GLOBAL-LOCAL AWARENESS NETWORK FOR IMAGE SUPER-RESOLUTION

Pin-Chi Pan[⋆†] *Tzu-Hao Hsu*[⋆§] *Wen-Li Wei*[⋆] *Jen-Chun Lin*[⋆]

[⋆] Institute of Information Science, Academia Sinica
[†] Department of Electrical Engineering, National Chung Cheng University
[§] Graduate Institute of Electrical Engineering, National Taiwan University

ABSTRACT

Deep-net models based on self-attention, such as Swin Transformer, have achieved great success for single image super-resolution (SISR). While self-attention excels at modeling global information, it is less effective at capturing high frequencies (*e.g.*, edges etc.) that deliver local information primarily, which is crucial for SISR. To tackle this, we propose a global-local awareness network (GLA-Net) to effectively capture global and local information to learn comprehensive features with low- and high-frequency information. First, we design a GLA layer that combines a high-frequency-oriented Inception module with a low-frequency-oriented Swin Transformer module to simultaneously process local and global information. Second, we introduce dense connections in-between GLA blocks to strengthen feature propagation and alleviate the vanishing-gradient problem, where each GLA block is composed of several GLA layers. By coupling these core designs, GLA-Net achieves SOTA performance on SISR.

Index Terms— Super-resolution, self-attention, transformer, inception

1. INTRODUCTION

Single image super-resolution (SISR) is a classical problem in the field of low-level computer vision. It aims to reconstruct a natural and sharp detailed high resolution (HR) image from a low resolution (LR) one. Various applications, including surveillance [1], medical imaging [2], and object recognition [3], benefit from SISR. Due to the ill-posed nature of SISR, many deep-net models have been proposed to learn mappings between LR and HR image pairs for tackling such an inverse problem.

Recently, self-attention-based deep-net models [4, 5, 6, 7] have been extensively studied on SISR with remarkable success. For example, Zhang *et al.* [4] proposed a residual non-local attention network (RNAN) that combines non-local and local attention blocks to extract features by capturing long-range dependencies between pixels and paying more attention to challenging parts. Mei *et al.* [5] devised a cross-scale non-local (CS-NL) attention module that explicitly formulates the pixel-to-patch and patch-to-patch similarities inside the image to demonstrate that additionally mining cross-scale self-similarities improves SISR. Later, Mei *et al.* [6] proposed a non-local sparse attention (NLSA) module to retain long-range modeling capability from non-local operation while enjoying robustness and high-efficiency of sparse representation. Recently, Liang *et al.* [7] followed the Swin Transformer [8] to propose SwinIR for image restoration, which introduces a Swin Transformer layer (STL) with shifted window scheme for computing self-attention in the residual Swin Transformer block (RSTB) to extract deep features with greater efficiency and effectiveness. SwinIR achieves current state-of-the-art SISR results.

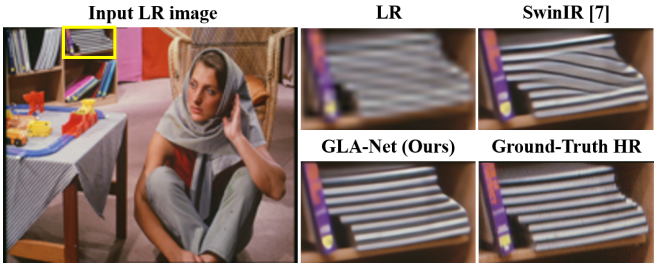

Fig. 1. Visual comparison with state-of-the-art SwinIR [7] at 4× super-resolution. Our GLA-Net is more effective at recovering local details, such as edges, than SwinIR.

Despite the great success that self-attention mechanisms have achieved in SISR due to their high effectiveness in capturing global (low-frequency) information [9], such as global shapes and structures of a scene or object, they are less effective at learning local (high-frequency) information [9, 10], such as edges and textures, which is essential for SISR.

