$

Table 3. Evaluation on COCO's validation set (16 images) using uniformly sampled, unordered FF frequencies, a decreasing FF frequency order and the original, increasing order.

the number of levels and the sampled frequency density manually and provide empirical studies for the best values of these hyperparameters in Sections 5.4 and 5.5. However, it would be interesting to evaluate to what extent these parameters can be learnt as part of the representation. 2) Similarly to (Tancik et al., 2020) we find the best value for σ through parameter search, but our experiments in Appendix B indicate a correlation between the frequency composition of a scene and the frequencies of the input mapping that lead to the best reconstruction. We believe that it could be beneficial to infer a scene-specific input mapping based on this correlation. 3) We did not experiment with generalisation to novel scenes within the scope of this work but believe this to be an interesting extension. 4) The proposed architecture is in principle applicable to novel-view synthesis; e.g. for every sample point, the individual residuals could be integrated along their respective ray to yield the corresponding pixel-residual. It would be interesting to extend PINs for novel-view synthesis and compare it to related approaches such as (Barron et al., 2021).

7. Conclusion

We introduce a novel multi-scale implicit representation based on progressive positional encoding. Through conditioning a hierarchical MLP structure on incremental Fourier Features, our method learns to decompose a scene into progressive levels of detail without level-specific supervision. We achieve higher reconstruction accuracy for 2D images and 3D shapes compared to baselines, in particular, for scenes with wide frequency spectra. The modularity of the architecture allows to only use part of the network at inference time, if a coarser representation is sufficient. Overall, our method provides a flexible, continuous and multi-scale implicit representation with a simple, end-to-end training scheme.

References

3D Warehouse. https://help.sketchup.com/en/3d-warehouse. last accessed: 25-01-2022.

Agustsson, E. Challenge on single image super-resolution: Dataset and study. 2017.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., and Srinivasan, P. P. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ArXiv*, abs/2103.13415, 2021.

Basri, R., Galun, M., Geifman, A., Jacobs, D. W., Kasten, Y., and Kritchman, S. Frequency bias in neural networks for input of non-uniform density. In *ICML*, 2020.

Bochner, S. *Vorlesungen über Fouriersche Integrale*. Leipzig : Akademische Verlagsgesellschaft, 1932.

Chabra, R., Lenssen, J. E., Ilg, E., Schmidt, T., Straub, J., Lovegrove, S., and Newcombe, R. A. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *ECCV*, 2020.

Chen, Z. and Zhang, H. Learning implicit fields for generative shape modeling. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5932–5941, 2019.

Chen, Z., Zhang, Y., Genova, K., Fanello, S., Bouaziz, S., Haene, C., Du, R., Keskin, C., Funkhouser, T., and Tang, D. Multiresolution deep implicit functions for 3d shape representation, 2021.

Deng, B., Genova, K., Yazdani, S., Bouaziz, S., Hinton, G. E., and Tagliasacchi, A. Cvxnet: Learnable convex decomposition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31–41, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Fathony, R., Sahu, A. K., Willmott, D., and Kolter, J. Z. Multiplicative filter networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OmtmcPkkhT.

Genova, K., Cole, F., Sud, A., Sarna, A., and Funkhouser, T. Local deep implicit functions for 3d shape. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4856–4865, 2020.

Hertz, A., Perel, O., Giryes, R., Sorkine-Hornung, O., and Cohen-Or, D. Sape: Spatially-adaptive progressive encoding for neural optimization, 2021.

Huang, J., Huang, S.-S., Song, H., and Hu, S. Di-fusion: Online implicit 3d reconstruction with deep priors. *ArXiv*, abs/2012.05551, 2020.

Jang, W. and Agapito, L. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12949–12958, October 2021.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Li, K., Rünz, M., Tang, M., Ma, L., Kong, C., Schmidt, T., Reid, I. D., de Agapito, L., Straub, J., Lovegrove, S., and Newcombe, R. A. Frodo: From detections to 3d objects. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14708–14717, 2020.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. Towards a unified analysis of random fourier features. In *ICML*, 2019.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015.

