NL-FFC: Non-Local Fast Fourier Convolution for Image Super Resolution

Abhishek Kumar Sinha, S. Manthira Moorthi, Debajyoti Dhar
Signal and Image Processing Group
Space Applications Centre, Ahmedabad
{aks,smmoorthi,deb}@sac.isro.gov.in

Abstract

Deep neural networks have shown promising results in image super-resolution by learning a complex mapping from low resolution to high resolution image. However, most of the approaches learns to upsample by using convolution in spatial domain and are confined to local features. This results into restricting the receptive field of the network and therefore deteriorates the overall quality of the high-resolution image. To alleviate this issue, we propose an architecture that learns both local and global features, and fuses them together to generate high quality images. The network uses a non-local attention aided Fast Fourier Convolutions (NL-FFC) to widen the receptive field and learn long-range dependencies. The analyses further show that these Fourier features implicitly provide faster convergence on low frequency components only to learn prior for unobserved high frequency components. The model generalizes well to different datasets. We further investigate the role of non-local attention, and the ratio of local and global features to maximize the performance gain in the ablation study.

1. Introduction

Deep learning has been a major player in the domain of computer vision and image processing for solving many real world problems, including classification [9, 18], object detection [6, 23], in-painting [31], and so on. The image super-resolution is a crucial task in image processing. The recent growth in image super resolution (SR) can significantly improve the digital media content for quality experience. While the conventional interpolation algorithms, such as nearest neighbour, bilinear upsampling, bicubic upsampling, are capable of solving the challenge to some extent, deep learning approaches outperform them by huge margin.

Image super-resolution is an ill posed problem since the Low Resolution (LR) image can be mapped to many different High Resolution (HR) images. Various architectures and improved training strategies have been proposed to continuously improve the SR images. Initial pioneer works in deep image-super resolution are based on the application of convolutional neural networks such as SRCNN [5] and LapSRN [12]. Subsequently, Generative Adversarial Networks (GAN) inspired architectures, including SRGAN [13] and ESRGAN [29], replaced the conventional CNN to produce the realistic textures in the high resolution images. Other way to improve the performance is to use very deep architecture to increase the capacity for high more complex nonlinear mapping. Lim *et al*. [17] used the residual blocks in EDSR to form a very wide model and a very deep model (MDSR). DBPN [8] uses a deep projection network that exploits the iterative sampling mechanism providing an error feedback mechanism.

Although deep Single Image Super-Resolution (SISR) methods have been contributing to prosperous developments, they still ignore the global and long-range features to improve the quality of the images. The convolution layers locally derive the features to generate high frequency components that may not suitably fit in terms of global perspectives. As a result, these methods perform well in terms of Peak Signal-to-Noise Ratio (PSNR) but miserably fail to satisfy the human perceptual quality. To explore beyond the localised vision, non-local mean filtering based method [22, 32] globally searches for the similar patches in the LR image. Other methods include non-local attention based approaches. Mei *et al*. [20] proposed the cross scale non local attention module to exploit pixel-to-patch and patch-to-patch based image similarity and a Self Exemplar Mining cell to fuse all the information recurrently. Subsequently, Mei *et al*. [19] suggests to enforce sparsity on the non-local attention module by adopting Locality Sensitive Hashing (LSH) for grouping and assign each group a Hash code.

Another way to widen the receptive field or capture the global features is to process the image features in frequency domain. Spectral networks gained lots of attention to process the features in frequency domain. For example, Rippel *et al*. [24] proposed spectral pooling to perform dimensionality reduction in frequency domain and Zhong *et al*. [36]

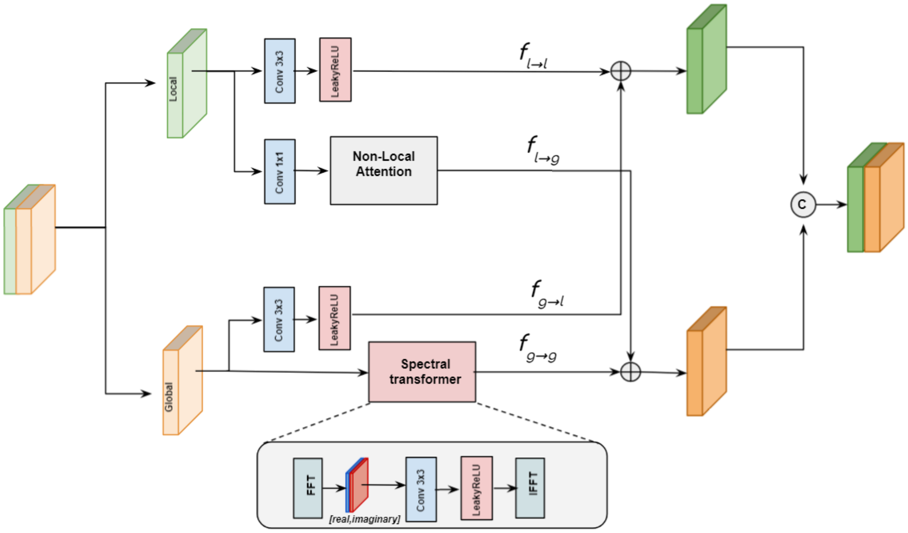


Figure 1. Illustration of Non Local-Fast Fourier Convolution (NL-FFC) layer. The image also shows the exploded view of spectral transformer in $f_{g \longrightarrow g}$ branch.

utilizes wavelet transform for restoration of high resolution images. The Fourier Transform is known to project the features of same frequency along the same basis vector. Fast Fourier Convolutions [3] (FFC) used the concept of FFT based feature learning for global to global feature mapping. Although this approach shows promising results, a major drawback can be derived from its local to global feature projection. The theory of effective receptive field says that convolutions tend to shrink to the central regions. FFC simply convolves the local features to learn the corresponding global mapping ignoring the limited receptive field of the convolution layer.

