3D Shape Completion using Multi-resolution Spectral Encoding

Pallabjyoti Deka Saumik Bhattacharya Debashis Sen Prabir Kumar Biswas
Indian Institute of Technology Kharagpur
pallabjyotidk@gmail.com, {saumik, dsen, pkb}@ece.iitkgp.ac.in

Abstract

Reconstruction of intricate local patterns and large missing regions during 3D shape completion has the contradictory requirements of computation over a wider context and operations for finer detail restoration. To this end, we propose a multi-resolution spectral encoding based 3D shape completion approach to work on truncated Signed Distance Field (SDF) based shape representations. Our novelty lies in judiciously integrating multi-resolution 3D convolutional blocks that encode the input shape and a spectral module (SM) that captures the shape-wide context, thus addressing the contradictory requirements. SM acts on the features extracted from both partial input scans and shape priors using the multi-resolution convolutional blocks. Our SM contains a 3D convolutional block placed between fast Fourier transform (FFT) and inverse FFT operations, which results in the expansion of the receptive field for the appropriate context computation. Our approach has an attention-based encoder-decoder architecture, where the encoding of a partial scan is acted upon by shape prior encodings to produce attention maps. These attention maps are leveraged differently in pretraining, and in the later training and inference stages of our approach to produce the reconstructed 3D shape. A surface gradient-based loss function is used in addition to the L1 loss, both in the pretraining and training stages for emphasizing the differences in minute details. These along with an attention refinement operation often leads to complete reconstruction while restoring finer details. Experiments using standard synthetic and real datasets demonstrate the superiority of our approach over the state-of-the-art.

1. Introduction

Many areas such as robotics [40], archaeology [11, 35], medical imaging [26] and augmented reality [14] rely heavily on accurate 3D reconstruction, especially since modern RGB-D sensors such as Microsoft Kinect or LiDAR sensors on phones are able to acquire high-quality scans [6, 9, 21, 32, 33]. Despite the recent advances in 3D scanning technology, there still remains a significant possibility of obtaining incomplete and noisy scans. Thus, a shape completion approach is essential, which not only generates the missing parts more realistically but also does so in a manner where the finer details are reconstructed faithfully. Moreover, such a model should generalize well to novel object categories, which were not considered during training.

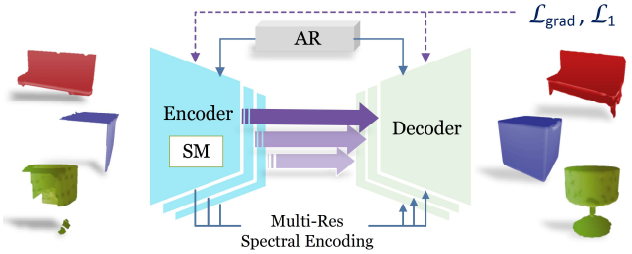

Figure 1. The overview of our proposed 3D shape completion approach. It has a multi-resolution attention-based encoder-decoder architecture, performs Fourier convolution as a part of spectral encoding and uses Attention Refinement (AR) block and $\mathcal{L}_{grad}$ loss.

A considerable amount of effort has been dedicated to 3D shape completion utilizing diverse 3D shape representations, where unique challenges associated with each representation have been addressed. These challenges include the sparsity of point clouds [7, 27, 30], voxelization artifacts in voxel-based methods [20, 46], intricate topology constraints in mesh-based methods [1, 23, 39], and loss of details in implicit function-based representations [2, 16, 49].

Recently, the use of implicit functions based on truncated Signed Distance Field (SDF) has been found to demonstrate superior shape modeling performance compared to the other three representations at least in a few cases [12, 36, 44, 47]. Modern 3D shape completion approaches have been proposed based on various investigations, particularly, in the areas of multi-modal processing [31, 45], single image guidance [28, 41], and single /multi-view inputs [10, 17, 18, 34]. A brief overview of the state-of-the-art shape completion methods is given in Sec. 2.

In cases of shapes with complex patterns or with larger missing areas, high-quality shape completion generally

requires computations over a substantial receptive field. While multi-resolution architectures can compute over a hierarchically formed large receptive field, we find that such a receptive field can be inadequate for capturing *shape-wide* characteristics. Spectral Convolution Theorem [24] states that modifying a single value in the spectral domain has a global impact on the original data. Thus, computation over non-local receptive fields performed using Fourier convolution [4] can be used to capture the shape-wide attributes. However, multi-resolution architectures with local computations are more adept at appropriately learning the relation between local structures, which is crucial for reconstructing finer details. Incorporating both these contradictory requirements is central to achieving satisfactory 3D shape completion in different scenarios, which forms our motivation.

Therefore, in this paper, we propose a shape completion model that leverages a judicious integration of local multi-resolution computation and Fourier convolution to work with a shape-wide context while preserving the model's ability to learn from local structures. Spectral module (SM), one of the key components of our model, computes channel-wise real Fast Fourier Transform (FFT) and its inverse to capture a shape-wide global context. The spectral module works on learnt multi-resolution features to produce an output, which is fused with the learnt features to perform encoding. This novel configuration allows an adaptation of the features into the Fourier convolution, which is unprecedented. To the best of our knowledge, this is the first attempt at designing a 3D shape completion network based on Fourier convolution that capture shape-wide context integrated with multi-resolution modeling of intricate details.

As shown in Fig. 1, our approach has an attention-based multi-resolution encoder-decoder architecture, where the use of SM causes only a minor increase in the training time (See *Supplementary*). Our method encodes both the partial input scan and shape priors, where the set of multi-resolution prior encodings are learnt from the shape priors to represent common substructures [34]. The prior encodings act on the partial input encodings through an attention mechanism at multiple resolutions to perform effective shape completion. During the attention map computation, intricate variations are highlighted using an Attention Refinement (AR) block that performs attention mixing using convolutional blocks.

