arXiv:2103.12716v1 [cs.CV] 23 Mar 2021

UltraSR: Spatial Encoding is a Missing Key for Implicit Image Function-based Arbitrary-Scale Super-Resolution

Xingqian Xu[1], Zhangyang Wang[2], Humphrey Shi[3,1,4]

[1]University of Illinois at Urbana-Champaign, [2]University of Texas at Austin, [3]University of Oregon, [4]Picsart AI Research (PAIR)

Abstract

The recent success of NeRF and other related implicit neural representation methods has opened a new path for continuous image representation, where pixel values no longer need to be looked up from stored discrete 2D arrays but can be inferred from neural network models on a continuous spatial domain. Although the recent work LIIF has demonstrated that such novel approach can achieve good performance on the arbitrary-scale super-resolution task, their upscaled images frequently show structural distortion due to the faulty prediction on high-frequency textures. In this work, we propose ***UltraSR****, a simple yet effective new network design based on implicit image functions in which spatial coordinates and periodic encoding are deeply integrated with the implicit neural representation. We show that spatial encoding is indeed a missing key towards the next-stage high-accuracy implicit image function through extensive experiments and ablation studies. Our UltraSR sets new state-of-the-art performance on the DIV2K benchmark under all super-resolution scales comparing to previous state-of-the-art methods. UltraSR also achieves superior performance on other standard benchmark datasets in which it outperforms prior works in almost all experiments. Our code will be released at https://github.com/SHI-Labs/UltraSR-Arbitrary-Scale-Super-Resolution.*

1. Introduction

Image data has long been stored and computed with discrete 2D arrays, which is a compromised solution between flexibility and complexity, to some extent. The popular convolutional layers have also strengthened such discretization in which convolutional kernel parameters are scattered on fixed spatial locations. Despite some efforts to break such constraints (*e.g.* RoI alignment [38], deformable convolution [10], etc.), a majority of research works in computer vision accepted and adapted such discretization without question. Yet, the great potential of utilizing the continuity of spatial domain must not be underestimated. The new state-of-the-art (SOTA) work LIIF [9] on arbitrary-scale super-resolution (SR) suggests that high-resolution images can be learned as a continuous function of low-resolution (LR) features and pixel coordinates, and a simple multilayer perceptron (MLP) can be an adequate implementation of such functions and achieve decent performance. Despite the initial success of LIIF, we find plenty of issues in the high-frequency domain of the high-resolution (HR) images generated by LIIF, especially structural distortions and noticeable artifacts. This observation motivates us to examine the current ideas and methodology for arbitrary-scale SR approaches more carefully.

Figure 1: The quality comparison between LIIF [9] and our UltraSR at resolution scale $\times 18$. As shown in the figure, our model can avoid structural distortions at extreme scales.

As suggested by recent neural rendering works [32, 37, 50, 44], using spatial encoding is critical for recovering high-frequency details in 3D scenes. It is natural to ask the following questions: Does spatial encoding also matter when we render 2D images in the arbitrary-scale SR task? What factors are critical to making implicit functions accurate, and what network adjustment shall we make to reflect these critical factors? To what levels can these implicit functions recover high-frequency patterns such as sharp edges and fine texture? To answer these questions, in this paper, we will introduce a new arbitrary-scale super-resolution method, namely UltraSR, in which the implicit image functions are learned by a more carefully-designed network to address the aforementioned issues. With the help of our periodic spatial encoding and deep coordinate

fusion, we are able to provide a very stable performance boost in arbitrary-scale SR, and our UltraSR surpasses the previous SOTA method LIIF on all resolution scales.

Meanwhile, arbitrary-scale SR is a raising research topic with tremendous application potentials. Prior CNN-based SR approaches usually apply to only one fixed resolution scale and cannot adjust their output dimension without changing the low-resolution input. Such design creates a huge gap between academic research and practical usage. Even sensitive to precision, most image up-sampling applications still heavily rely on bicubic interpolation despite its poor quality. Empowered by the rapidly advancing techniques in implicit neural representation, images and scenes can now be generalized by network-learned implicit functions on various vision topics (*e.g*. free-viewpoint scene reconstruction, 3D video generation, etc.). Specifically for our SR task, the idea that uses one trained network for all zoom-in scales on any input image will bring both convenience and accuracy to downstream users in the near future. In summary, the three main contributions of this article are as follows:

- We reveal the importance of spatial encoding for implicit functions on 2D images through analysis and experiments.
- We introduce a series of architecture designs such as deep coordinate fusion and residual MLP that work well with spatial encoding. Without these designs, the effectiveness of applying spatial encoding in implicit functions would be largely reduced.
- Our UltraSR sets the new SOTA on arbitrary-scale super-resolution. The performance of UltraSR surpasses LIIF on the DIV2K validation set under all resolution scales and beats prior arts on other 5 benchmark datasets under a majority of the testing schemes.

2. Related Work

This section will briefly introduce recent works on implicit neural representation using spatial encoding and various super-resolution methods related to our work.

