Residual Multiplicative Filter Networks for Multiscale Reconstruction

Shayan Shekarforoush[1, 2]
shayan@cs.toronto.edu

David B. Lindell[1, 2]
lindell@cs.toronto.edu

David J. Fleet[1, 2, 3]
fleet@cs.toronto.edu

Marcus A. Brubaker[1, 2, 4]
mab@eecs.yorku.ca

[1]University of Toronto [2]Vector Institute [3]Google Research [4]York University

Abstract

Coordinate networks like Multiplicative Filter Networks (MFNs) and BACON offer some control over the frequency spectrum used to represent continuous signals such as images or 3D volumes. Yet, they are not readily applicable to problems for which coarse-to-fine estimation is required, including various inverse problems in which coarse-to-fine optimization plays a key role in avoiding poor local minima. We introduce a new coordinate network architecture and training scheme that enables coarse-to-fine optimization with fine-grained control over the frequency support of learned reconstructions. This is achieved with two key innovations. First, we incorporate skip connections so that structure at one scale is preserved when fitting finer-scale structure. Second, we propose a novel initialization scheme to provide control over the model frequency spectrum at each stage of optimization. We demonstrate how these modifications enable multiscale optimization for coarse-to-fine fitting to natural images. We then evaluate our model on synthetically generated datasets for the the problem of single-particle cryo-EM reconstruction. We learn high resolution multiscale structures, on par with the state-of-the art. Project webpage: https://shekshaa.github.io/ResidualMFN/.

1 Introduction

Coordinate networks have emerged as a powerful way to represent and reconstruct images, videos, and 3D scenes [1, 2, 3], and for solving challenging inverse problems such as 3D molecular reconstruction for cryo-electron microscopy (cryo-EM) [4, 5, 6]. They typically take as input a point defined on a continuous, low-dimensional domain (e.g., a 2D position for images), and they output the signal value at that point (e.g., the color). In recently proposed architectures [7, 8], the frequency content of the represented signal can also be explicitly controlled. Such networks are motivated in part by the effectiveness of multiscale methods in image processing and 3D reconstruction.

Nevertheless, while current scale-aware coordinate network architectures offer some control over scale [8, 9, 10], they are not readily compatible with classical multiscale methods used for coarse-to-fine optimization, like pyramids [11, 12] or multigrid solvers [13]. State-of-the-art cryo-EM models, for instance, use *frequency-marching* to progressively estimate 3D density, beginning with a coarse structure, then adding finer structure at each iteration [14, 15]. Some networks use heuristics to empirically constrain the scale of network output for coarse-to-fine optimization, but they have no explicit constraints on the represented frequencies [9, 16, 17]. Others represent signals at multiple scales at inference time, but do not enable control of the frequency spectrum during training [8, 10].

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

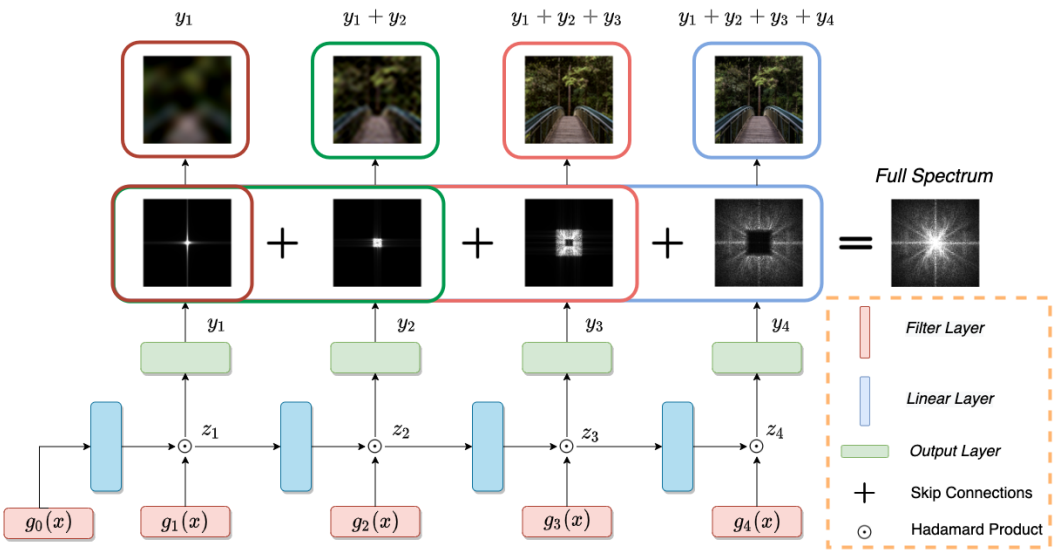


Figure 1: Overview of residual multiplicative filter networks. Skip connections efficiently combine low-frequency information from earlier layers with high-frequency information in later layers. With the proposed initialization scheme, each layer learns separate, increasing frequency bands.

Here, we introduce a new coordinate network architecture and training scheme that enables one to effectively control the frequency support of the learned representations during multiscale optimization, building on recent neural networks with an analytical and controllable Fourier spectra [7, 8]. For coarse-to-fine reconstruction, one can divide the training process into stages, each of which learns a single scale. Training starts at the coarsest resolution and then progressively adds finer scales. Naively employing this procedure with existing architectures will damage the signal reconstruction at coarser scales when fitting finer ones (e.g., see Fig. 3). To mitigate this, we modify MFNs by adding skip connections [18], so reconstruction at one scale naturally incorporates the learned structure at coarser scales. This alleviates the need to adapt previous layers in favor of fitting the finer grained reconstructions, effectively obviating the need to re-learn signal structure captured by coarser scales. Second, we derive an initialization scheme which explicitly introduces hierarchical gaps in the frequency spectrum of a new layer so that, by construction, its frequency support has a controllable degree of overlap. This requires the network to use the skip connections in order to fill the holes in the spectrum and ensures lower levels remain faithful to coarse scale representations of the signal.

We apply our technique to the inverse problem of single-particle cryo-EM reconstruction [6, 4, 14, 19] to determine the 3D structure of macromolecular complexes, like proteins, from collections of noisy 2D tomographic projections. This is a challenging non-linear inverse problem, closely related to multi-view reconstruction, requiring estimation of the pose and 3D structure of bio-molecules. Successful 3D reconstruction relies on coarse-to-fine methods, sometimes called frequency marching [20], to reduce the risk of becoming trapped in poor local minima. Our results demonstrate effective and efficient *ab initio* reconstruction of 3D structures with cryo-EM to high resolution.

This paper proposes a new architecture and training approach to enable powerful multiscale estimation techniques with coordinate-based networks. Specifically, we make the following contributions:

- We develop residual multiplicative filter networks, a new coordinate network architecture with a tailored initialization scheme for mulitscale signal reconstruction and representation.
- We design fast and efficient coarse-to-fine training strategies for residual multiplicative filter networks that leverage the multiscale representation.
- We apply the proposed architecture and training strategy to cryo-EM reconstruction on two synthetic datasets and achieve final reconstructions competitive with the cryoSPARC [14], a state-of-the-art method.

2 Related Work

Coordinate-based Networks. Coordinate-based networks (e.g., [2, 3]) offer a memory-efficient, continuous function parameterization that can be flexibly trained to reconstruct 3D appearance [1, 21, 22, 23, 24, 25] and structure [26, 27, 28, 29, 30, 31, 32], including applications in biomedical imaging [33, 34] and cryo-EM [5, 4, 6]. Although originally formulated as fully-connected MLP architectures, new models have been proposed to improve performance and interpretability. For instance, one can use multiple small fully-connected networks to improve representational capacity [35, 36, 37], or combine coordinate-based networks with explicit feature grids to improve efficiency [38, 39], or facilitate generalization across shapes or scenes [40, 41, 42, 43, 29, 44, 45]. Multiplicative Filter Networks (MFNs) replace the conventional MLP architecture with successive layers of Hadamard products and sine non-linearities [7]. Closest to our own work, MFNs have an analytical Fourier spectrum whose bandwidth can be explicitly constrained [8], improving the controllability and interpretability of the representation. Inspired by this work, we establish *Residual MFNs* (rMFNs) with specialized control over the Fourier spectrum to enable coarse-to-fine optimization.

Multiscale Reconstruction. Multiscale representations and reconstruction methods are fundamental concepts in signal and image processing. For example, wavelets are a fast and efficient multiscale signal representation [46, 47] useful for denoising, compression, communications, and optical flow estimation [48]. Multigrid frameworks are used for many differential equation solvers for physics simulations [13]. Gaussian and Laplacian pyramids [49, 11] have broad application to image processing and have, in part, inspired modern deep learning architectures [50].

Recent coordinate-based network architectures build on these fundamental concepts of multiscale signal representation and reconstruction. Some architectures use explicit feature grids at multiple scales to enable efficient training and inference for representing shape [39, 51] or rendering scenes [52, 38]. Multiscale fully-connected coordinate network architectures have been realized by progressively increasing the frequencies of positional encodings during optimization, for example, for bundle adjustment and neural scene representation [9, 17]. It is also possible to optimize separate networks with inductive biases towards low or high frequencies to improve shape representation [16]. Still, while these methods introduce multiscale training techniques, the architectures rely on inductive biases rather than explicit constraints on the scale or Fourier spectrum of the representation.

Other methods use coordinate networks for multiscale representation rather than efficient optimization. For instance, scale-aware positional encodings are trained at all scales simultaneously so that different output scales can be queried at inference time [10]. Band-limited coordinate networks (BACON) extend MFNs to control the network bandwidth, but require training on multiple output scales simultaneously for multiscale representation at inference time [8]. Our work is inspired by BACON, but significantly modifies the architecture, initialization scheme, and training techniques to make efficient multiscale optimization techniques compatible with coarse-to-fine reconstruction.

Cryo-EM Reconstruction. While coordinate networks have previously been used for cryo-EM reconstruction [5, 4, 6], they differ from our method in that they explicitly represent signals in the Fourier domain. Our goal is to demonstrate coarse-to-fine reconstruction methods with coordinate networks in the real domain, and we demonstrate this for cryo-EM reconstruction. The proposed method may also prove promising for advanced cryo-EM methods that resolve intrinsic motion of structures since spatial motion can be directly modeled in the primal domain [53].

3 Background

Coordinate neural networks have emerged as effective tools for approximating complex spatial data (e.g., see [1, 2, 7, 8, 10]). In the case of images, for example, they provide a mapping from continuous positions on the image plane to RGB values. Mapping 3D coordinates to volumetric density allows modeling 3D geometry. In particular, these networks have been shown to be readily trained to accurately approximate low-dimensional, complex signals in a memory-efficient way. Among this broad family of network architectures, in this paper we focus on Multiplicative Filter Networks (MFNs) [7] such as BACON [8], as they provide explicit control over the Fourier spectra of the function approximations.