To address the above issue, in this paper, we aim to learn comprehensive features with low- and high-frequency information, which enables the recovered super-resolution images to not only retain global shapes and structures but also more accurately preserve local details (see Figure 1). Specifically, we propose a novel global-local awareness (GLA) layer to extend the STL in SwinIR to form a global-local awareness network (GLA-Net), as shown in Figure 2. Taking advantage of the fact that convolutional neural network (CNN) can cover more high-frequency information through local convolution within the receptive fields [9, 10], in the GLA layer, we introduce an Inception module [11] in parallel with a win-

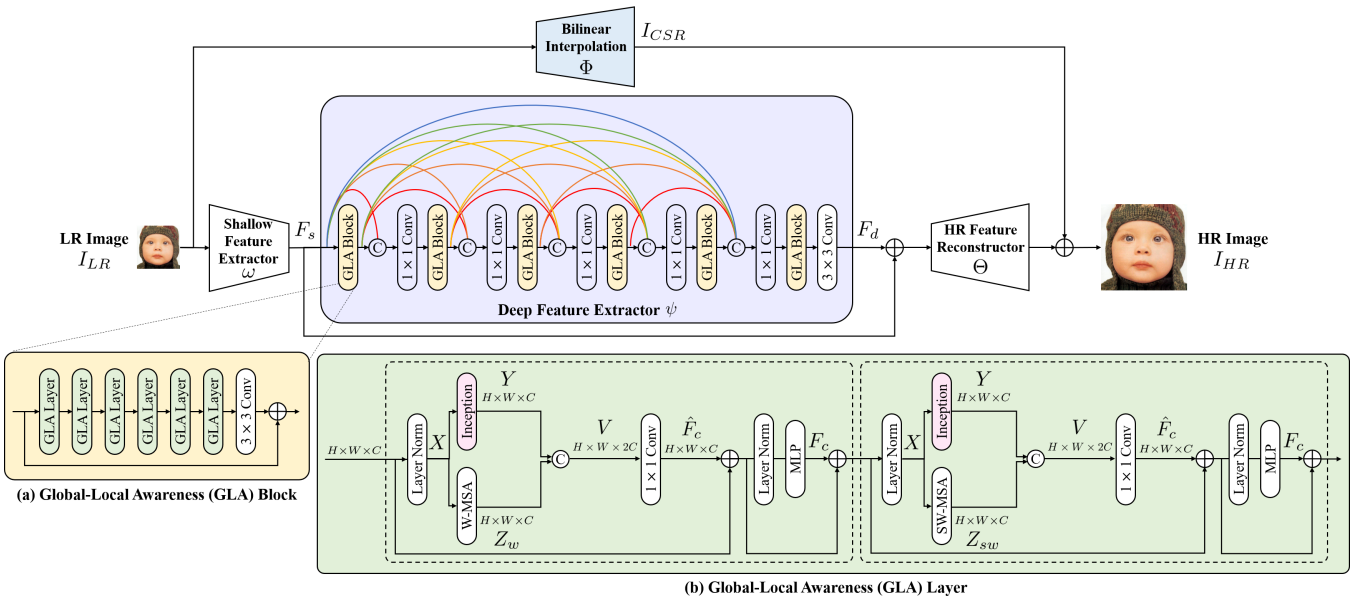

Fig. 2. The architecture of the proposed GLA-Net for single image super-resolution. Where © and ⊕ denote concatenation and element-wise sum, respectively.

dow or shifted window multi-head self-attention (W-MSA or SW-MSA) module [7, 8] to capture high-frequency and low-frequency information on the corresponding channel individually, and then combine them to obtain comprehensive features. Then we concatenate several GLA layers to form a GLA block, and introduce dense connections [12] in-between GLA blocks to strengthen feature propagation and alleviate the vanishing-gradient problem. By coupling these core designs, our GLA-Net can achieve superior SISR results (*e.g.*, in Figure 1) against recent leading methods.