2.1. Rendering with Spatial Encoding

Learning 3D scene representation with a parameterized neural network has been largely explored by recent works from various angles such as implicit signed distance function [21, 5, 34], occupancy [31, 35, 14], volume rendering (*i.e*. radiance field) [32, 26, 33, 37, 50, 44], and shapes [2, 15, 14]. Such implicit neural representation also started to influence traditional 2D tasks such as image representation [23, 40], super-resolution [9], and medical image analysis [49]. Among these works, spatial encoding has played a critical role in 3D scene reconstruction, whose effectiveness has not been fully explored in the 2D domain. In NeRF [32], Mildenhall *et al*. expanded the input 3D coordinates with periodic functions before fed them into NeRF. They showed clear improvements on rendered images which were free from blurry details and structural distortion. Such technique has been used as a default setting in later 3D works such as [26, 50, 39, 44], whose major focus were to reduce the rendering speed and improve output quality. The missing theory part on why spatial encoding tremendously boosts rendering quality has been partially analyzed in [36, 8, 40]. In [36], Rahaman *et al*. highlighted that there existed strong learning biased toward low-frequency spectral using neural nets. A similar conclusion has also be summarized by Chen [8] that generators tended to distort images in the high-frequency domain. Recently in SIERN [40], images and their gradients and laplacians were near-flawlessly reconstructed by replacing ReLU with periodic functions in neural networks.

2.2. Single Image Super-Resolution

Single image super-resolution (SISR) is one of the low-level vision tasks that has been studied for decades in our vision community. Traditional approaches can be roughly divided into patch-based [6, 13], edge-based [12, 41], and statistic-based [54] methods. The first CNN-based SISR work was SRCNN [11], in which three convolutional layers were used for patch extraction, feature mapping, and image reconstruction. Later, larger residual structures such as VDSR [22], IRCNN [51], and SRResNet [24] were proposed. Lim *et al*. [25] proposed EDSR, in which they improved the residual blocks by removing BN layers. Yu *et al*. [47] further enhanced EDSR into WDSR with an even wider channel before ReLU. Meanwhile, RDN [53] proposed residual dense block (RDB) with three convolutional layers densely connected to each other that boosted SR quality to the next level. Recently, the non-local attention module [43, 42] becomes rather popular. Zhang *et al*. [52] adopted the non-local concept in RCAN, proposed a high-performing SR model with residual channel attention. Mei *et al*. [30, 29] performed cross-scale attention to exhaustively search repeated patterns for super-resolution.

2.3. Arbitrary-Scale Super-Resolution

Arbitrary-scale super-resolution methods significantly surpass previous SISR works in terms of practice and convenience. The idea that uses one neural network for any super-resolution scale could be dated back to MDSR [25] introduced by Lim. To achieve SR in different scales, MDSR proposed the pre-processing modules and the scale-specific up-sampling modules and placed them in the network's front and back. Nevertheless, MDSR could only handle scales $\times 2$, $\times 3$, and $\times 4$, and it was not a true arbitrary-scale method. MetaSR [18] was the first CNN-

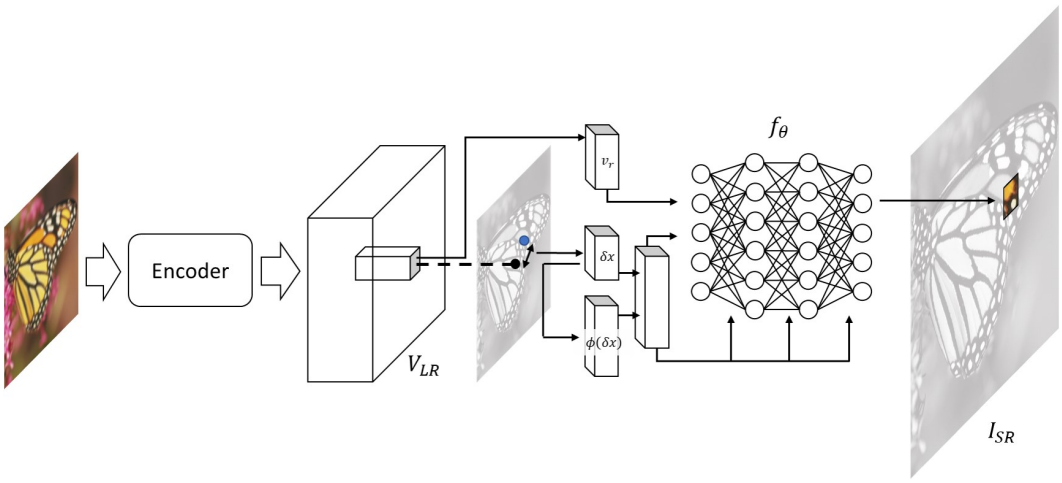


Figure 2: This figure shows the overall structure of our UltraSR. The blue point refers to the location of our target HR pixel that needs to be rendered. The black point shows the location of the nearby feature in the LR feature map V_{LR}. δx is the location difference and $\phi(\cdot)$ is the spatial encoding function. f_θ is our rendering network in which coordinates and encoding are deeply fused to each network's hidden layer.