Multiplicative Filter Networks. Most coordinate-based networks, like SIREN [2] and Random Fourier Features [3], use an MLP architecture consisting of the successive composition of linear

transformations and element-wise non-linearities. In contrast, Multiplicative Filter Networks (MFNs) use a Hadamard product (i.e., element-wise multiplication) instead of composition [7]. Concretely, in an L-layer MFN, the d_{in} dimensional input coordinates, $\mathbf{x} \in \mathbb{R}^{d_{in}}$, are transformed using $L+1$ non-linear filter modules, denoted by $g^{(i)}(\cdot) : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_h}$, $i = 0, 1, \ldots, L$, and d_h is the hidden layer dimension. In [7], either *sinusoidal* or *Gabor* filters are used for this transformation. The sinusoidal case is given by $\mathbf{g}^{(i)}(\mathbf{x}) = \sin(\boldsymbol{\omega}^{(i)}\mathbf{x} + \boldsymbol{\phi}^{(i)})$, where $\boldsymbol{\omega}^{(i)} \in \mathbb{R}^{d_h \times 3}$ and $\boldsymbol{\phi}^{(i)} \in \mathbb{R}^{d_h}$ are referred to as frequencies and phases. At layer i, the intermediate representation of previous layer, $\mathbf{z}^{(i-1)}$, after being linearly transformed, is multiplied by the non-linear filter of input $\mathbf{g}^{(i)}(\mathbf{x})$. Formally,

$$\begin{aligned} \mathbf{z}^{(0)} &= \mathbf{g}^{(0)}(\mathbf{x}), \\ \mathbf{z}^{(i)} &= \mathbf{g}^{(i)}(\mathbf{x}) \odot \left(\mathbf{W}^{(i)}\mathbf{z}^{(i-1)} + \mathbf{b}^{(i)}\right), \quad i = 1, \ldots, L\,, \end{aligned} \tag{1}$$

The behaviour of MFNs can be understood by analyzing the frequency of intermediate layers. Using the trigonometric identity

$$\sin(a)\sin(b) = \frac{1}{2}\sin(a + b - \pi/2) + \frac{1}{2}\sin(a - b + \pi/2)\,, \tag{2}$$

closed-form expressions for intermediate representations can be derived [7, 8]. Dropping the bias terms $\mathbf{b}^{(i)}$ for simplicity, the individual components of the intermediate representation $\mathbf{z}^{(i)}$ is equal to a weighted sum of an exponential number of sine terms [7]:

$$\mathbf{z}^{(i)}_{n_i} = \sum_{\substack{\mathbf{n}=(n_0,\cdots,n_{i-1},n_i) \\ \mathbf{s}=(s_1,\cdots,s_i)}} \overline{\alpha}^{(i)}(n) \sin\left(\overline{\boldsymbol{\omega}}^{(i)}(\mathbf{n},\mathbf{s})\,\mathbf{x} + \overline{\boldsymbol{\phi}}^{(i)}(\mathbf{n},\mathbf{s})\right) \tag{3}$$

where $\mathbf{n} = (n_0, \cdots, n_{i-1}, n_i)$ is a tuple of indices with $n_j \in \{1, \ldots, d_h\}$ and $s_j \in \{-1, 1\}$. Each term in the summation comprises an amplitude, frequency and phase shift, given by

$$\overline{\alpha}^{(i)}(\mathbf{n}) = \frac{1}{2^i}\prod_{l=1}^{i} \mathbf{W}^{(l)}_{n_l,n_{l-1}}, \qquad \begin{cases} \overline{\boldsymbol{\omega}}^{(i)}(\mathbf{n},\mathbf{s}) &= \boldsymbol{\omega}^{(0)}_{n_0} + \sum_{l=1}^{i} s_l \boldsymbol{\omega}^{(l)}_{n_l} \\ \overline{\boldsymbol{\phi}}^{(i)}(\mathbf{n},\mathbf{s}) &= \boldsymbol{\phi}^{(0)}_{n_0} + \sum_{l=1}^{i} s_l(\boldsymbol{\phi}^{(l)}_{n_l} - \frac{\pi}{2})\,. \end{cases} \tag{4}$$

BACON. This analysis of the frequency content of an MFN shows that the bandwidth of $\mathbf{z}^{(i)}(\mathbf{x})$ is the sum of the input bandwidths. This was leveraged in BACON [8] to band-limit the representation of each individual layer. Additionally, an output was created for each layer to produce intermediate reconstructions; i.e.,

$$\mathbf{y}^{(i)} = \mathbf{W}^{(i)}_{\text{out}}\mathbf{z}^{(i)} + \mathbf{b}^{(i)}_{\text{out}}\,. \tag{5}$$

where $\mathbf{W}^{(i)}_{\text{out}} \in \mathbb{R}^{d_{\text{out}} \times d_h}$ and $\mathbf{b}^{(i)}_{\text{out}} \in \mathbb{R}^{d_{\text{out}}}$. However, to encourage $\mathbf{y}^{(i)}$ to preserve its signal representation at scale i, one requires extra scale-specific losses during training, without which, subsequent training will corrupt the lower resolution representation.

Motivating Example. Although BACON does provide control over spectral bandwidth, when adopted in the context of a coarse-to-fine optimization procedure, it fails to preserve and carry the learned coarse-scale representations when one moves to the next finer-scale stage of the optimization. Consequently, at each round, one must re-optimize the entire representations at all previous scales, hindering efficient coarse-to-fine optimization. As an example, in Fig. 3 we examine the representation learned by BACON within a coarse-to-fine training strategy, where successive output layers are trained to fit an image at finer scales. We apply a staged training procedure such that a loss is applied to outputs at each scale in separate rounds of optimization. Unfortunately, when fitting finer scales BACON completely forgets the learned representations of coarse scale outputs from previous optimization rounds (see supplement). For a fairer comparison, given any round, we keep all scale-specific losses, but let only the linear output layers for other scales to be updated. Still, the results of BACON (Fig. 3, top row) show that the representations at coarser scales yet become corrupted while training at finer scales. In fact, reconstructions at different scales are highly coupled in BACON, since new layers can result in new frequencies anywhere within the band limit. In the following, we modify the architecture and its initialization and training scheme (Fig. 1) to make coordinate networks applicable to coarse-to-fine multiscale reconstruction.

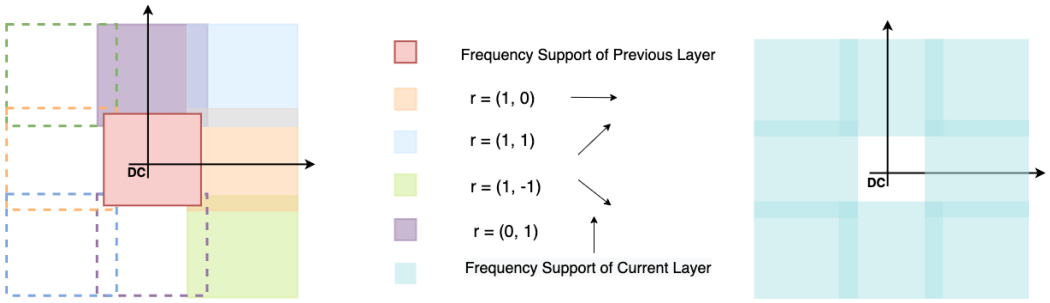


Figure 2: Illustration of the initialization scheme. (Left) The central red square depicts the base spectrum comprising frequencies of the previous layer $\overline{\boldsymbol{\omega}}^{(i-1)}(\mathbf{n}, \mathbf{s})$. Other coloured regions depict copies of the base spectrum, shifted versions by $\lambda_2 B_{i-1}$ in four directions with $\lambda_2 \approx 2$. The copies are slightly larger due to the perturbation by the random $\mathbf{v}_j^{(i)}$. Dashed regions are naturally introduced because of the sign factors $s_i = \pm 1$. (Right) The new frequency spectrum of $\overline{\boldsymbol{\omega}}^{(i)}(\mathbf{n}, \mathbf{s})$ is the union of the shaded regions. The initialization controls the overlap with the spectrum of the previous layer.

4 Residual Multiplicative Filter Networks

From Eq. 4, we can express frequencies in layer i in terms of frequencies in previous layers with the recursion

$$\overline{\boldsymbol{\omega}}^{(i)}(\mathbf{n}, \mathbf{s}) = \overline{\boldsymbol{\omega}}^{(i-1)}(\tilde{\mathbf{n}}, \tilde{\mathbf{s}}) + s_i \boldsymbol{\omega}_{n_i}^{(i)}, \tag{6}$$

where $\tilde{\mathbf{n}}$ and $\tilde{\mathbf{s}}$ are formed from $\mathbf{n}$ and $\mathbf{s}$ by removing last entry, e.g., $\tilde{\mathbf{n}} = (n_0, \cdots, n_{i-1}) = \mathbf{n}_{<i}$. Thus $\mathbf{z}_{n_i}^{(i)}$ contains frequencies from the previous layer, shifted by $\boldsymbol{\omega}_{n_i}^{(i)}$ and $-\boldsymbol{\omega}_{n_i}^{(i)}$, where $\boldsymbol{\omega}_{n_i}^{(i)}$ is the frequency of hidden unit n_i in layer i. More specifically, assuming that frequencies in the previous layer, i.e., $\overline{\boldsymbol{\omega}}^{(i-1)}(\tilde{\mathbf{n}}, \tilde{\mathbf{s}})$, are band-limited to $[-B_{i-1}, +B_{i-1}]$, those in $\mathbf{z}^{(i)}$ lie in the union of several of its shifted clones. From this perspective, the frequencies at layer i, $\boldsymbol{\omega}^{(i)}$, characterize two important concepts: 1) the growth in bandwidth from the previous layer, and 2) the extent of overlap between the previous layer frequency spectrum and each new individual clone. For instance, given entries of $\boldsymbol{\omega}^{(i)}$ in $[-\delta, +\delta]$, the range of frequencies represented in $\mathbf{z}^{(i)}$ expands to $[-B_{i-1} - \delta, B_{i-1} + \delta]$. Also, if a specific $\boldsymbol{\omega}_{n_i}^{(i)}$ is set to a relatively large value of $2B_{i-1}$ this results in a clone with frequency support in the range $[B_{i-1}, 3B_{i-1}]$, and it avoids overlap with the original bandwidth $[-B_{i-1}, B_{i-1}]$.

4.1 Frequency Initialization

Motivated by this interpretation of $\boldsymbol{\omega}^{(i)}$, we propose a new initialization scheme which explicitly controls the overlap and expansion of frequencies in successive layers. For simplicity, we describe this initialization for the 2D case, though it can be extended to higher dimensions straightforwardly. First, we follow a procedure similar to BACON [8] to initialize and fix frequencies of filters which allows limiting outputs to specified bandwidths. Then, given the band-limited support $[-B_{i-1}, +B_{i-1}]$ of frequencies in layer $i-1$, we initialize the 2D vectors $\boldsymbol{\omega}_j^{(i)}$ as follows:

$$\begin{aligned} \boldsymbol{\omega}_j^{(i)} &= \lambda_2 B_{i-1} \mathbf{r}_j^{(i)} + \mathbf{v}_j^{(i)}, \\ \mathbf{r}_j^{(i)} \sim U(\mathcal{R}), \qquad \mathbf{v}_j^{(i)} &\sim U(-\lambda_1 B_{i-1}, \lambda_1 B_{i-1})^2 \end{aligned} \tag{7}$$

where $\mathcal{R} = \{(0,1), (1,0), (1,1), (1,-1)\}$, and U denotes either continuous or discrete uniform distributions. Here, $\boldsymbol{\omega}_j^{(i)}$ is composed of two terms. The first term mainly specifies the direction ($r_j^{(i)}$) and the proportional amount ($\lambda_2 B_{i-1}$) by which frequencies of layer $i-1$ will be shifted. The second term is a relatively small random perturbation to break the symmetry. An intuitive illustration of these operations is provided in Fig. 2. In 2D, there are 8 different directions to expand, half of which are redundant since $s_i = \pm 1$ entails two shifts in opposite directions $\pm(r_{j,1}, r_{j,2})$. Thus we only consider 4 possible values R. Using this initialization $\mathbf{z}^{(i)}$ is thus band-limited to $\overline{\boldsymbol{\omega}}^{(i)} \in [-(1 + \lambda_1 + \lambda_2)B_{i-1}, (1 + \lambda_1 + \lambda_2)B_{i-1}]$.