While training on SDF values for 3D shape completion, the $\mathcal{L}_1$ loss function has been predominantly used in literature [10, 34]. However, we find that the use of $\mathcal{L}_1$ loss alone does not provide the required emphasis on the discrepancies around finer details. Hence, we propose the use of a local SDF surface gradient-based loss $\mathcal{L}_{grad}$ as well to address the issue.

In our experiments, we observe that our contributions ensure a detailed reconstruction of 3D shape and our method outperforms the current state-of-the-art in the popularly used ShapeNet [3] and ScanNet [8] datasets on Intersection-over-Union (IoU), Chamfer Distance (CD) and F_1 measures. In summary, the main contributions of our paper are as follows:

- We propose a spectral module based on Fourier convolution and multi-resolution processing for 3D shape completion. This facilitates the availability of a shape-wide receptive field for fetching the context without compromising the reconstruction of local structures.
- We propose the use of the SDF surface gradients as an additional loss along with $\mathcal{L}_1$, and the use of an attention refinement block on multi-resolution encodings to emphasize intricate details and local patterns.

Rest of the paper is organized as follows. In Sec. 2, we provide a comprehensive overview of the related works. Sec. 3 details our methodology for the 3D shape completion task. In Sec. 4, we evaluate the performance of our work compared to current state-of-the-art methods. Finally, we conclude the paper in Sec. 5 summarizing our key contributions and discussing the significance of our findings. The implementation of our code will be available at github.com/pjd96/MSSC.

2. Related Work on 3D Shape Completion

In the context of 3D real-world objects represented as $S \in \mathbb{R}^3$, wherein certain regions may be damaged or missing, the fundamental objective revolves around estimating the most plausible geometric structure for these incomplete areas with fine-grained details. In existing work, Sahay et al. [35] discussed a method rooted in Dictionary Learning to address the issue of filling the missing sections in 3D mesh objects. Another noteworthy technique, Poisson surface reconstruction [25], is aligned with this surface reconstruction paradigm. This approach entails a hierarchy of locally supported basis functions, leading to a sparse linear system.

Recent advancements have witnessed a notable shift towards deep learning-based solutions, which can harness data-driven knowledge for 3D shape completion. Notably, there are point cloud-based methods that utilize GAN inversion [12] [47], where a pre-trained GAN is employed to seek one or multiple latent codes that can best reconstruct the given partial input. Hu et al. [18] on the other hand focussed on shape completion using multi-view consistent inference. Additionally, other point cloud methodologies [19] [43], have demonstrated remarkable performance on synthetic datasets. However, it is essential to note that these successes are often observed in scenarios where the testing data belongs to known categories. In contrast, some works like ours consider both seen and unseen categories.

Much progress towards shape completion can be seen in the methods with encoder-decoder architecture. 3D-EPN [10] proposed such an architecture to complete partial 3D shapes. IF-Nets [5] leveraged continuous implicit functions for this task, whereas Auto-SDF [31] took an autoregressive modeling approach for the multimodal shape completion by utilizing a VQ-VAE backbone. Wallace and Hariharan [41] addressed few-shot reconstruction challenges by leveraging category priors. PatchComplete [34] also utilized priors for 3D shape completion in novel categories, employing a multi-resolution architecture. In our work, the primary focus lies in incorporating the global context of the shape along with surface gradients that emphasizes local structures to achieve detailed and reliable shape completion.

2.1. Fourier Transforms in Neural Networks

Given the efficiency of Fourier transforms in neural networks with minimal computational overhead, they have been incorporated into diverse architectures, including RNNs, CNNs, and transformers. For instance, [37] worked on reconstructing 3D from 2D images that is based on Fourier projection-slice theorem. It tries to learn a projection from 2D image to a 2D slice of 3D shape by predicting a thickness map through deep neural network. We, on the other hand, use a multi-resolution Fast Fourier Convolution (FFC) based method for 3D shape completion. Fast Fourier Convolution (FFC) [4] conducts convolution in the frequency domain to effectively provide non-local receptive fields to capture wider contexts. In the realm of image inpainting, works like [38] and [22] utilize FFC, significantly improving the quality of reconstructed images by introducing global context in the early layers of the network. However, none of the existing Fourier convolutional operations support multi-resolutional processing, which is one of the key contributions of our work. Moreover, to the best of our knowledge, Fourier convolution-based operations are not yet explored for 3D reconstruction tasks.

3. The Proposed Method

We use the volumetric truncated Signed Distance Field (SDF) [21] to represent a 3D shape inside a box $\mathcal{B}$ of size D^3 with grid cells of equal size, and consider $D = 32$. Building on the motivation elaborated in Sec. 1, in our 3D shape completion work, we focus on computing from a reasonable context for the reconstruction to achieve faithful (to the entire shape) and realistic results. We also emphasize on the reconstruction of the finer details in the missing portions along with the overall structure. As evident from Fig. 1, our approach has a pretraining stage involving encoding processes, which is discussed in Sec. 3.1, and a later training stage involving encoding and decoding, which is discussed in Sec. 3.2.

3.1. The Model Encoders and their Pretraining

For the model of our 3D shape completion approach, we adopt an attention-based multi-resolution encoder-decoder architecture as shown in Fig. 2, where the partial scan input as a truncated SDF is denoted as Q and a shape prior by K. Similar to Q, the ground truth scan S_{gt} is represented as a truncated SDF of size D^3.

Our approach employs two parallel encoders to encode the partial scan input and the shape priors, respectively. These two encoders have similar structures, where 3D ResNet blocks [15] compute features from the respective inputs. This feature computation is performed at three SDF resolutions $(D/R)^3$ with $R = 32,\ 8,\ 4$, employing downsampling.

