FCIR: RETHINK AERIAL IMAGE SUPER RESOLUTION WITH FOURIER ANALYSIS

Yan Zhang[1,2]*, Pengcheng Zheng*[1]*, Jianan Jiang*[1]*, Pu Xiao*[1]*, and Xinbo Gao*[1]

[1]Chongqing University of Posts and Telecommunications
[2]State Key Laboratory of Integrated Services Networks (Xidian University)

ABSTRACT

Recent years, deep-learning-based methods achieve remarkable improvements on the super-resolution (SR) task. However, recovering high-quality (HQ) texture from the low-quality (LQ) aerial image is still challenging due to the limited contextual modeling ability of current deep-learning methods as well as the sharp artificial texture of aerial images. In this paper, we rethink aerial image super resolution (AISR) task with the perspective of Fourier analysis. Firstly, we build the Fourier Global Convolution (FGC) inspired by the convolution theorem of the Fourier Transform to extract the shadow features from the LQ image. Then, following the Gabor Transform, a carefully designed oriented Texture Contextual Block (OTCB) is proposed to enhance the oriented texture representation. By stacking FGC and OTCB, we propose a simple but effective straight-forward network named Fourier Consistency Image Reconstruction Model (FCIR) to restore HQ aerial image. Moreover, we design a gradient consistency loss (GC Loss) to enhance the quality of reconstructed high-frequency details. Compared with very recent state-of-the-art super-resolution methods, experimental results demonstrate promising SR performance boosts from FCIR on 3 typical aerial image datasets.

Index Terms— Super Resolution, Aerial Image, Fourier Analysis, Deep Learning, Remote Sensing

1. INTRODUCTION

Recently, with the rapid development of aerial imaging technology, the number of accessible aerial images are explosively increased. However, the imaging resolution is limited by the under-sampling effect of sensors and complex degradation processing. How to efficiently recover high-quality (HQ) aerial images from low-quality (LQ) images has drawn significant attention from remote sensing community.

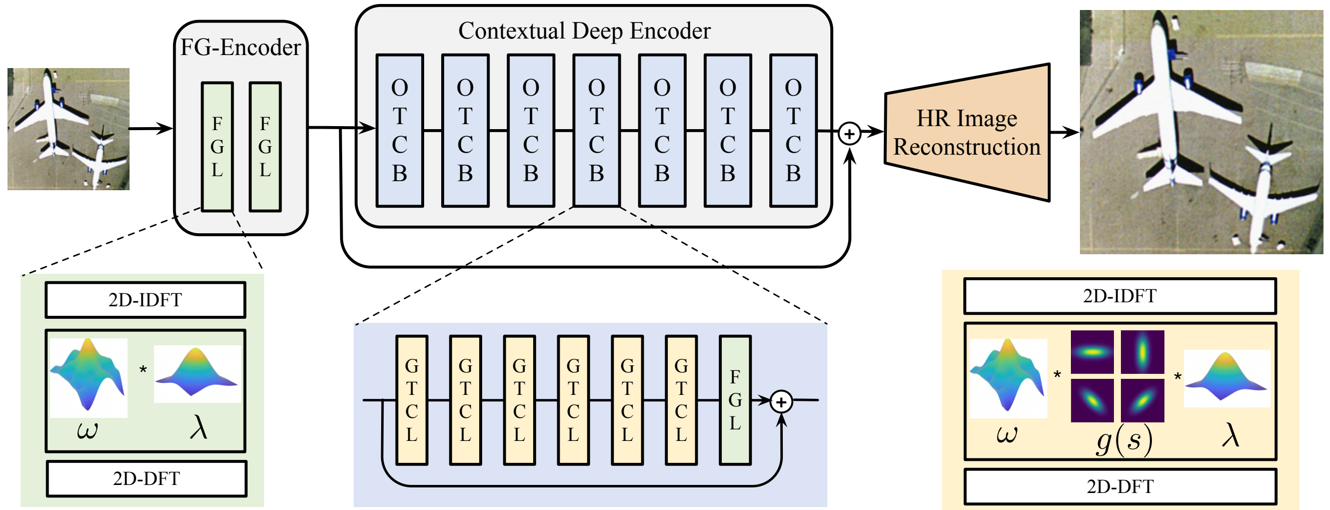
Along with deep-learning developing prosperously, numbers of CNN-based methods on image SR task emerge in endlessly. A variety of deep-learning-based methods have been applied to SR tasks, ranging from the first CNN-based SRCNN[1] method to its promising mutations, such as RCAN[2], DRCN[3], DRRN[4], VDSR[5], RDN[6] and other sequence of approaches[7, 8, 9, 10]. In recent years, with the success of self-attention mechanism[11] and Transformer[12] on image processing tasks, researchers adopted the popular Swin Transformer[13] into SR task and proposed the SwinIR[14] method. The above mentioned methods not only show great success on general image SR, but also have competitive results on aerial images. Meanwhile, considering the rich and sharp texture of aerial images, researchers extended from SwinIR[14] and proposed ARSRN[15] method, which firstly employed the self-adaption difference convolution block to exploit the artifact information remove in AISR task. It has been proven that the ARSRN[15] method shows superior performance than the previously MHAN[16] method and achieved state-of-the-art (SOTA) results on aerial image SR task.