Lindell, D. B., Van Veen, D., Park, J. J., and Wetzstein, G. Bacon: Band-limited coordinate networks for multiscale scene representation. *arXiv preprint arXiv:0000.00000*, 2021.

Liu, S., Saito, S., Chen, W., and Li, H. Learning to infer implicit surfaces without 3d supervision. In *NeurIPS*, 2019.

Martel, J. N. P., Lindell, D. B., Lin, C. Z., Chan, E., Monteiro, M., and Wetzstein, G. Acorn: Adaptive coordinate networks for neural scene representation. *ACM Trans. Graph.*, 40:58:1–58:13, 2021.

Mescheder, L. M., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4455–4465, 2019.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Niemeyer, M. and Geiger, A. Giraffe: Representing scenes as compositional generative neural feature fields. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11448–11459, 2021.

Niemeyer, M., Mescheder, L., Oechsle, M., and Geiger, A. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

Niemeyer, M., Mescheder, L. M., Oechsle, M., and Geiger, A. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3501–3512, 2020b.

Ost, J., Mannan, F., Thuerey, N., Knodt, J., and Heide, F. Neural scene graphs for dynamic scenes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2855–2864, 2021.

Park, J. J., Florence, P. R., Straub, J., Newcombe, R. A., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 165–174, 2019.

Rahaman, N., Baratin, A., Arpit, D., Dräxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. C. On the spectral bias of neural networks. In *ICML*, 2019.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *NIPS*, 2007.

Reiser, C., Peng, S., Liao, Y., and Geiger, A. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. *ArXiv*, abs/2103.13744, 2021.

Ricker, D. W. *Echo Signal Processing*. The Springer International Series in Engineering and Computer Science. Springer, Boston, MA, 2003. ISBN 978-1-4613-5016-3.

Robert Fisher, Simon Perkins, A. W. and Wolfart, E. Fourier transform. https://homepages.inf.ed.ac.uk/rbf/HIPR2/fourier.htm, 2003. Accessed: 2022-01-25.

Saito, S., , Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019.

Schwarz, K., Liao, Y., Niemeyer, M., and Geiger, A. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sitzmann, V., Martel, J. N., Bergman, A. W., Lindell, D. B., and Wetzstein, G. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

Skorokhodov, I., Ignatyev, S., and Elhoseiny, M. Adversarial generation of continuous images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10748–10759, 2021.

Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020.

Tang, J.-H., Chen, W., Yang, J., Wang, B., Liu, S., Yang, B., and Gao, L. Octfield: Hierarchical implicit functions for 3d modeling, 2021.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

Wang, P.-S., Liu, Y., Yang, Y.-Q., and Tong, X. Spline positional encoding for learning 3d implicit signed distance fields. In *IJCAI*, 2021a.

Wang, Y., Rahmann, L., and Sorkine-Hornung, O. Geometry-consistent neural shape representation with implicit displacement fields. *ArXiv*, abs/2106.05187, 2021b.

Yu, A., Li, R., Tancik, M., Li, H., Ng, R., and Kanazawa, A. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021a.

Yu, A., Ye, V., Tancik, M., and Kanazawa, A. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021b.

Zhong, E. D., Bepler, T., Davis, J. H., and Berger, B. Reconstructing continuous distributions of 3d protein structure from cryo-em images. In *ICLR*, 2020.

We provide the following additional content: (A) Complementary graphs illustrating our analysis on learning the base component **c** *(B) An analysis relating encoding frequencies to signal frequencies (C) Additional qualitative examples for 1D and 2D regression tasks (D) Qualitative examples for reconstructing at different values of* σ.

