

Filtering In Neural Implicit Functions

Yixin Zhuang

yixin.zhuang@gmail.com

Abstract

Neural implicit functions are highly effective for data representation. However, the implicit functions learned by neural networks usually include unexpected noisy artifacts or lose fine details if the input data has many scales of detail or contains both low-frequency and high-frequency bandwidths. Removing artifacts while preserving fine-scale contents is challenging and usually comes out with over-smoothing or noisy issues. To solve this dilemma, we propose a new framework (FINN) that integrates a filtering module into the MLPs to perform data reconstruction while adapting regions containing different frequencies. The filtering module has a smoothing operator acting on intermediate results of the network that encourages the results to be smooth and a recovering operator bringing high frequencies to regions overly smooth. The two counteractive operators play consecutively in all MLP layers to adaptively influence the reconstruction. We demonstrate the advantage of FINN on several tasks and showcase significant improvement compared to state-of-the-art methods. In addition, FINN also yields better performance in both convergence speed and network stability.

Introduction

Filtering techniques are widely applied to many applications to enhance the content of interest or remove unexpected artifacts from the data, such as images and 3D shapes. For the data with many scales of details or a wide range of frequencies, it is challenging to recover the data while maintaining free of unwanted content using neural networks. Therefore, filtering methods are usually necessary for post-processing, otherwise additional strategies, e.g., coarse to fine generation, hierarchical representations, or scale-aware operations, are needed to adapt different scales of details. This paper addresses built-in learnable filters in a fully connected neural network to achieve both reconstruction and filtering purposes for neural implicit representation.

An implicit function is a continuous function that maps spatial coordinates to the corresponding values under varying representations (e.g., RGB, signed distance, etc.). It can be learned using simple network structures, i.e., Multi-Layer Perceptrons (MLPs) with ReLU activation functions (Mescheder et al. 2019; Park et al. 2019; Chen and Zhang 2019). Due to the spectral bias of neural networks (Basri et al. 2020; Rahaman et al. 2019), conventional MLPs are limited to

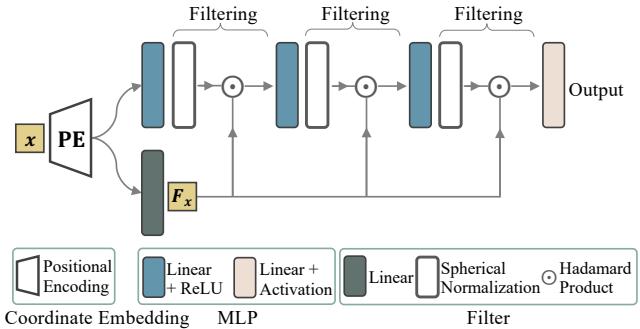


Figure 1: We present a new neural network for implicit functions. Reconstructing a signal containing low-frequency and high-frequency regions usually suffers from either underfitting or overfitting issues, i.e., losing fine-scale contents or adding noises. To alleviate such a dilemma, we propose a filtering module for MLPs. Our filter performs smoothing and recovering consecutively at all layers, such that the smoothing operator globally removes high frequencies and the recovering operator restores high frequencies to regions over-smoothed. Two counteractive filtering operators reduce both issues and lead to better training and generalization.

low-frequency signals, and large networks and much training time are required to fit complex signals. Recently, MLPs can easily reconstruct high-frequency by encoding the coordinates to Fourier features (FFN) (Tancik et al. 2020) or replacing ReLU activation functions with sinusoidal functions (SIREN) (Sitzmann et al. 2020). The sinusoidal function increases local variations to the signal by tuning up the frequency parameters and reduces the spectral bias.

Despite the impressive progress, FFN and SIREN usually create over-smoothed regions or noise in the results. It is unavoidable since there is no global effective frequency parameter for both low- and high-frequency regions and each spatial location or region need to be individually adapted. Towards learning adaptive frequency, based on FFN, SAPE (Hertz et al. 2021) progressively encodes the input with increasing frequencies for individual spatial coordinates to avoid consuming excessive frequencies at smooth regions. And based on SIREN, ModSIREN (Mehta et al. 2021) subdivides the input coordinates into grids and modu-

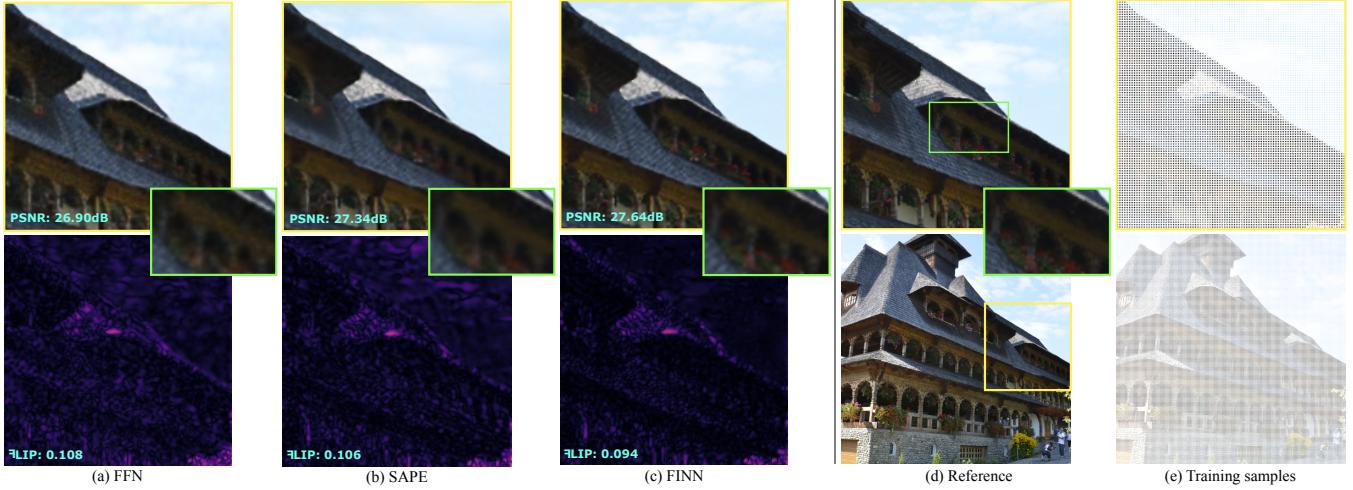


Figure 2: (a) An image generated by Fourier Features Network (FFN) contains noises (Top) highlighted in the error map from FLIP (Bottom). (b) SAPE smoothes out noises by learning adaptive Fourier features, but at the expense of losing small structures and textures, e.g., clouds and roofs. (c) Our method (FINN) incorporates a filtering module that can effectively remove artifacts while recovering distinct features, containing fewer errors in flat and sharp regions. We utilize a training image (e) at 256×256 resolution sampled from a reference image (d) and predicts images at $4 \times$ resolutions for all methods.

lates sinusoidal activation functions according to individual grids. An example is shown in Figure 2, where the explicit spatial adaptation method (i.e., SAPE) provides results with less noisy artifacts compared to the baseline method (i.e., FFN). Nevertheless, SAPE tends to over-smooth the result (e.g., the sky and the regions under roofs). We observe that directly adapting frequency parameters leads to a dilemma – lowering the frequency to remove noisy artifacts is usually at the expense of over smooth while increasing the frequency to reveal fine details may introduce noises. We show that a general-purpose filter in MLPs can significantly alleviate this dilemma and improve both regression and generalization. Unlike modulating sinusoidal activations (SIREN-based) or adapting frequency parameters of Fourier features (FFN-based), our network (FINN) manipulates intermediate outputs of MLPs with filters, leaving frequency fixed.

FINN is a simple network function composed of filtering functions and MLPs. Following FFN, we also incorporate the Fourier Features for the representation of high frequencies. The filtering function is decoupled into two operators, a smoothing operator, and a recovering operator, as shown in Figure 1. The two operators play consecutively at all layers of MLPs. The smoothing operator is a spherical normalization function that constrains the features (i.e., intermediate results) on the hypersphere whose local structures are smooth everywhere, thus encouraging the feature space and final result to be smooth. The recovering operator is a Hadamard product that multiplies the features with learned scaling vectors to manipulate feature spaces, thus enabling it to represent complex local structures. Thus the inductive bias of the smoothing operator is global smoothing regularization that reduces overfitting, and the recovering operator controls local sharpness that helps to reduce underfitting. Like FiLM (Perez et al. 2018), the recovering operator mod-

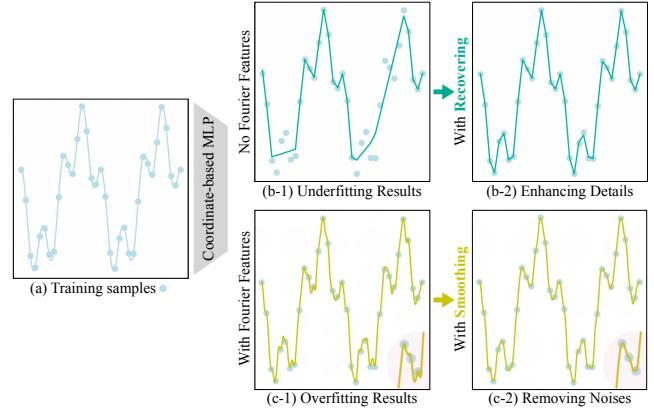


Figure 3: Reconstructing a 1D signal in (a) with coordinate-based MLPs could suffer from underfitting (b-1) when the capacity of MLPs is insufficient or overfitting (c-1) when using overly-high frequency for the coordinate embedding of Fourier Features. To overcome individual issues, we present a recovering filter that restores local structures to the underfitting results (b-2) and a smoothing filter that suppresses local high frequencies from the overfitting results (c-2).

ulates features with scaling vectors generated conditioning on coordinates or their Fourier features. It scales the features up or down or negates them to change their shape locally to better fitting the target. The two operators enable learning adaptive feature embedding at spatial locations containing different scales of variations.