2. METHOD

Figure 2 shows the overall pipeline of our GLA-Net. We elaborate each module in GLA-Net as follows.

2.1. Shallow and Deep Feature Extraction

Given an LR image as input $I_{LR} \in \mathbb{R}^{H\times W\times C_{in}}$, where H, W, and C_{in} represent the height, width and channel of the LR image, respectively, following SwinIR [7], we use a 3×3 convolutional layer $\omega(\cdot)$ to extract shallow feature $F_s \in \mathbb{R}^{H\times W\times C}$ as

$$F_s = \omega(I_{LR}). \tag{1}$$

Then, the extracted F_s is sent to the deep feature extractor to calculate the deep feature $F_d \in \mathbb{R}^{H\times W\times C}$ as

$$F_d = \psi(F_s), \tag{2}$$

where $\psi(\cdot)$ is the deep feature extractor consisting of six GLA blocks, five 1×1 and one 3×3 convolutional layers. In-between GLA blocks, we add dense connections to strengthen feature propagation and alleviate the vanishing-gradient problem [12], where the 1×1 convolution applied in-between GLA blocks aims to retain the same channel dimension as the input of the previous GLA block. For each GLA block, as shown in Figure 2(a), six GLA layers are applied to extract comprehensive features with low- and high-frequency information, followed by a convolutional layer to bring the inductive bias of the convolutional operation into the network, laying a better foundation for subsequent GLA blocks.

GLA Layer Inspired by the recent success of iFormer [10] in image classification, object detection and instance segmentation, we modify the Inception mixer proposed in iFormer [10] to extend the STL [7] to form a GLA layer (see Figure 2(b)) for the SISR task. Specifically, to better learn the high- and low-frequency information in the corresponding channels, unlike the Inception mixer, we do not split the channel information. The Inception module and W-MSA (or SW-MSA) module in the proposed GLA layer are configured to use the same full-channel information produced from the previous (Layer Norm) module. In addition, to preserve joint feature learning across channels, we use the classical convolution operation in the Inception module, instead of the depthwise convolution used in the Inception mixer [10].

To build the high-frequency-oriented "Inception module," as shown in Figure 3, we utilize a parallel structure to learn high-frequency components by leveraging the sharp sensitivity of the maximum filter and the detail perception of the convolutional operation [10]. For the branch that utilizes the maximum filter, given the input feature map $X \in \mathbb{R}^{H\times W\times C}$, we use a max pooling with padding trick and a 1×1 convolutional layer to capture high-frequency component as

$$Y_{max} = Conv(MaxPool(X)). \tag{3}$$

Then, other high-frequency components $Y_{conv} = \{Y^i_{conv}\}^N_{i=1}$ are captured by classical convolution branches with various

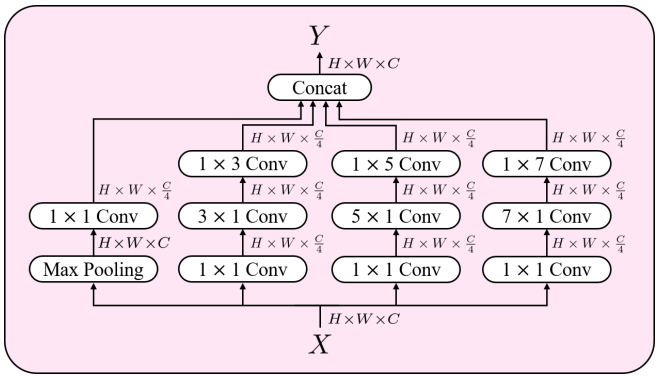

Fig. 3. The proposed Inception module.

kernel sizes, where N denotes the total number of convolution branches. The final high-frequency feature Y for the Inception module is created by concatenating Y_{max} and Y_{conv}.