Following the success of FFC, we seek to further improve the local to global mapping for Single Image Super Resolution task. Specifically, we incorporate the idea of non-local attention networks [2] to learn the long-range dependencies of the local features. After that, we integrate the NL-FFC blocks in the network. The major contributions of this work are summarized as follows:

1. We propose a SISR network based on the idea of Non-local Fast Fourier Convolutions (NL-FFC). We further present mathematical arguments to show that learning in frequency domain converges faster for lower frequency components, which then implicitly acts as prior for unobserved high frequency components.

2. To overcome the limitation of local to global feature mapping in FFC, we incorporate the non-local attention module to learn the long-range dependencies of query pixels in the local feature and investigate the performance gain in the network with improvement in local to global mapping.

3. The model is trained separately for evaluation in terms of both distortion and perceptual quality. The distortion based model is trained using reconstruction loss whereas the perceptual quality based model is trained using VGG based perceptual loss and adversarial loss.

4. The proposed method outperforms on multiple benchmark datasets. We further ablate the model to explore the effectiveness and contribution of different modules.

2. Related Works

2.1. Non-local attention

Non-local networks have set a milestone in capturing long-range dependency. Initial prior work is reported by Wang *et al*. [28] in non-local neural networks that performs non-local operation by computing the position specific response as a weighted sum of features at all features. This work further proposed different functional embeddings for the position and the feature to perform non-local operation. Zhang *et al*. [34] studied residual non-local attention model for image restoration that uses local and non-local mask branches to extract local and non-local features respectively by adaptively rescaling the hierarchical features with mixed attention. Cross scale non-local attention [20] combines CS-NL prior with local and in-scale non-local priors in a powerful recurrent fusion cell to find more cross-scale feature correlations within a single low-resolution (LR) image.

2.2. Deep super resolution

Application of Convolutional Neural Networks (CNNs) in super-resolution is now a well-established approach. Af-

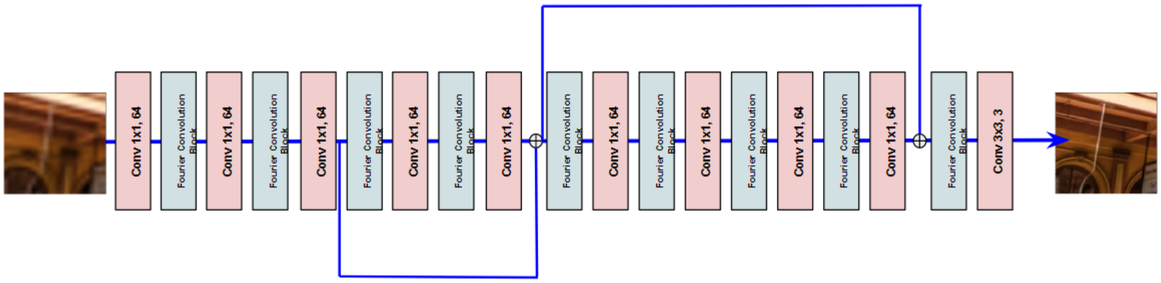

Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 1x1, 64
Fourier Convolution Block
Conv 3x3, 3

Figure 2. Single image super resolution architecture. Two skip connections are also used in the network to avoid vanishing gradient problem.

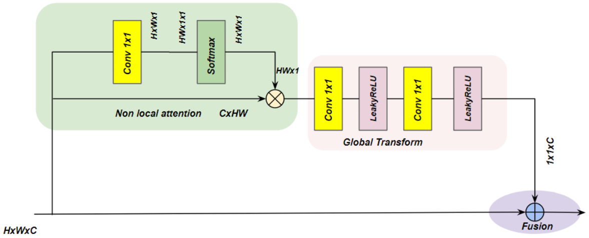


Figure 3. Non-local attention network

ter Dong *et al.* [5], several works have advanced the accuracy and shown promising results. Following [5], Kim *et al.* [10] increased the number of layers with small filters and high learning rate to improve the accuracy. In the following work, Kim *et al.* [11] used a deep recursive network with skip connections. Lim *et al.* [17] proposed a carefully designed ResNet by removing unnecessary modules and expanding the model size. Tai *et al.* [26] presented a very deep memory persistent network (MemNet) that introduced a memory block to mine the persistent memory through adaptive learning process. Several works [4,8,35] continued to marginally improve the image quality. While all of these approaches may appear to move forward in terms of PSNR, they could not further improve the visual quality beyond a certain threshold.