However, existed AISR methods ignore 2 critical characters of aerial images: (1) The spatial resolution of the aerial image (near 0.1m/pixel) is much more blur than general images, which requires the help of long-range context for accurately reconstructing the details of HQ images. (2) Due to the top-down view of drones, aerial image contains complex artificial texture with specific local orientation (shown in Fig 1). Also, we found these 2 characters of aerial image is highly related to the character of Fourier Transform and its mutations. For this, we propose to analyze AISR task with the perspective of Fourier Transform and propose a novel Fourier Consistency Image Reconstruction (FCIR) model. Unlike the existing SR methods based on Fourier Transform, such as NL-FFC[17] and SwinFIR[18], FCIR sufficiently integrates the characters of Fourier Transform to enhance the reconstructed performance of aerial image. The main contributions of this paper are listed as below:

(1) With the convolution theorem and global characteristics of Fourier Transform, we propose to apply the global Fourier convolution to efficiently exploring global contextual representation from the shadow layers.

(2) Inspired by the Gabor Transform (Short-Time Fourier Transform with Gaussian Window), we propose a novel Windowed Gabor Convolution to learn multi-orientation and multi-level features in deeper layers, which is critical for recovering the local details in aerial image.

(3) A novel gradient consistency loss is introduced to enhance reconstructed high-frequency consistency between the reconstructed image and HQ image.



(a) Fourier Global Layer (FGL) (b) Orientational Texture Contextual Layer (OTCB) (c) Gabor Texture Convolution Layer (GTCL)

Fig. 1. The architecture of the proposed FCIR for aerial image super resolution. $g(x)$ in figure (c) indicates Gabor template with orientation

2. METHOD

As shown in Fig. 1, the proposed FCIR is designed following the simple encoder-decoder fashion. The LQ image is firstly fed into the global encoder to model the global contextual information. Then the output of global encoder is transformed into Gabor Contextual encoder, which is consisted of several oriented Texture Contextual Blocks (OTCB). Finally, we use the sub-pixel convolution[19]as the Reconstruction Decoder to efficiently reconstruct HR image. In this paper, we use LQ, HQ and HR to demonstrate the original low-quality image, ground-truth high-quality image, and reconstructed high-resolution image, respectively. Moreover, we design a gradient consistency loss to enhance the high-frequency information of the reconstructed HR image.