3.1.1 Spectral Module:

Here, we propose the use of Fourier convolution [4] for the shape completion problem. The features computed from the ResNet blocks in both the encoders are acted upon by our Spectral Module (SM) independently to aggregate comprehensive global cues from the entire 3D shapes.

As shown in Fig. 2, our SM contains a 3D convolutional block with ReLU and batch normalization acting on the concatenated real and imaginary spectral components obtained through 3D Fast Fourier Transform (FFT). The output of the 3D convolutional block is split into two parts representing the transformed real and imaginary spectral components, which are then subjected to 3D inverse FFT. The output from the inverse FFT is added as a residual to the input into the FFT operation, which is obtained by applying a $1 \times 1 \times 1$ convolutional block on the input features where the feature dimension is squeezed to reduce computation. The residual-added output is then subjected to a $1 \times 1 \times 1$ convolutional kernel to restore the input feature dimension.

3.1.2 Encoding:

We denote the outputs from the two SMs of the two encoders as $g_{partial}$ and g_{prior}, and the feature outputs from the corresponding ResNet blocks as $f_{partial}$ and f_{prior}, which are shown in Fig. 2. While $g_{partial}$ and g_{prior} carries global cues, $f_{partial}$ and f_{prior} predominantly represents local features. Hence, we obtain encodings laden with both global and local attributes as follows:

$$Q' = f_{partial} + g_{partial} \tag{1}$$

$$K' = f_{prior} + g_{prior} \tag{2}$$

where $+$ represents element-wise addition. We refer Q' as the input encoding and K' as the prior encoding, and their corresponding encoders as $e_{partial}$ and e_{prior}, respectively.

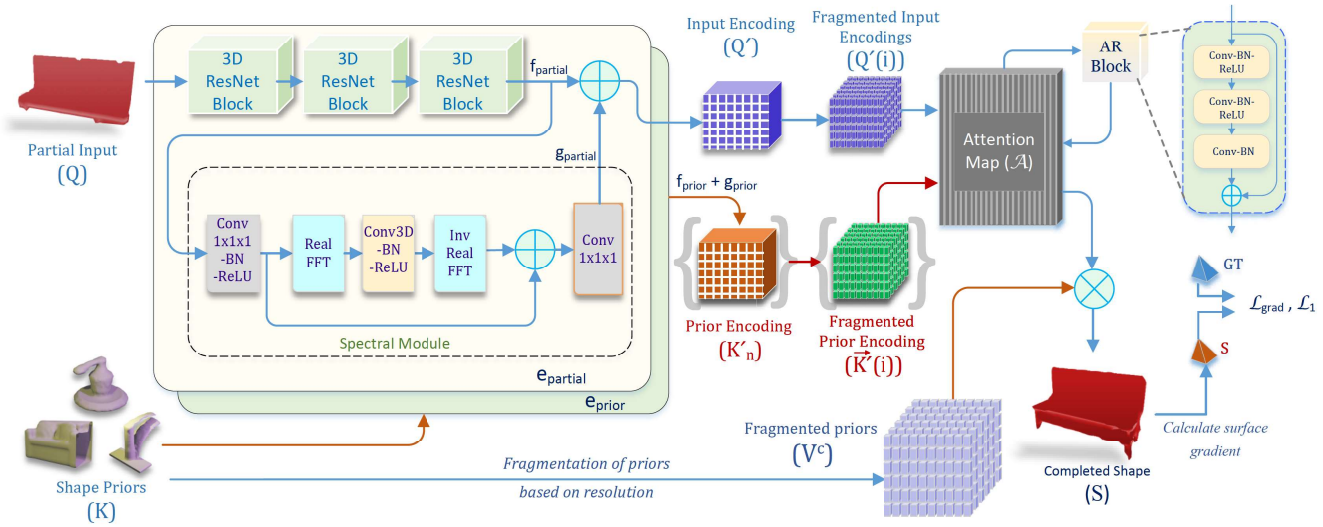

Figure 2. Overview of the proposed encoding process at the pretraining stage. Here, $e_{partial}$ and e_{prior} encoders have same architecture. The spectral module implements the Fourier convolution which allows for shape-wide context computation. While the use of AR block enables emphasis on local patterns through attention mixing, the employment of the $\mathcal{L}_{grad}$ loss provides sufficient penalty on the intricate detail discrepancies.

The computations resulting in the Q' and K' encodings consider a global shape-wide receptive field to capture the context through the use of SM, while carrying the local details.

3.1.3 Attention maps and shape reconstruction:

In our approach, we employ the publicly available 112 learnable shape priors provided by [34], which are created using mean-shift clustering of objects within each training category. We denote the multiple shape priors as $K_n, n = 1, 2, \ldots, 112$. Let us represent the corresponding 112 prior encodings from e_{prior} as $K'_n, n = 1, 2, \ldots, 112$, and consider them in an array $(K'_n, \forall n)$. Consider the fragmentation of Q' and all K'_n into $Q'(i)$ and $K'_n(i)$, respectively, where $i = 1, 2, \ldots, (D/R)^3$. Here, fragmentation yields 3D local regions in the input and prior encodings, and we use it to map local partial inputs to local shape prior regions. We also have the array $(K'_n(i), \forall n)$ and let us represent it as $\overrightarrow{K'}(i)$. We compute attention maps by querying using $Q'(i)$ for the relevant encoding in $\overrightarrow{K'}(i)$ through cross-attention computation as follows:

$$\mathcal{A}(i) = softmax_i(Q'(i)\overrightarrow{K'}(i)^T/(d/2)) \tag{3}$$

where d is the feature dimension of $Q'(i)$ and $\overrightarrow{K'}(i)$.