In terms of W-MSA and SW-MSA modules, in Swin Transformer [8], it is still based on the classical multi-head self-attention (MSA) mechanism [13], but introduces a new shifted window scheme that restricts MSA computation to non-overlapping local windows and allows cross-window connections to draw long-range dependencies (global information), increasing effectiveness and efficiency. Taking advantage of self-attention in capturing global (low-frequency) information [9, 10], we use the W-MSA and SW-MSA, as the core modules in the GLA layer to capture low-frequency components $Z = \{Z_w, Z_{sw}\}$ of the image.

Finally, as shown in Figure 2(b), the outputs of low- and high-frequency components are concatenated along the channel dimension as $V = Concat(Y, Z_\bullet)$. The comprehensive feature $\hat{F}_c$ is then obtained by using a 1×1 convolution (the number of channels is half of V) to weightedly combine the channels of V under a residual learning fashion. F_c denotes the output comprehensive feature after operation by the multilayer perceptron (MLP) module.

2.2. HR Image Reconstruction

We compensate for the coarse super-resolution image I_{CSR} produced by bilinear interpolation by aggregating shallow and deep features to generate the final reconstructed HR image I_{HR}:

$$I_{HR} = \Theta(F_s + F_d) + I_{CSR}, \tag{4}$$

$$I_{CSR} = \Phi(I_{LR}), \tag{5}$$

where $\Phi(\cdot)$ represents the function of bilinear interpolation, and $\Theta(\cdot)$ is the function of the HR feature reconstructor, which is implemented by sub-pixel convolution layer [14, 7] to upsample the feature.

2.3. Loss Function

Following SwinIR [7], we supervise GLA-Net training by imposing an $\mathcal{L}_1$ loss between the reconstructed HR image I_{HR} and the ground-truth HR image I_{GT}:

$$\mathcal{L}_1(I_{HR}, I_{GT}). \tag{6}$$

2.4. Implementation Details

The parameters of GLA-Net are optimized by the Adam solver [15] at a learning rate of 1×10^{-4}. The batch size is set to 32. The window size and the number of attention heads for W-MSA and SW-MSA are set to 8 and 6, respectively. We use an NVIDIA RTX A6000 to train the entire network for 5×10^5 iterations. PyTorch [16] is used for code implementation.

3. EXPERIMENTS

Following SwinIR [7], we use the DIV2K dataset [17] as training set and adopt benchmark datasets including Set5 [18], Set14 [19], BSD100 [20], Urban100 [21], and Manga109 [22] for evaluation. We compare the proposed GLA-Net with several leading methods including, RCAN [23], SAN [24], IGNN [25], HAN [26], NLSA [6], and SwinIR [7]. PSNR and SSIM are adopted as standard evaluation metrics. The values of the comparison method in Table 1 are from SwinIR [7], but we validated them independently.

3.1. Comparison with State-of-the-Art Methods

The results in Table 1 show that our GLA-Net outperforms the existing SISR methods in almost all metrics and datasets. The maximum PSNR gain reaches 0.2dB on Urban100 for scale factor $\times 2$. This demonstrates that by learning comprehensive features with low- and high-frequency information in the proposed GLA layer, and employing dense connections to strengthen feature propagation in-between GLA blocks (see Figure 2), performance can indeed be improved. To further verify our claims, we conducted an ablation study for GLA-Net on the Urban100 dataset at $2\times$ super-resolution (*i.e.*, scale factor $\times 2$). The results in Table 2 demonstrate that both the dense connection and the inception module have a positive impact on GLA-Net, *i.e.*, removing any of them has a significant impact on the performance of GLA-Net. By combining both, GLA-Net (Ours) gets the best results, *i.e.*, the rich feature representations (low- and high-frequency information) extracted from the window or shifted window multi-head self-attention module and inception module can be passed on to the following GLA block through dense connections (strengthen feature propagation). Although leading methods apply different attention mechanisms, such as channel and spatial attention for RCAN [23] and HAN [26], or self-attention for NLSA [6] and SwinIR [7], they still lack the ability to simultaneously consider high-frequency and low-frequency information to learn comprehensive features.