The two filtering operators can play standalone to deal with either overfitting or underfitting issues or combined with the other when both occur. A 1D toy example in Fig-

ure 3 demonstrates the effects of the two filtering modes. The smoothing operator removes unexpected local bumps on the curve in Figure 3, (c-1)→(c-2), and the recovering operator restores missing components and fine details, (b-1)→(b-2). Note that (b-1) and (b-2) have similar network capacities, and so as (c-1) and (c-2). For a complex signal containing both high-frequency and low-frequency, applying both filtering modes helps to construct both spectrum bandwidths, in a way that the former first overly smooths the entire regions, and the latter brings fine details to the regions over-smoothed.

To demonstrate the effectiveness of the filtering module, we evaluate our network on several tasks. Experimental results show that our method performs better than STOA methods. Moreover, our filtering design yields a faster convergence speed and better network stability under SGD optimization. Our contributions can be summarized as follows.

- We propose a new filtering scheme for MLPs that learns adaptive reconstruction on data containing low and high frequencies. Our filter can remove noisy artifacts while enhancing distinct details that significantly alleviate the dilemma of overfitting and underfitting issues.
- Our filter design is simple and flexible to control the sharpness and smoothness of the reconstruction. We show the superiority of our method in several tasks against STOA methods.

Related Work

Implicit Neural Representation. Deep neural network has been shown to be highly effective for learning implicit function that represents images (Tancik et al. 2020; Sitzmann et al. 2020; Mildenhall et al. 2020; Chen, Liu, and Wang 2021) and 3D objects and scenes (Park et al. 2019; Mescheder et al. 2019; Chen and Zhang 2019; Atzmon and Lipman 2020; Gropp et al. 2020; Sitzmann, Zollhöfer, and Wetzstein 2019). They incorporate coordinates as inputs to MLPs that could be sampled to an arbitrarily high spatial resolution. Thus such representation can directly be applied to super-resolution tasks. Other applications include view-synthesis (Mildenhall et al. 2020; Martin-Brualla et al. 2021; Yu et al. 2021; Bemana et al. 2020), point cloud based 3D reconstruction (Atzmon and Lipman 2020; Gropp et al. 2020; Williams et al. 2021, 2019) and single-image 3D reconstruction (Mescheder et al. 2019; Xu et al. 2019, 2020; Saito et al. 2019). Besides generative tasks, implicit representation is also applicable and effective for other tasks, such as feature matching (Chen et al. 2021) and scene understanding (Zhi et al. 2021).

As evidenced by (Basri et al. 2020; Rahaman et al. 2019), ReLU-based MLP networks are struggling to capture highly detailed signals, due to the spectral bias trait. FFN uses positional encoding to map input coordinates of signals to a high dimensional space using sinusoidal functions. Meanwhile, SIREN (Sitzmann et al. 2020) replaces the ReLU activations in the MLP network with sinusoidal functions. They share a similar spirit that manipulates the input or intermediate results in the frequency domain to capture a wider frequency bandwidths in the output. Further, to better fit individual

data, the sinusoidal functions in positional encoding or MLP networks are devised to be learnable (Fathony et al. 2021; Mehta et al. 2021). In contrast to using sinusoidal functions, Spline Positional Encoding (SPE) (Wang et al. 2021) studies learnable spline functions for coordinate embedding. With sufficient local supports of splines, the SPE can also approximate the signal with high frequencies. However, when using a small number of local supports, the boundaries between patches of the spline become notable that introduces stripe noises and significantly reduce the visual quality.

Reconstructing high-frequency details is commonly at the expense of introducing visual artifacts to the results. Some recent advances have led to structured or hierarchical designs that can further close the gap between generated results and the target function. They divide the complex functions of 3D object and scenes (Chabra et al. 2020; Genova et al. 2020; Jiang et al. 2020) or image (Mehta et al. 2021) into regular subregions, and fitting each subregions individually while considering globally consistency.

Region subdivision can hardly take effect when a subregion itself is complex. More recently, spatial adaptation on frequencies appears as an improved solution. SAPE (Hertz et al. 2021) presents a progressive optimization strategy to encode signals with increasing frequencies at individual spatial locations. The method reduces the noisy artifacts but generates over-smoothed results since it resists overfitting and leans toward underfitting.

This paper seeks an alternative scheme that uses filters to adapt the intermediate outputs of the MLPs. Existing methods, such as MFN (Fathony et al. 2021) and BACON (Lindell et al. 2022), pi-GAN (Chan et al. 2020), and ModSIREN (Mehta et al. 2021) also adapts the intermediate results of MLPs by modulating the activation functions. Particularly, they modulate frequency, phase, or amplification parameters of sinusoidal functions. Tuning those parameters may cause global and possibly large changes in the results, hence those parameters are difficult to optimize and sensitive to initialization. Rather than modulating sinusoidal functions, FINN runs the filtering after each MLP layer, outside sinusoidal functions, thus allowing more local controls on the intermediate results and more stable performance.

Image Filtering with Deep Neural Networks. Image filters usually compute the weighted average of neighboring pixels of the image as output or sometimes leverage regularization terms for image optimization. Recently, a line of research presents filters with neural networks that can be learned from a large number of datasets for various applications (Jianpani, Kiefel, and Gehler 2016; Zhang et al. 2017; Yoon et al. 2015; Gharbi et al. 2017; Chen, Xu, and Koltun 2017; Li et al. 2019), such as image super-resolution, denoising, and deblurring, etc. They are mainly based on the convolutional network from which the local neighborhood of the query point is determined. For continuous functions, it is difficult to be aware of the neighbors except to discretize the input domain with prior feature scales or resolutions. Thus, we attempt to devise a general filtering scheme for MLPs dealing with image functions and other continuous signals and handling different frequency bandwidths.

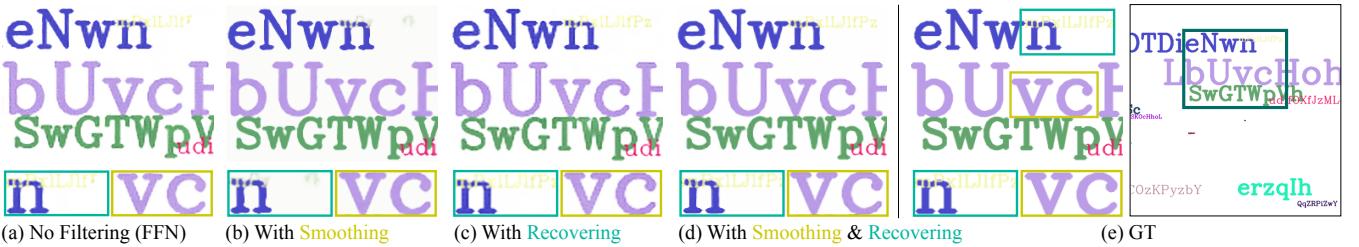


Figure 4: Effects of filtering operators. The target (e) contains homogenous regions everywhere and some tiny-scale texts each referred to low-frequency or high-frequency respectively. (a) is generated by FFN with no filtering and observes random noises (yellow rectangle). With smoothing (b), smoother texts are reconstructed but at the expense of smoothing out some small-scale structures (green rectangle). In contrast, the recovering operator (c) restores fine details but is incapable of reducing noise. Combining two filtering operators enables FINN smooths noises and recovers small structures (d).

Method

In this section, we introduce a new fully-connected neural network with filtering modules, as illustrated in Figure 1. We first present the neural implicit functions and the positional encoding. Next, we describe the the filtering functions.

Neural Implicit Functions

An implicit function is a continuous function $f_\theta : \mathbb{R}^a \rightarrow \mathbb{R}^n$ which takes as input a coordinate of any query point from the Euclidean space $x \in \mathbb{R}^a$ and predicts a value in the target domain \mathbb{R}^n . Learning f_θ with a neural network requires a set of coordinate samplings as input and corresponding values as output. Examples of the f_θ include mapping pixels to intensity values for image regression or projecting 3D coordinates to scalar values for 3D surface reconstruction.