2.1. Global Shadow Encoder

Given the LQ image $I_{LQ} \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$, where H, W and C are the height, width, and channel of LQ image, respectively, we use the Global Encoder to extract shallow feature F $\in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}_1}$. In Global Encoder, we apply the global Fourier convolution (GF-Conv) to model the shadow feature. In GF-Conv, we first use the 2-D Discrete Fourier Transform[20] to map I_{LQ} into frequency domain $f_{LQ} \in \mathrm{C}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$. Due to the convolution theorem of DFT, a trainable parameter $\lambda \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ is multiplied with I_{LQ} to simulate global convolution with the kernel size of [H, W]. Then, the feature in frequency domain is transformed back into spatial domain via Inverse Discrete Fourier Transform[21] (IDFT). After GF-Conv, we use a FFN to expand the channel into C_1 . Global Encoder can be computed through Eq .1:

$$f_{ge} = FFN\left(IDFT\left(DFT\left(I_{LQ} * \lambda\right)\right)\right) \tag{1}$$

where f_{ge} is the output of Global Encoder. In this part, we mainly utilize the convolution theorem and the global character of DFT to efficiently learn the shadow global feature from I_{LQ}.

2.2. Contextual Deep Encoder

As mentioned before, both of the vanilla convolution and self-attention are isotropic and not sensitive to the local orientation of texture in aerial images. Hence, we propose the oriented Texture Contextual Block (OTCB) to enhance the oriented texture representation. As shown in Fig 1, the critical component of OTCB is the Gabor Texture Convolution Layer (GTCL). In GTCL, we firstly split the feature map into m non-overlapping windows $\{f_0, ..., f_m\}$, the choice of window size is studied in experiment section. Then we define a classical 2D Gabor transform templates with 4 different orientations and 3 scales $\{g(0), ..., g(11)\}$. Compared with DFT, Gabor transform is sensitive to local texture with specified orientation. The 2D Gabor template $g(s)$ is computed by Eq .2:

$$\begin{aligned}\psi\left(x, y; f_0, \theta\right) &= \frac{f_0^2}{\pi\gamma\eta} e^{-\frac{f_0^2}{\gamma^2}x'^2 + \frac{f_0^2}{\eta^2}y'^2} e^{j2\pi f_0 x'} \\ x' &= x\cos\theta + y\sin\theta \\ y' &= -x\sin\theta + y\cos\theta\end{aligned} \tag{2}$$

where $f_0 \in [2, 5, 7]$, $\theta \in [0°, 45°, 90°, 135°]$, and other hyper parameters are initialized following the default setting of OpenCV[22]. Next, the windowed feature f and $g(s)$ are transformed into the frequency domain via DFT. The oriented texture is acquired as Eq .3:

$$f = \underset{i,s}{cat}(IDFT(DFT(f_i) * g(s) * \lambda)) \tag{3}$$

Just like FG-Conv, λ is the learnable filter template. Then, we use FFN to aggregate the multi-scale and multi-orientation contextual representation. In each OTCB, there are 6 GTCL with a residual connection. We append a Global Fourier Convolution at the tail of OTCB for gathering and exchanging information over spatial dimension.

2.3. Reconstruction and Loss Function

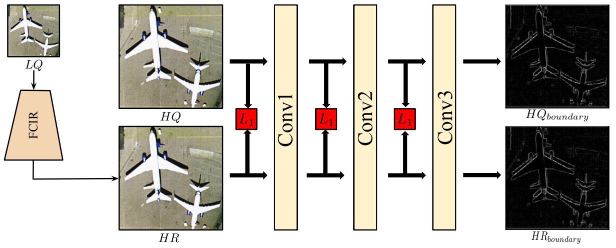


Fig. 2. Computational flow of Gradient Consistency Loss.