based SR method on arbitrary scales. The proposed Meta-Upscale module mapped SR pixels onto the LR domain using the nearest-neighbor rule. During training, all mapped values were multiplied by dynamically learned weights based on scales and coordinates. The output images were then generated via the Meta-Upscale module with some additional convolutional layers. The recent SOTA work LIIF [9] proposed a novel framework in which RGB values were computed by multilayer perceptron using coordinates, cell size (*i.e.* scale factors), and LR features as its inputs. LIIF was also recognized for its robust performance under extreme SR scales up to $\times 30$. The new evaluation standard in LIIF that one network is tested on a large range of scales, will also bring us more reliable and universal SR methods for practical usage.

3. Methods

In this session, we introduce UltraSR, a new arbitrary-scale SR model capable of generating any scales of high-resolution images from low-resolution images using implicit neural representation. Our work is strongly motivated by the recent works NeRF [32] and LIIF [9]. The former demonstrated that combining neural rendering with spatial encoding can synthesize free-viewpoint 3D scenes with fine details. And the latter proved that one properly learned implicit image function could restore HR images with pleasing quality at arbitrary SR scales.

Like LIIF, we formulate the implicit form of any image in the HR domain with the following equation:

$$\underset{\text{target pixel}}{s} = f_\theta(\underset{\text{feature vector}}{v_r}, \underset{\text{normalized distance}}{\delta x}), \quad \delta x \propto x - x_r \tag{1}$$

where s is the target pixel value need to be reconstructed, v_r is the reference feature vector, and δx is the normalized distance between target pixel location x and the reference feature location x_r. The reference feature vector v_r is extracted from the LR feature map $V_{LR} \in R^{C \times H \times W}$ in which its spacial location x_r is near to x. f_θ is then the implicit image function simulated by a network with parameter θ.

3.1. Periodic Spatial Encoding

Learning how to reconstruct the high-frequency part of the image is the key in the SR task. Many recent works [32, 37, 50, 44] have shown that a carefully designed spatial encoding can help the network recover fine details in 3D scenes. LIIF [9] overlooked the importance of these spatial encoding by directly feeding coordinates into the implicit image function represented by a vanilla MLP. We empirically notice that, without spatial encoding, these neural representations tend to generate images with structural distortions and noticeable artifact. Moreover, these artifacts appear across different resolution scales (see Figure 3), which may create many troubles if applied in areas requiring accuracy, such as medical applications. Such phenomenon also aligns with the discovery that neural networks are biased towards low-frequency signals and are insensitive to high-frequency signals [36, 8]. Therefore, in order to minimize the structural distortion and to enhance the network in the high-frequency domain. Our UltraSR expands the 2D linear spatial input into the 48D encoding using the following equations:

$$\begin{aligned}\phi(x) = (&sin(w_1 x), cos(w_1 x), \\ &sin(w_2 x), cos(w_2 x), ...)\end{aligned} \tag{2}$$

$$s = f_\theta\left(v_r, \delta x, \phi(\delta x)\right) \tag{3}$$

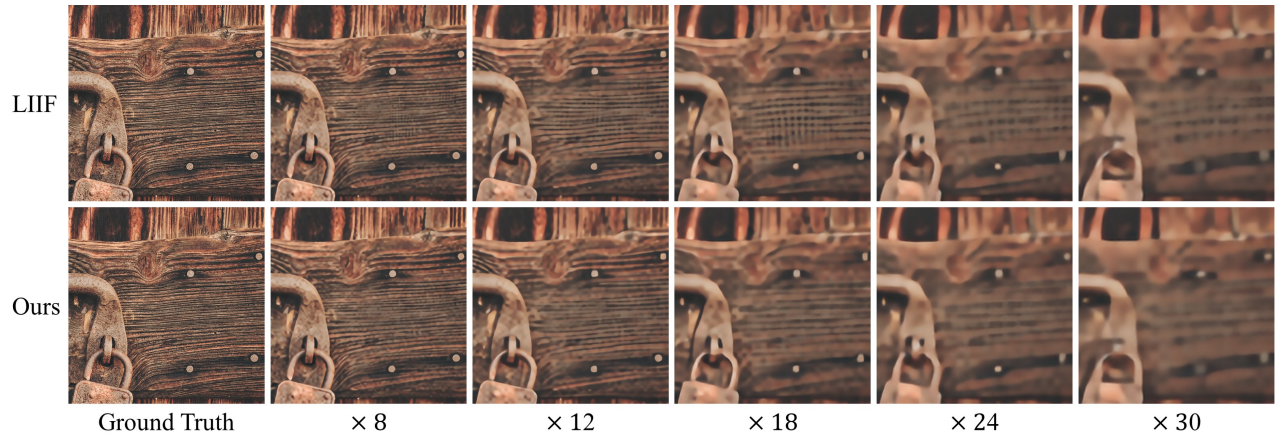


Figure 3: An example shows a group of persistent grid-like artifacts on all large resolution scales generated by LIIF. Such artifacts do not appear in our UltraSR. Please zoom in for details.