Qualitative comparisons between SwinIR (the current state-of-the-art method [7]) and GLA-Net also confirm that GLA-Net can better preserve local details, especially edges and textures, in recovering super-resolution images, as shown in Figure 4. Taking Figure 4(a) as an example, the results show that GLA-Net can effectively preserve the edges of

Table 1. Quantitative comparison (average PSNR/SSIM for scale $\times 2$, $\times 4$) with state-of-the-art methods on benchmark datasets (Set5, Set14, BSD100, Urban100, and Manga109). Bold denotes the best and underlined denotes the second best performance.

Method	Scale	Training Dataset	Set5 [18]		Set14 [19]		BSD100 [20]		Urban100 [21]		Manga109 [22]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN [23]	$\times 2$	DIV2K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [24]	$\times 2$	DIV2K	38.31	**0.9620**	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	**0.9792**
IGNN [25]	$\times 2$	DIV2K	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
HAN [26]	$\times 2$	DIV2K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
NLSA [6]	$\times 2$	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
SwinIR [7]	$\times 2$	DIV2K	38.35	**0.9620**	34.14	0.9227	32.44	0.9030	33.40	0.9393	39.60	**0.9792**
GLA-Net (Ours)	$\times 2$	DIV2K	**38.37**	**0.9620**	**34.34**	**0.9236**	**32.48**	**0.9034**	**33.62**	**0.9412**	**39.61**	0.9789
RCAN [23]	$\times 4$	DIV2K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [24]	$\times 4$	DIV2K	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IGNN [25]	$\times 4$	DIV2K	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
HAN [26]	$\times 4$	DIV2K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
NLSA [6]	$\times 4$	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
SwinIR [7]	$\times 4$	DIV2K	32.72	0.9021	28.94	0.7914	27.83	0.7459	27.07	0.8164	31.67	0.9226
GLA-Net (Ours)	$\times 4$	DIV2K	**32.77**	**0.9026**	**29.00**	**0.7931**	**27.86**	**0.7469**	**27.13**	**0.8178**	**31.77**	**0.9236**