It is common to model implicit function f_θ with MLPs. However, ReLU-based MLPs can hardly fit high-frequency signals due to the spectral bias trait, leading to the severe underfitting problem. Hence, to avoid favoring the low-frequency spectrum, FFN projects the input to the frequency domain with a controllable frequency bandwidth. Specifically, the coordinate x is represented by a d -dimensional Fourier feature vector, $\gamma(x) \in \mathbb{R}^d$, as follows:

$$\gamma(x) = \frac{s}{\sqrt{d}} [\cos(2\pi x B^T) \| \sin(2\pi x B^T)] \quad (1)$$

where $\|$ is the concatenation of two vectors, and B is a $\frac{d}{2} \times a$ matrix randomly sampled from Gaussian distribution with standard deviation σ . s is a constant for global scaling.

Fourier feature embedding facilitates learning complex functions with wider frequency bandwidths, but at the expense of overfitting issue: introducing unexpected dramatic local variations (e.g., noises, local bumps) to the reconstructed function. Lowering the global σ reduces the noise but towards underfitting. Adapting local σ for individual coordinates also leads to over-smoothing (i.e., SAPE). Thus we present a filtering scheme to adaptively influence the reconstruction to alleviate the dilemma of overfitting and underfitting issues.

Filtering Functions

Let $\theta_i, i = 1, 2, \dots, k$ denote layers of a k -layer MLPs, then the coordinate x passing through the MLPs will produce a

sequence of outputs, such as $y_i, i = 1, 2, \dots, k$. To reconstruct the target function, we apply supervision on the final output function y_k . And to control the regional sharpness in y_k , we modulate all the intermediate features of MLPs with a filtering function. It is also possible to filter the final output directly, but more sophisticated designs are needed to trade off the data fidelity and the filtering purpose. The formulation of the new network function can be written in recurrence as follows:

$$\begin{aligned} F_x &= \gamma_s(x) M^T \\ y_1 &= \theta_1(\gamma(x)) \\ y_i &= o(\theta_i(y_{i-1})) \odot F_x, i = 2, \dots, k-1 \\ f_\theta &= \theta_k(y_{k-1}) \end{aligned} \quad (2)$$

where $o(\cdot)$ is the spherical normalization function, such that $o(v) = \frac{v}{\|v\|}$, and \odot is the Hadamard product. The filter learns to adaptively influence the output by applying normalization and a linear transformation to the MLP's intermediate features based on spatial locations. Note that F_x are generated from input Fourier features $\gamma_s(x)$ that can be different to $\gamma(x)$, the input of MLPs. This enables more flexibility in modeling wide frequency bandwidths by setting different σ for $\gamma(x)$ and $\gamma_s(x)$. For example, for the color field and signed distance field, $\gamma(x)$ and $\gamma_s(x)$ can be the same, but for mixing fields, like neural radiance field containing color and density field, better to be different.

Figure 3 shows the filtering effects of the two operators on a 1D example. And a 2D example in Figure 4 also demonstrates the effects of the two operators and their combination. The recovering operator in Figure 4 (c) reconstructs more details than the original FFN (a) but is unable to avoid noise, and the smoothing operator (b) smoothes out noise, but some distinct details are missing. The joint operator (d) recovers fine-grained details without introducing more noise.

As mentioned earlier, the inductive bias of the smoothing and recovering operators provides a way of modulating the feature spaces to influence the output. Specifically, their combination first smooths everywhere and then modulates features in a selective manner, scaling down features to remain smooth in homogenous regions and turning them up or negating them to increase variations in sharp regions.

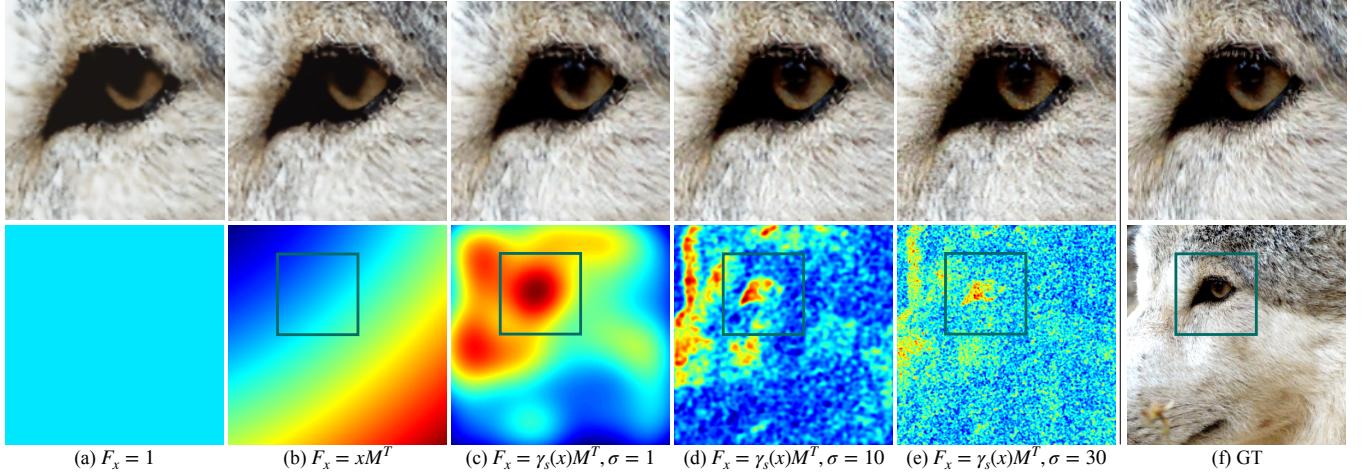


Figure 5: (Top) The reconstruction results reveal a different degree of sharpness depending on scaling vectors F_x (Bottom), which is a constant (a) or generated from the coordinates x (b) or Fourier features $\gamma_s(x)$ (c-e). F_x directly generated from x usually leads to underfitting. In contrast, Fourier features have more flexibility in controlling the fitting capability. In particular, larger σ adds more variations to the scaling vectors, resulting in sharper reconstruction, and an overly-high σ creates more randomness and noise that reduces generalization. (d) has the best PSNR that balances sharpness and smoothness well.

Filtering Effects Besides tuning the frequency parameter in $\gamma(x)$, one can also fix $\gamma(x)$ and modify $\gamma_s(x)$ to control the sharpness of reconstructions, since $\gamma_s(x)$ controls the complexity of scaling vector space. For example, smoothing reconstruction requires simpler scaling vector space that can be generated conditioning on a less complex input, e.g., a constant, x , or a smaller σ in $\gamma_s(x)$. In Figure 5, we draw F_x in the bottom row, including $F_x = 1$, xM^T , and $\gamma_s(x)M^T$ ($\sigma = 1, 10$ and 30). The scaling vectors and their corresponding reconstruction results reveal a different degree of sharpness. In practice, $\gamma(x)$ and $\gamma_s(x)$ can be different for adapting hybrid fields or the same for single fields.

Experiments

We validate the benefits of applying filtering in learning 2D images and 3D shapes, i.e., single fields, and novel synthesis supported by color and density field, i.e., hybrid field.

Image Regression and Generalization

Image functions map a 2D coordinate to RGB color. Our network has 512-dimensional Fourier features ($\sigma = 10, s = 80$ for $\gamma(x)$ and $\gamma_s(x)$), MLPs with 3 hidden layers, 256 channels, ReLU activation and Sigmoid on output. In addition, our filtering module contains a 512×256 matrix for generating scaling vectors. We use MSE loss to train the network for 2k epochs with a learning rate of $1e-3$.

We compare our method to state-of-the-art methods, including FFN (Tancik et al. 2020), SIREN (Sitzmann et al. 2020), SPE (Wang et al. 2021) and SAPE (Hertz et al. 2021), on the datasets of “natural” and “text” images from FFN. The training pixels are sampled on a regularly-spaced grid containing 25% of the pixels in the image. Then, we use all the pixels for testing. We report PSNR and FLIP (Andersson et al. 2020) in Table 1 and qualitative results in Fig-

Model	PSNR↑		$\text{FLIP} \downarrow$	
	Natural	Text	Natural	Text
FFN	25.57 ± 4.19	30.47 ± 2.11	0.131 ± 0.041	0.096 ± 0.043
SIREN	27.03 ± 4.28	30.81 ± 1.72	0.114 ± 0.040	0.070 ± 0.020
SPE	26.49 ± 3.89	31.12 ± 2.18	0.130 ± 0.038	0.065 ± 0.022
SAPE	28.09 ± 4.04	31.84 ± 2.15	0.118 ± 0.026	0.083 ± 0.041
FINN	28.51 ± 4.35	33.09 ± 1.97	0.100 ± 0.037	0.042 ± 0.016

Table 1: Quantitative results of image generalization.