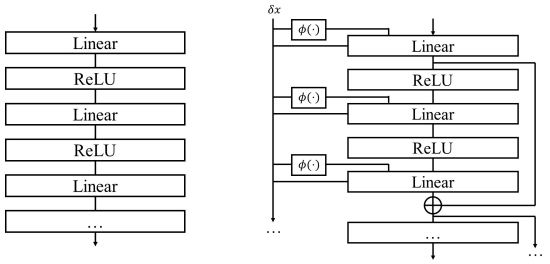
where the frequency parameters $w_1, w_2, ...$ are initially set to $2e^n, n \in 1, 2, ...$, and are later fine-tuned in training. We follow the convention by using cos and sin as our encoding functions. At last, the input coordinates δx together with their spatial encoding $\phi(\delta x)$ are feed into the implicit image function in UltraSR as shown in Equation 1.

3.2. Deep Coordinate Fusion

Nevertheless, a simple concatenation on coordinates and spatial encoding does not provide us a solution. In fact, we find no obvious evidence that the quality of the output images can be improved if we directly feed these spatial encoding into our implicit image function simulated by MLP. It is also misleading because recent neural rendering articles use vanilla MLP for various tasks, yet few have investigated what network structure is the optimal one. In fact, we have found that MLP is rather sub-optimal because it cannot prioritize input parameters based on their importance.

As shown in Equation 3, our implicit function's three inputs are feature vectors, coordinates, and spatial encoding. Among these inputs, we believe that the 2D coordinates and their encoding are more important than the feature vectors. This is an intuitive assumption, and the reason is the following. If UltraSR is asked to up-sample an image with some large-scale k, the input feature vector from the low-resolution domain does not change when rendering the nearby k^2 pixels in the high-resolution domain. Therefore, the detail textures in our output must highly depend on the coordinates and the spatial encoding. So we need to adjust our network to make these inputs integrated in a much tighter way. Figure 2 shows the overall structure of UltraSR, in which we concatenate the 2D coordinates with the 48D encoding and feed them to all hidden layers. Such fusion ensures that all hidden layers have direct access to the target pixel's critical spatial information and can utilize the high-frequency hints inside the spatial encoding off-the-shelf.

could not understand



(a) MLP (b) ResMLP

Figure 4: This figure shows the structure of the vanilla MLP used by LIIF [9] (left) and our modified MLP with coordinate fusion and residual links (right).

3.3. Network details

The overall network structure of UltraSR is shown in Figure 2. Besides the spatial encoding and the coordinate fusion, we also add residual links in MLP (ResMLP) to further strengthen its ability to generate the image's high-frequency part. The detailed structure of our residual MLP follows the convention in [17, 24, 4], in which hidden features with the same dimension are connected between every two layers before activation (see Figure 4). We have noticed that the residual link can help the network restore images with high fidelity. It frees the network from processing low-frequency information by passing them directly to a later layer. Besides, we also adopt the same feature unfold, local ensembling, and cell decoding strategies mentioned in [9]. We will show results and comparisons on our UltraSR with more details in the next session.

Method	×2	×3	×4	×6	×12	×18	×24	×30
Bicubic [25]	31.01	28.22	26.66	24.82	22.27	21.00	20.19	19.59
EDSR-baseline [25]	34.55	30.90	28.92	–	–	–	–	–
MetaSR-EDSR [18, 9]	34.64	30.93	28.92	26.61	23.55	22.03	21.06	20.37
LIIF-EDSR [9]	34.67	30.96	29.00	26.75	23.71	22.17	21.18	20.48
UltraSR-EDSR (ours)	**34.69**	**31.02**	**29.05**	**26.81**	**23.75**	**22.21**	**21.21**	**20.51**
MetaSR-RDN [18, 9]	**35.00**	31.27	29.25	26.88	23.73	22.18	21.17	20.47
LIIF-RDN [9]	34.99	31.26	29.27	26.99	23.89	22.34	21.31	20.59
UltraSR-RDN (ours)	**35.00**	**31.30**	**29.32**	**27.03**	**23.93**	**22.36**	**21.33**	**20.61**

Table 1: PSNR (dB) comparison between MetaSR [18], LIIF [9], and UltraSR (ours) on the DIV2K validation set with different SR scales. All three methods use one model for all scales. The bold numbers indicate the best results.