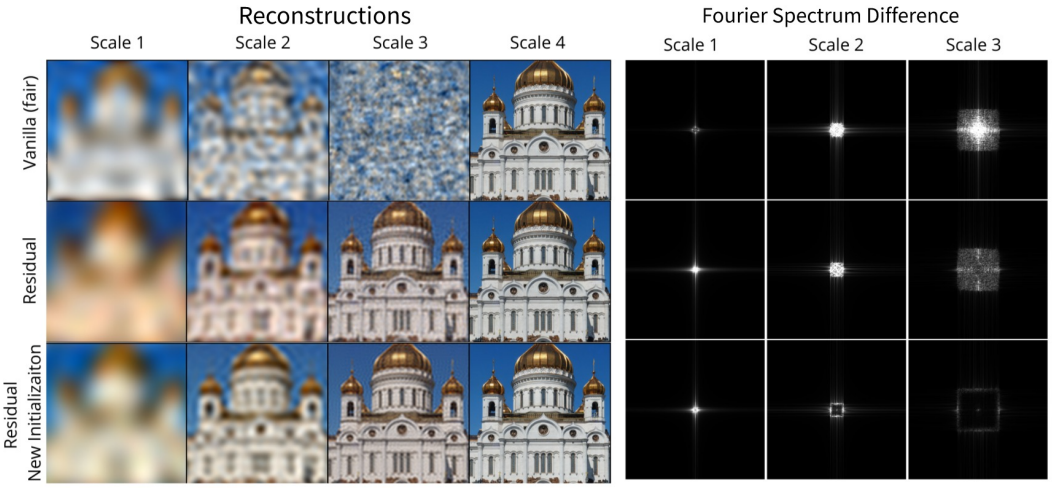


Figure 3: Coarse-to-fine image fitting experiments. We compare fair BACON (Vanilla), a residual MFN with the BACON initialization scheme, and a residual MFN with the proposed initialization. The proposed method performs best at retaining image quality at each scale during the staged training regiment. We also compare the difference between the Fourier spectra at each output scale immediately after the training stage and at the end of training (Right). In the proposed method, residual connections allow information at each scale to be re-used while other methods overwrite lower frequency information during coarse-to-fine training.

4.2 Residual Connections

Consider setting $\lambda_2 = 2 + \lambda_1$ during initialization. In this case, by design, the frequency spectrum for sine terms in layer i would thus not overlap at all with that of layer $i-1$; in at least one direction, the frequencies would either end up to left or right of interval $[-B_{i-1}, B_{i-1}]$. Thus the output layer would include no sine terms with frequencies in the central interval $[-B_{i-1}, B_{i-1}]$. To obtain the output of layer i, we sum it with the output of layer $i-1$. That is,

$$\mathbf{y}^{(i)} = \mathbf{y}^{(i-1)} + (\mathbf{W}_{\text{out}}^{(i)}\mathbf{z}^{(i)} + \mathbf{b}_{\text{out}}^{(i)}) \,. \tag{8}$$

This can be interpreted as a residual connection that ensures we keep an identical copy of frequencies from the previous layer; i.e., $\mathbf{y}^{(i-1)}$ will fill the gap in the frequency space. In practice, rather than the setting $\lambda_2 = 2 + \lambda_1$, one can use other less strict combinations which allow for different amounts of overlap and expansion. We found empirically setting $\lambda_2 = 2, \lambda_1 = 0.3$ to be effective.

5 2D Image Fitting

We first evaluate each of the proposed modifications to band-limited networks in the context of 2D image fitting. This task is useful for didactic purposes, allowing one to clearly see the benefits of skip connections and our new initialization scheme in the context of coarse-to-fine image reconstruction. We adopt a staged training procedure and fit a series of band-limited coordinate networks to natural images provided by [1]. The dataset consists of 32 RGB images, each of size 512×512. We first downsample the images by a factor of 2 to obtain images of size 256×256. These images are fit during training using a Mean Squared Error (MSE) training objective, applied to the output at the scale corresponding to the stage of training. As previously described, for the fairest comparison, the "Vanilla" BACON baseline is optimized with losses at all scales, though, at the stage of a given scale, only the linear output layers of other scales can be updated. Still, this keeps the optimization of the internal representations staged. All networks use a four-layer architecture and are trained for $10,000$ iterations, divided into chunks of $500, 500, 1000, 8000$ for optimizing output scales $\{1, 2, 3, 4\}$, respectively. The frequencies for all methods are initialized to have the same band-limit for consistency. All methods were implemented in PyTorch [54] and trained on a single NVIDIA GeForce RTX 2080.

Table 1: Coarse-to-fine image fitting results. We show the mean absolute difference of Fourier spectrum of outputs at each scale immediately after optimization and after the entire training regiment and peak signal-to-noise ratio for the image generalization task. Residual connections and the new initialization help to preserve the spectrum during training, especially for the second and third scales. The final reconstruction achieves competitive generalization results in terms of PSNR metrics.

Method	Mean Abs. Difference			PSNR (dB)
	Scale 1	Scale 2	Scale 3	Scale 4
SAPE [9]	-	-	-	27.02 ± 3.29
BACON [8]	-	-	-	28.39 ± 3.66
(Staged) BACON (fair)	$\mathbf{1.61 \pm 0.45}$	7.36 ± 1.84	17.97 ± 4.27	$\mathbf{28.46 \pm 3.61}$
(Staged + Res) BACON	4.33 ± 0.97	5.16 ± 1.22	9.23 ± 2.73	27.90 ± 3.27
(Staged + Res + Init) BACON	3.60 ± 0.95	$\mathbf{3.42 \pm 0.85}$	$\mathbf{4.32 \pm 1.35}$	28.33 ± 4.42

Figure 3 visualizes the outputs at each scale for BACON, the residual MFN with the band-limited BACON initialization, and the residual MFN with our proposed initialization. Since the coarse-to-fine training scheme can potentially disturb previously trained outputs when applying the loss at the finer scales, we visualize the differences between the Fourier spectra at each output scale after the corresponding training iterations and after the full training regiment (Fig. 3, right). The proposed method shows the least disturbance after training. For the entire image dataset, we report the Fourier spectrum mean absolute difference and standard deviation in Table 1.

Finally, we examine all networks on the image generalization task by querying them on a finer grid of 512×512 and computing the Peak Signal-to-Noise Ratio (PSNR) between reconstructed signal and the original ground truth. The average PSNR values over all images is summarized in the last column of Table 1. We also benchmark SAPE [9] as a baseline for coarse-to-fine reconstruction, on the image generalization task. SAPE gradually exposes higher frequencies as a function of time and space, while providing only a single reconstruction at a time, so only their full-scale reconstruction can be evaluated for comparison. While our approach improves interpretability of the reconstructions, by enabling reuse of previous learned representations, it achieves empirical performance competitive with baselines. Additional experiments on a textual dataset are provided in Appendix A.

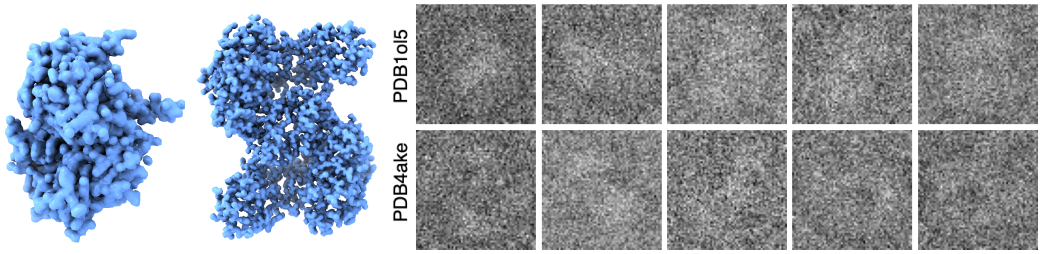
6 Cryo-EM 3D Reconstruction

Single particle cryo-EM is an increasingly popular experimental technique to recover the 3D structure of macromolecular complexes, such as proteins and viruses. Cryo-EM has gained widespread attention in the last decade with pioneering advances in hardware and data processing techniques. This has enabled atomic or near-atomic resolution reconstruction of challenging structures [55]. Multiscale, coarse-to-fine reconstructions has long played a critical role in cryo-EM estimation, e.g., [20], making it an ideal testbed for our proposed methods.

Image Formation Model. Based on a simplified model [56], the protein structure is represented as a function $V : \mathbb{R}^3 \to \mathbb{R}_{\geq 0}$, which maps points in 3D space to non-negative real values, expressing the Coulomb potential induced by the atoms. Images $\{I_i\}_{i=1}^n$ are captured with a transmission electron microscope which are 2D orthogonal projections of randomly oriented potential maps. Each image I_i has a corresponding latent pose (orientation) $R_i \in SO(3)$ representing the 3D rotation of the molecule in the image. In practice, to ensure sufficient contrast, the microscope is defocused when images are captured, resulting in a point-spread function (PSF) g_i. To minimize radiation damage to delicate biological molecules, samples are exposed to low dosages of electrons resulting notoriously noisy images. Taken together, the image formation model can be expressed as

$$I_i^*(x, y) = g_i \star \int_{\mathbb{R}} V(R_i^T \mathbf{x})\, dz + \epsilon, \quad \mathbf{x} = (x, y, z)^T \tag{9}$$

where the projection is, by convention, assumed to be along the z-direction after rotation, $\star$ corresponds to convolution and ϵ is Gaussian noise [14, 4]. The PSF can be estimated in a preprocessing step and is assumed here to be known; for more details see [57]. (A more detailed formulation for cryoEM reconstruction is given in Appendix B.)



(a) PDB1ol5. (b) PDB4ake. (c) Sample images.

Figure 4: Ground truth potential maps and cryo-EM measurements. (Left) Potential maps generated using the corresponding atomic models. (Right) Noisy example images from a synthetically generated particle stack after applying the PSF and adding white Gaussian noise.

3D Reconstruction Problem. The goal of a cryo-EM experiment is to estimate the unknown 3D potential map of the biomolecule present in the observed images $\{I_i\}_{i=1}^n$. This task has some intrinsic challenges. First, the images are extremely noisy (see Fig. 4c). Second, the ground-truth poses are typically unknown, or only a rough estimate is available, so it is not possible to perform 3D reconstruction using the forward model of Eq. 9 directly. Previous work [6, 4, 19] estimate pose parameters during reconstruction with various techniques including amortized inference, variational bounds or direct marginalization. Here we use a stochastic gradient-based method, which optimizes the representation of the 3D structure V, parameterized by our proposed residual MFN, together with the unknown pose variables $\{R_i\}_{i=1}^n$. The loss function is the squared error between the observed image I_i and that generated by the model, i.e.,

$$\mathcal{L} = \sum_{x,y} \left(I_i(x,y) - I_i^*(x,y)\right)^2 \tag{10}$$

where $I_i^*(x,y)$ is given above in Eq. 9. See Appendix D for more detail on the pose parameterization and optimization.