	FINN _{so}	FINN _{ro}	FINN _{in}	FINN _{fxs}
Natural	27.91 ± 4.21	28.09 ± 4.06	28.26 ± 4.35	28.59 ± 4.39
Text	31.19 ± 1.86	31.99 ± 1.73	32.22 ± 2.58	33.17 ± 2.01

Table 2: FINN variants on image generalization.

ure 6. FLIP provides an error map and a global measurement (i.e., the weighted median of errors on all pixels). Our method consistently outperforms all compared methods on all datasets and metrics with considerable gains. The visual results show that all the methods can generate realistic images. Nevertheless, SAPE produces over-smoothed images. FNN and SIREN usually include random noisy artifacts, and SPE contains stripe noise. FINN reconstructs better in both backgrounds (low-frequency) and small structures (high-frequency), handling over smooth and noise well.

Design Choices. We validate each component of our filter and show the numerical results of the variants in Table 2.

Using smoothing or recovering operator only. FINN_{ro} and FINN_{so} incorporate the recovering and smoothing operator to FFN, respectively. The two variants perform worse

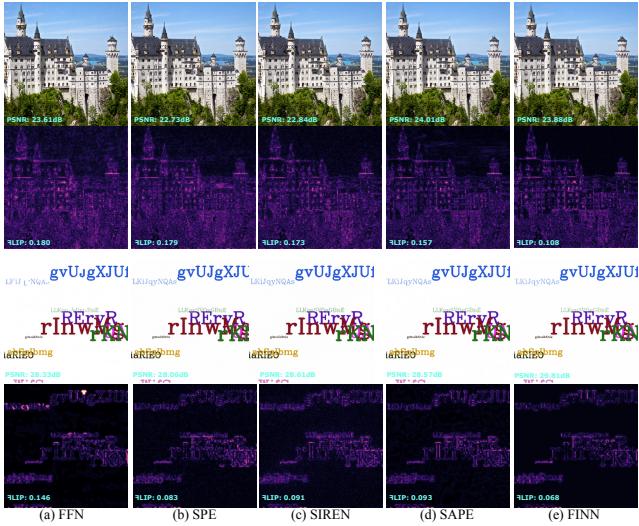


Figure 6: Qualitative results of image generalization.

than FINN but both outperform FFN. It indicates that either suppressing noises or enhancing details helps to improve the reconstruction. Figure 4 shows the filtering effects of individual operators. Their combination reaches an adaptive reconstruction for different levels of detail.

Filtering on inputs instead of features of MLPs. We apply the filter to input Fourier features with a variant denoted by FINN_{in} . Although FINN_{in} also learns adaptive feature embedding and achieves better performance than FFN, its performance is inferior to FINN since the no filtered layers of MLPs can still overfit the results, making the smoothing filter less effective.

Using layer-wise scaling vectors F_x^i instead of a global F_x . FINN_{fx} means scaling vectors passed to different MLP layers are generated independently. Even adding more network parameters, its performance gain over the global F_x is negligible. The main reason is that F_x and all F_x^i are derived directly from the same input Fourier features by a simple linear transformation, which limits their capability of generating diverse scaling vectors. It is possible to use a more complex scaling vector generator instead of a simple linear mapping to boost the performance. But it is not encouraged since a stronger recovering operator will take over the smoothing operator and thus reduce the smoothing effects. In order to control the sharpness, one can manipulate the frequency of Fourier features for scaling vector generator. The filtering effects of various F_x are shown in Figure 5.

Network Convergence and Stability. Figure 7 shows test PSNR curves of FFN (Left) and FINN (Right), trained on 32 images from the “natural” and “text” datasets for 2k iterations. FINN converges after 300 iterations and keeps stable PSNR scores for individual images since then. In contrast, FFN converges slower and experiences more randomness during optimization, yielding slightly curvy and nonmonotonic PSNR curves. The stability suggests filtering operators serve as a regularization to resist over-fitting during opti-

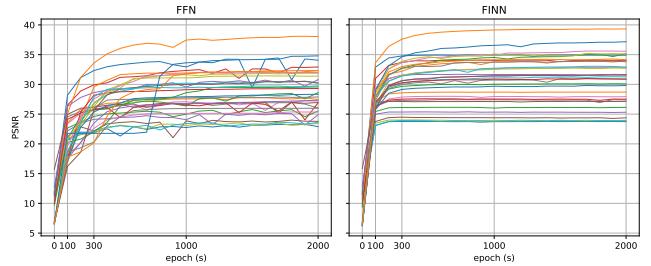


Figure 7: Convergence and stability comparison of networks with and without filters.

mization, leading to better generalization.

3D Inverse Rendering for View Synthesis

The view synthesis task receives a set of 2D images with known camera poses and generates images from novel viewpoints. We construct a vector field on 3D space from input images to represent the scene using Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) that utilizes MLPs to map from 3D coordinates to RGB colors and volume densities. Then one can render images from the field with given viewpoints via volume rendering. NeRF is optimized via MSE loss between the input images and their re-rendered images from the predicted field. Following FFN, we use a “simplified NeRF” that removes hierarchical sampling and view dependence from the original NeRF framework for evaluation.

We let FINN and FFN have the same number of network parameters: FFN, 512-dimensional Fourier features, MLP with 6 layers, 256 channels, ReLU activation, sigmoid on RGB output; and FINN, similar to FFN, except that a 4 layers MLP and a scaling vector generator with 512×256 parameters. Both models are trained for 50k iterations using the Adam optimizer with a learning rate of 5×10^{-4} . We use two datasets for evaluation, a synthesized scene “Lego” and a real scene “Fern”. The number of training and testing images is (100, 25) for “Lego” and (17, 3) for “Fern”.

FFN sets σ to 6.05 for Lego and 5.0 for Fern by grid search. Different from the parameter setting in image regression, we uses a different σ for $\gamma(x)$ and $\gamma_s(x)$ in learning NeRF, such that 12, 1 for Lego and 10, 1 for Fern. The main reason is NeRF contains a density field and an RGB color field that have different complexity and require different frequencies. For example, images (color fields) expect a higher σ , 10, and signed distance fields (similar to density fields) only use $\sigma \approx 1$ described in the next section. Thus a better encoding scheme will use different frequency parameters for individual fields. However, this will need more sophisticated designs and is beyond our research. FINN provides a simple and flexible way to model such hybrid fields by setting high frequency in $\gamma(x)$ and low frequency in $\gamma_s(x)$, allowing the network to adjust the local variations of the features and final output in a wide range.

The results in Table 3 show that FINN performs better than FFN in both training and testing on the two datasets. Figure 8 shows FINN also provides better rendering quality,

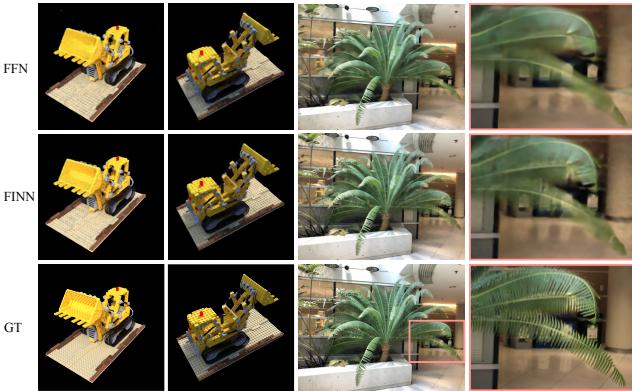


Figure 8: Qualitative results on novel view synthesis.

Model	Lego		Fern	
	Train PSNR	Test PSNR	Train PSNR	Test PSNR
FFN	26.25 ± 1.00	25.74 ± 1.06	25.06 ± 0.44	24.30 ± 0.67
FINN	26.70 ± 1.00	26.02 ± 1.10	25.99 ± 0.53	25.28 ± 0.63

Table 3: Numerical results on novel view synthesis.

revealing more distinct details with fewer noises in the results. For sparse training images from Fern, the performance gain and visual improvement of FINN are significant, evidencing that FINN can fit a wider frequency bandwidth and generalizes better to unknown viewpoints.

3D Shape Reconstruction

Further, we evaluate our method in the 3D surface reconstruction task. We consider the 3D surface as a zero level-set of the signed distance field (SDF) from which every 3D query point has a corresponding value telling its distance and whether inside or outside from the surface. We train the network to fit the SDF on point clouds with additional point normals information. In addition to SDFs and normals for supervision, we also apply a regularization term to constrain the gradient of the learned field to be a unit vector at any spatial locations, as suggested by (Atzmon and Lipman 2020; Gropp et al. 2020). The loss function is defined as

$$\begin{aligned} \mathcal{L}_{\text{sdf}} = & \sum_{x \in \Omega} ||\nabla f_\theta(x)|| - 1| + \sum_{x \in \Omega \setminus \Omega_0} \exp(-|f_\theta(x)|) \\ & + \sum_{x \in \Omega_0} (|f_\theta(x)| + \|\nabla f_\theta(x) - n(x)\|) \end{aligned} \quad (3)$$

where Ω is the point set and Ω_0 contains only on-surface points. $f_\theta(x)$ and $\nabla f(x)$ are the fitted SDFs and gradients, and $n(x)$ is the ground truth point normal. Each loss term is weighted by a constant for optimization.