The adversarial learning in the super resolution paved a new way to learn synthetic and realistic features for HR image, and therefore significantly improves the visual quality. The works of SRGAN [13] and ESRGAN [29] made pioneering attempts to apply GAN for super resolution. Besides, ESRGAN employed the idea of relativistic GAN that not only increases the probability that fake data is real but also decreases the probability that real data is real. This helps the discriminator to realise the probability that how real images are relatively realistic than fake images. Although GANs help to generate visually appealing results, their performance drop in terms of distortion. In our work, we study the performance of the model in both regimes by training the network with and without adversarial loss.

2.3. Fourier features

Recently, Fourier features have gained a lot of attention to improve the performance of deep networks and its application has been reported in various domains of image processing. Tanick *et al.* [27] showed that the Fourier features are capable of learning high frequency features in lower dimensions. Using Neural Tangent Kernel, they showed that the Fourier feature mapping turns the NTK into a stationary kernel with tunable bandwidth. A recent work by Suvorov *et al.* [25] applied the idea of FFC in image inpainting for larger missing areas. FFC widens the receptive field, and therefore draws inference for very large patches from the neighbouring pixels. FALCON [15] showed the utility of Fourier Transform to perform secured CNN based predictions. The CNN is secured using homomorphic encryption and uses FFT based cipher text encryption for efficiency.

3. Methodology

In this section, we describe the Non local Fast Fourier Convolution layer and the non-local attention model that we use to perform image super resolution. The overall architecture of the network is also presented in this section.

Figures and 1 and 2 show the NL-FFC and the network architecture to perform image super resolution respectively. The model takes the bicubic upsampled image and generates the corresponding HR image. It primarily consists of alternating convolution and NL-FFC blocks. As the network grows deeper, skip connections are used to mitigate the vanishing gradient problems. The super resolution network acts a generator in GAN based training setup. The discriminator used for adversarial regularization is same as SRGAN's discriminator [13].

3.1. Non-Local Fast Fourier Convolution

The Figure 1 presents the architecture of NL-FFC. The input features are split into the set of local and global features to learn local-to-local ($f_{l \to l}$), local-to-global ($f_{l \to g}$), global-to-local ($f_{g \to l}$), and global-to-global ($f_{g \to g}$) mappings. Since the local features do not require long-range dependency, they are directly mapped to new local features using a vanilla convolution layer. For local-to-global mapping, it explores the global dependency for each query pixel using non-local attention model [2]. The input global features are mapped to the local features by applying simple 3×3 convolution operation since a convolution kernel estimates the local feature of each pixel by using neighbour pixels in 3×3 window. The global-to-global mapping requires that the features are first transformed to frequency domain to widen the receptive field, and then transformation is applied to generate new global features. The new learnt features are mapped back to spatial domain to generate new set of global features. A major advantage of using FFT comes in the form of prior for unobserved high frequency components. Since super resolution is an one to many mapping, learning features in frequency domain converges faster for low frequency components, which later act as the prior information to learn the unobserved features for high frequency, and therefore reduces the size of possible mappings. The relevant analytical details are provided in the next section.

For update procedure formulation, the input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ is divided into local and global features, $\mathbf{X}^l$ and $\mathbf{X}^g$, in the ratio of $(1-\alpha)/\alpha$. The local and global transformations are computed as,

$$f_{l\rightarrow l}(\mathbf{X}^l) = Conv_{3\times 3}^{l\rightarrow l}(\mathbf{X}^l) \tag{1}$$

$$f_{l\rightarrow g}(\mathbf{X}^l) = \mathcal{G}(Conv_{1\times 1}^{l\rightarrow g}(\mathbf{X}^l)) \tag{2}$$

$$f_{g\rightarrow l}(\mathbf{X}^g) = Conv_{3\times 3}^{g\rightarrow l}(\mathbf{X}^g) \tag{3}$$

$$f_{g\rightarrow g}(\mathbf{X}^g) = \mathcal{T}(\mathbf{X}^g) \tag{4}$$

$$\mathbf{Y}^l = f_{l\rightarrow l}(\mathbf{X}^l) + f_{g\rightarrow l}(\mathbf{X}^g) \tag{5}$$

$$\mathbf{Y}^g = f_{l\rightarrow g}(\mathbf{X}^l) + f_{g\rightarrow g}(\mathbf{X}^g) \tag{6}$$

Here, $\mathcal{G}$ and $\mathcal{T}$ denote non-local attention model and spectral transformation, respectively.

3.2. Non-Local Attention Model

Non-local attention model involves residual transform learning approach accompanied by a context mechanism as shown in Figure 3. The non-local attention module generates a spatial map for different query positions by aggregating the features from different positions. The attention coefficients are multiplied to the input features and passed to a transformation module to learn the residual features for each spatial location. Figure 4 compares the non local attention maps generated for different query positions. Though all these maps look very similar, they are slightly different due to difference in the global context of query pixels.

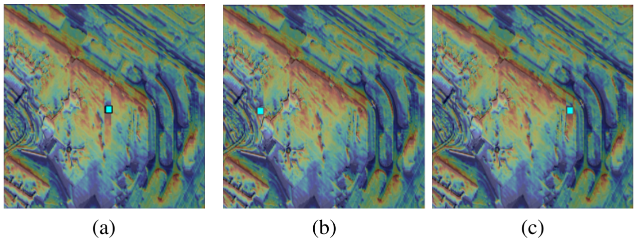

(a) (b) (c)

Figure 4. The figures show the non-local attention map generated for different query pixels (shown in blue). Based on the pixel dependencies, there are slight variations in all the maps.