In practice, naively optimizing for both the poses and potential map suffers from poor local minima. To avoid this, coarse-to-fine reconstruction approaches are commonly employed. At early stages, low frequency content of the map is estimated as the SNR is higher so estimation is easier, which then allows for more reliable subsequent pose estimation as higher frequencies are included. Previous approaches [4, 14] employ this technique, called *Frequency Marching* [20], as they model the map in the Fourier domain where individual frequencies are readily accessible. In contrast, we address the reconstruction problem in a coarse-to-fine manner while expressing the potential map in real space.

6.1 Experiments

We evaluated our approach on two synthetic datasets which allow us to directly evaluate against ground truth structures. The datasets are based on atomic models of 1OL5 (`https://www.rcsb.org/structure/1ol5`) and 4AKE (`https://www.rcsb.org/structure/4ake`) from Protein Data Bank (PDB). We first convert atomic models to $64 \times 64 \times 64$ discrete potential maps, with voxel sizes 1 Å and 1.2 Å respectively, using *pdb2mrc.py* module from EMAN2 [58]. Then we simulate the image formation model as in Equation 9, at $50,000$ rotations uniformly sampled on $SO(3)$. To generate each image, we rotate the map V, resample it on the original grid, project it along canonical z-direction and apply the PSF. Zero-mean white Gaussian noise is added to these simulated clean images produce a realistic SNR of 10%. The structures and sample images are visualized in Fig. 4.

Experimental Settings. We use a 3-layer network with 128 hidden units to represent the 3D structure. The network outputs the structure in 3 hierarchical scales and, accordingly, the training process is divided into 3 stages, one for each scale. We set the hyperparameters as $\lambda_1 = 0.3, \lambda_2 = 2.0$ and train the network for a total of 100 epochs. For PDB1ol5, we spend respectively 15, 15, and 70 epochs optimizing the first to the third scales. For PDB4ake, we find a better reconstruction if we dedicate more epochs on the first and second scales, namely 25 epochs on each of the first and second scales. When computing the loss, images are band-limited to the corresponding scale using a Gaussian filter whose scale parameter is set to match the bandwidth of the current stage

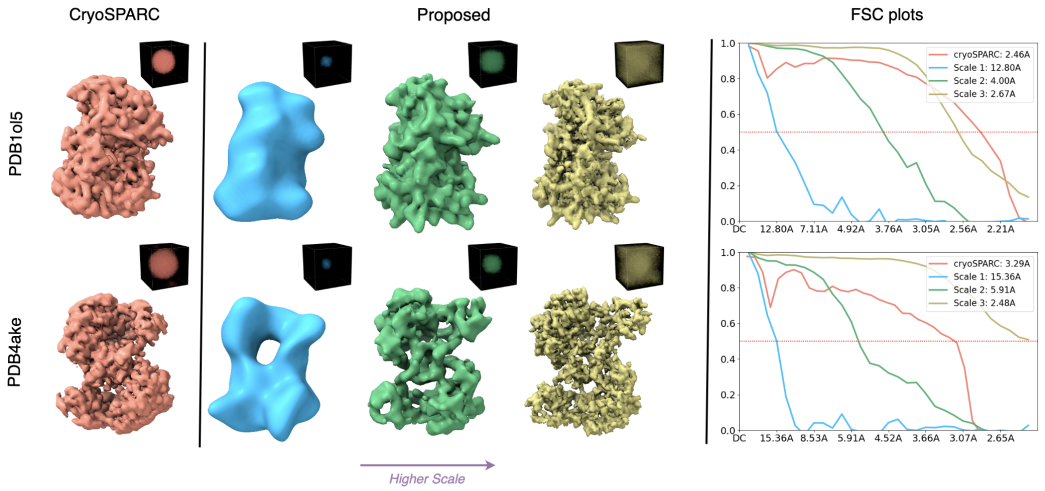


Figure 5: (Left) CryoSPARC reconstructions. (Middle) Our reconstruction at multiple scales. (Right) Fourier Shell Correlation plot measuring correlation in Fourier spectrum between reconstructions and ground-truth structure. Threshold 0.5 is the criterion to compute the resolutions in Angstroms.

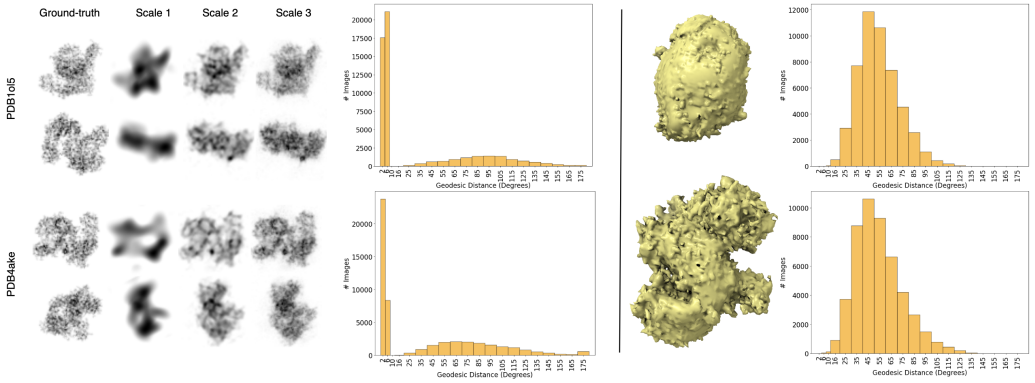


Figure 6: (Left) Histogram of geodesic distances between estimated and ground-truth poses. Some example reconstructed images are illustrated as well. The first row for each dataset contains a particle with the estimated pose within at most 4 degrees away from ground-truth while the second row corresponds to another with totally wrong pose estimate. (Right) Getting stuck at local minima when optimizing the full-scale reconstruction from scratch (not coarse-to-fine). Although, the overall shape is well captured, no significant improvement is achieved during several iterations and poses are stuck at local minima, mostly concentrated around 45 degrees away ground-truth.

scale. Mini-batches of images are used to update poses and weights of the network representing the structure. For each mini-batch, we alternate between optimizing the structure and poses for 5 and 20 iterations, respectively. We use Adam optimizer [59] with a learning rate 0.001 for the network parameters and a higher learning rate of 0.01 for pose parameters. Each pose parameter is initialized to random value uniformly distribution over a geodesic distance of 45 to 90 degrees from the ground truth. This can be considered as a very crude initial estimate for poses. In Appendix E we consider even coarser initial pose values, showing that they achieve competitive qualitative and quantitative performance to those here. Experiments are implemented in PyTorch [54], models are trained on 4 NVIDIA TITAN Xp GPUs, and ChimeraX [60] is used to visualize the 3D maps.

Results. Qualitative comparison of reconstructions with those obtained by homogeneous refinement of cryoSPARC (as the baseline) are illustrated in Fig. 5. In the middle, we observe that, from low to high scales, increasingly more structural features are resolved. The low-scale reconstructions capture only the global and low-resolution features, whereas the full-scale ones contain fine-grained details. Our approach is able to well-resolve some peripheral parts of the structure, especially for the more

complex structure of PDB4ake. Also there are some structural features learned by our reconstruction yet totally missed by cryoSPARC. All in all, our final reconstruction visually looks competitive or better when compared with the baseline.

As shown on the right side of Fig. 5, to quantitatively evaluate reconstructions, we compute the Fourier Shell Correlations (FSCs) [61] between the reconstructed map and the ground-truth. FSC is the gold-standard metric in cryo-EM area measuring the normalised cross-correlation coefficient between two 3D volumes in Fourier domain along shells of increasing radius (see Appendix C for more background on FSC). We measure FSC between ground-truth and reconstructions at all scales, following the convention of the threshold 0.5 to compute final resolution, as ground-truth is known *a priori* in our experiments. We also evaluate how well the poses are optimized and converge to the ground-truth. To this end, the histogram showing number of images with various geodesic distances between estimated and ground-truth poses are visualized as in Fig. 6. For both datasets, there is a concentration of images on left hand side corresponding to slight error in estimation of pose parameters. Thus, for most of the images, predicted pose parameters are within few degrees of the corresponding ground truth, verifying that poses are optimized well along with the reconstruction. Furthermore, per dataset, we select two sample images, one with almost the correct pose estimate, and the other with one that is far from ground truth. Then we demonstrate their reconstruction at all scales as in Fig. 6. Finally, we conduct an experiment in which we supervise the network only at the full-scale reconstruction throughout the entire training. The output structures are depicted in the right panel of Fig. 6, realizing a situation where poses are trapped in local minima, and we thus observe no substantial improvement in the structure for many iterations.

7 Conclusions, Limitations and Future Work

In this paper, we build on Band-limited Coordinate Network (BACON) and propose a new architecture and training strategy providing fine-grained control over spectral characteristics of network outputs, yielding a multiscale representation suitable for coarse-to-fine reconstruction. The combination of skip connections and a consistent initialization scheme allows the proposed network to ensure lower levels of the network represent the coarser scales of the signal which are then reused at finer scales. We empirically show that these modifications enable coordinate networks to fit natural images in coarse-to-fine fashion. They are also shown to be effective for cryo-EM reconstruction.

Limitations. We find that the proposed multi-scale optimization procedure is effective in avoiding local minima in the Cryo-EM reconstruction problem. We have yet to explore whether residual MFNs are also effective on other problems with local minima due to the existence of nuisance variables. Although beyond the scope of this paper, it is also possible that minor modifications of the architectural and hyper-parameter settings would improve the robustness of the optimization; e.g., increasing the number of layers for more fine-grained scales, or a more diverse inspection of λ hyperparameters. Moreover, the cryo-EM experiments in this paper used synthetic data, to enable more thorough the evaluation of the method. Real datasets are likely to require the use of other bases as well, e.g., *Gabor*, which are better able to represent the spatial variability of resolution in cryo-EM reconstructions. We have also focused on homogeneous and rigid structures. An advantages of our approach is that the efficient multiscale representation of the signal in the real domain opens promising doors for modeling flexible structures, a major open problem in cryo-EM [4, 53].

Societal Impact. Cryo-EM for macromolecular structure determination is one of the foremost exciting areas in molecular biology with enormous social impact. Cryo-EM tools were key to determining the structure of the SARS-COV2 spike protein, the discovery of its pre-fusion conformation, and the assessment of potential medical countermeasures. Advances in this field enable the discovery and exploration of large protein complexes and viruses more generally [62]. On the other hand, we strongly condemn any usage of our proposed coordinate networks for generating malicious representations, improperly modifying signals, or spreading misinformation.

Acknowledgments and Disclosure of Funding

We thank Ali Punjani for numerous valuable discussions and for use of the cryoSPARC software package. This research was supported in part by the Province of Ontario, the Government of Canada, through NSERC, CIFAR, and the Canada First Research Excellence Fund for the Vision: Science to Applications (VISTA) programme, and by companies sponsoring the Vector Institute.

References

[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020.

[2] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

[3] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proc. NeurIPS*, 2020.

[4] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. CryoDRGN2: Ab initio neural reconstruction of 3D protein structures from real cryo-EM images. In *Proc. ICCV*, 2021.