In order to perform the shape reconstruction during the pretraining stage, each input shape prior K_n is reshaped to reduce its dimension from D^3 to $(D/R)^3$ by transferring content to the channel dimension yielding K^c_n. This is done so that the reshaped K^c_n can also be fragmented into $K^c_n(i)$ with $i = 1, 2, \ldots, (D/R)^3$. Let us denote the array $(K^c_n(i), \forall n)$ as $V^c(i)$. The fragmented part $S^c(i)$ of the reconstructed shape is then obtained by attending the input shape priors as:

$$S^c(i) = \mathcal{A}(i)V^c(i) \tag{4}$$

and then $S^c(i), \forall i, c$, are recomposed to get the reconstructed shape S^c, which is further reshaped to revert the earlier content transfer to the channel dimension in order to get the full reconstructed shape S as a truncated SDF of D^3 dimension.

3.1.4 Attention Refinement (AR) block:

To achieve more refined representations of local patterns, we design an Attention Refinement (AR) block that acts on computed attention map $\mathcal{A}$. Our AR block processes the 3D attention maps to put preference on the attention values computed corresponding to certain shape priors over the others. Such an operation not only provides the opportunity to mix attention values related to different priors but also allows emphasis on certain attention maps with intricate local variations in them. We implement our AR block on the $\mathcal{A}$ computed for $R = 32$.

On the attention map $\mathcal{A}$, a convolution kernel w of size $k \times k \times k$ is applied as follows:

$$\mathcal{A}_{l,m,n} = \sum_{a,b,c=-k/2}^{k/2} w_{a,b,c} \cdot \mathcal{A}_{l+a,m+b,n+c} \tag{5}$$

We apply a sequence of channel-wise *Conv-BatchNorm-ReLU* processes [50] to effectively mix the attention values. In our experiments, we observe that AR block enables the network to significantly improve the reconstructed regions, as evident from the sample in Tab. 5 given in Sec. 4.4.

3.1.5 Surface Gradient-based Loss Function:

Although $\mathcal{L}_1$ loss has been predominantly used for 3D shape completion, we find that utilizing only $\mathcal{L}_1$ loss fails to capture intricate discrepancies in details and results in smoothed-out reconstructed surfaces [35]. So, we propose to use an additional loss based on surface gradient $\mathcal{L}_{grad}$ to allow seamless reconstruction of finer details. The surface gradients are computed using the finite differentiation method [48] [13]. We use raw surface gradients without any kind of normalization so that they can replicate the surface normals [42]. The proposed $\mathcal{L}_{grad}$ is computed as:

$$\mathcal{L}_{grad} = ||\nabla S_{gt} - \nabla S||^2 \tag{6}$$

where ∇S_{gt} and ∇S are the surface gradients of the ground truth SDF S_{gt} and predicted SDF S, respectively. We use different weights on the $\mathcal{L}_1$ losses to penalize false predictions based on grid occupancy.

$$\mathcal{L}_{one} = w_a\mathcal{L}_1(S^{FN}, S_{gt}) + w_b\mathcal{L}_1(S^{FP}, S_{gt}) + w_c\mathcal{L}_1(S^{correct}, S_{gt}) \tag{7}$$

where S^{FN}, S^{FP} and $S^{correct}$ represent false negative, false positive, and correct sign predictions [34], whereas w_a, w_b and w_c are their penalty weights, respectively. Therefore, the overall loss used in our pretraining is:

$$\mathcal{L}_{total} = \mathcal{L}_{grad} + \mathcal{L}_{one} \tag{8}$$

The detailed architectures of the various parts of our partial input and shape prior encoders are given in the *supplementary*.

3.2. Encoder-Decoder Model Training and Inference

While the use of AR block enables refined representation of local patterns through attention mixing, the employment of the $\mathcal{L}_{grad}$ loss provides additional penalty on discrepancies in intricate details.

From Sec. 3.1.1, we obtain the pretrained encoder $e_{partial}$, which takes a partial scan Q as the input and works on it at the different resolutions $(D/R)^3$ with $R = 32, 8, 4$. The encoding Q' of the input Q is obtained from $e_{partial}$, and the different encodings K'_n of all the shape priors learned in our pretraining stage are then considered here in the training and inference stages of our approach. The attention maps $\mathcal{A}$ are then computed using Q' and K'_n as depicted in Eq. (3) and explained in Sec. 3.1.3. The AR block is also incorporated to modify $\mathcal{A}$ as discussed in Sec. 3.1.4.

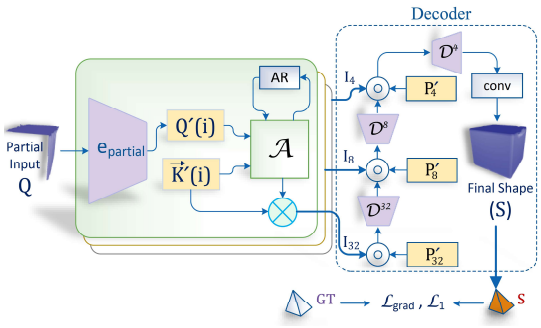

Figure 3. The complete training and inference methodology including the decoder module. The $\mathcal{L}_{grad}$ loss is used here as well along with $\mathcal{L}_1$ loss. The diagram depicts the working of the encoder-decoder framework at three different resolutions.