3.3. Loss Function

There are three loss functions in the proposed approach. The overall loss is computed by,

$$\mathcal{L}_{overall} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{percep}\mathcal{L}_{percep} \tag{7}$$

Reconstruction loss: The reconstruction error is computed using Mean Absolute Error (MAE) since it has shown sharper performance and converges faster compared to the blurry results of Mean Squared Error. The reconstruction loss is given by,

$$\mathcal{L}_{rec} = \frac{1}{HWC}||I_{HR} - I_{SR}||_1, \tag{8}$$

where (H,W,C) are the height, width and the number of channels in the image.

Perceptual loss: Reconstruction loss does not account for the loss in perceptual quality of the reconstructed image. For this purpose, the perceptual loss is described using feature loss from pre-trained VGG-19 network. The perceptual loss is computed by,

$$\mathcal{L}_{percep} = \sum_i \frac{1}{H_i W_i}||\phi_i(I_{SR}) - \phi_i(I_{HR})||_2^2, \tag{9}$$

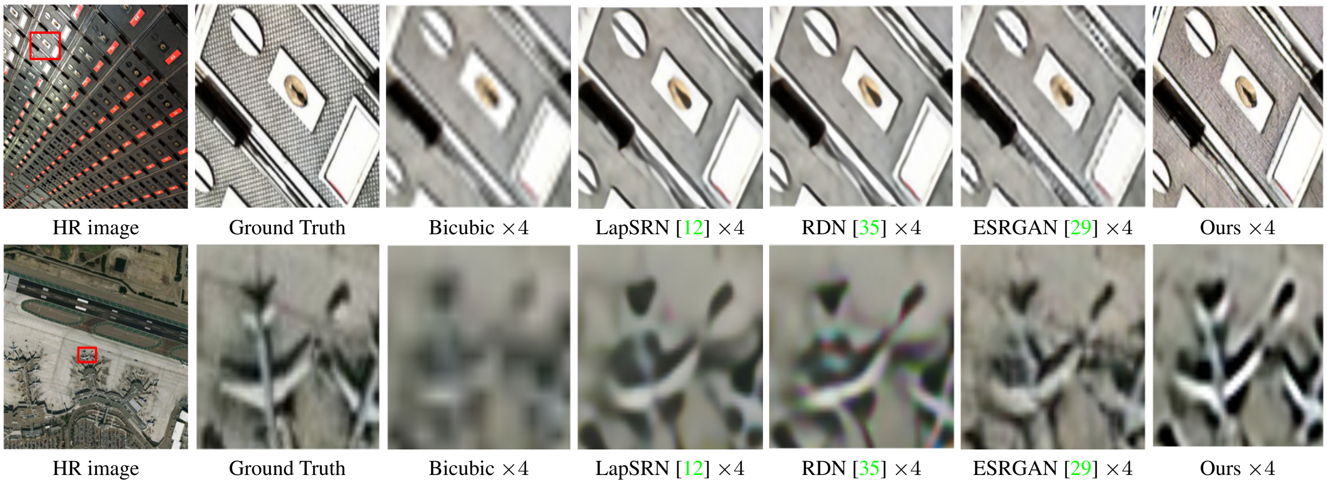
where ϕ_i indicates the i^{th} feature layer of VGG-19 network.

Adversarial loss: Generative adversarial nets [7] have already shown superior performance in terms of visually clear and perceptually enhanced image. Therefore, we employ adversarial loss to perform min-max optimization and is given by,

$$\mathcal{L}_G = -\mathbb{E}_{\hat{x}\sim X_{SR}}[log(D(\hat{x}))] \tag{10}$$

$$\mathcal{L}_D = -\mathbb{E}_{\hat{\hat{x}}\sim X_{SR}}[log(1 - D(\hat{x})] - \mathbb{E}_{x\sim X_{HR}}[log(D(x))] \tag{11}$$

The performance is evaluated in two tracks, one in terms of PSNR, and other using perceptual quality.Blau *et al.* [1] showed that the improvement in perceptual quality induces



HR image | Ground Truth | Bicubic ×4 | LapSRN [12] ×4 | RDN [35] ×4 | ESRGAN [29] ×4 | Ours ×4

HR image | Ground Truth | Bicubic ×4 | LapSRN [12] ×4 | RDN [35] ×4 | ESRGAN [29] ×4 | Ours ×4

Figure 5. Qualitative comparison of super resolved images for different models.

the distortion in the image. For distortion based evaluation (PSNR), we set λ_{rec}, λ_{percep} and λ_{adv} as 1, 0 and 0.0001 respectively. A very small weight is assigned to the adversarial loss so that it does not add significant distortion, and enhances the sharpness of image. For perceptual quality assessment, λ_{rec}, λ_{percep} and λ_{adv} are set to 0.007, 1, and 0.01 respectively.