[5] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods*, 18(2):176–185, 2021.

[6] Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. CryoAI: Amortized inference of poses for ab initio reconstruction of 3D molecular volumes from real cryo-EM images. In *Proc. CVPR*, 2022.

[7] Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. Multiplicative filter networks. In *Proc. ICLR*, 2021.

[8] David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proc. CVPR*, 2022.

[9] Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. SAPE: Spatially-adaptive progressive encoding for neural optimization. In *Proc. NeurIPS*, 2021.

[10] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proc. ICCV*, 2021.

[11] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41, 1984.

[12] Eero P. Simoncelli and William T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. ICIP*, 1995.

[13] James H Bramble. *Multigrid methods*. Chapman and Hall/CRC, 2019.

[14] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, 2017.

[15] Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012.

[16] Wang Yifan, Lukas Rahmann, and Olga Sorkine-Hornung. Geometry-consistent neural shape representation with implicit displacement fields. In *Proc. ICLR*, 2022.

[17] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *Proc. ICCV*, 2021.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[19] Youssef SG Nashed, Frédéric Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. CryoPoseNet: End-to-end simultaneous learning of single-particle orientation and 3D map reconstruction from cryo-electron microscopy data. In *Proc. ICCV*, 2021.

[20] Alex Barnett, Leslie Greengard, Andras Pataki, and Marina Spivak. Rapid solution of the cryo-em reconstruction problem by frequency marching. *SIAM Journal on Imaging Sciences*, 10(3):1170–1195, 2017.

[21] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *Proc. IROS*, 2021.

[22] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *Proc. CVPR*, 2021.

[23] David B. Lindell, Julien N. P. Martel, and Gordon Wetzstein. AutoInt: Automatic integration for fast neural volume rendering. In *Proc. CVPR*, 2021.

[24] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Proc. NeurIPS*, 2021.

[25] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. ICCV*, 2021.

[26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proc. CVPR*, 2019.

[27] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local SDF priors for detailed 3D reconstruction. In *Proc. ECCV*, 2020.

[28] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *Proc. CVPR*, 2020.

[29] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proc. CVPR*, 2022.

[30] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T. Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proc. ICCV*, 2019.

[31] Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proc. CVPR*, 2021.

[32] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, 2019.

[33] Albert W Reed, Hyojin Kim, Rushil Anirudh, K Aditya Mohan, Kyle Champley, Jingu Kang, and Suren Jayasuriya. Dynamic CT reconstruction from limited views with implicit neural representations and parametric motion fields. In *Proc. ICCV*, 2021.

[34] Yu Sun, Jiaming Liu, Mingyang Xie, Brendt Wohlberg, and Ulugbek S Kamilov. Coil: Coordinate-based internal learning for imaging inverse problems. *arXiv preprint arXiv:2102.05181*, 2021.

[35] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3D shape. In *Proc. CVPR*, 2020.

[36] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3D scenes. In *Proc. CVPR*, 2020.

[37] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proc. ICCV*, 2021.

[38] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Proc. NeurIPS*, 2020.

[39] Julien N. P. Martel, David B. Lindell, Connor Z. Lin, Eric R. Chan, Marco Monteiro, and Gordon Wetzstein. ACORN: Adaptive coordinate networks for neural scene representation. *ACM Trans. Graph. (SIGGRAPH)*, 40(4), 2021.

[40] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proc. ECCV*, 2020.

[41] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, 2019.

[42] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proc. CVPR*, 2021.

[43] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *Proc. CVPR*, 2021.

[44] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3D-aware image generation. In *Proc. CVPR*, 2022.

[45] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3D-consistent image and geometry generation. In *Proc. CVPR*, 2022.

[46] Martin Vetterli and Cormac Herley. Wavelets and filter banks: Theory and design. *IEEE Trans. Signal Process.*, 40(9):2207–2232, 1992.

[47] Olivier Rioul and Martin Vetterli. Wavelets and signal processing. *IEEE Signal Process. Mag.*, 8(4):14–38, 1991.

[48] Stéphane Mallat. *A wavelet tour of signal processing*. 1999.

[49] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in Computer Vision*, pages 671–679. 1987.

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, 2015.

[51] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In *Proc. CVPR*, 2021.

[52] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *Proc. ICCV*, 2021.

[53] Ali Punjani and David J Fleet. 3D flexible refinement: structure and motion of flexible proteins from cryo-EM. *BioRxiv*, 2021.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Proc. NeurIPS*, 2019.

[55] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[56] Earl J Kirkland. Theory of calculation of images of thick specimens. In *Advanced Computing in Electron Microscopy*, pages 143–195. 2020.

[57] Joseph A Mindell and Nikolaus Grigorieff. Accurate determination of local defocus and specimen tilt in electron microscopy. *Journal of Structural Biology*, 142(3):334–347, 2003.

[58] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. EMAN2: an extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, 2007.

[59] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015.

[60] Thomas D Goddard, Conrad C Huang, Elaine C Meng, Eric F Pettersen, Gregory S Couch, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Meeting modern challenges in visualization and analysis. *Protein Science*, 27(1):14–25, 2018.

[61] Marin Van Heel and Michael Schatz. Fourier shell correlation threshold criteria. *Journal of Structural Biology*, 151(3):250–262, 2005.

[62] John H Van Drie and Liang Tong. Cryo-EM as a powerful tool for drug discovery. *Bioorganic & medicinal chemistry letters*, 30(22):127524, 2020.

[63] Pawel A Penczek. Three-dimensional spectral signal-to-noise ratio for a class of reconstruction algorithms. *Journal of Structural Biology*, 138(1-2):34–46, 2002.

[64] Ali Punjani, Haowei Zhang, and David J Fleet. Non-Uniform Refinement: Adaptive regularization improves single particle cryo-em reconstruction. *Nature Methods*, 17:1214–1221, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The arguments reflect the actual contribution and scope.
   (b) Did you describe the limitations of your work? [Yes] We have described the limitations of our work in 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed both positive and negative potentials of our work in 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have reviewed the guidelines and the paper comply with it.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code for generating synthetic data is provided along with the implementation of networks and training schemes in supplementary.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Training details are provided in experimental settings sections.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] When relevant, error bars are provided as in Table 1
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Number of GPUs is stated individually for each experiment.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] All used softwares and codes cited.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code for data generation and visualizations are provided in supplementary.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Residual Multiplicative Filter Networks for Multiscale Reconstruction

A 2D Image Fitting.

We provide additional natural image fitting results in Figs. 8 and 9. We also include the case of training "unfair" vanilla BACON in a staged fashion which, unlike the proposed fair variant, does not continue to update the linear output layer during the subsequent optimization stages of each scale. For the generalization task, we visualize the final reconstructions of SAPE [9], BACON [8], and our proposed method and compare it with ground truth as in Fig. 10

We further investigate the proposed modifications on a new dataset of 32 RGB text images, each of size 512×512, provided by [1]. Similar to natural images, we first obtain downsampled 256×256 images. For all networks, we use the same architecture as above, in the main body of the paper. The models are fit to images using Mean Squared Error (MSE) as the objective loss. The only difference is that we also visualize learned reconstructions for the unfair case of staged training of vanilla BACON. We keep all the experimental settings same, including the number of iterations dedicated to each stage and the learning rates. Further, the frequencies for all networks are initialized to have the same band-limit.

In Fig. 11, the outputs at each scale for both unfair and fair staged BACON, the residual MFN with the band-limited BACON initialization, and the residual MFN with our proposed initialization, are visualized. The differences between the Fourier spectra at each output scale after the corresponding training iterations and after the full training regiment are illustrated too (Fig. 11, right). We provide additional examples of fitting text images in Fig. 12, comparing reconstruction at coarser scales of two and three. We also report the Fourier spectra mean absolute difference and standard deviation in Table 2 for the text dataset. Finally all networks are examined on the same image generalization task and the Peak Signal-to-Noise Ratio (PSNR) between reconstructed signal and the original ground truth is computed. We include SAPE [9] as a baseline for coarse-to-fine reconstruction. Average PSNR values over all images is summarized in the last column of Table 2. Similarly, visualization of the final reconstruction on this task is provided in Fig. 13.

Table 2: Coarse-to-fine image fitting results on Text dataset [1]. The mean absolute difference of Fourier spectra of outputs at each scale immediately after optimization and after the entire training regiment is reported. With skip connections and the new initialization, we observe better conservation of spectra on scales two and three. Also the peak signal-to-noise ratio (PSNR) on the image generalization task is reported indicating that the final reconstruction is on par with baselines.

Method	Mean Abs. Difference			PSNR (dB)
	Scale 1	Scale 2	Scale 3	Scale 4
SAPE [9]	-	-	-	30.99 ± 2.32
BACON [8]	-	-	-	31.54 ± 2.31
(Staged) BACON (fair)	$\mathbf{1.06 \pm 0.29}$	3.57 ± 1.35	13.79 ± 4.60	31.57 ± 2.29
(Staged + Res) BACON	3.22 ± 1.06	3.77 ± 1.04	8.11 ± 2.60	31.17 ± 2.77
(Staged + Res + Init) BACON	3.01 ± 0.51	$\mathbf{2.99 \pm 0.99}$	$\mathbf{3.56 \pm 1.16}$	$\mathbf{32.04 \pm 2.34}$

B Cryo-EM

Single particle cryo-EM is a revolutionary imaging technique used by structural biologists to discover the 3D structure of macromolecular complexes, including proteins and viruses. Recovering 3D structure is a major step towards understanding how these miniature biological machines perform actions in our body cells. During sample preparation for a cryo-EM experiment, one first obtains a purified sample in solution, containing many instances of a specimen. This sample is then plunged into a cryogenic liquid, like liquid ethane, which freezes the molecules at random orientations. Once frozen in vitreous ice, the sample is loaded into a transmission electron microscope and exposed to

parallel electron beams which are recorded underneath by special detectors. The majority of these electrons pass through the structure without interaction, causing shot noise. The rest are scattered either in an elastic or inelastic fashion. Elastic scattering positively contributes to image formation by introducing a relatively weak phase contrast while the inelastic scattering damages the structure leading to garbage particles.

The image formation in cryo-EM can be well-approximated using the weak-phase object model [56]. This model assumes the 3D structure is a density map $V : \mathbb{R}^3 \to \mathbb{R}_{\geq 0}$ represented under a canonical orientation. Ideally, without any corruption, as depicted in Fig. 14, one can formalize clean images $\{I_i^*\}_{i=1}^n$ as 2D orthogonal projections of the map under random orientations $R_i \in SO(3)$,

$$\begin{aligned} I_i^*(x,y) &= (\mathcal{P}V_{R_i})(x,y) \ , \\ \mathcal{P}V_{R_i} &= \int_{\mathbb{R}} V_{R_i}(x,y,z)dz \ , \\ V_{R_i}(x,y,z) &= V(R_i^T(x,y,z)^T) \ . \end{aligned} \tag{11}$$

Here, $\mathcal{P}$ is a linear operator performing the orthogonal projection and V_{R_i} corresponds to a structure rotated by R_i. However, in practice, images are intentionally captured under defocus in order to improve the contrast. This is modeled by convolving the clean images with an image-specific point-spread functions (PSF) g_i. We express all sources of noise with an additional term and, taken together, the observed image can be formally described as

$$I_i(x,y) = (g_i \star I_i^*)(x,y) + \epsilon(x,y). \tag{12}$$

Here, $\star$ denotes convolution, ϵ is the additive noise, and $V_{R_i}(x,y)$ corresponds to clean (noiseless) projection of the structure rotated by R_i, as in Fig. 14. Statistically, as several factors contribute to the noise, it is common practice to assume ϵ is white (or colored) Gaussian noise.