Our network structure for this task is similar to image regression, except that the output layer is linear, and hidden layers increase from 3 to 4. SDFs are *scalar* fields, usually less complex than vector fields, e.g., RGB color fields, thus we use a small frequency, $\sigma \in [1, 2]$ for FFN and FINN, and

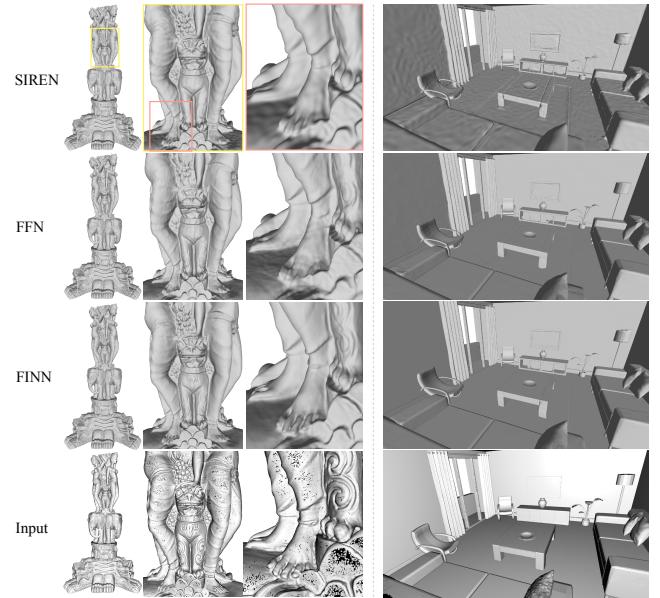


Figure 9: Large-scale surface reconstruction. All methods show good reconstructions. With filtering, FINN suppresses high frequencies on homogenous regions and restores sharp geometric features.

reduce the dimension of the Fourier features to 256. Training SDF requires dense samples on the underlying surface and sparse samples randomly scattered in a bounding cube. After trained, we sample a set of regular grid points to obtain the SDFs and use the Marching Cubes algorithm (Lorensen and Cline 1987) to extract the triangular mesh.

We compare our method with FFN and SIREN on large-scale 3D meshes containing tens of millions of triangles. The examples include an object “Thai Statue”, containing very complex geometric features, and a scene “Interior Room”, having large flat patches and small shapes, as shown in Figure 9. We set σ to 1 for those two examples for FFN and FINN and use 1600^3 grid points to extract high-resolution triangular meshes. The results show that FINN reconstructs both flat (low-frequency) and sharp (high-frequency) surface regions with high fidelity. The results of FFN and SIREN are also plausible, but introduce unexpected local bumps and noises and lose many fine surface details. We will present numerical results in the appendix.

Conclusion

We present a novel filtering scheme in MLPs for implicit functions. The filter improves the reconstruction of complex signals containing a wide range of frequencies. Our filter combines a smoothing and a recovering operator that first globally over-smooths the result and then adaptively scales local variations to restore fine details smoothed. Such a filtering scheme helps to fit both high-frequencies and low-frequencies with high fidelity. We demonstrate the effectiveness of our filter on several tasks and show significant improvement against state-of-the-art methods.

References

- Andersson, P.; Nilsson, J.; Akenine-Möller, T.; Oskarsson, M.; Åström, K.; and Fairchild, M. D. 2020. FLIP: A Difference Evaluator for Alternating Images. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 3(2): 15:1–15:23.
- Atzmon, M.; and Lipman, Y. 2020. SAL: Sign Agnostic Learning of Shapes From Raw Data. In *CVPR*, 2562–2571. Computer Vision Foundation / IEEE.
- Basri, R.; Galun, M.; Geifman, A.; Jacobs, D.; Kasten, Y.; and Kritchman, S. 2020. Frequency Bias in Neural Networks for Input of Non-Uniform Density. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 685–694. PMLR.
- Bemana, M.; Myszkowski, K.; Seidel, H.-P.; and Ritschel, T. 2020. X-Fields: Implicit Neural View-, Light- and Time-Image Interpolation. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2020)*, 39(6).
- Chabra, R.; Lenssen, J. E.; Ilg, E.; Schmidt, T.; Straub, J.; Lovegrove, S.; and Newcombe, R. A. 2020. Deep Local Shapes: Learning Local SDF Priors for Detailed 3D Reconstruction. In *ECCV* (29), volume 12374 of *Lecture Notes in Computer Science*, 608–625. Springer.
- Chan, E.; Monteiro, M.; Kellnhofer, P.; Wu, J.; and Wetzstein, G. 2020. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *arXiv*.
- Chen, Q.; Xu, J.; and Koltun, V. 2017. Fast Image Processing with Fully-Convolutional Networks. In *ICCV*, 2516–2525. IEEE Computer Society.
- Chen, Y.; Fernando, B.; Bilen, H.; Mensink, T.; and Gavves, E. 2021. Neural Feature Matching in Implicit 3D Representations. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 1582–1593. PMLR.
- Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8628–8638.
- Chen, Z.; and Zhang, H. 2019. Learning Implicit Fields for Generative Shape Modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Fathony, R.; Sahu, A. K.; Willmott, D.; and Kolter, J. Z. 2021. Multiplicative Filter Networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Genova, K.; Cole, F.; Sud, A.; Sarna, A.; and Funkhouser, T. A. 2020. Local Deep Implicit Functions for 3D Shape. In *CVPR*, 4856–4865. Computer Vision Foundation / IEEE.
- Gharbi, M.; Chen, J.; Barron, J. T.; Hasinoff, S. W.; and Durand, F. 2017. Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph.*, 36(4): 118:1–118:12.
- Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit Geometric Regularization for Learning Shapes. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 3789–3799. PMLR.
- Hertz, A.; Perel, O.; Giryes, R.; Sorkine-Hornung, O.; and Cohen-Or, D. 2021. SAPE: Spatially-Adaptive Progressive Encoding for Neural Optimization. In *Thirty-Fifth Conference on Neural Information Processing Systems*.
- Jampani, V.; Kiefel, M.; and Gehler, P. V. 2016. Learning Sparse High Dimensional Filters: Image Filtering, Dense CRFs and Bilateral Neural Networks. In *CVPR*, 4452–4461. IEEE Computer Society.
- Jiang, C. M.; Sud, A.; Makadia, A.; Huang, J.; Nießner, M.; and Funkhouser, T. A. 2020. Local Implicit Grid Representations for 3D Scenes. In *CVPR*, 6000–6009. Computer Vision Foundation / IEEE.
- Li, Y.; Huang, J.; Ahuja, N.; and Yang, M. 2019. Joint Image Filtering with Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8): 1909–1923.
- Lindell, D. B.; Van Veen, D.; Park, J. J.; and Wetzstein, G. 2022. BACON: Band-limited coordinate networks for multiscale scene representation. In *CVPR*.
- Lorensen, W. E.; and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics*, 21(4): 163–169.
- Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S. M.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 7210–7219. Computer Vision Foundation / IEEE.
- Mehta, I.; Gharbi, M.; Barnes, C.; Shechtman, E.; Ramamoorthi, R.; and Chandraker, M. 2021. Modulated Periodic Activations for Generalizable Local Functional Representations. In *IEEE International Conference on Computer Vision (ICCV)*.
- Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proc. CVPR*.
- 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 *ECCV* (1), volume 12346 of *Lecture Notes in Computer Science*, 405–421. Springer.
- Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *CVPR*.
- Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- 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 Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5301–5310. PMLR.
- Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision*, 2304–2314.

- Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Sitzmann, V.; Martel, J. N.; Bergman, A. W.; Lindell, D. B.; and Wetzstein, G. 2020. Implicit Neural Representations with Periodic Activation Functions. In *Proc. NeurIPS*.
- Sitzmann, V.; Zollhöfer, M.; and Wetzstein, G. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. In *NeurIPS*, 1119–1130.
- Tancik, M.; Srinivasan, P. P.; Mildenhall, B.; Fridovich-Keil, S.; Raghuvaran, N.; Singhal, U.; Ramamoorthi, R.; Barron, J. T.; and Ng, R. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *NeurIPS*.
- Wang, P.; Liu, Y.; Yang, Y.; and Tong, X. 2021. Spline Positional Encoding for Learning 3D Implicit Signed Distance Fields. In *IJCAI*, 1091–1097. ijcai.org.
- Williams, F.; Schneider, T.; Silva, C. T.; Zorin, D.; Bruna, J.; and Panozzo, D. 2019. Deep Geometric Prior for Surface Reconstruction. In *CVPR*, 10130–10139. Computer Vision Foundation / IEEE.
- Williams, F.; Trager, M.; Bruna, J.; and Zorin, D. 2021. Neural Splines: Fitting 3D Surfaces With Infinitely-Wide Neural Networks. In *CVPR*, 9949–9958. Computer Vision Foundation / IEEE.
- Xu, Q.; Wang, W.; Ceylan, D.; Mech, R.; and Neumann, U. 2019. DISN: Deep Implicit Surface Network for High-quality Single-view 3D Reconstruction. In *NeurIPS*, 490–500.
- Xu, Y.; Fan, T.; Yuan, Y.; and Singh, G. 2020. Ladybird: Quasi-Monte Carlo Sampling for Deep Implicit Field Based 3D Reconstruction with Symmetry. In *ECCV(1)*, volume 12346 of *Lecture Notes in Computer Science*, 248–263. Springer.
- Yoon, Y.; Jeon, H.; Yoo, D.; Lee, J.; and Kweon, I. S. 2015. Learning a Deep Convolutional Network for Light-Field Image Super-Resolution. In *ICCV Workshops*, 57–65. IEEE Computer Society.
- Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixel-NeRF: Neural Radiance Fields from One or Few Images. In *CVPR*.
- Zhang, J.; Pan, J.; Lai, W.; Lau, R. W. H.; and Yang, M. 2017. Learning Fully Convolutional Networks for Iterative Non-blind Deconvolution. In *CVPR*, 6969–6977. IEEE Computer Society.
- Zhi, S.; Laidlow, T.; Leutenegger, S.; and Davison, A. 2021. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

More Results

Surface Reconstruction

We report numerical results of 3D shape reconstruction on a collection of medium-size point sets in Table 4. The input point clouds contain tens of thousands of points and include holes in the bottom and valleys, as shown in Figure 10 (training samples). FINN generally outperforms FFN on all testing shapes except ‘‘Bimba’’, whose bottom part is empty and reconstruction has large uncertainty. The qualitative results show both FINN and FFN can fit points and interpolate holes well. FINN provides slightly better local details mainly because of the recovering operator. The underlying shapes of input point sets are generally smooth, especially compared with RGB and its hybrid fields. Thus a smaller frequency parameter is enough for reconstruction and is free of noisy artifacts. To further validate the effectiveness of removing noises, we perform 3D reconstruction on noisy point clouds.

Robustness to Noise. We randomly perturb the point samples by adding each point with a vector drawn from gaussian noise, $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.02$. And we use the original points and perturbed points as training samples. FINN can deal with noisy point sets without requiring explicit regularizer terms. As shown in Figure 11, FINN reconstructs smooth and clean shapes, while without filtering, FFN overfits noisy points, creating random patches along underlying surfaces.

Network Generalization

Interpolation Results. We fit an image at 128×128 resolution and upsample to 512×512 pixels to validate the learned interpolants, FINN and FFN, equipped with or without filtering. The results in figure 12 (c) show that FFN and FINN can both fit the training pixels faithfully. Then we upsample the fit images by a factor of 4 along vertical and horizontal directions using bilinear interpolation (d) and the learned interpolants (e). The results based on bilinear interpolation (d) are similar, with very few variations within individual characters. The interpolation via networks (e) without filtering adds many artifacts by randomly deeper and shallower the colors. That indicates when FFN allows ReLU-MLP to fit high-frequency training samples, it also introduces high-frequency noises between them, leading to overfitting. FINN lowers the variations between nearby training pixels, thus generalizing better to unseen pixels. Further, we will compare the interpolations at various sampling rates.

Impact of Sampling Rate. To understand the behaviors of interpolation at dense and sparse training sampling, we acquire a set of images by downsampling a given image with decreasing sampling rates, ranging from 100% to $1/256$, as shown in Figure 13 top row. The networks with or without filtering both interpolate well at unseen pixels (a-c) and fails (d-e) when the distances between neighboring training pixels are too long. Although failed at sparse reconstruction (e), the experiment shows that FINN generates smooth interpolation between training pixels, while FFN overfits individual pixels and experiences randomness at unseen pixels.

	Bunny	Armadillo	Dragon	Fandisk	Bimba	Gargoyle
FFN	2.04	2.92	1.97	1.77	139.06	4.58
FINN	1.90	2.84	1.92	1.69	138.11	4.24

Table 4: Numerical results of 3D shape reconstruction. Chamfer distance is multiplied by 10^5 .

Experimental Details

Networks

The detailed networks for 2D images, 3D surface, and NeRF are shown in Figure 14. The source code can be found in the attached file.

Baselines include FFN (Tancik et al. 2020), SIREN (Sitzmann et al. 2020), SPE (Wang et al. 2021) and SAPE (Hertz et al. 2021). Their codes can be downloaded from the following links,

- FFN: <https://github.com/tancik/fourier-feature-networks>.
- SIREN: <https://github.com/vsitzmann/siren>.
- SAPE: <https://github.com/amirhertz/SAPE>.
- SPE: <https://github.com/microsoft/SplinePosEnc>.

Simplified NeRF can be found at <https://github.com/tancik/fourier-feature-networks>. We use a PyTorch implementation of NeRF from <https://github.com/yenchenlin/nerf-pytorch/> and modify it for our network.

Compared to FFN, FINN usually needs more network parameters because of the scaling vector generator. For d -dimensional Fourier features and m -dimensional hidden layers of MLP, our method need $d \times m$ more parameters, i.e., 512×256 , 256×256 , 512×256 for image, shape and NeRF respectively. By increasing the number of hidden layers for FFN, we can set their network capacities to be the same, e.g., we use a 6-layer MLP in FFN and a 4-layer MLP in FINN for the NeRF task.

Additional Training Details

Surface Reconstruction. The batch size for training large shapes is 200k, which contains 100k on-surface points sampled from the input and 100k off-surface points randomly generated within the box. For medium-size shapes, we set the batch size to 32^3 for training and 256^3 for extracting meshes. To compute Chamfer distance, we uniformly sample 64^3 surface points. Loss terms in Equation 3 are weighted as follows,

$$\begin{aligned} \mathcal{L}_{\text{sdf}} = & \sum_{x \in \Omega} a_1 \|\nabla f_\theta(x)\| - 1 \\ & + \sum_{x \in \Omega \setminus \Omega_0} a_2 \cdot \exp(-50|f_\theta(x)|) \\ & + \sum_{x \in \Omega_0} (a_3 |f_\theta(x)| + a_4 \|\nabla f_\theta(x) - n(x)\|) \end{aligned} \quad (4)$$

where a_1, a_2, a_3 and a_4 are set to 1000, 10, 5 and 20 respectively. σ is set to 1 for all examples.