4. Indirect Prior for High Frequency features

Based on the idea of Neural Tangent Kernel (NTK) and approach used in Lee *et al.* [14], we attempt to show that the features learned in Fourier domain helps to learn the prior for unobserved high frequency components of HR image. While the neural networks are known to perform worse in NTK regime, it is worth studying the training and convergence of over-parameterized system using simplified NTK based linear model. For simplicity, we assume a $L+1$-layered network $F(.)$ such that,

$$F(\mathbf{x}) = h_{L+1} = f(\mathbf{A}^{-1}g(\mathbf{A}h_L(\mathbf{x}))), \tag{12}$$

where $\mathbf{A}$ and $\mathbf{A}^{-1}$ are DFT and inverse DFT matrices respectively, and $f(.)$ and $g(.)$ are leaky ReLU layers and are mathematically expressed as,

$$f(\mathbf{x}) = LReLU(\mathbf{x}) = \Gamma_1^T\mathbf{x} \tag{13}$$

$$g(\mathbf{x}) = LReLU(\mathbf{x}) = \Gamma_2^T\mathbf{x}, \tag{14}$$

where Γ_{1i} and Γ_{2i} are equal to 1 for $\mathbf{x}_i \geq 0$, else $0 < \Gamma_{1i}, \Gamma_{2i} \leq 1$ for $\mathbf{x}_i \leq 0$. Equation 12 can be rewritten as,

$$F = h_{L+1} = \Gamma_1^T\mathbf{W}_f\mathbf{A}^{-1}\Gamma_2^T\mathbf{W}_g\mathbf{A}h_L \tag{15}$$

Further derivation closely follows Lee *et al.* [14] and proceeds by replacing h_L with its linearization around the initial parameters θ_0:

$$h_L^t = h_L^0 + \nabla_\theta h_L^0|_{\theta=\theta_0}(\theta_t - \theta_0), \tag{16}$$

where t denotes time in continuous-time gradient flow dynamics. Following [14], the gradient flow is given by,

$$\dot{h}_L^t = -\eta\Theta_0(\mathbf{x},\mathbf{X})\nabla_{h_L(\mathbf{x})}\mathcal{L}(h_L), \tag{17}$$

where $\Theta_t(.,.) = \nabla h_L(.)\nabla h_L(.)^T$ is the NTK matrix at time t. For convenience, we rewrite the loss function in terms of h_L instead of h_{L+1} as $\mathcal{L} = \frac{1}{2}||h_{L+1} - \mathbf{y}||_2^2 = \frac{1}{2}||\Gamma_1^T\mathbf{W}_f\mathbf{A}^{-1}\Gamma_2^T\mathbf{W}_g\mathbf{A}h_L - \mathbf{y}||_2^2$. The gradient of the loss is further given by,

$$\nabla_{h_L}\mathcal{L} = \mathbf{\Omega}^T(\mathbf{\Omega}h_L - \mathbf{y}), \tag{18}$$

where $\mathbf{\Omega} = \Gamma_1^T\mathbf{W}_f\mathbf{A}^{-1}\Gamma_2^T\mathbf{W}_g\mathbf{A}$. Substituting and solving the equation 17, we get,

$$h_L^t = \mathbf{\Omega}^{-1}(\mathbf{I} - e^{-\eta\Theta_0\mathbf{\Omega}^\mathbf{T}\mathbf{\Omega}t})\mathbf{y} + h_L^0 e^{-\eta\Theta_0\mathbf{\Omega}^\mathbf{T}\mathbf{\Omega}t} \tag{19}$$

Substituting equation 19 in equation 15, we get,

$$F^t = h_{L+1}^t = (\mathbf{I} - e^{-\eta\Theta_0\mathbf{\Omega}^\mathbf{T}\mathbf{\Omega}t})\mathbf{y} + \mathbf{\Omega}^\mathbf{T}h_L^0 e^{-\eta\Theta_0\mathbf{\Omega}^\mathbf{T}\mathbf{\Omega}t} \tag{20}$$

Following again [14], decomposing F^t and substituting the initialization $h_L^0 \approx 0$, we finally get,

$$F^t = h_{L+1}^t = \Theta_0(\mathbf{x},\mathbf{X})\Theta_0^{-1}(\mathbf{X},\mathbf{X})(\mathbf{I} - e^{-\eta\mathbf{K}\mathbf{\Omega}^\mathbf{T}\mathbf{\Omega}t})\mathbf{y} \tag{21}$$

Here, $\mathbf{K}$ is the kernel to estimate Θ_0. The weight matrices $\mathbf{W}_f$ and $\mathbf{W}_g$ in $\mathbf{\Omega}$ represent the neurons in the neural network. If Ω is unitary, the training is alone governed by $\mathbf{K}$ that is equivalent to learning only direct measurements. For an infinite width network, these weights are never full rank due to redundant neurons. These weights become non-full rank matrices with finite probability in the finite width over-parameterized neural network. This potentially leads to the finite probability that $\mathbf{\Omega}$ becomes non-full rank matrix. However, if $\mathbf{\Omega}$ is not a full rank matrix, the training only affects the features with non-zero eigenvalues in