C Fourier Shell Correlation (FSC) and Resolution

Fourier shell Correlation (FSC) is the standard way to measure the resolution of a 3D density map. In a nutshell is is the correlation between the Fourier coefficients of two aligned maps, as a function of frequency [61]. Let S_r be a spherical shell with radius r centered at the origin of the Fourier domain (i.e., containing 3D frequencies whose wavelengths are close to $1/r$). Then, given two maps in the Fourier domain, F_1 and F_2, FSC(r) is defined as the normalized cross-correlation between the Fourier coefficients of F_1 and F_2 within S_r; i.e.,

$$\text{FSC}(r) = \frac{\sum_{k \in S_r} F_1(k)\, F_2^*(k)}{\sqrt{\sum_{k \in S_r} |F_1(k)|^2 \sum_{k \in S_r} |F_2(k)|^2}} \ . \tag{13}$$

One can show that FSC(r) is real-valued, and ranges from -1 to 1. Once correlations are computed for all shells, we plot FSC as a function of r. In practice, one defines $D/2$ shells for maps with D^3 voxels. Typpically, one of F_1 or F_2 is the grount truth and the other is an estimated map. Or the dataset is randomly partitioned into two halves, used to independently estimate the two maps F_1 or F_2. At small r (i.e., low frequencies), where observed images have strong signal, maps are higher quality and FSC is close to 1. With higher frequencies the maps become noisier and FSC therefore decreases (see the right panel of Fig. 5). When a reconstruction is compared to a ground-truth map, one usually defines the resolution of the estimated map as the r for which FSC$(r) = 0.5$. Under mild assumptions this corresponds to a SNR of 1 [63].

It is worth noting that FSC also plays a key role in a more advanced version of frequency marching. In experimental settings, as the true map is not available, FSC is computed between two independent reconstructions. During frequency marching, as we increase the frequency support, the FSC curve is used to cross-validate the refinement to each reconstruction. In fact, due to high frequency noise in the data, refinements are likely to overfit independently, yielding inconsistencies in the subsequent high frequency regime. This is often observed as a sudden drop in the FSC curve. And when this is detected, we adaptively limit the frequency support [64].

In our work, we achieve high quality reconstructions without this adaptive regularization. On the other hand, we only have few scales for computational efficiency and this adaptive version is justified when more scales are dedicated. We leave further investigations of this issue to the future work.

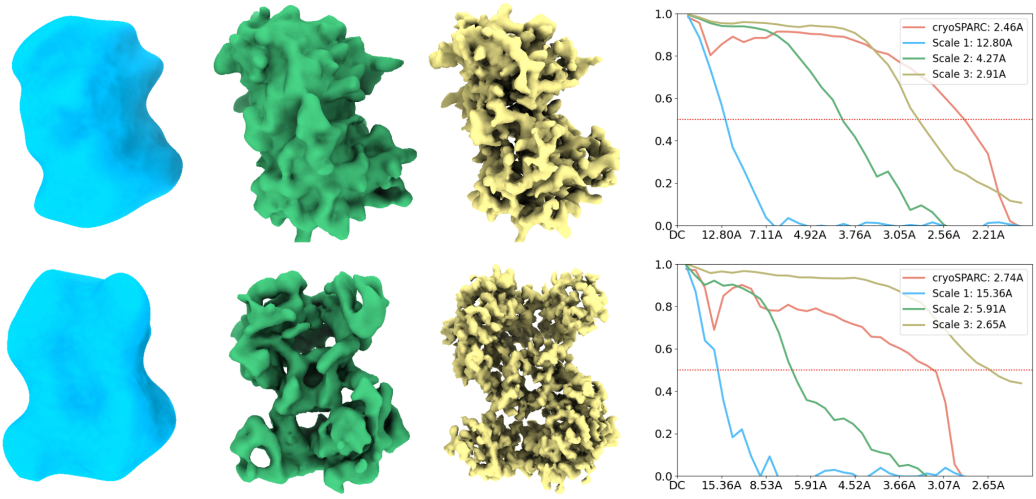


Figure 7: Results on coarser initialization of poses. (Left) Our reconstruction at multiple scales. (Right) Fourier Shell Correlation plot measuring correlation in Fourier spectrum between reconstructions and ground-truth structure. Threshold 0.5 is the criterion to compute the resolutions in Angstroms.

D Pose Optimization

Here, we provide the details of learning pose parameters, describing the 3D orientation of the molecule in an observed given particle image. Let us denote current estimate of pose parameters by $\hat{R}_i \in SO(3)$. By applying a small perturbation $\delta R_i \in SO(3)$, one can obtain a new estimate for poses by multiplication $(\delta R_i)\hat{R}_i$. Since, $SO(3)$ is a group closed under multiplication, the perturbed pose $(\delta R_i)\hat{R}_i$ remains a valid rotation. To find a better estimate, we iteratively find the optimal perturbation, and accordingly update pose estimates at each iteration. To parameterize the perturbation δR_i, similar to [21], we adopt the exponential map which is a mapping from the tangent space to the points on the $SO(3)$ manifold. One way to realize the exponential map is through axis-angle representation $(\theta_i, \mathbf{u}_i)$, where θ_i is a scalar angle between zero and π and $\mathbf{u}_i \in \mathbb{R}^3$ is a unit vector. One can simply use Rodrigues formula to get back to the matrix form

$$\delta R_i = I + \sin(\theta_i)[\mathbf{u}_i]_\times + (1 - \cos(\theta_i))[\mathbf{u}_i]_\times^2 \,, \tag{14}$$

where $[\mathbf{u}_i]_\times$ denotes the skew-symmetric 3×3 matrix of $\mathbf{u}_i$. We rely on gradient-based optimization to solve for optimal perturbation parameters $(\theta_i, \mathbf{u}_i)$

$$(\hat{\theta}_i, \hat{\mathbf{u}}_i) = \underset{(\theta_i, \mathbf{u}_i)}{\arg\min}\, \mathcal{L}\left((\theta_i, \mathbf{u}_i)|V, I_i\right), \tag{15}$$

where the image-specific loss is conditioned on the fixed potential map V and the corresponding image I_i. We iteratively compute gradients with respect to the pose parameters and then update them. Note that, after such an update, the resulting pose angle may no longer lie in the valid interval $[0, \pi]$, or the new axis $\hat{\mathbf{u}}_i$ may no longer have unit length. For this reason we re-establish these constraints by truncating angles back to $[0, \pi]$, and normalizing $\mathbf{u}_i$ to be a unit vector.

E Coarser Initialization of Poses

We conduct further experiments where poses are initialized to be uniformly distribution with angles within $[0, \pi]$ of the ground-truth. In this more challenging coarser initialization of pose, we dedicate more epochs to optimize the first and second scales, namely 30 epochs each. Qualitatively, we obtain reconstructions on par with those in the main paper. In terms of our FSC resolution metric, we achieve full-scale resolutions of 2.91A and 2.65A for PDB1ol5 and PDB4ake, respectively, as shown in 7.

F λ hyperparameters

The λ hyperparameters control the growth and disjointness of frequency support in the residual MFN (see Sec. 4.1). Our first intuition was to avoid settings that create gaps in coverage of the Fourier domain, i.e., when $\lambda_2 > 2 + \lambda_1$. Beyond this constraint, we investigated a range of hyper-parameter settings in our experiments. In all cases, the network fits the signal well; but we found that for $\lambda_2 = 2$, $\lambda_1 = 0.3$, the model yields a lower test loss (based on held out data for image fitting experiments). For cryoEM experiments (3D regression) we found the same hyperparameters perform well, based on FSC with ground truth 3D density. We therefore used the same hyperparameters for cryo-EM experiments with 2D images, without further hyper-parameter search.

G Videos

For further visualization of cryo-EM experimental results, we use Chimera [60] to create a set of videos, provided under *videos* folder, showing the results obtained by our method. In videos *pdb1ol5_all_scales.mp4* and *pdb4ake_all_scales.mp4*, we visualize final 3D structures learned at different scales, respectively for pdb1ol5 and pdb4ake. Reconstructions are rotated in the visualization to show the resolved structural features from several views.

We further visualize the coarse-to-fine evolution of the structure for each dataset during the entire training in videos *pdb1ol5_evolution.mp4* and *pdb4ake_evolution.mp4*. In both cases, we start from a random structure. We then obtain a coarse reconstruction and later stages capture more fine-grained (i.e. high-resolution) details of the estimated structure.

Finally, we use Chimera to fit the available atomic model of pdb4ake into the corresponding final reconstruction. As shown in video *model_fit.mp4*, various secondary structures, such as α-helices, are well aligned with the reconstructed 3D density map.