To perform the shape reconstruction in the training and inference stages, K'_n is fragmented into $K'_n(i)$ with $i = 1, 2, \ldots, (D/R)^3$ and the array $\overrightarrow{K'}(i) = (K'_n(i), \forall n)$ is considered to compute an intermediate quantity $I(i)$ as follows:

$$I(i) = \mathcal{A}(i)\overrightarrow{K'}(i) \tag{9}$$

I is then subjected to a concatenation process as follows:

$$J = I \circ P' \tag{10}$$

where $P' = Q' \circ g_{partial}$ (obtained from $e_{partial}$) and $\circ$ denotes concatenation. Let us denote the three different J obtained for $R = 32, 8, 4$ as J_R. All J_Rs are then used in a decoding process as follows:

$$S = conv(\mathcal{D}^4[J_4 \circ \mathcal{D}^8[J_8 \circ \mathcal{D}^{32}[J_{32}]]]) \tag{11}$$

where S is the desired full reconstructed shape of D^3 dimension during the training and inference. In the above, $\mathcal{D}^{32}[\cdot]$ is a decoder with transposed convolution and residual block [15] that maps the input embedding of $(3d, 1, 1)$ to $(d, 1, 1)$, and $\mathcal{D}^8[\cdot]$ and $\mathcal{D}^4[\cdot]$ are similar decoders where each of them maps the input from $(4d, 1, 1)$ to $(d, 1, 1)$. The $conv$ block represents a 3D convolution kernel.

The entire pipeline for the training and inference stage is shown in Fig. 3. Comprehensive architecture details of the various blocks in our encoder-decoder model are given in the *supplementary*.

4. Experiments

4.1. Setting up everything

Datasets: Our model is trained using a processed dataset [34], which includes a combination of the synthetic ShapeNet dataset [3] and real-world data from the ScanNet dataset [8].

During training, the model is exposed to the synthetic ShapeNet dataset encompassing 3202 objects across 18 categories. Subsequently, the model is evaluated on a testing set consisting of 1325 objects across 8 categories, all of which were unseen during the training phase.

In case of the real-world ScanNet dataset, fine-tuning of the ShapeNet-based trained model is performed using 8 categories containing 7537 samples. The model is then tested on 6 of the unseen categories comprising 1191 samples. Results of the seen categories in the real-world dataset are provided in the *supplementary material*.

For uniformity, all objects are represented as 32^3 SDFs, with truncated values set at 2.5 for ShapeNet data and 3 for ScanNet data, which are commonly used.

SOTA Comparison: We evaluate our work against different relevant state-of-the-art shape completion methods, namely, 3D-EPN [10], IF-Net [5], the few-shot approach of [41], and the recent AutoSDF [31] and PatchComplete [34]. While 3D-EPN and IF-Net perform shape completion tasks using voxel grid and continuous implicit function representations, respectively, AutoSDF [31] and PatchComplete [34] both use truncated SDF representations as considered in our approach as well.

Evaluation Measures: To assess the quality of the reconstructed shape, we employ two measures: the L_1 Chamfer Distance (CD) and Intersection over Union (IoU). These measures are evaluated on objects in the canonical system. We report the Chamfer Distance by sampling 10,000 points and scaling it by a factor of $\times 10^2$. For the occupancy grid-based methods [41] and [5], we use 0.4 and 0.5 as the occupancy thresholds, respectively. We extract the iso-surface at zero level for the SDF-based methods using the marching cubes technique [29]. We also use $F1$ score to evaluate the contributions of $\mathcal{L}_{grad}$, AR block and the Spectral Module in our shape completion task.

Implementation Details: Our model is trained on an NVIDIA A40 GPU using the Adam optimizer with a batch size of 32 for both the ShapeNet [3] and ScanNet [8] datasets. The initial learning rate is set to 10^{-3}, and the training is performed for 120 epochs, with a learning rate reduction by half after every 50 epochs to facilitate convergence. This configuration remains consistent throughout the training process using the synthetic dataset. The encoder pretraining durations for R=32, 8, and 4 are 1.8 hours, 2.8 hours, and 6.7 hours, respectively. The duration for the training stage of the multiresolution encoder-decoder model is 22.8 hours, which is later finetuned using the real-world ScanNet dataset.

4.2. Evaluation on Synthetic Data (ShapeNet)

In Tab. 1, we assess the performance of our approach using the previously discussed measures in the categories that were not part of the model's training (unseen categories). The results demonstrate that our approach outperforms the state-of-the-art in most cases. On average, both on an instance level and across categories, our method exhibits substantial superiority. The scores suggest that our method excels in reconstructing finer details compared to the existing approaches, and the overall quality aligns more closely with ground truth data. This improvement may be attributed to our model's ability to consider a wider global context while carrying local details during the shape completion process. The visual results can be seen on Fig. 4. As can be seen, the proposed method has higher fidelity in the reconstruction task compared to the others.

4.3. Evaluation on Real Data (ScanNet)

In Tab. 2, we evaluate our approach on the real scanned objects of the ScanNet dataset from unseen categories. The shapes here are more noisy compared to the ShapeNet dataset. As seen on Fig. 5, the predicted complete shapes by our approach are closer to the ground truth objects compared to the other methods.

4.4. Ablation Studies

How does $\mathcal{L}_{grad}$, SM, and AR block affect the model performance? Considering the unseen categories of ShapeNet dataset, we analyze the importance of the components of our method that mainly target to improve the overall global context required along with detailed reconstruction of the missing regions for the shape completion task. Compared to using $\mathcal{L}_{one}$ alone (**Base**), the use of $\mathcal{L}_{grad}$ along with $\mathcal{L}_{one}$ ($\mathcal{L}_{\mathbf{total}}$) mainly contributes to the improvement of Intersection-over-Union (IoU), which can be observed in Tab. 4. Further, when the Attention Refinement (AR) block is used with $\mathcal{L}_{total}$, improvement is observed in both IoU and Chamfer Distance (CD) (Tab. 3 and Tab. 4). Finally, when the Spectral Module (SM) is also used along with AR block and $\mathcal{L}_{total}$ (**Ours**), we achieve the best performance on an average in terms of both the measures. It is also observed from the table, the SM model with $\mathcal{L}_{one}$ (**SM**) performs significantly better than Base.