A. Regressing the base component

In the main paper, we introduced the experiment which validates our choice for setting the base component **c** of our scene composition to the image mean: we let the implicit network regress the base component as part of the optimisation. We initialise **c** at different values 0.9, 0.1 and 0.5 and observe that it naturally converges to the image mean (Figure 16):

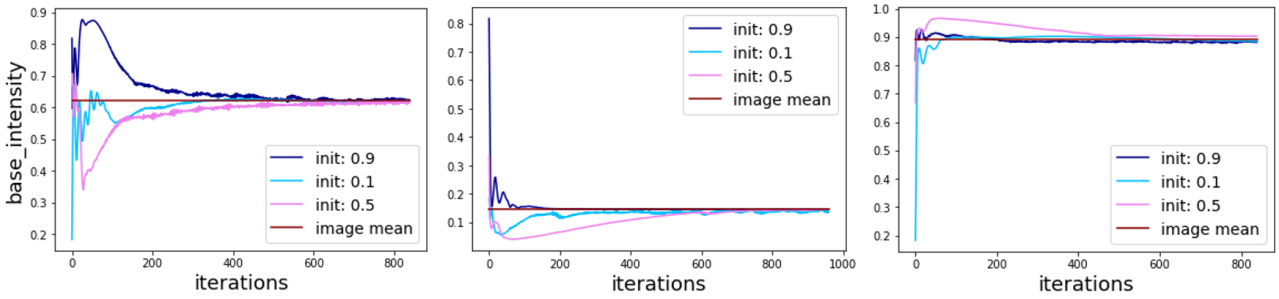

Figure 16. Regressing the base intensity for 3 different images. We set the base intensity as a learnable parameter and initialise it at different values. We then optimise together with the network weights.

B. Encoding Frequencies & Signal Frequencies

The results presented in section 5.4 demonstrate that $\sigma = 15$ is a good value to reconstruct natural images. We believe that a more principled approach to select positional encoding frequencies will leverage the frequency composition of the target signal. To pave the road for future work, we explore the relationship between the frequencies in the positional encoding and the frequencies of the signal with the following experiments: 1) For a set of synthetic 1D signals with one specific frequency (see examples in Figure 17), we test for which value of σ the highest fitting accuracy is achieved. Not surprisingly, for higher frequency signals, PSNR peaks at higher σ values (Figure 19, left). However, for these simple signals, the maximum frequency present inside the Fourier Features yielding the highest PSNR is overall lower than the signal frequency itself (Figure 19, right). 2) For a 2D regression task, we compute the DFT and visualise how the sampled Fourier basis from the Fourier Feature encoding relate to the Fourier basis functions present in the images' DFT. For a natural image, the (discrete) Fourier Transform can be written as:

$$F(k,l) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f[m,n] e^{-j2\pi(\frac{k}{M}m + \frac{l}{N}n)} \tag{4}$$

where M and N are the image dimensions. Intuitively, the image is transformed into a set of $M \times N$ basis functions: every pixel represents one basis. Frequencies in the DFT of an image of dimension $M \times N$ range from 0 to $\frac{m}{M}$ and $\frac{n}{N}$ (Robert Fisher & Wolfart, 2003); the position m, n of a particular Fourier Feature frequency f_{FF} in the DFT domain can be computed as $m = f_{FF}M$ and $n = f_{FF}N$. For a natural image from the DIV2K dataset we plot its DFT along with the sampled Fourier Feature frequencies (Figure 22, top row) and display the respective reconstruction quality obtained from a simple MLP with FF positional encoding (Figure 22, bottom row). As the plots illustrate, when sampling frequencies from within the DFT of the image, the reconstruction is lacking high frequency detail. To obtain a detailed reconstruction, the sampled frequencies in the positional encoding need to be up to 40 times higher than the highest pixel frequency present in the image. A complete analysis is beyond the scope of this work, but we hope to motivate future work that will establish a quantitative relationship between a signal's frequency decomposition and the positional encoding required to reconstruct it.