Method	Scale	Set-5		Urban100		Set-14		B100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LapSRN [12]	×2	37.50	0.9590	30.41	0.9101	33.08	0.9130	31.08	0.8950
MemNet [26]	×2	37.78	0.9542	31.31	0.9115	33.28	0.9142	32.08	0.8978
EDSR [17]	×2	38.11	0.9584	32.93	0.9353	33.92	0.9195	32.32	0.9013
DBPN [8]	×2	38.04	0.9610	32.56	0.9342	33.85	0.9190	32.27	0.9000
RDN [35]	×2	38.21	0.9612	32.87	0.9355	34.01	0.9212	32.34	0.9017
SAN [4]	×2	38.31	0.9621	33.10	0.9360	34.07	0.9213	32.42	**0.9028**
CSNLN [20]	×2	38.28	0.9616	33.25	0.9386	**34.12**	0.9223	32.40	0.9024
SwinIR [16]	×2	38.14	0.9611	32.76	0.9340	33.86	0.9206	32.31	0.9012
NLSN [19]	×2	**38.34**	0.9618	**33.42**	**0.9394**	34.08	**0.9231**	32.43	0.9027
Ours	×2	38.21	**0.9622**	33.21	0.9366	34.09	0.9218	**32.44**	0.9019
LapSRN [12]	×3	33.82	0.9227	27.07	0.8280	29.87	0.8320	28.82	0.7980
MemNet [26]	×3	32.09	0.9249	27.54	0.8375	30.00	0.8350	28.96	0.8001
EDSR [17]	×3	34.58	0.9282	28.84	0.8641	30.52	0.8462	29.25	0.8093
RDN [35]	×3	34.71	0.9296	28.80	0.8655	30.57	0.8468	29.26	0.8093
SAN [4]	×3	34.75	0.9300	28.93	0.8671	30.59	0.8476	29.33	0.8112
CSNLN [20]	×3	34.74	0.9300	29.13	0.8712	30.66	0.8482	29.33	0.8105
SwinIR [16]	×3	34.62	0.9289	28.66	0.8624	30.54	0.8463	29.20	0.8082
NLSN [19]	×3	34.85	0.9306	**29.25**	**0.8726**	**30.70**	**0.8485**	**29.34**	**0.8117**
Ours	×3	**34.86**	**0.9356**	29.11	0.8655	30.63	0.8478	**29.34**	0.8116
LapSRN [12]	×4	31.41	0.8839	25.25	0.7549	28.19	0.7720	27.32	0.7270
MemNet [26]	×4	31.62	0.8887	25.11	0.7618	28.26	0.7723	27.40	0.7281
EDSR [17]	×4	32.22	0.8892	26.54	0.7994	28.80	0.7876	27.71	0.7420
DBPN [8]	×4	32.41	0.8975	26.37	0.7942	28.82	0.7860	27.72	0.7400
RDN [35]	×4	32.48	0.8987	26.66	0.8032	28.81	0.7871	27.72	0.7419
SAN [4]	×4	32.64	0.9003	26.79	0.8068	28.95	0.7888	**27.80**	0.7436
CSNLN [20]	×4	32.68	0.9004	**27.22**	**0.8168**	28.95	0.7888	**27.80**	0.7439
SwinIR [16]	×4	32.44	0.8976	26.47	0.7980	28.77	0.7858	27.69	0.7406
NLSN [19]	×4	32.59	0.9000	26.69	0.8109	28.87	**0.7891**	27.78	**0.7444**
Ours	×4	**32.76**	**0.9018**	27.04	0.8081	**28.96**	0.7888	**27.80**	0.7438

Table 1. Performance comparison of deep super resolution models on Set-5, Set-14 Urban-100 and BSD-100 datasets. The results are presented in terms of Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [30] for ×2, ×3, and ×4 scales. The best ones are shown in **bold** and second best ones are underlined.

$\mathbf{K\Omega^T\Omega}$. In that case, the composed kernel provides large eigenvalues to the features that $\mathbf{\Omega}$ can represent (these representative features are low frequency components of HR image), so that the network converges quickly on these observable low frequency features and acts as a prior for the unobserved high frequency components.

5. Experiments

5.1. Datasets and Study area

We train the model using DIV-2K dataset. The trained network is then evaluated using Set-5, Set-14, Urban-100, and BSD-100 datasets. In this study, the resolution between LR image and HR image is set to ×2, ×3, and ×4. Following Dong *et al.* [5], the corresponding LR images are obtained by downsampling the HR image using bicubic kernel. However, we upsample the LR image to the original size before feeding into the model.

5.2. Training and Evaluation strategy

The results are evaluated using Peak Signal-to-Noise Ratio and SSIM using the model trained with reconstruction loss oriented loss function. For fair comparison, the PSNR and SSIM are evaluated on Y-channel of the images. Furthermore, adversarially trained model is used to compare the performance in terms of perceptual quality. Regarding training details, the cropped patches of size 256×256 are fed into to the network in the batch size of 16. The Adam optimizer is used to update the parameters with $\beta_1 = 0.5$ and $\beta_2 = 0.5$. The learning rate is initially set to 0.001 and gradually reduced to 10^{-6} as the training reaches sub-optimality. All the experiments are performed using PyTorch library and trained using Nvidia Quadro P4000 GPU.

5.3. Results

	DBPN	EDSR	RDN	SAN	CSNLN	Ours
Param.	10M	43M	22.3M	16M	3M	**0.423M**
FLOPS (G)	5209.4	1338.8	801.1	3835.9	2245.2	**423.2**
PSNR	38.09	38.11	38.24	**38.31**	38.28	38.21

Table 2. Model size and performance comparison on Set-5 (x2).