In the end of CDE, we use a simple sub-pixel convolution [19] as the reconstruction decoder to reconstruct the HR image from the deep feature. As to the loss function, we design a novel Boundary Consistency loss (BC Loss) to measure the high-frequency difference between the reconstructed HR image and HQ image. From the perspective of Fourier analysis, boundary detection operator is a high-pass filter to enhance the texture of images. In BC loss, we first train a 3-layer CNN θ to simulate the Sobel boundary detection operator[23]. Then, the parameter of θ is fixed during training FCIR. The BC loss is computed as below:

$$BC \text{ Loss } = \sum_{i=1}^{3} L_1\left(\theta^i(I_{HR}), \theta^i(I_{HQ})\right) \tag{4}$$

where $\theta^i(HR)$ and $\theta^i(HQ)$ indicates the ith feature map of θ with the inputs of HR image and HQ image, respectively. The proposed BC loss can keep the boundary consistency between HR image and real HQ image, and therefore enhance the quality of high-frequency texture. Together with BC loss, we also apply the L_1 loss as the loss function:

$$L_1 = |I_{HQ} - I_{HR}| \tag{5}$$

The entire loss function is computed as Eq. 6:

$$L_{sr} = L_1 + \alpha * L_{bc} \tag{6}$$

where α is a balanced parameter (0.2 in our experiments).

3. EXPERIMENT

3.1. Dataset and Implement Details

Dataset. We use 3 classical aerial image datasets in our experiments (UC Merced dataset[24], RSSCN7 dataset[25] and UCAS-AOD dataset[26]). We crop and resize these images into slices with fixed resolution (240 × 240). Bicubic function is applied to down-sample aerial images and get x2, and x4 LR images. We randomly choose 80% images as the train set and the rest images as the test set.

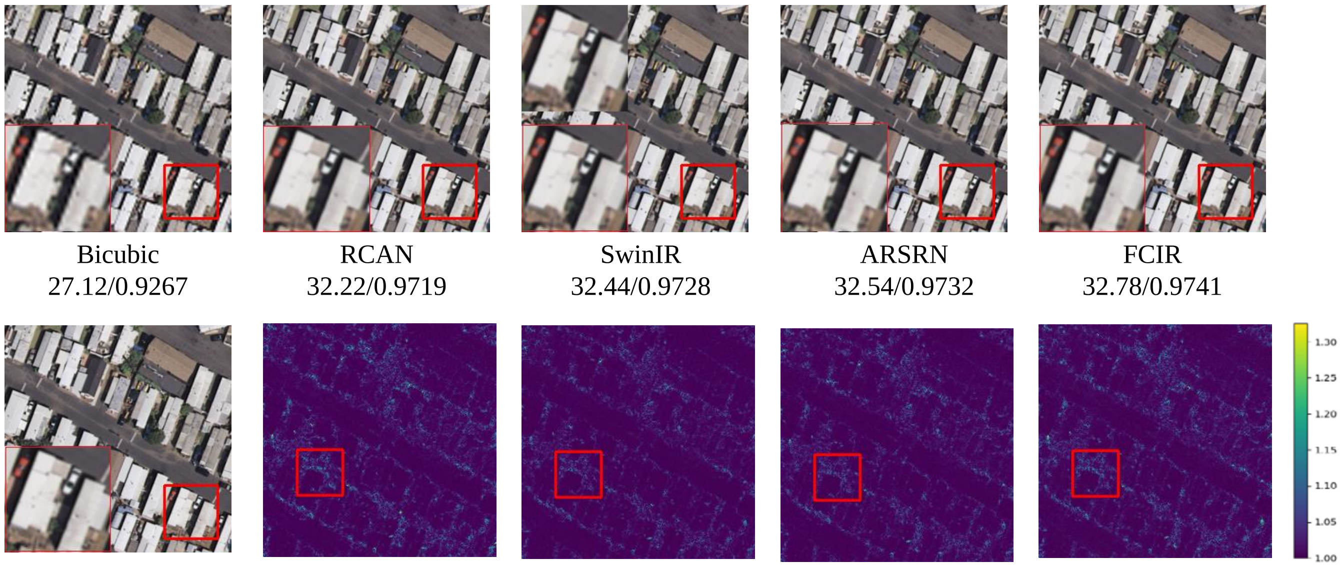
Implement Details. We compare the performance of FCIR with latest SOTA SR methods (RDN[6], RCAN[2], and SwinIR[14]) and specifically designed AISR methods (MHAN [16] and ARSRN[15]). Two widely used image quality assessment metrics (PSNR, SSIM) are employed to quantify AISR performances. Adam optimizer [27] is employed for optimization with default setting of Pytorch. The learning rate is initialized to 0.0001 and decreases by polygon learning rate adjustment schedule. All models are trained on 1 NVIDIA GTX3090.