Dataset	Method	×2	×3	×4	×6	×8	×12
Set5 [3]	RDN [53]	**38.24**	**34.71**	32.47	–	–	–
	MetaSR-RDN [18, 9]	38.22	34.63	32.38	29.04	29.96	–
	LIIF-RDN [9]	38.17	34.68	**32.50**	29.15	27.14	**24.86**
	UltraSR-RDN (ours)	38.21	34.67	32.49	**29.33**	**27.24**	24.81
Set14 [48]	RDN [53]	**34.01**	30.57	28.81	–	–	–
	MetaSR-RDN [18, 9]	33.98	30.54	28.78	26.51	24.97	–
	LIIF-RDN [9]	33.97	30.53	28.80	26.64	25.15	23.24
	UltraSR-RDN (ours)	33.97	**30.59**	**28.86**	**26.69**	**25.25**	**23.32**
B100 [27]	RDN [53]	32.34	29.26	27.72	–	–	–
	MetaSR-RDN [18, 9]	32.33	29.26	27.71	25.90	24.83	–
	LIIF-RDN [9]	32.32	29.26	27.74	25.98	24.91	23.57
	UltraSR-RDN (ours)	**32.35**	**29.29**	**27.77**	**26.01**	**24.96**	**23.59**
Urban100 [19]	RDN [53]	32.89	28.80	26.61	–	–	–
	MetaSR-RDN [18, 9]	32.92	28.82	26.55	23.99	22.59	–
	LIIF-RDN [9]	32.87	28.82	26.68	24.20	22.79	21.15
	UltraSR-RDN (ours)	**32.97**	**28.92**	**26.78**	**24.30**	**22.87**	**21.20**
Manga109 [28]	RDN [53]	39.18	34.13	31.00	–	–	–
	MetaSR-RDN [18]	–	–	–	–	–	–
	LIIF-RDN [9]	**39.26**	34.21	31.20	27.33	25.04	22.36
	UltraSR-RDN (ours)	39.09	**34.28**	**31.32**	**27.42**	**25.12**	**22.42**

Table 2: PSNR (dB) comparison between RDN [53], LIIF [9], and UltraSR (ours) on 5 benchmark datasets. The performance of UltraSR surpasses RDN and LIIF in the majority of the table entries. Specifically, when the dataset is large (*e.g.* B100, Urban100, Manga109), our results surpass prior works on all scales that are larger than 2.

4. Experiment

In this section, we will go through the dataset, metric, and training details of our experiments. We will then discuss our model performance by comparing UltraSR with other methods. Lastly, we will further analyze UltraSR via several ablation studies.

4.1. Dataset and Metrics

The main dataset we use to train and evaluate our UltraSR is the DIV2K dataset [1] from NTIRE 2017 Challenge. DIV2K consists of 1000 2K high-resolution images together with the bicubic down-sampled low-resolution images under scale ×2, ×3 and ×4. We maintain its original train validation split, in which we use the 800 images from the train set in training and the 100 images from the validation set for testing. Follows many prior works [25, 24, 30, 53, 18, 9], we also report our model performance on 5 benchmark datasets: Set5 [3], Set14 [48], B100 [27], Urban100 [19] and Manga109 [28].

Meanwhile, we use the widely adopted Peak Signal-to-Noise Ratio (PSNR) as our evaluation metric on UltraSR. With pixel values ranging from 0 to 1, PSNR is computed

as 10 multiplies the log10 of one over the mean square error between two images. Please also see Equation 4 for PSNR computation in detail.

$$PSNR = 10 \log_{10} \left(\frac{1}{\frac{1}{H \times W} \|I_{SR} - I_{HR}\|_2^2} \right) \quad (4)$$

4.2. Training Details

To train UltraSR, we first create LR training images from ground truth HR images through bicubic interpolation. The down-sampling scales of these LR images are uniformly sampled from $\times 2$ to $\times 4$. We then randomly crop 48x48 patches from these LR images and feed them into UltraSR's encoder. After that, we randomly render 2304 pixels in the HR domain and back-propagate its l1 loss against the ground truth. For the optimizer, we choose ADAM with betas 0.9 and 0.999. We train the entire pipeline for 1000 epochs, and each epoch has 1000 iterations. Our initial learning rate is set to 10^{-4}, and it decays by one half at epoch 200, 400, 600, and 800. Except for some minor changes, the entire setting closely follows the convention in [9, 18]. Like LIIF [9], we also choose EDSR [25] and RDN [53] excluding up-sampling layers as two choices of our encoder in UltraSR. EDSR is a compact-sized model, so we train our UltraSR-EDSR on one RTX 2080 Ti GPU with batch-size 16. Meanwhile, RDN contains more layers, so we train UltraSR-RDN on two RTX 2080 Ti GPUs with batch-size 8 per GPU.

4.3. Results and Comparison

Table 1 compares the performances of MetaSR [18], LIIF [9] with our UltraSR under different resolution scales on DIV2K. We have noticed that our model's PSNR results are consistently higher than MetaSR and LIIF using both encoders and on all scales. We have also noticed that these increments are maximized around scale $\times 3$ to $\times 12$, reach a bold 0.05. We guess that the magnitude of improvement is more quickly saturated at extreme scales when too much or too little LR information is provided.