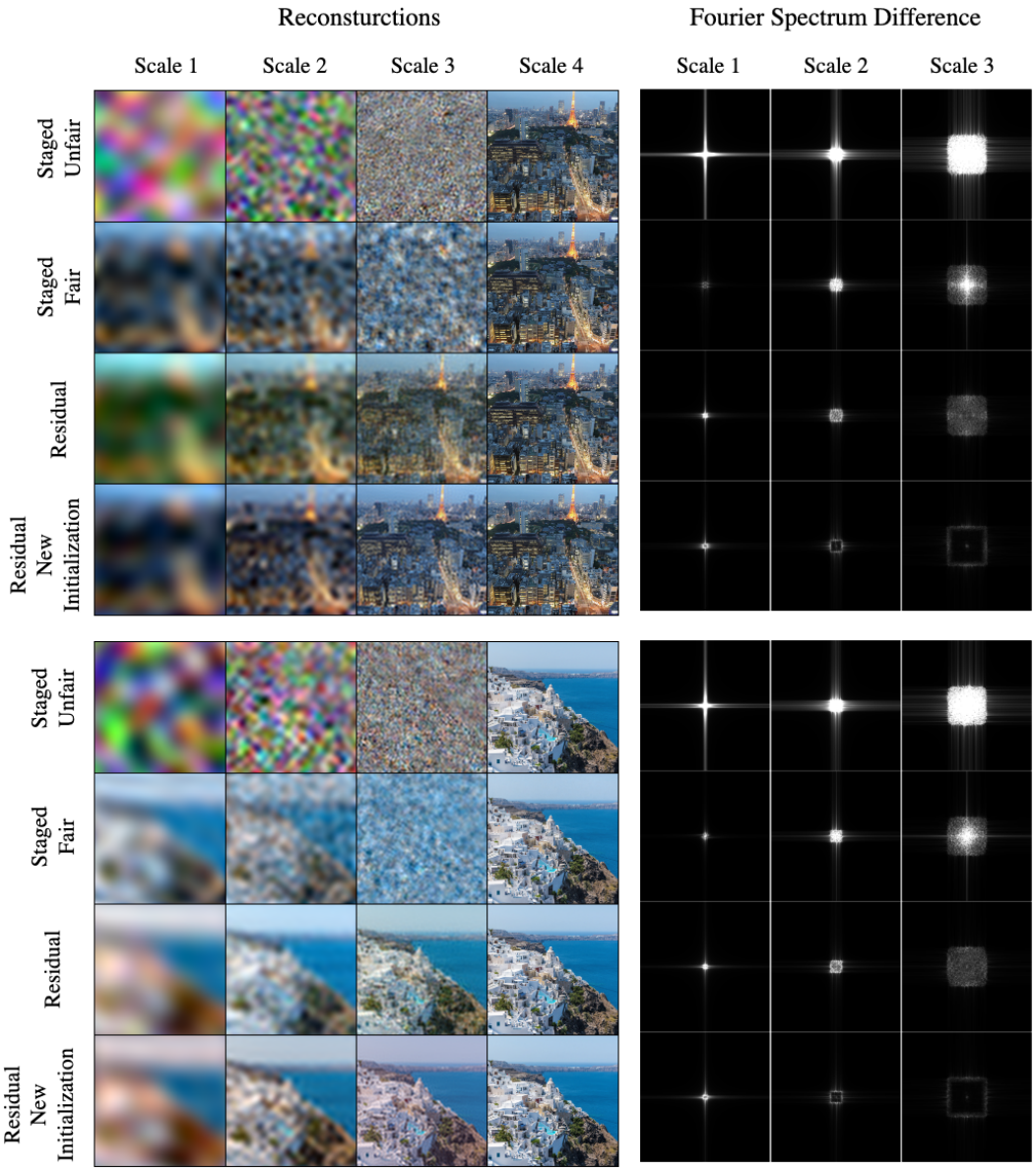


Figure 8: Coarse-to-fine fitting experiment on two example Natural images. (Left) Reconstructions depicted at all scales. (Right) Disruption to Fourier spectra of coarser scales caused by later training stages.

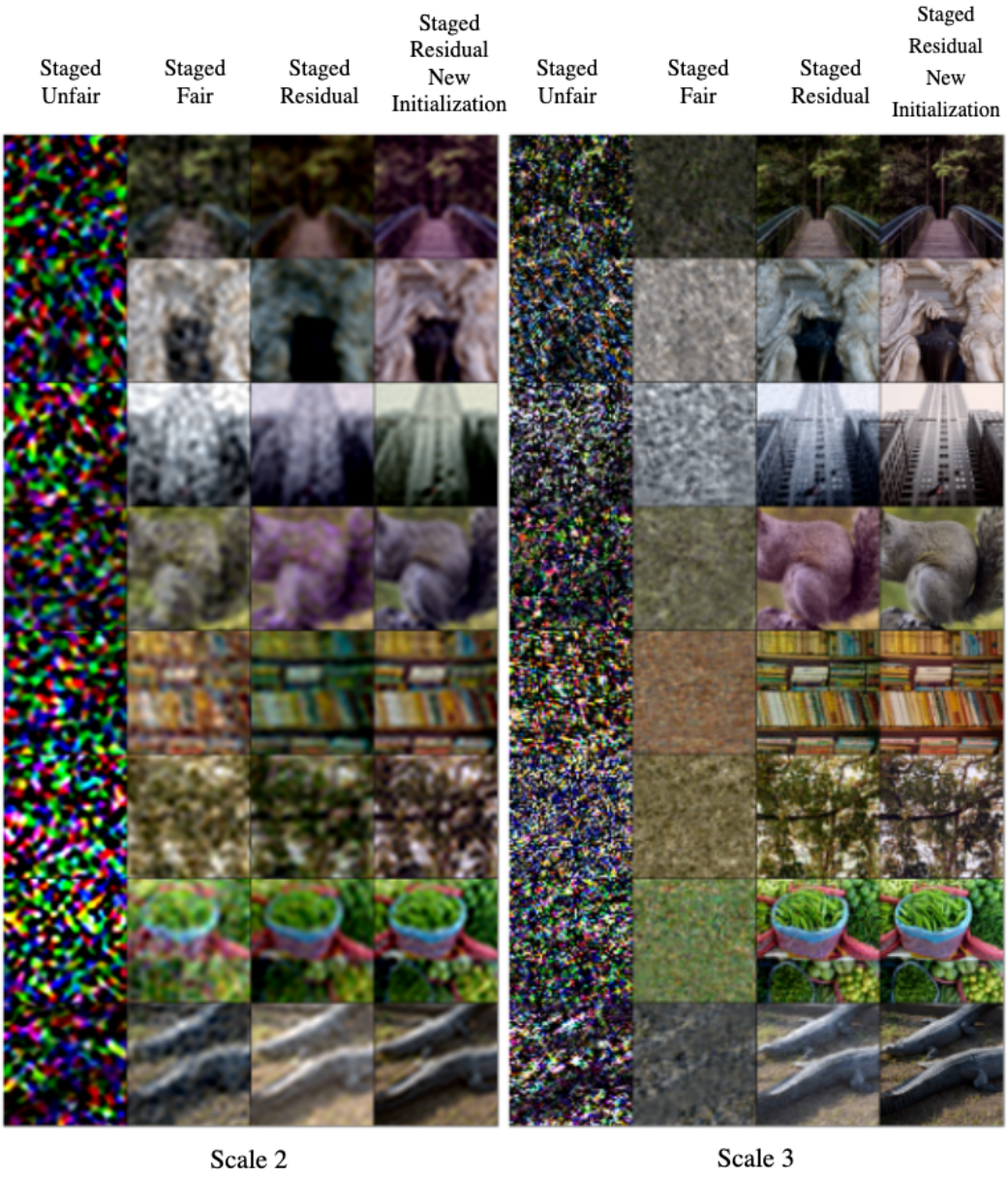


Figure 9: Coarse-to-fine fitting experiment on more examples of Natural images. Qualitative comparison of reconstructions at scales 2 (Left) and 3 (Right) for different methods imply that both residual connections and the new initialization noticeably improve the preservation of spectra at lower scales, especially scale 3.

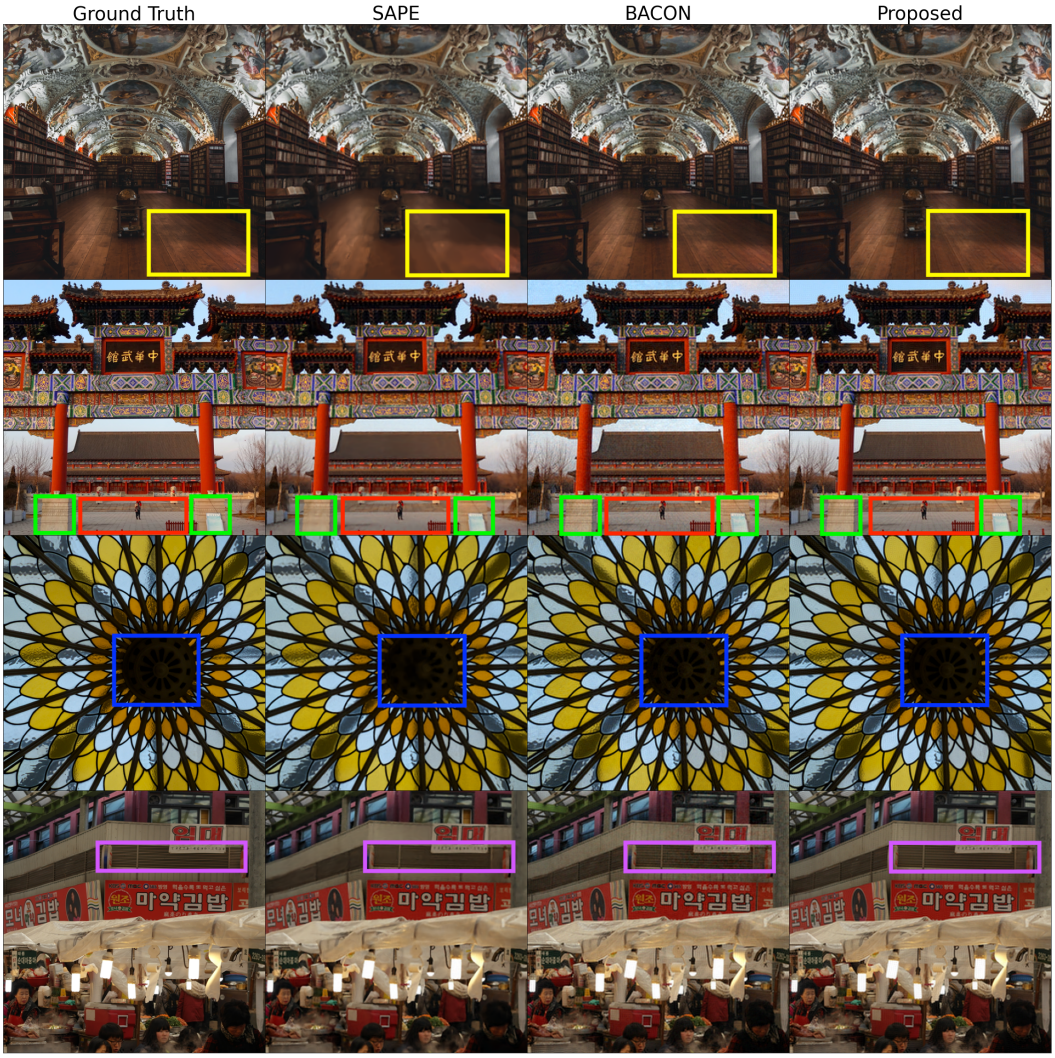


Figure 10: Examples on Image Generalization on Natural dataset. Some areas are highlighted with colored boxes showing regions where differences among reconstructions or from the ground truth can be observed.