3.2. SOTA Comparisons

		UC Merced		UCAS-AOD		RSSCN-7	
Method	Scale	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RDN	×2	34.44	0.9298	38.17	0.9537	35.14	0.9298
RCAN		35.07	0.9359	38.39	0.9556	35.54	0.9353
MHAN		34.97	0.9329	38.29	0.9548	35.35	0.9329
SwinIR		35.09	0.9352	38.77	0.9680	35.96	0.9349
ARSRN		35.16	0.9360	38.51	0.9562	35.98	0.9394
FCIR		**35.97**	**0.9386**	**38.98**	**0.9612**	**36.11**	**0.9431**
RDN	×4	28.48	0.7588	31.44	0.8299	29.86	0.7742
RCAN		28.36	0.7614	31.70	0.8359	29.75	0.7731
MHAN		28.42	0.7626	31.36	0.8273	29.83	0.7765
SwinIR		28.57	0.7694	**32.13**	**0.8441**	29.95	0.7794
ARSRN		28.79	0.7719	31.81	0.8385	30.15	0.7835
FCIR		**29.19**	**0.7907**	32.09	0.8435	**30.67**	**0.8038**

Table 1. SOTA performance comparisons. Best and second scores are **highlighted** and underlined.

The objective results of FCIR are listed in Table 1. On the UC Merced dataset. The average PSNR scores of FCIR are 0.81 and 0.4 higher than that of the second-best ARSRN[15] with scale factors t = 2 and t = 4, respectively. The similar tendency can also be found from SSIM score. The average SSIM of FCIR are 0.02 and 0.2 higher than ARSRN[15] with different scale factors. We visualize the reconstructed HR image on Fig 3, the HR image from FCIR has better texture on the outline of vehicles and eaves, which strongly proves the proposed method is able to obtain accurate contour and reconstruct texture information. FCIR also indicates promised performance on RSSCN7 dataset. On x2 test set, FCIR achieves the SOTA results of 36.11 and 0.9431 on PSNR and SSIM, respectively. On x4 test set, FCIR obtains 30.67 PSNR scores and is 0.5dB


Bicubic
27.12/0.9267

RCAN
32.22/0.9719

SwinIR
32.44/0.9728

ARSRN
32.54/0.9732

FCIR
32.78/0.9741

Mobilehomepark17
PSNR/SSIM

RCAN

SwinIR

ARSRN

FCIR

Fig. 3. Qualitative comparison of the proposed method with four counterparts on a typical satellite image pair from the UC Merced dataset[24] with an upsampling factor of ×2. Images in the second row visualize the MSE between the results and the ground truth.

higher than previous SOTA method (ARSRN[15]). As to the UCAS-AOD dataset, FCIR still surpasses most of competitors with a big margin. However, SwinIR shows great ability on UCAS-AOD on x4 test set. This is mainly because the original spatial resolution is higher than UC Merced and RSSCN7, which depresses the importance of learning the local oriented representation.

3.3. Ablation Studies

In this section, we ablate the importance of elements in the proposed FCIR, all the ablation results are conducted on UC Merced dataset.

Window Size. In Table 2, we first investigate the impact of window size towards the performance of FCIR. In the OTCB block, window size determines of performing oriented texture learning, different window size results in different SR performance. In our experimental results, w=4 is too short to measure the oriented texture, while w=16 gains no distinct boost but takes much higher latency. w=12 achieve the best trade-off between efficiency and effectiveness.