C. Additional Qualitative Results (1D and 2D regression)

In Figure 23 we present additional qualitative results for regressing 1D periodic signals and Figure 20 shows additional qualitative results on the 2017 COCO validation set for our method and baselines. In Figure 21 we display more examples of the predicted residuals $R_{1...3}$ by the network and the reconstructions $S_{1...3}$ at progressive level of detail.

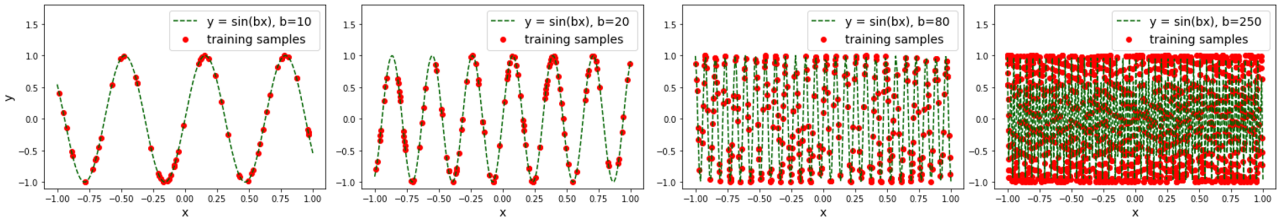

Figure 17. Synthetic single frequency signals used in our analysis on relating positional encoding to signal frequency (Appendix B)

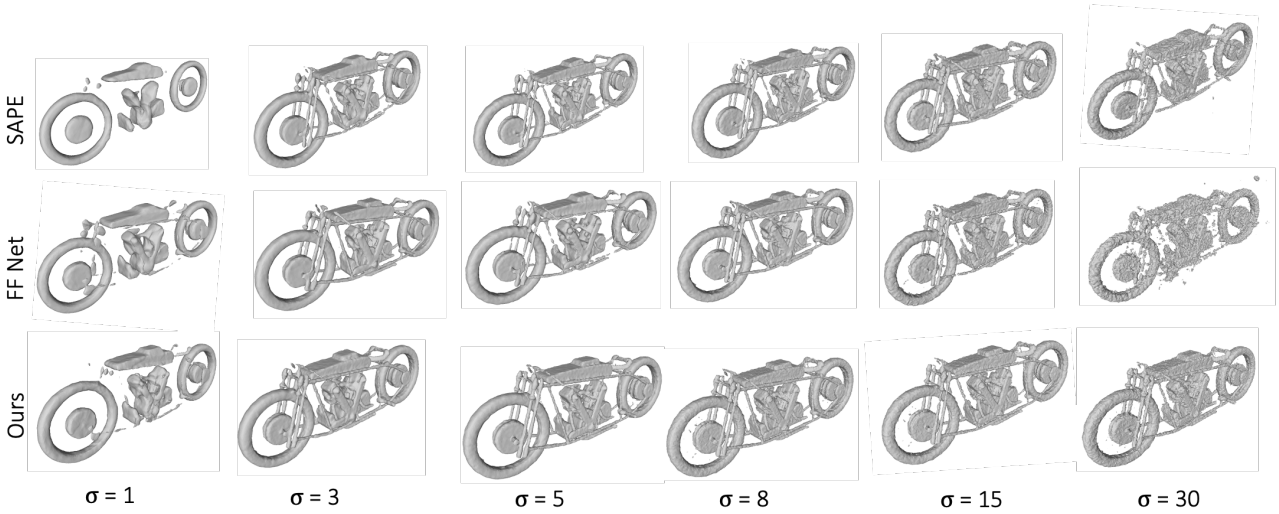

Figure 18. Fitting a 3D shape (3D Warehouse) with Fourier Features sampled at different values of σ.