How does AR block improve the details in the predicted shape? From Tab. 3 and Tab. 4, we analyze the importance of the AR block. The table clearly shows its contribution in improving the quality of the reconstructed shape as explained earlier. Further looking into a single resolution, the positive impact of AR block is clearly evident from Tab. 5, the results of which are for unseen categories on the ShapeNet dataset.

More ablation results related to the components of our model, computational efficiency and effects of train-test splits are given in the *supplementary*.

	Chamfer Distance ↓ ($\times 10^2$)						IoU ↑					
	IFN	**3DEPN**	**FShot**	**PC**	**ASDF**	**Ours**	**IFN**	**3DEPN**	**FShot**	**PC**	**ASDF**	**Ours**
Laptop	6.47	3.90	10.35	3.77	4.81	**3.51**	0.583	0.620	0.313	0.638	0.511	**0.668**
Bathtub	4.72	4.21	7.05	3.78	5.17	**3.52**	0.550	0.579	0.457	0.663	0.410	**0.695**
Lamp	5.70	8.07	15.10	**4.68**	6.57	**4.68**	0.508	0.472	0.254	0.564	0.391	**0.587**
Bench	5.03	4.54	8.11	3.70	4.31	**3.58**	0.497	0.483	0.272	0.539	0.395	**0.558**
Printer	5.83	5.15	9.26	4.63	7.52	**4.47**	0.705	0.736	0.567	0.776	0.499	**0.780**
Basket	**4.44**	7.90	8.72	5.15	6.70	5.03	0.502	0.540	0.406	0.610	0.398	**0.635**
Bag	4.77	5.01	8.00	**3.94**	5.81	**3.94**	0.698	0.738	0.561	0.776	0.563	**0.780**
Bed	5.34	5.84	10.03	4.49	6.01	**4.35**	0.607	0.584	0.396	0.668	0.446	**0.678**
Inst. Avg.	5.37	5.48	9.75	4.23	5.76	**4.08**	0.574	0.582	0.386	0.644	0.446	**0.664**
Cat. Avg.	5.29	5.58	9.58	4.27	5.86	**4.13**	0.581	0.594	0.403	0.654	0.452	**0.673**

Table 1. Evaluation of the different approaches on unseen categories of ShapeNet [3] (synthetic data). The comparison table contains state-of-the-art methods IF-Nets [5], 3D-EPN [10], Few-Shot [41], PatchComplete [34] and AutoSDF [31].

	Chamfer Distance ↓ ($\times 10^2$)						IoU ↑					
	IFN	**3DEPN**	**FShot**	**PC**	**ASDF**	**Ours**	**IFN**	**3DEPN**	**FShot**	**PC**	**ASDF**	**Ours**
Bathtub	7.19	7.56	7.77	6.77	7.84	**6.41**	0.395	0.410	0.382	0.480	0.366	**0.501**
Lamp	10.16	14.27	11.88	9.42	11.17	**9.15**	0.249	0.207	0.196	0.284	0.244	**0.314**
Printer	8.28	8.36	8.30	6.84	9.66	**6.69**	0.607	0.630	0.622	0.705	0.499	**0.717**
Basket	6.74	7.74	8.02	**6.60**	7.54	**6.60**	0.427	0.365	0.343	**0.455**	0.361	0.447
Bed	8.24	7.76	9.07	**7.24**	7.91	7.26	0.449	0.478	0.349	0.484	0.380	**0.489**
Bag	8.96	8.83	9.10	**8.23**	9.30	8.31	0.442	0.537	0.449	0.583	0.487	**0.587**
Inst. Avg.	8.12	8.60	8.83	7.38	8.56	**7.24**	0.426	0.441	0.387	0.498	0.386	**0.508**
Cat. Avg.	8.26	9.09	9.02	7.52	8.90	**7.40**	0.426	0.440	0.386	0.495	0.389	**0.510**

Table 2. Evaluation of the different approaches on unseen categories of ScanNet [8] (real data)

	Chamfer Distance ↓ ($\times 10^2$)				
	Base	**SM**	$\mathcal{L}_{total}$	**AR,** $\mathcal{L}_{total}$	**Ours**
Bag	**3.94**	3.99	4.21	4.25	**3.94**
Lamp	4.68	4.65	4.65	**4.62**	4.68
Bathtub	3.78	3.61	3.76	3.69	**3.52**
Bed	4.49	4.42	4.57	4.48	**4.35**
Basket	5.15	**4.72**	5.22	5.17	5.03
Printer	4.63	**4.45**	4.66	4.68	4.47
Laptop	3.77	3.55	3.59	**3.45**	3.51
Bench	3.70	3.72	3.70	3.62	**3.58**
Inst. Avg.	4.23	4.11	4.24	4.18	**4.08**
Cat. Avg.	4.27	4.14	4.30	4.25	**4.13**

Table 3. Ablation study on the components of our method using Chamfer Distance (CD) as the evaluation metrics. 'Base' denotes the performance of the method when only $\mathcal{L}_{one}$ loss is used, $\mathcal{L}_{total}$ denotes the use of $\mathcal{L}_{grad}$ along with $\mathcal{L}_{one}$, 'SM' shows the performance when the Spectral Module is used with only $\mathcal{L}_{one}$. Inst. Avg. → Instance Average, Cat. Avg. → Category Average. The above results are based on the unseen categories of the ShapeNet dataset. The second-best scores are underlined.