Component Ablations. In this section, we conduct a series of controlled experiments to investigate the importance of different modules in FCIR. As shown in Table 2, we start with RCAN and achieve 35.07dB PSNR score. When apply the proposed GF-Conv as shadow encoder, 0.08dB PSNR can be observed. Next, replacing the deep encoder of RCAN with OTCB brings remarkable improvement of 0.69dB PSNR. Extra 0.14dB PSNR bonus can be obtained via our carefully designed GC Loss. When these modules are combined, we can get the best AISR result of 35.98dB PSNR on UC Merced dataset. It demonstrates that all components are indispensable for the proposed AISR model.

RCAN	GF-Conv	OTCB				GC-Loss	PSNR
		w=4	w=8	w=12	w=16		
√	×	×	×	×	×	×	35.07
√	√	×	×	×	×	×	35.15
√	√	√	×	×	×	×	35.62
√	√	×	√	×	×	×	35.8
√	√	×	×	√	×	×	35.84
√	√	×	×	×	√	×	35.77
√	√	×	×	√	×	√	35.98

Table 2. Ablation studies for the proposed FCIR

4. CONCLUSION

In this paper, we rethink the existed detects of AI SR task, and give an analysis with the perspective of Fourier Transform and its mutations. We utilize the global character and convolution theorem of FT to extract global representation on shadow encoder. Also, we apply Gabor transform to measure oriented texture in deep encoder. Combining with a carefully designed gradient consistency loss, a new SOTA is reported by the proposed FCIR model, which strongly demonstrates the potential of integrating classical Fourier analysis with deep neural networks.

5. REFERENCES

[1] 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*, vol. 38, no. 2, pp. 295–307, 2015.

[2] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[3] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Deeply-recursive convolutional network for image super-resolution," *CoRR*, vol. abs/1511.04491, 2015.

[4] Ying Tai, Jian Yang, and Xiaoming Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

[5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," *CoRR*, vol. abs/1511.04587, 2015.

[6] 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*, 2018, pp. 2472–2481.

[7] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," *CoRR*, vol. abs/1704.03915, 2017.

[8] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang, "Multi-scale residual network for image super-resolution," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 517–532.

[9] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang, "Image super-resolution via dual-state recurrent networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1654–1663.

[10] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4539–4547.

[11] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021.

[14] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[15] Jiaming Wang, Zhenfeng Shao, Xiao Huang, Tao Lu, Ruiqian Zhang, and Yong Li, "From artifact removal to super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[16] Dongyang Zhang, Jie Shao, Xinyao Li, and Heng Tao Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5183–5196, 2020.

[17] Abhishek Kumar Sinha, S Manthira Moorthi, and Debajyoti Dhar, "Nl-ffc: Non-local fast fourier convolution for image super resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 467–476.

[18] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin, "Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution," *arXiv preprint arXiv:2208.11247*, 2022.

[19] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[20] MR Smith, ST Nichols, RT Constable, and RM Henkelman, "A quantitative comparison of the tera modeling and dft magnetic resonance image reconstruction techniques," *Magnetic resonance in medicine*, vol. 19, no. 1, pp. 1–19, 1991.

[21] Safa Isam and Izzat Darwazeh, "Simple dsp-idft techniques for generating spectrally efficient fdm signals," in *2010 7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010)*. IEEE, 2010, pp. 20–24.

[22] Gary Bradski and Adrian Kaehler, "Opencv," *Dr. Dobb's journal of software tools*, vol. 3, pp. 120, 2000.

[23] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker, "Design of an image edge detection filter using the sobel operator," *IEEE Journal of solid-state circuits*, vol. 23, no. 2, pp. 358–367, 1988.

[24] Yi Yang and Shawn Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[25] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

[26] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3735–3739.

[27] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.