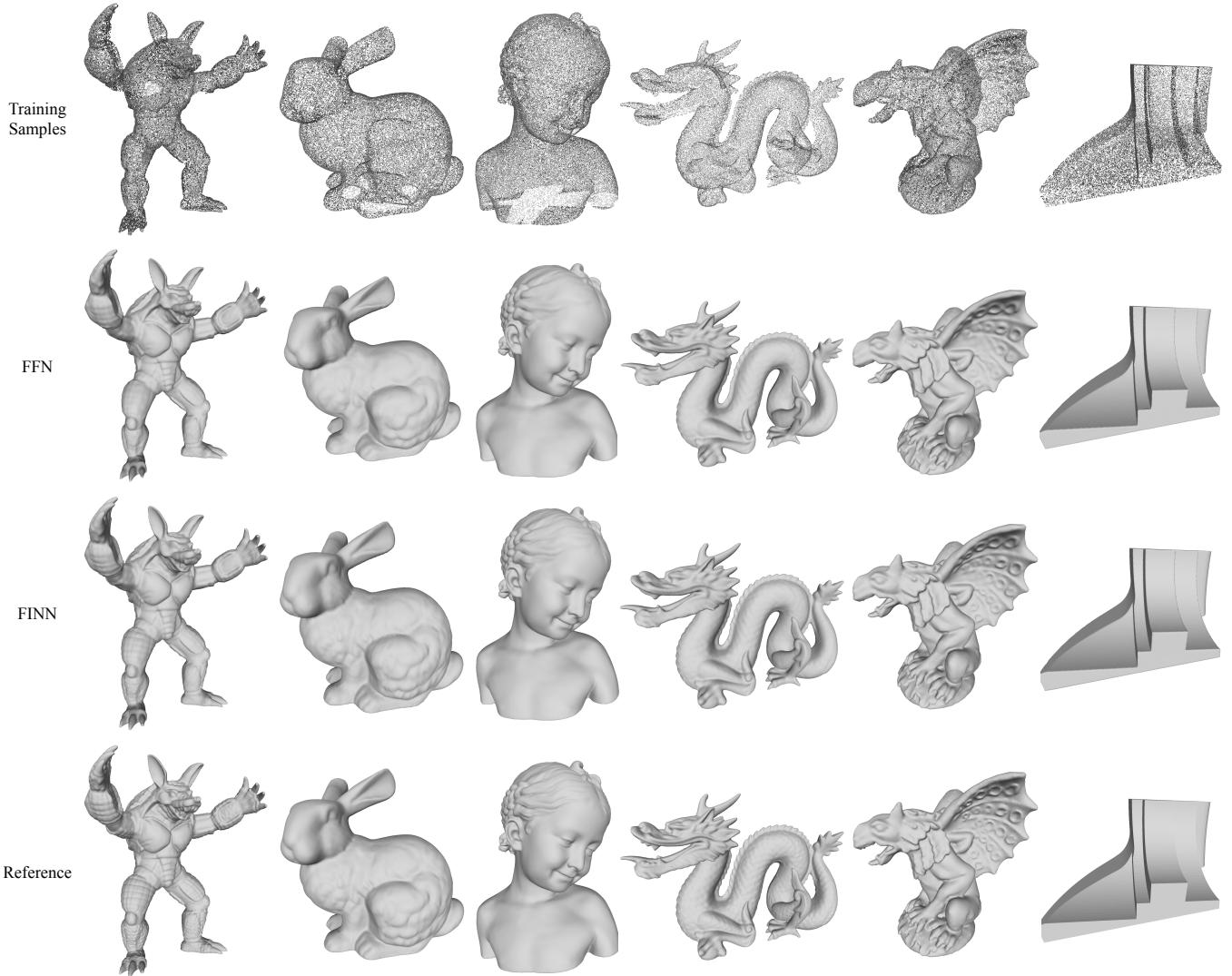


Figure 10: 3D shape reconstruction from point clouds.

NeRF. We use the “Lego” and “Fern” scene datasets. Lego contains 100 training images rendered from a virtual scene with black background colors. Fern has 17 training images captured in the real world. The ground truth camera poses are computed via structure from motion (SfM) that optimizes the extrinsic parameters of input images and the corresponding 3D locations of selected pixels. It is common to open source COLMAP (Schönberger and Frahm 2016) for SfM. We sample 1024 rays with 128 points along each ray in a batch for training. The training takes around an hour for FFN and FINN for 50k epochs.

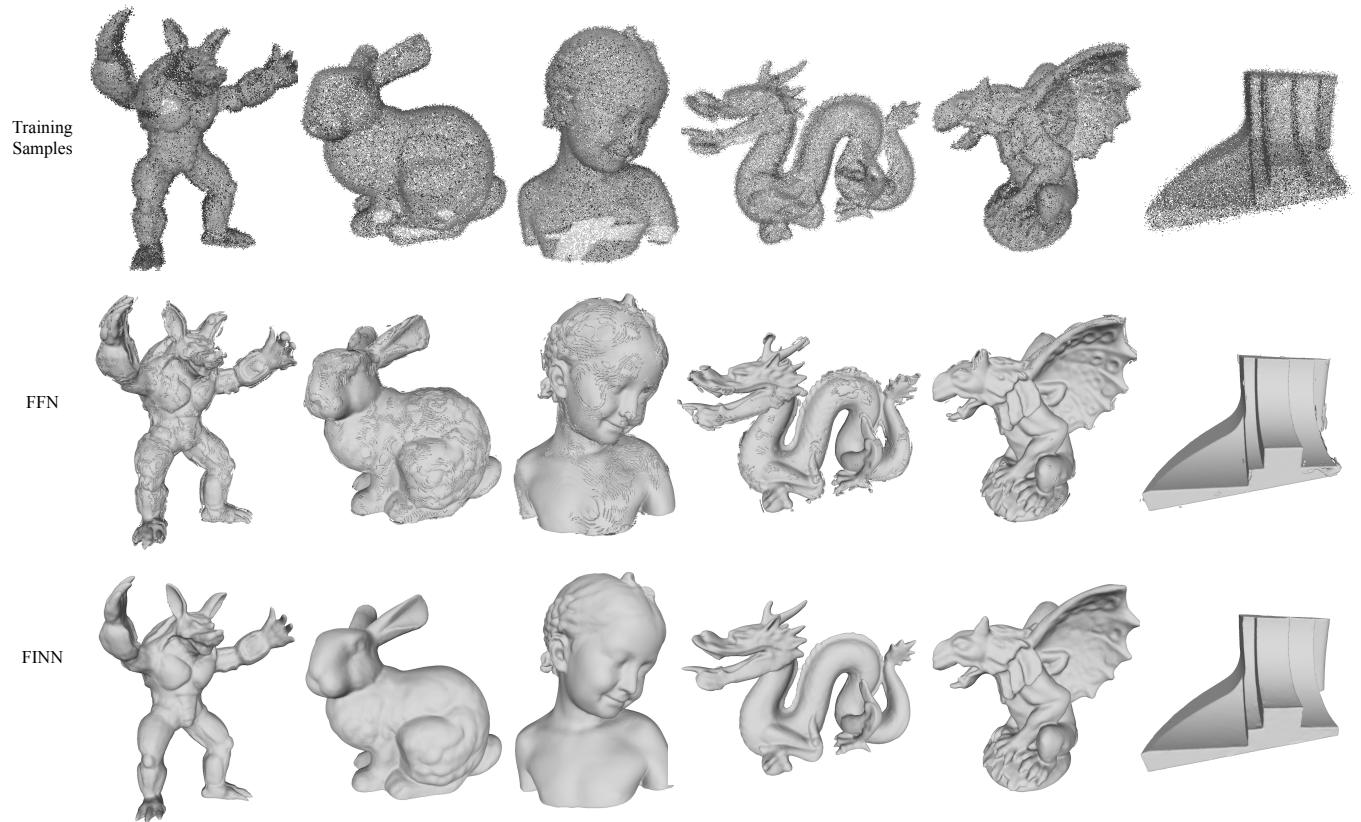


Figure 11: 3D shape reconstruction from noisy point clouds.

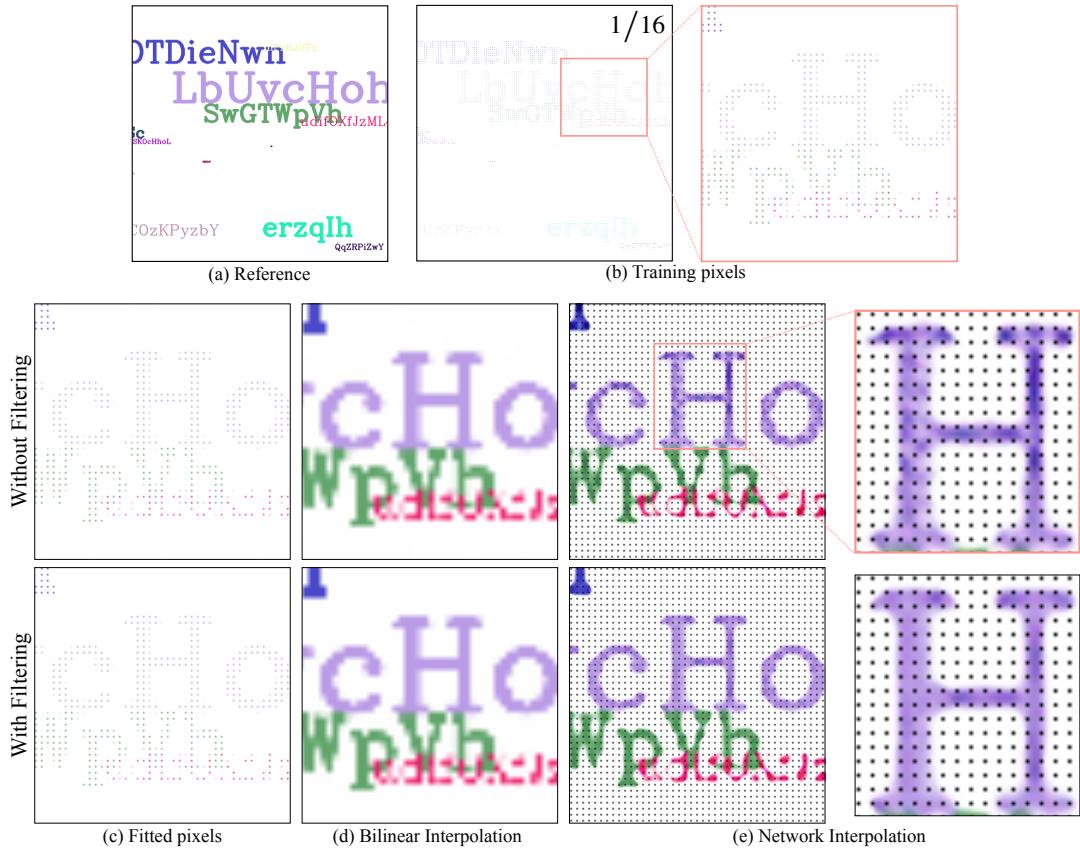


Figure 12: Generalization validation via image super-resolution. A fitting image (c) is upsampled by 16 via bilinear interpolation (d) or networks (e). While the results of bilinear interpolation are similar, the results of networks are very different. With filtering, FINN consumes high-frequency Fourier features for fitting, similar to FFN but at the same time suppresses high-frequency variations at the neighbors of training samples, thus interpolating pixels with fewer artifacts. Training pixels in (e) are painted in black.