		SR	ESR	RankSR	NLFFC(D)	NLFFC(P)
Set-14	NIQE	3.82	**3.28**	**3.28**	4.65	**3.28**
	PSNR	26.68	28.99	26.57	**28.96**	26.61
BSD100	NIQE	6.43	3.29	3.21	5.21	**3.16**
	PSNR	25.67	25.85	25.57	**27.80**	25.26

Table 3. Perceptual quality comparison for ×4 super-resolution. The proposed method is compared for both NLFFC(D) (PSNR oriented) and NLFFC(P) (perceptual quality oriented) models. In the Table, SR, ESR, and RankSR stand for SRGAN [13], ESRGAN [29], and RankSRGAN [33].

A comprehensive analysis is performed to observe the performance of the proposed method. The model is compared with LapSRN [12], MemNet [26], EDSR [17], DBPN [8], RDN [35], ESRGAN [29], RankSRGAN [33], SAN [4], CSNLN [20], SwinIR [16], and NLSN [19] in Table 1. Based on the comparative results, it is worth mentioning that the proposed method has relatively comparable performance on the benchmark datasets. Moreover, it even outperforms these methods on few datasets. Furthermore, Table 2 compares the sizes of multiple benchmark models, corresponding FLOPS and the achieved PSNR on Set-5 dataset. FLOPS is computed for 256×256 sized images for $\times 2$ scale. The proposed approach shows reasonable performance even with fewer parameters when compared to other very deep models. Table 3 compares the trade-off in terms of NIQE score [21] and PSNR. It is observed that NL-FFC at least shows comparable performance for relatively same distortion. Figure 5 qualitatively compares the super resolved images on different class of images.

6. Ablation Study

In this section, we study the roles played by different of the proposed non-local FFT module. We ablate two fundamental properties of the non-local Fast Fourier Convolution, including the ratio of global-to-local features and non-local network. The detailed analysis are presented in the subsequent sub-sections.

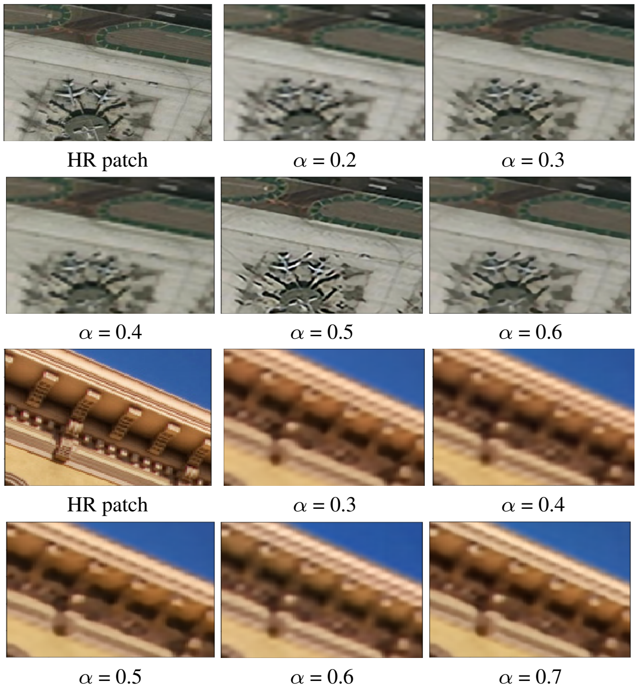

HR patch
α = 0.2
α = 0.3
α = 0.4
α = 0.5
α = 0.6
HR patch
α = 0.3
α = 0.4
α = 0.5
α = 0.6
α = 0.7

Figure 6. Qualitative results for different fractions of global features (α) for ×4 resolution. The best results are attained for $\alpha = 0.5$. Excessive local features create staggered edges whereas dominance of global features cause over-smoothing.

6.1. Effect of Global-to-Local features ratio

To this end, we have observed the performance of the network with equal split-up between global and local features. While both of these features have their own role to play, the overall performance has straight relation to the ratio of the split-up. To realize the outcome, we perform an experiment by varying the ratio of these two features. We utilize the same network architecture trained under same settings to maintain consistency in the observation. As per the visual analysis in Figure 6, one can conclude that the dominance of local features is biased towards the local frequency components, and therefore leads to pronounced staggered or zigzag like pattern in the super resolved images. At the same time, additional global features may miss the finer details of the image due to wider receptive field and cause over-smoothing.

Table 4 even provides some interesting insights through quantitative analysis of images. As the contributions from both the sides approach equality, overall performance of the network boosts up. In addition to distortion measures, we also include a no reference image quality measure, called Naturalness in Image Quality Evaluator (NIQE), to observe

α	**PSNR**	**SSIM**	**NIQE**
0.2	28.21	0.8676	6.14
0.3	28.97	0.8719	5.88
0.4	29.68	0.8801	5.04
0.5	32.76	0.9018	4.89
0.6	31.60	0.8961	5.01
0.7	30.16	0.8921	5.69

Table 4. Quantitative analysis for different values of α on Set-5 dataset. PSNR is compared using the distortion oriented model whereas NIQE score is evaluated using adversarially trained model.

the changes in the natural quality of the image which shows that the best quality is observed for $\alpha = 0.5$.

6.2. Effect of Non-Local Attention

Learning local to global feature mapping is one of the key ingredients of this study. To gain further insight into the role of non-local attention block, we train a similar network with vanilla convolution for local to global feature mapping. In addition, we also evaluate the merits of generative modelling over supervised learning by training the model without adversarial regularization.