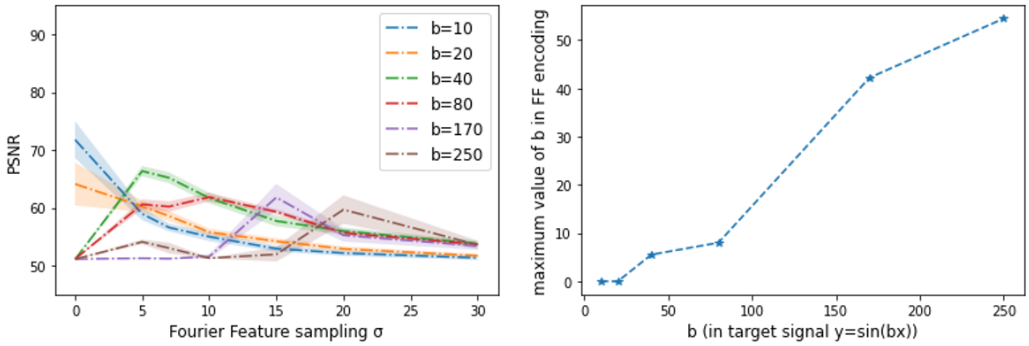

Figure 19. **Left**: PSNR achieved for Fourier Features (FF) sampled at different σ for 1D synthetic signals ($y = sin(bx)$) of varying frequencies. **Right**: highest frequency present in the FF encoding yielding the best PSNR vs. frequency of the 1D synthetic signal.

D. Qualitative Results for reconstructing with Fourier Features sampled at different σ

We provide qualitative results for reconstructing 2D and 3D scenes with Fourier Features sampled at different values of σ. Note how our model provides overall crisper natural image reconstructions over a wide range of σ (Figure 24). For 3D shapes (Figure 18), our model has a reconstruction quality similar to that of SAPE, even for encodings with very high frequencies (sampled at $\sigma > 15$). Compared to FF Net, the reconstruction does not diverge for high encoding frequencies (see middle row at $\sigma = 30$). The reconstructions for high σ values remain noisy however and we experimentally find that the best reconstructions are obtained for σ values between 3 and 5.

Figure 20. Qualitative results on the 2017 COCO validation set.