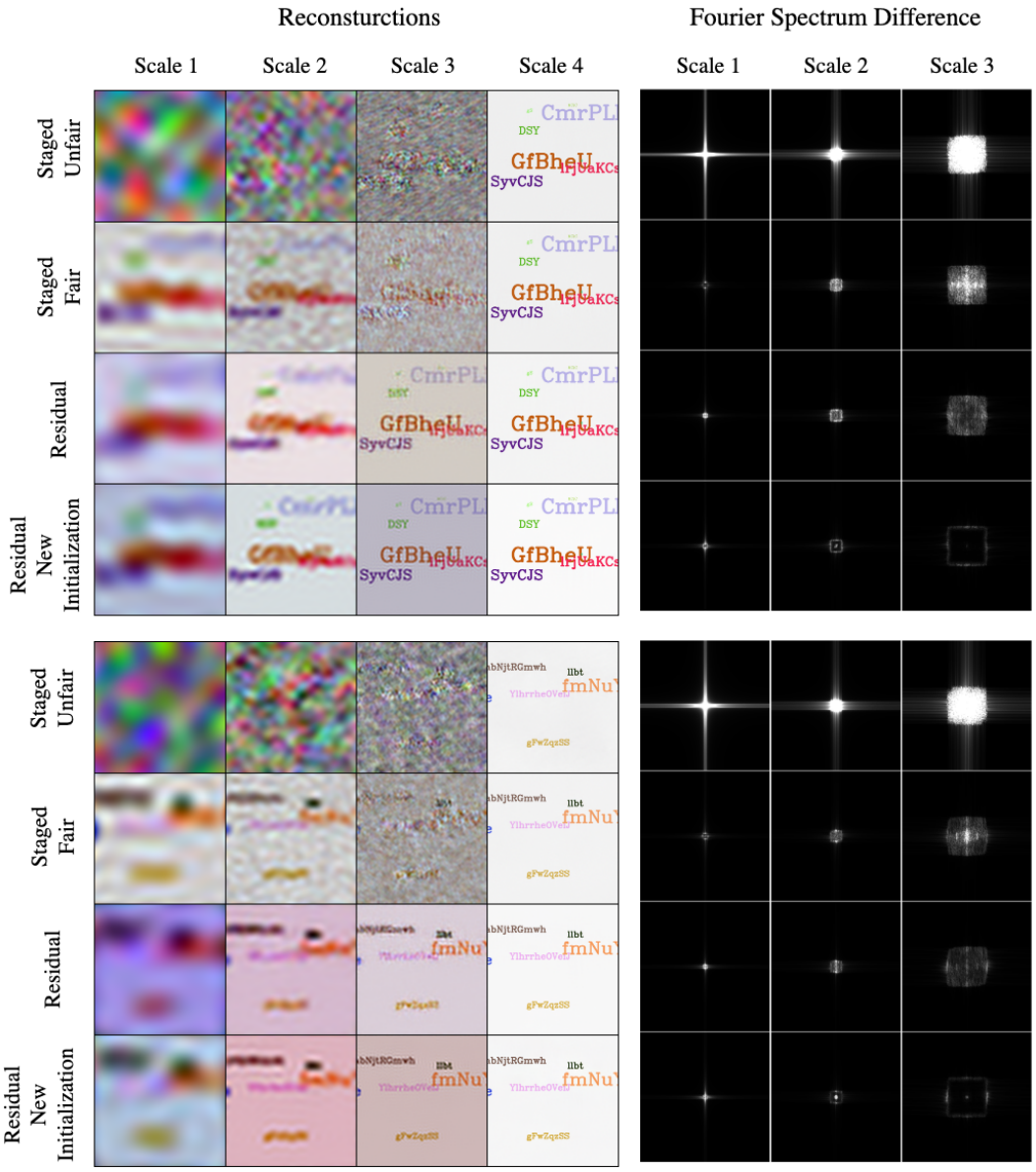


Figure 11: Coarse-to-fine fitting experiments on two example Text images. (Left) Reconstructions depicted at all scales. (Right) Disruption to Fourier spectra of coarser scales caused by later training stages.

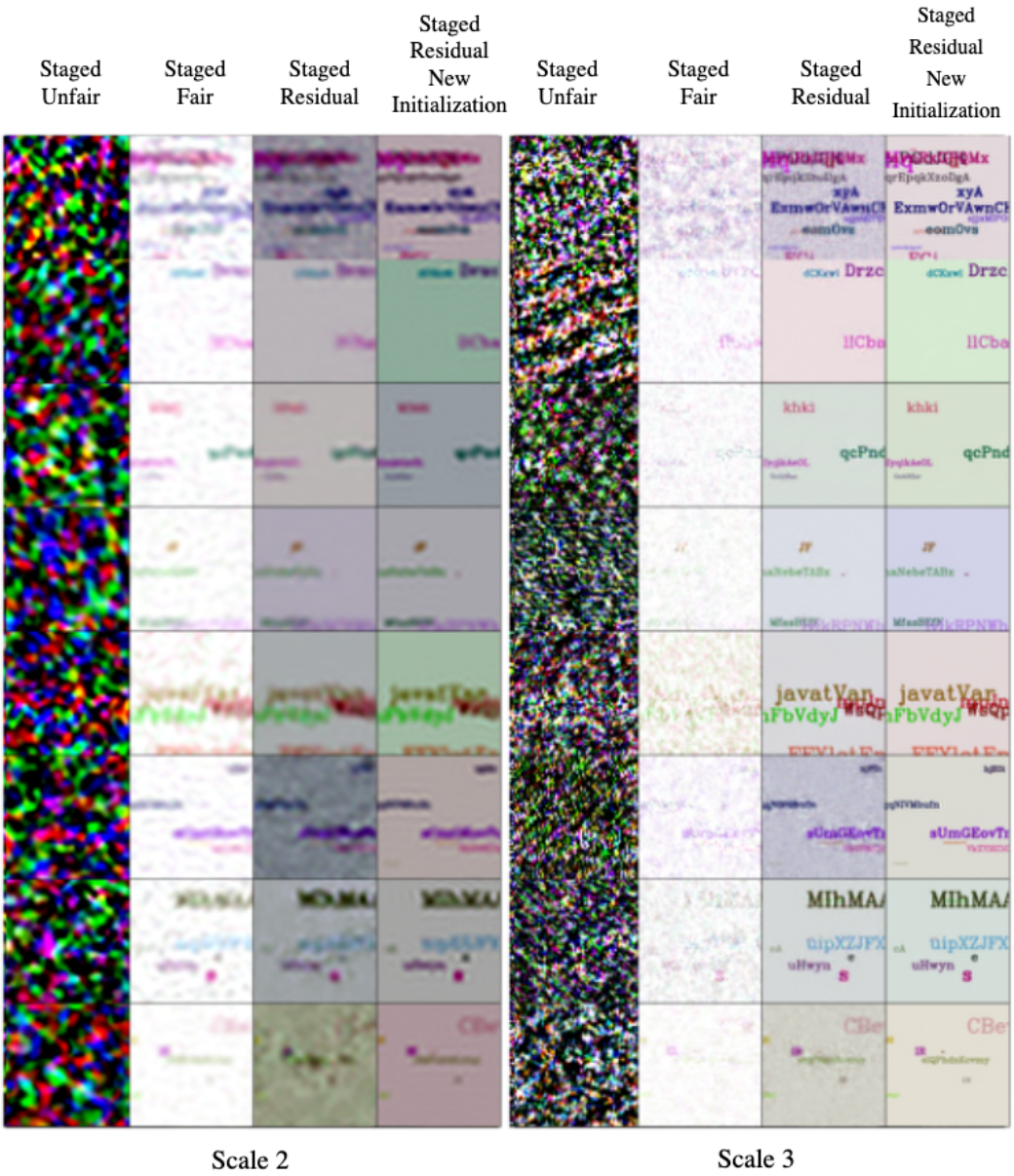


Figure 12: Coarse-to-fine fitting experiment on more examples of Text images. Qualitative comparison of reconstructions at scales 2 (Left) and 3 (Right) for different methods imply that both residual connections and the new initialization noticeably improve the preservation of spectra at lower scales, especially scale 3.

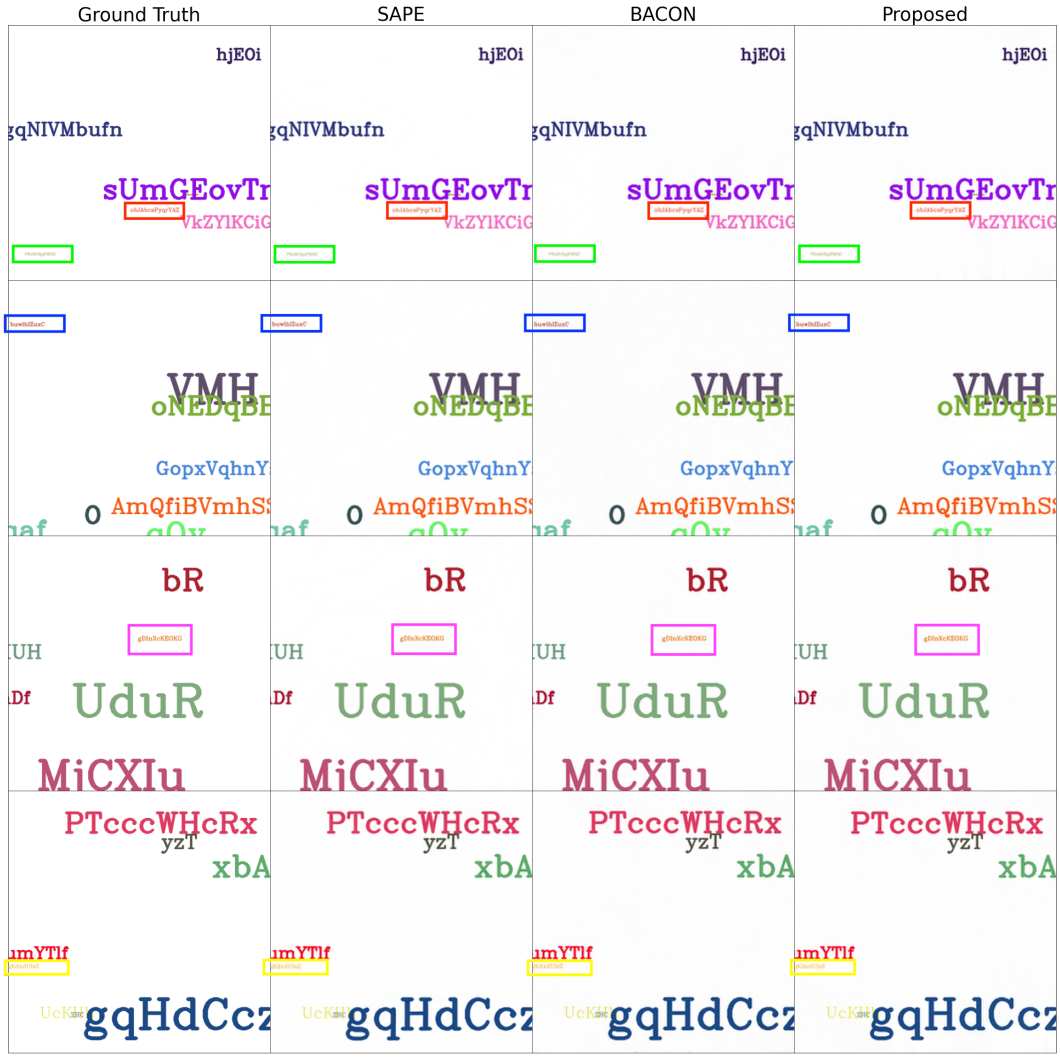


Figure 13: Examples on Image Generalization on Text dataset. Some areas are highlighted with colored boxes showing regions where differences among reconstructions or from the ground truth can be observed.

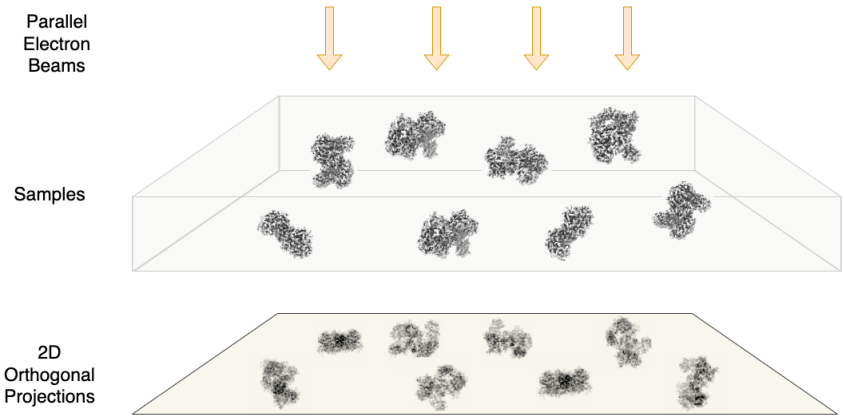


Figure 14: Ideal image formation model generating clean 2D orthogonal projections of samples.