	IoU ↑				
	Base	**SM**	$\mathcal{L}_{total}$	**AR,** $\mathcal{L}_{total}$	**Ours**
Bag	0.776	0.773	0.756	0.757	**0.780**
Lamp	0.564	0.578	0.578	0.579	**0.587**
Bathtub	0.663	0.682	0.665	0.676	**0.695**
Bed	0.668	0.669	0.661	0.664	**0.678**
Basket	0.610	0.629	0.623	0.625	**0.635**
Printer	0.776	**0.783**	0.770	0.768	0.780
Laptop	0.638	0.669	0.658	**0.674**	0.668
Bench	0.539	0.545	0.541	0.555	**0.558**
Inst. Avg.	0.644	0.657	0.648	0.654	**0.664**
Cat. Avg.	0.654	0.666	0.657	0.662	**0.673**

Table 4. Ablation study on the components of our method using Intersection-over-Union (IoU) as the evaluation metrics.

	Inst. IoU (%)↑	**Cat. IoU (%)↑**
Base (32^3 priors only)	35	37
Base with AR (with 32^3 priors)	**44.4**	**46.67**

Table 5. Effect of using AR block on the 32^3 encoding considering ShapeNet data

Figure 4. Comparison with state-of-the-art methods on synthetic ShapeNet [3] dataset in unseen categories

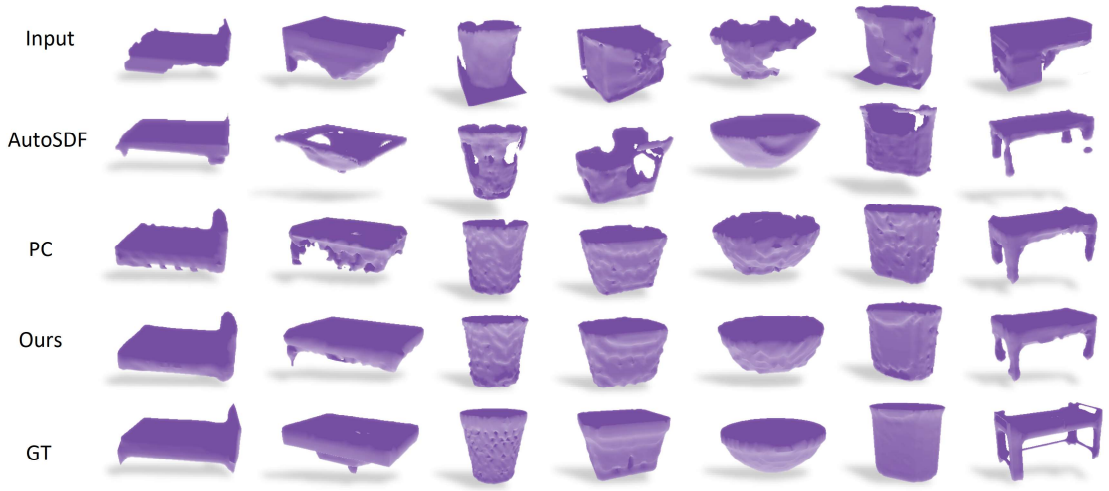

Figure 5. Comparison with state-of-the-art methods on the real-world ScanNet [8] dataset in unseen categories. The input data here is usually more noisy in comparison to the synthetic dataset.

5. Conclusion

In this study, we present a new 3D shape completion approach that excels in effectively learning shape priors and input encodings, ensuring robust shape completion. We emphasize the significance of our spectral module, which works along with multi-resolution convolutional blocks for the encoding. The spectral module helps the model to use the global context of a shape during reconstruction. Additionally, we address the challenge of reconstructing intricate details by introducing a gradient-based loss, $\mathcal{L}_{grad}$, on top of the $\mathcal{L}_1$ loss. Further, we discuss the enhancement of local patterns through the application of our attention refinement block, which operates on the attention maps computed from the encodings of local shape priors and the partial input. We find empirically through ablation studies that the aforesaid contributions of ours are indeed helpful in improving shape completion performance. Our objective evaluation employs the Chamfer Distance (CD), Intersection-over-Union (IoU) and the $F1$ measures, which also reveals that our model consistently outperforms the state-of-the-art methods in the field of 3D shape completion.

References

[1] Matthew Berger, Joshua A. Levine, Luis Gustavo Nonato, Gabriel Taubin, and Claudio T. Silva. A benchmark for surface reconstruction. *ACM Trans. Graph.*, 32(2), apr 2013. 1

[2] Matthew Berger, Joshua A Levine, Luis Gustavo Nonato, Gabriel Taubin, and Claudio T Silva. A benchmark for surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(2):1–17, 2013. 1

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 5, 6, 7, 8

[4] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4479–4488. Curran Associates, Inc., 2020. 2, 3

[5] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 3, 6, 7

[6] Sungjoon Choi, Qian-Yi Zhou, and V. Koltun. Robust reconstruction of indoor scenes. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[7] Yubo Cui, Jiayao Shan, Zuoxu Gu, Zhiheng Li, and Zheng Fang. Exploiting more information in sparse point cloud for 3d single object tracking. *IEEE Robotics and Automation Letters*, 7(4):11926–11933, 2022. 1

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 5, 6, 7, 8

[9] Angela Dai and M. Nießner. Scan 2 mesh : From unstructured range scans to 3 d meshes. 1

[10] Angela Dai, Charles Ruizhongtai Qi, and Matthias NieBner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. *computer vision and pattern recognition*, 2017. 1, 2, 3, 6, 7

[11] Luca Di Angelo, Paolo Di Stefano, and Emanuele Guardiani. A review of computer-based methods for classification and reconstruction of 3d high-density scanned archaeological pottery. *Journal of Cultural Heritage*, 56:10–24, 2022. 1

[12] Krishnendu Ghosh, Aupendu Kar, Saumik Bhattacharya, Debashis Sen, and Prabir Kumar Biswas. Multi-latent gan inversion for unsupervised 3d shape completion. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3460–3464, 2022. 1, 2