In Table 2, we also compare UltraSR with other methods on 5 standard benchmark datasets: Set5 [3], Set14 [48], B100 [27], Urban100 [19] and Manga109 [28]. Again, we have proved the supreme of our UltraSR by surpassing both RDN and LIIF in most of our experiments. Specifically, on large datasets (*e.g.* B100, Urban100, and Manga109) with scales larger than 2, our model surpasses RDN and LIIF in all experiments, even that RDN was trained one dedicated model for each scale.

Lastly, we show the qualitative comparisons between LIIF and UltraSR. As shown in Figure 3, despite it can interpolate images on extreme scales, LIIF is more likely to distort high-frequency patterns, downgrading its reliability on large scales. Our UltraSR, on the contrary, has been protected from these unwanted distortions by using spatial encoding and coordinate fusion. More examples of these comparisons can be found in Figure 6.

4.4. Ablation Studies

In this session, we performed a series of experiments to justify the spatial encoding effectiveness and our other designs.

In our first ablation study, we trained multiple models, among which we progressively added our new designs. For a fair comparison, we used EDSR as our encoder in all models. The training settings were kept unchanged, in which the same ADAM optimizer, learning rate, scheduler, and training length were used. The overall result is shown in Table 5, where we demonstrate that the simple spatial encoding does not work very well, but spatial encoding with coordinate fusion gives the magic. We also show that ResMLP is a simple but better design than MLP in terms of accuracy.

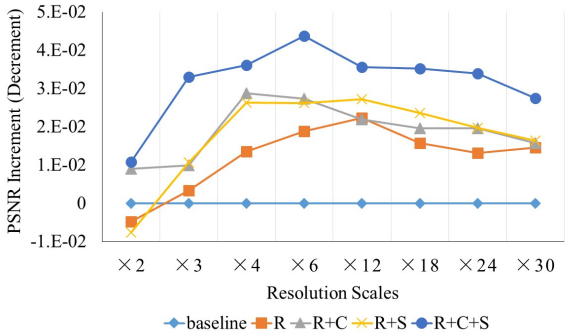


Figure 5: This plot shows PSNR increment/decrement on our model's performance against baseline under different settings. The baseline is plain UltraSR without spatial encoding(S), coordinate fusion(C), or residual links(R). R+C+S is the full UltraSR-EDSR shown in Table 1, and the rest are with parts of our design added to the baseline.

Our second study was to investigate the relationship between the dimension number of our spatial encoding and our model's performance. Like the first study, we maintained all other hyper-parameters besides the dimension we used in spatial encoding. We tested a total of three variations: 12, 24, 48. The final result is shown in Figure 7. We noticed that besides $\times 2$ and $\times 4$, our performance stably increased with the dimension number. The irregular behavior at $\times 2$ and $\times 4$ might due to the 2 to 4 random scale sampling in training, which made a high result at $\times 3$.

Our second study investigated the relationship between the dimension number of our spatial encoding and our model's performance. Like the first study, we maintained all other hyper-parameters besides the dimension we used in the spatial encoding. We tested a total of three variations:

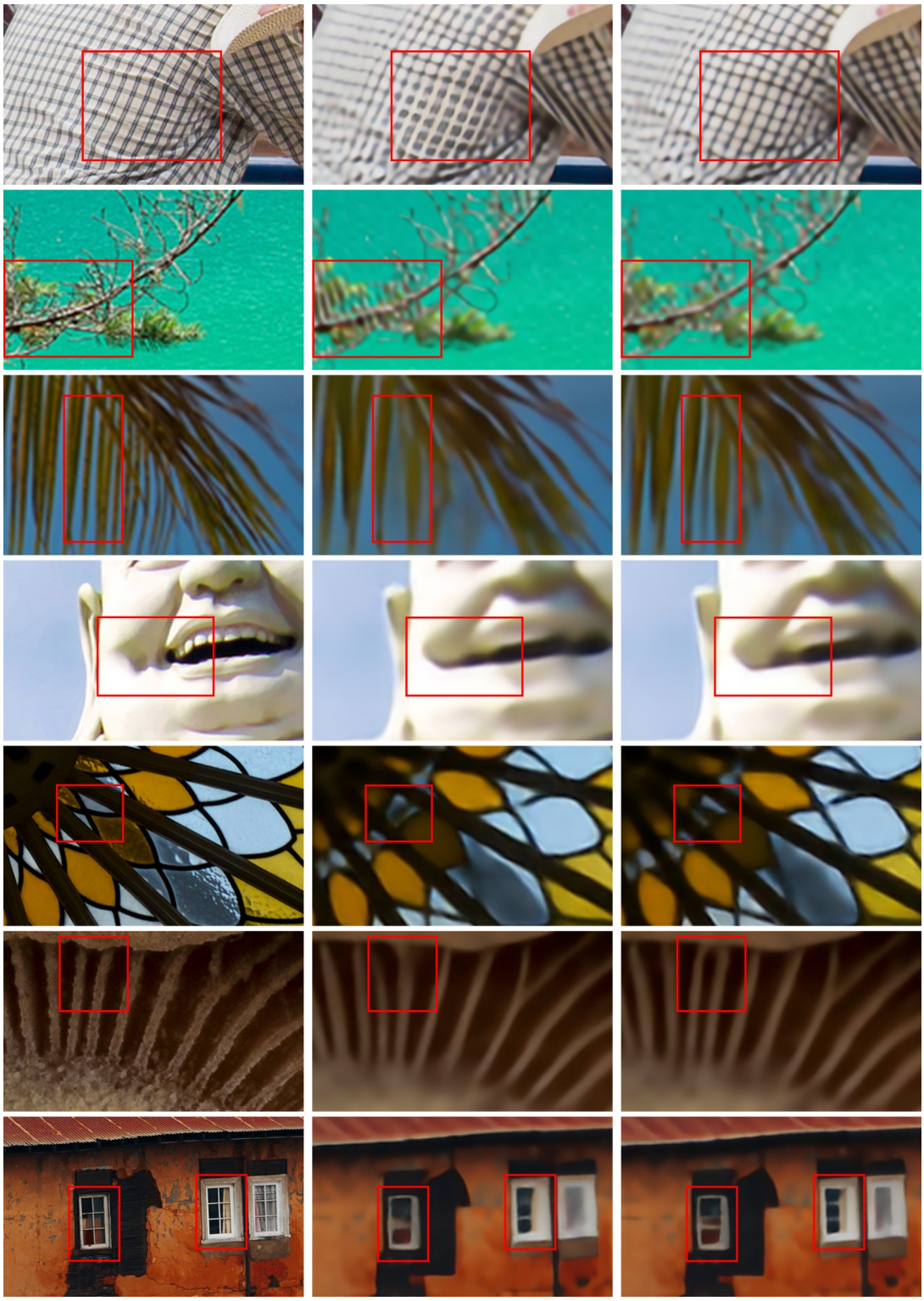


Figure 6: Additional SR sample images generated by LIIF [9] and by our UltraSR. The left column shows ground truth images; the middle column shows images from LIIF, and the right column shows images from UltraSR. As shown in this figure, our model can avoid structural distortions, reduce artifacts and generate sharp edges in various types of scenes. Please zoom in for details.

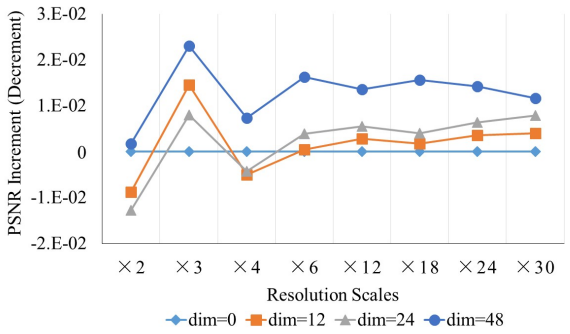


Figure 7: This plot shows the performance changes using different dimensions in our spatial encoding. The baseline is the model with dimension equals 0 (*i.e.* the R+C model in Figure 5). As shown, though there are some fluctuations in small scales, the growing encoding dimension can robustly increase the performance on large scales.

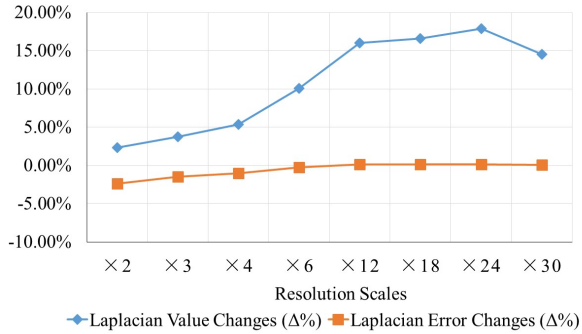


Figure 8: This plot shows the delta percentage changes of the mean absolute value/error of image Laplacian when use/no-use our spatial encoding. As shown, more sharp edges are generated when using spatial encoding because the mean absolute value increase by averagely 10% (blue line), while fewer errors are made because the mean absolute error against ground truth decreased by averagely 1% (orange lines). In conclusion, spatial encoding can help to create more clear and reliable SR images.

12, 24, 48. The final result is shown in Figure 7. We noticed that besides $\times 2$ and $\times 4$, our performance stably increased with the dimension number. The irregular behavior at $\times 2$ and $\times 4$ might be due to the 2 to 4 random scale sampling in training, which made a high result at $\times 3$.

Lastly, we proved that period spatial encoding could largely improve the sharpness of the generated images without losing fidelity. We analyzed the Laplacian of the images generated by two models, one without spatial encoding (UltraSR-S) and the other with spatial encoding (UltraSR+S). These models were the same R+C and R+C+S models in Figure 5. Laplacian filter is a well-known edge detector whose output represents the likelihood of an image patch contains an edge. Our study compared the delta percentage of UltraSR+S over UltraSR-S on two values across different scales: a) mean absolute value on Laplacian, and b) mean absolute error on Laplacian against ground truth. The former tells how willing the model generated sharp edges, and the latter tells how correct these edges were. In Figure 8, we showed that when comparing UltraSR+S with UltraSR-S, its mean absolute value on Laplacian is roughly 10% higher, the but error is roughly 1% lower. This proved that UltraSR with spatial encoding was capable of generating sharper images with fewer errors.