Model	NL	Adv. Train	PSNR (dB)	SSIM	NIQE
NL-FFC	✓	✓	26.814	0.8651	4.89
FFC	×	✓	25.890	0.8524	5.10
NL-FFC	✓	×	32.76	0.9018	5.97
FFC	×	×	30.41	0.8876	6.14

Table 5. Ablation study to compare the performance with/without non-local attention and adversarial learning on Set-5 dataset. Here, NL stands for non-local attention , and third column checks whether model is trained using adversarial learning or supervised learning.

Table 5 shows the quantitative analysis to study the effect of non-local attention and adversarial learning. It can be seen that the NL-FFC with adversarial learning have better perceptual quality than other ablated architectures. Furthermore, the model trained without adversarial loss achieves the best PSNR in the case of non-local attention added to it.

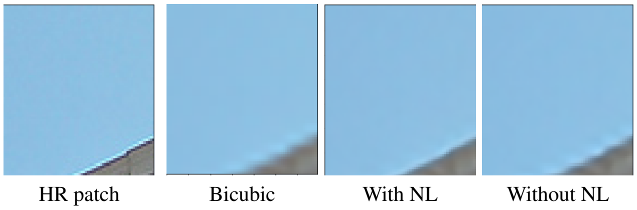


Figure 7. Qualitative results to observe the effects of non-local attention in ×4 super resolution. First and second row demonstrates the visual quality for Urban100-0004 image patch and RS-2A FCC patch.

Figure 7 compares the visual quality for images trained with and without Non-local attention (NL) module. In the case of no NL attention module, the image has discontinuous and rough edges. A plausible reason is that the network ignores the contextual relation of a query pixel to the surrounding pixels due to limited receptive field of convolution layer, and therefore the super-resolved pixel does not fit locally to the surrounding information. However, NL module extracts the necessary information for a given location that results into a continuous and visually pleasant quality in the resolved image.

7. Conclusion

In this work, we proposed an improved version of Fast Fourier Convolution by incorporating the idea of non-local attention in local to global feature mapping, and theoretically investigated the merits of learning features in the Fourier space. The proposed super resolution approach is studied for multiple benchmark datasets. The ablation study further sheds more light on the importance of different components of the proposed NL-FFC layer. Moreover, our approach outperforms or at least shows comparable performance with respect to a good number of existing state-of-the-art methods.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2018. 4

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE International Conference on Computer Vision Workshops*, 2019. 2, 4

[3] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *Proceedings of Advances in Neural Information Processing Systems*, 2021. 2

[4] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, page 11065–11074, 2019. 3, 6, 7

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015. 1, 3, 6

[6] Ross Girshick. Fast r-cnn. In *IEEE International Conf. Comput. Vis*, 2015. 1

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and

Yoshua Bengio. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems*, Dec. 2014. 4

[8] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Enhanced deep residual networks for single image super-resolution. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, page 136–144, 2017. 1, 3, 6, 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional network. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2016. 3

[11] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeplyrecursive convolutional network for image super-resolution. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2016. 3

[12] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2017. 1, 5, 6, 7

[13] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2017. 1, 3, 7

[14] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha SohlDickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Proceedings of Advances in Neural Information Processing Systems*, 2019. 5

[15] Shaohua Li, Kaiping Xue, Bin Zhu, Chenkai Ding, Xindi Gao, David Wei, and Tao Wan. Falcon: A fourier transform based approach for fast and secure convolutional neural network predictions. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2020. 3

[16] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE International Conference on Computer Vision Workshops*, 2021. 6, 7

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 1, 3, 6, 7

[18] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015. 1

[19] Yiqun Mei, Yuchen Fan, and Yuqian Zhou Ukita. Image super-resolution with non-local sparse attention. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 1, 6, 7

[20] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Humphrey Shi Ukita. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, pages 5690–5699, 2020. 1, 2, 6, 7

[21] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 22(3):209–212., 2013. 7

[22] Matan Protter, Michael Elad, Hiroyuki Takeda, and Peyman Milanfa. Generalizing the nonlocal-means to superresolution reconstruction. *IEEE Transactions on Image Processing*, 18(1):36–51, 2008. 1

[23] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 1

[24] Oren Rippel, Jasper Snoek, and Ryan P Adams. Spectral representations for convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, 2015. 1

[25] 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 Winter Conference on Applications of Computer Vision*, 2022. 3

[26] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of International conference on Computer Vision*, page 4539–4547, 2017. 3, 6, 7

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

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, 2018. 2

[29] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of European Conference on Computer Vision (ECCV) Workshop*, 2018. 1, 3, 5, 7

[30] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[31] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018. 1

[32] Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(1):4544–4556, 2012. 1

[33] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for

image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 7

[34] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 2

[35] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of IEEE Conference of Computer Vision and Pattern Recognition*, page 2472–2481, 2018. 3, 5, 6, 7

[36] Zhisheng Zhong, Tiancheng Shen, Yibo Yang, Zhouchen Lin, and Chao Zhang. Joint sub-bands learning with clique structures for wavelet domain super-resolution. In *Proceedings of Advances in Neural Information Processing Systems*, 2018. 1