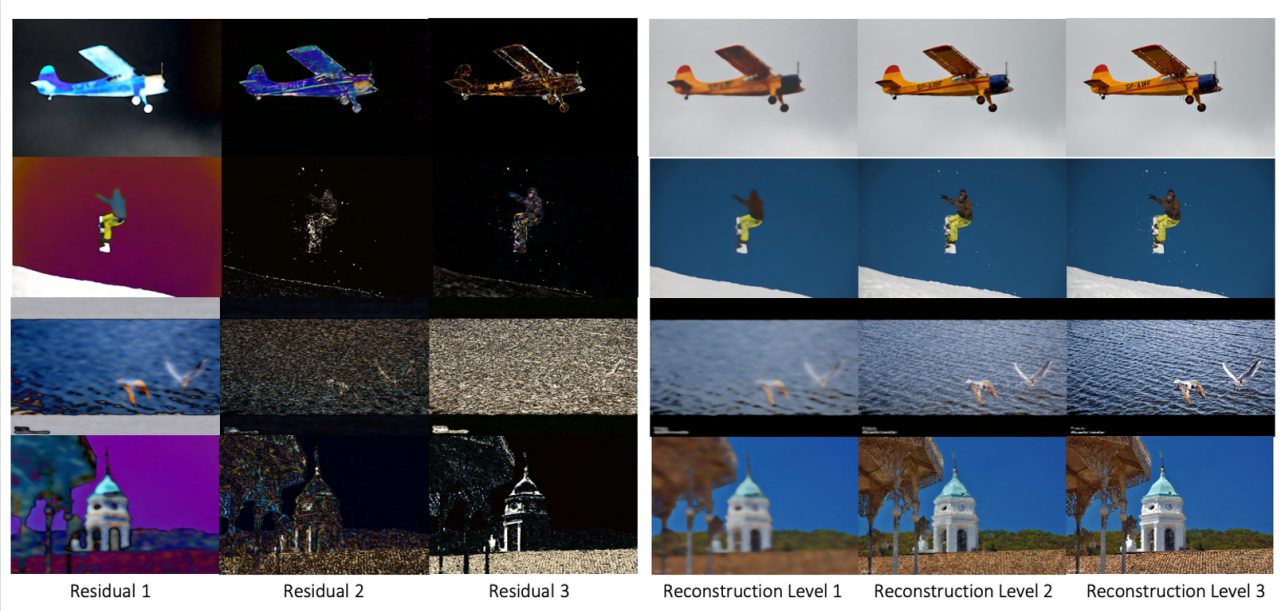

Figure 21. Additional qualitative examples showing our models residual output as well as the incremental level reconstructions. Images are taken from the COCO 2017 validation dataset (Lin et al., 2015).

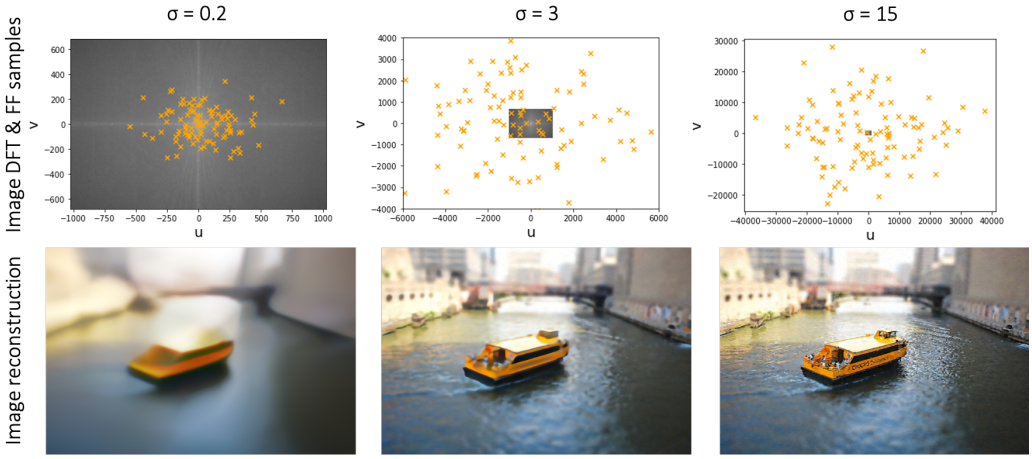

Figure 22. How sampled FF frequencies relate to the Discrete Fourier Transform (DFT) of a natural image. **Top**: the DFT and FF samples (orange). **Bottom**: the reconstruction .

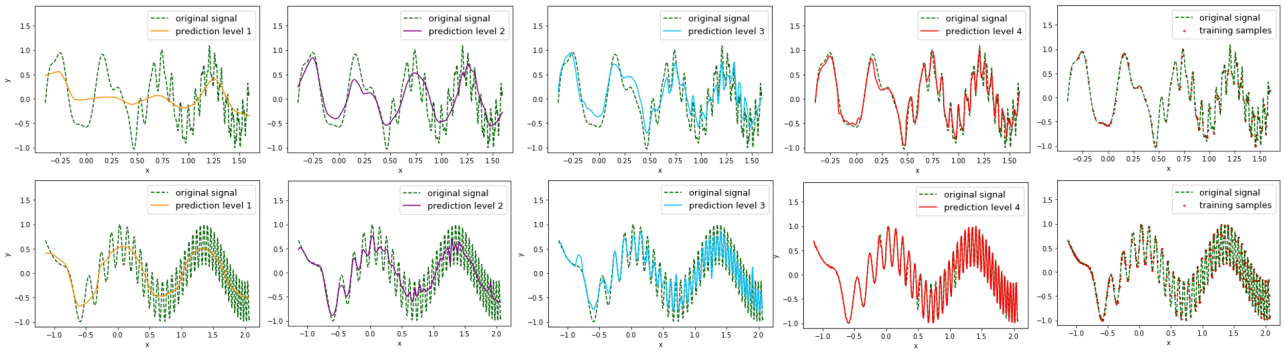

Figure 23. Regressing 1D periodic signals composed of multiple sinusoids. **Left** to **right**: Reconstruction Levels 1-4 and original signal with training samples.