5. Discussion

In this section, we will discuss some potential future directions that can be extended from our work.

Encoding function space. The Fourier basis (*i.e.* sin and cos) is just one type of many basis functions that can be applied to spatial encoding. Our community lacks research on whether there is another basis in function space that is more suitable than the Fourier basis. For example, radial basis functions and wavelet basis functions could be two candidates in which location information can also be properly encoded. Meanwhile, task-specific basis functions should also be strongly promoted.

Tasks beyond SR. The idea of implicit image functions may also be extended to other 2D vision tasks. For example, we may extend the framework to discriminative tasks such as classification [17], detection [38], and segmentation [7, 45, 46, 20]. By far, we do not know much about such extension, but the generality on implicit functions suggests that we should indeed expand the idea beyond SR.

Perceptual-orientated SR. Since SR becomes a rather ill-posed problem on extreme resolution scales, one should think that whether we can train a network to simulate image functions in a perceptual-orientated way, from which we can create photo-realistic images under extreme scales. The idea that combines UltraSR or any implicit neural representations with GAN [16] is a very attractive and visible next-stage for neural rendering.

6. Conclusions

We introduce a novel arbitrary-scale super-resolution model UltraSR that deeply combines spatial encoding with implicit image function. We also reveal the importance of spatial encoding and coordinate fusion through result comparisons and visual evidence in which structural distortions can be effectively reduced. In conclusion, the PSNR performance of our UltraSR surpasses all prior arts on DIV2K dataset under all resolution scales. We also show our results on other 5 benchmark datasets from which the superiority of using spatial encoding can be once again demonstrated.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 5

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 2

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 5, 6

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 4

[5] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*, pages 608–625. Springer, 2020. 2

[6] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004. 2

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 2018. 8

[8] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas Li. Ssd-gan: Measuring the realness in the spatial and spectral domains. *arXiv preprint arXiv:2012.05535*, 2020. 2, 3

[9] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 4, 5, 6, 7

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015. 2

[12] Raanan Fattal. Image upsampling via imposed edge statistics. In *ACM SIGGRAPH 2007 papers*. 2007. 2

[13] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011. 2

[14] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[15] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 8

[17] 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. 4, 8

[18] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1575–1584, 2019. 2, 5, 6

[19] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 5, 6

[20] Zilong Huang, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S. Huang, and Humphrey Shi. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 8

[21] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020. 2

[22] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 2

[23] Sylwester Klocek, Łukasz Maziarka, Maciej Wołczyk, Jacek Tabor, Jakub Nowak, and Marek Śmieja. Hypernetwork functional image representation. In *International Conference on Artificial Neural Networks*, pages 496–510. Springer, 2019. 2

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2, 4, 5

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–144, 2017. 2, 5, 6

[26] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020. 2

[27] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of IEEE International Conference on Computer Vision*, 2001. 5, 6

[28] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa.

Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 5, 6

[29] Yiqun Mei, Yuchen Fan, Yulun Zhang, Jiahui Yu, Yuqian Zhou, Ding Liu, Yun Fu, Thomas S Huang, and Honghui Shi. Pyramid attention networks for image restoration. *arXiv preprint arXiv:2004.13824*, 2020. 2

[30] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 5

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[32] 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 *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 1, 2, 3

[33] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[35] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *arXiv preprint arXiv:2003.04618*, 2, 2020. 2

[36] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 2, 3

[37] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. DeRF: Decomposed radiance fields. *https://arxiv.org/abs/2011.12490*, 2020. 1, 2, 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 1, 8

[39] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *arXiv preprint arXiv:2007.02442*, 2020. 2

[40] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020. 2

[41] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[44] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. *arXiv preprint arXiv:2103.05606*, 2021. 1, 2, 3

[45] Xingqian Xu, Mang Tik Chiu, Thomas S Huang, and Honghui Shi. Deep affinity net: instance segmentation via affinity. *arXiv preprint arXiv:2003.06849*, 2020. 8

[46] Xingqian Xu, Zhifei Zhang, Zhaowen Wang, Brian Price, Zhonghao Wang, and Humphrey Shi. Rethinking text segmentation: A novel dataset and a text-specific refinement approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 8

[47] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718*, 2018. 2

[48] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 5, 6

[49] Hang Zhang, Rongguang Wang, Jinwei Zhang, Chao Li, Gufeng Yang, Pascal Spincemaille, Thanh Nguyen, and Yi Wang. Nerd: Neural representation of distribution for medical image segmentation. *arXiv preprint arXiv:2103.04020*, 2021. 2

[50] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1, 2, 3

[51] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3929–3938, 2017. 2

[52] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018. 2

[53] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 2, 5, 6

[54] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 977–984. IEEE, 2011. 2