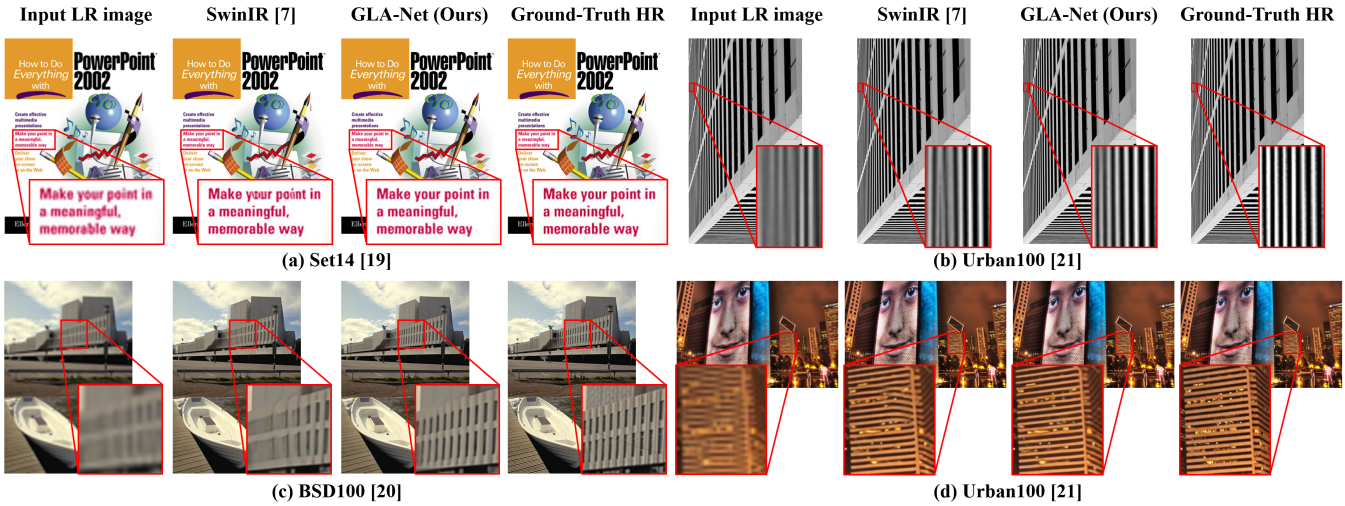

Fig. 4. Qualitative comparison of SwinIR [7] and our GLA-Net on Set14, Urban100, and BSD100 datasets. The first row is the result of $2\times$ super-resolution and the second row is the result of $4\times$ super-resolution. Best viewed by zooming.

Table 2. Ablation study for removing different modules of the GLA-Net on the Urban100 dataset at $2\times$ super-resolution.

Method	PSNR	SSIM
GLA-Net (w/o Dense Connection)	33.48	0.9405
GLA-Net (w/o Inception)	33.43	0.9396
GLA-Net (Ours)	**33.62**	**0.9412**

fonts to produce clearer text, while SwinIR struggles to do so. When it comes to running time, our GLA-Net achieves a speed of 2.46 seconds per image, while SwinIR achieves a speed of 1.91 seconds per image. Although GLA-Net is slightly slower, its performance is better than that of SwinIR. In addition, the comparisons on the five datasets also show the strong generalization property of our GLA-Net. Overall, our GLA-Net achieves the best performance.

4. CONCLUSION

We propose a GLA-Net for SISR. The main contributions lie in the design of the GLA layer and the introduction of dense connections in-between GLA blocks. The former combines an Inception module with a W-MSA (or SW-MSA) module to effectively learn comprehensive features with high- and low-frequency information, while the latter strengthens feature propagation and alleviates the vanishing-gradient problem. Extensive experiments on benchmark datasets have demonstrated that the proposed GLA-Net outperforms state-of-the-art SISR methods both quantitatively and qualitatively.

Acknowledgment: This work was supported in part by MOST under grant 110-2221-E-001-016-MY3 and Academia Sinica under grant AS-TP-111-M02.

5. REFERENCES

[1] Wilman W. W. Zou and Pong C. Yuen, "Very low resolution face recognition problem," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.

[2] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio Marvao, Tim Dawes, Declan O'Regan, and Daniel Rueckert, "Cardiac image super-resolution with global correspondence using multi-atlas patchmatch," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2013.

[3] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu, "Residual non-local attention networks for image restoration," in *International Conference on Learning Representations (ICLR)*, 2019.

[5] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] Yiqun Mei, Yuchen Fan, and Yuqian Zhou, "Image super-resolution with non-local sparse attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "SwinIR: Image restoration using swin transformer," *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[8] 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," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[9] Namuk Park and Songkuk Kim, "How do vision transformers work?," in *The Tenth International Conference on Learning Representations (ICLR)*, 2022.

[10] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan, "Inception transformer," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[14] Wenzhe Shi, Jose Caballero, Ferenc Huszar, 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[16] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," in *NeurIPS Workshop on Autodiff*, 2017.

[17] Eirikur Agustsson and Radu Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[18] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-line Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2012.

[19] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International Conference on Curves and Surfaces (ICCS)*, 2010.

[20] 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 Eighth IEEE International Conference on Computer Vision (ICCV)*, 2001.

[21] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] 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*, vol. 76, pp. 21811–21838, 2017.

[23] 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 (ECCV)*, 2018.

[24] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, "Second-order attention network for single image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[26] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen, "Single image super-resolution via a holistic attention network," in *European Conference on Computer Vision (ECCV)*, 2020.