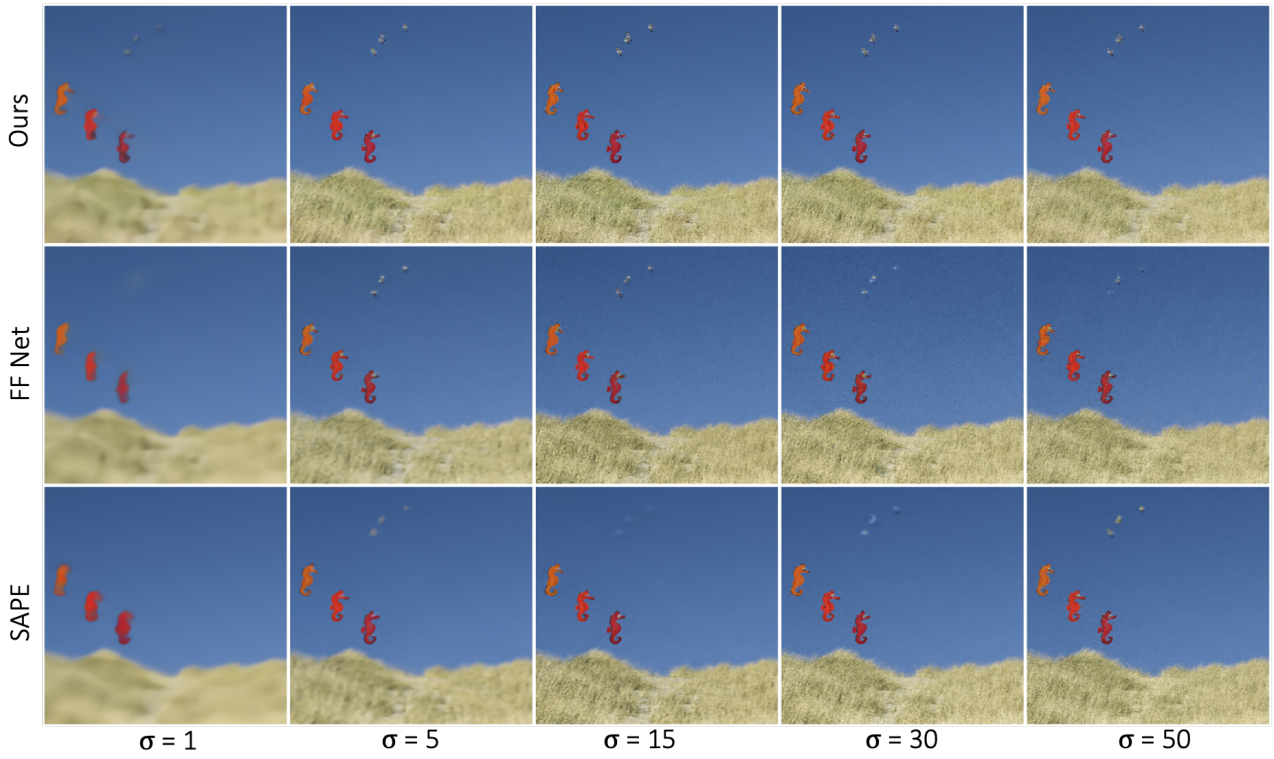

Figure 24. Regressing a 2D image (2017 COCO validation set) with Fourier Features sampled at different values of $sigma$