[13] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[14] Yun-Chih Guo, Tzu-Hsuan Weng, Robin Fischer, and Li-Chen Fu. 3d semantic segmentation based on spatial-aware convolution and shape completion for augmented reality applications. *Computer Vision and Image Understanding*, 224:103550, 2022. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 5

[16] Adrian Hilton, Andrew J Stoddart, John Illingworth, and Terry Windeatt. Implicit surface-based geometric fusion. *Computer Vision and Image Understanding*, 69(3):273–291, 1998. 1

[17] Tao Hu, Zhizhong Han, Abhinav Shrivastava, and Matthias Zwicker. Render4completion: Synthesizing multi-view depth maps for 3d shape completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[18] Tao Hu, Zhizhong Han, and Matthias Zwicker. 3d shape completion with multi-view consistent inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10997–11004, 2020. 1, 2

[19] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7662–7670, 2020. 2

[20] Patrick Hübner, Martin Weinmann, and Sven Wursthorn. Voxel-based indoor reconstruction from hololens triangle meshes. *arXiv preprint arXiv:2002.07689*, 2020. 1

[21] S. Izadi, David Kim, Otmar Hilliges, D. Molyneaux, Richard A. Newcombe, Pushmeet Kohli, J. Shotton, Steve Hodges, Dustin Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. *ACM Symposium on User Interface Software and Technology*, 2011. 1, 3

[22] Jitesh Jain, Yuqian Zhou, Ning Yu, and Humphrey Shi. Keys to better image inpainting: Structure and texture go hand in hand. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 208–217, 2023. 3

[23] Jurandir de Oliveira Santos Junior, Alexandre Vrubel, Olga RP Bellon, and Luciano Silva. 3d reconstruction of cultural heritages: Challenges and advances on precise mesh integration. *Computer Vision and Image Understanding*, 116(12):1195–1207, 2012. 1

[24] Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. 01 2004. 2

[25] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, page 61–70, Goslar, DEU, 2006. Eurographics Association. 2

[26] Usman Khan, AmanUllah Yasin, Muhammad Abid, Imran Shafi, and Shoab A Khan. A methodological review of 3d reconstruction techniques in tomographic imaging. *Journal of Medical Systems*, 42(10):190, 2018. 1

[27] Abderrazzaq Kharroubi, Florent Poux, Zouhair Ballouch, Rafika Hajji, and Roland Billen. Three dimensional change detection using point clouds: A review. *Geomatics*, 2:457–486, 10 2022. 1

[28] Dongping Li, Tianjia Shao, Hongzhi Wu, and Kun Zhou. Shape completion from a single rgbd image. *IEEE transactions on visualization and computer graphics*, 23(7):1809–1822, 2016. 1

[29] W. Lorensen and H. Cline. Marching cubes: a high resolution 3d surface construction algorithm. 1998. 6

[30] Baorui Ma, Yu-Shen Liu, and Zhizhong Han. Reconstructing surfaces for sparse point clouds with on-surface priors. *Computer Vision and Pattern Recognition*, 2022. 1

[31] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 1, 3, 6, 7

[32] Richard A. Newcombe, S. Izadi, Otmar Hilliges, D. Molyneaux, David Kim, A. Davison, Pushmeet Kohli, J. Shotton, Steve Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, 2011. 1

[33] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics*, 2013. 1

[34] Yuchen Rao, Yinyu Nie, and Angela Dai. Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. *Advances in Neural Information Processing Systems*, 35:34436–34450, 2022. 1, 2, 3, 4, 5, 6, 7

[35] Pratyush Sahay and A. N. Rajagopalan. Geometric inpainting of 3d structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015. 1, 2, 5

[36] Simon Schreiberhuber, Johann Prankl, Timothy Patten, and Markus Vincze. Scalablefusion: High-resolution mesh-based real-time 3d reconstruction. In *2019 International conference on robotics and automation (ICRA)*, pages 140–146. IEEE, 2019. 1

[37] Weichao Shen, Yunde Jia, and Yuwei Wu. 3d shape reconstruction from images in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[38] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 3

[39] Shoichi Tsuchie and Masatake Higashi. Surface mesh segmentation and reconstruction with smooth boundary curves. 2014. 1

[40] Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 2442–2447. IEEE, 2017. 1

[41] Bram Wallace and Bharath Hariharan. Few-shot generalization for single-image 3d reconstruction via priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3818–3827, 2019. 1, 3, 6, 7

[42] 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. *arXiv preprint arXiv:2106.10689*, 2021. 5

[43] Xiaogang Wang, Marcelo H Ang, and Gim Hee Lee. Point cloud completion by learning shape priors. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10719–10726. IEEE, 2020. 2

[44] Xiaogang Wang, Marcelo H Ang, and Gim Hee Lee. Voxel-based network for shape completion by leveraging edge generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13189–13198, 2021. 1

[45] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 281–296. Springer, 2020. 1

[46] Yusheng Xu, Xiaohua Tong, and Uwe Stilla. Voxel-based representation of 3d point clouds: Methods, applications, and its potential use in the construction industry. *Automation in Construction*, 126:103675, 2021. 1

[47] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *CVPR*, 2021. 1, 2

[48] Xinyang Zheng, Yang Liu, Pengshuai Wang, and Xin Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum*, volume 41, pages 52–63. Wiley Online Library, 2022. 5

[49] Deyun Zhong, Ju Zhang, and Liguan Wang. Fast implicit surface reconstruction for the radial basis functions interpolant. *Applied Sciences*, 9(24), 2019. 1

[50] Daquan Zhou, Yujun Shi, Bingyi Kang, Weihao Yu, Zihang Jiang, Yuan Li, Xiaojie Jin, Qibin Hou, and Jiashi Feng. Refiner: Refining self-attention for vision transformers, 2021. 5