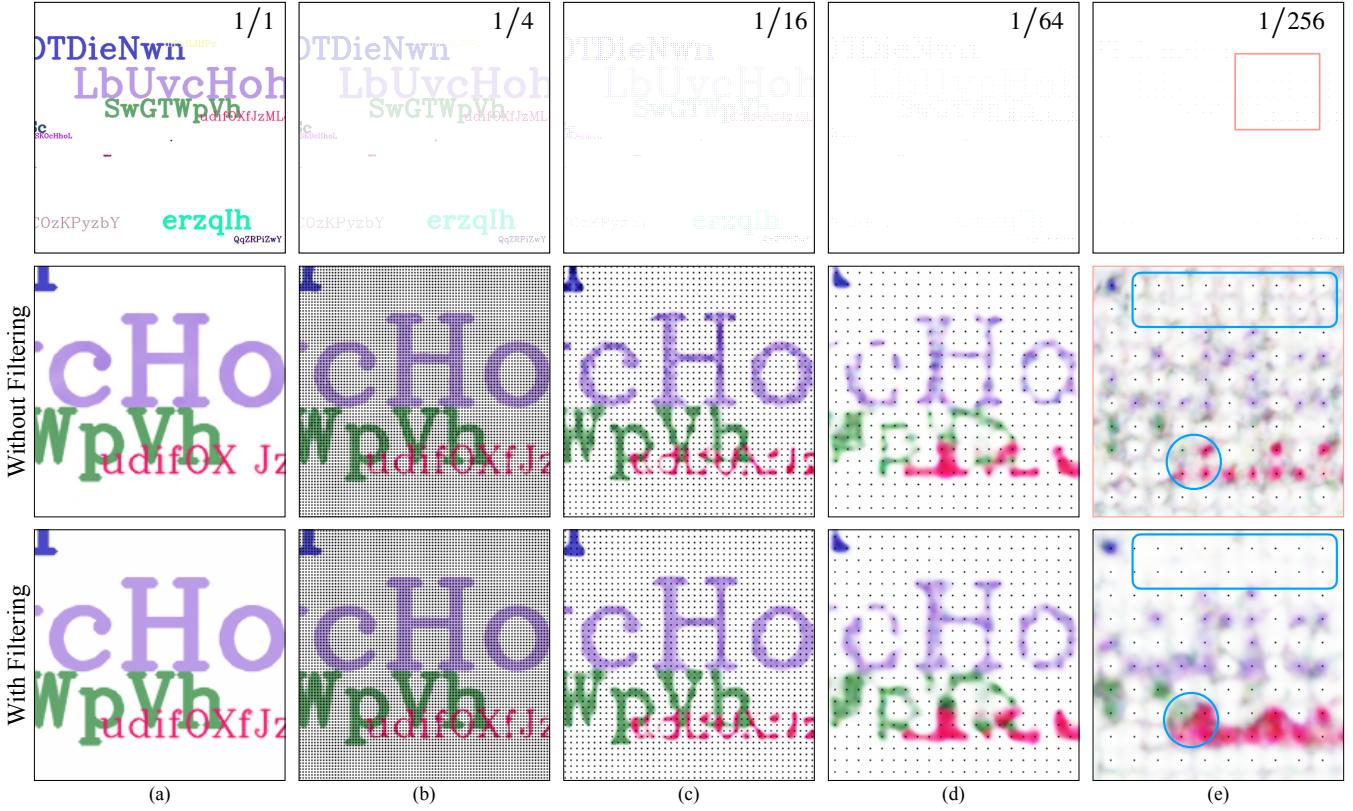
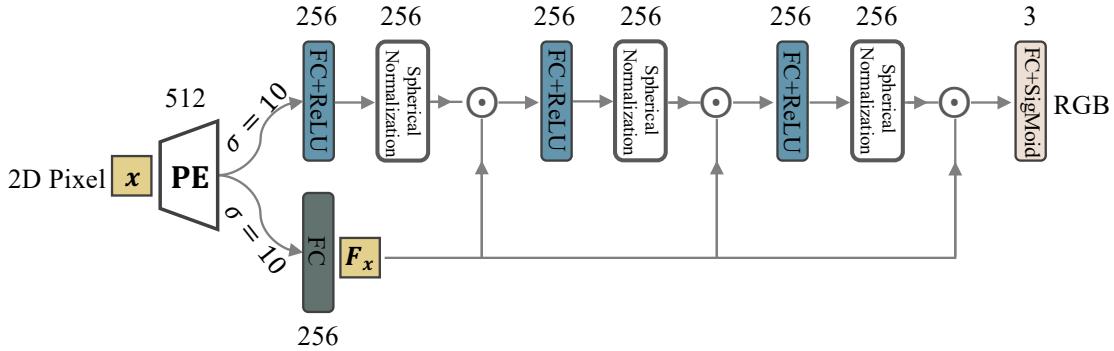
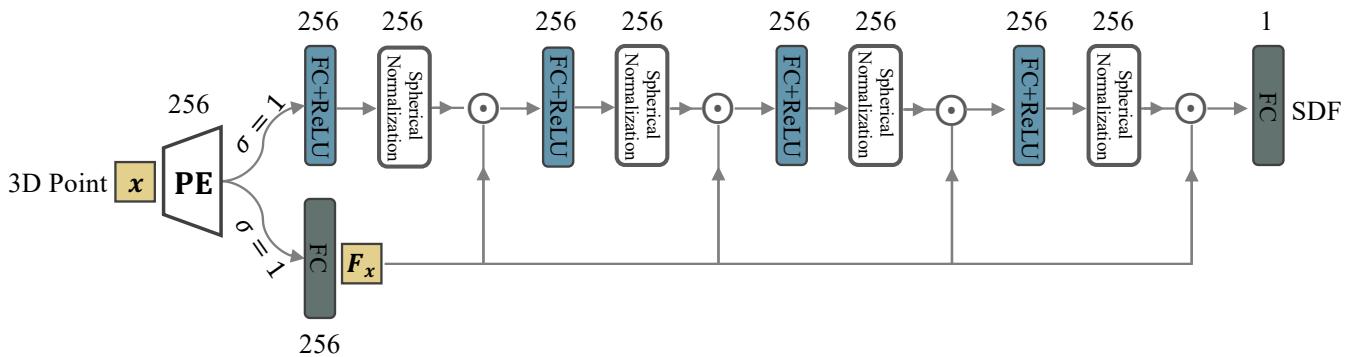


Figure 13: Impact of the sampling rate. We keep downsampling an image from 512×512 resolution to 32×32 , with sampling rates of $1/4, 1/16, 1/64$, and $1/256$. Then we use acquired images to train two networks, one with filtering and the other not, and generate new images all at 512×512 resolution. At the highest resolutions (a), the image quality of the two methods is similar. (b-c) shows FINN interpolates pixels with less noisy artifacts than FFN. When using too fewer training samples, both methods create isolated segments for a single character (d-e). Even though both methods fail in (d-e), we can see that FINN tends to connect isolated training pixels with smooth interpolation, while FFN introduces more randomness between training pixels. Training pixels in (b-e) are painted in black.

(A) Image Regression



(B) Implicit Surface Reconstruction



(C) Simplified Neural Radiance Field

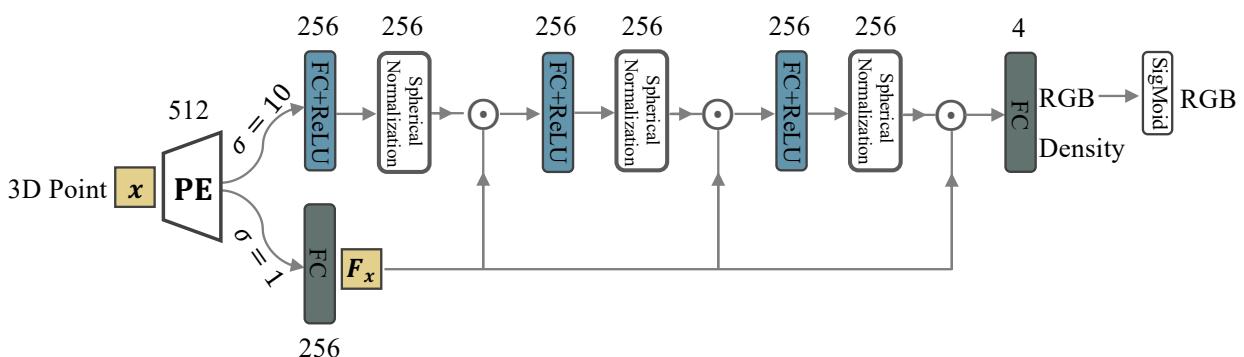


Figure 14: Our networks for three tasks.