TWO STRATEGIES TOWARD LIGHTWEIGHT IMAGE SUPER-RESOLUTION

Zongcai Du, Jie Liu, Jie Tang, Gangshan Wu*

State Key Laboratory for Novel Software Technology, Nanjing University, China
{151220022, jieliu}@smail.nju.edu.cn,
{tangjie, gswu}@nju.edu.cn

ABSTRACT

Recent convolution neural networks (CNNs) have achieved remarkable success in lightweight image super-resolution (LISR). The goal of LISR is to restore more accurate details with less model capacity. However, we observe two phenomena in current micro-architectures, one is the lack of consistent learning ability of high-frequency components, the other is large residual problem which does harm to the stability of residual learning. To tackle the two issues, we propose two strategies, namely global-guided attention strategy (GGAS) and channel-wise scaling strategy (CWSS), which can significantly improve the performance of the state-of-the-arts with negligible overheads.

Index Terms— image super-resolution, lightweight networks, global-guided attention, channel-wise scaling

1. INTRODUCTION

Image super-resolution (ISR) is a typical low-level computer vision problem, and the goal is to reconstruct a high-resolution (HR) image given its low-resolution (LR) counterpart. To find a satisfactory mapping function between LR and HR, many methods have been proposed, including image statistics-based [1], patch-based [2] and convolution neural network-based (CNN-based) methods [3, 4, 5, 6, 7, 8].

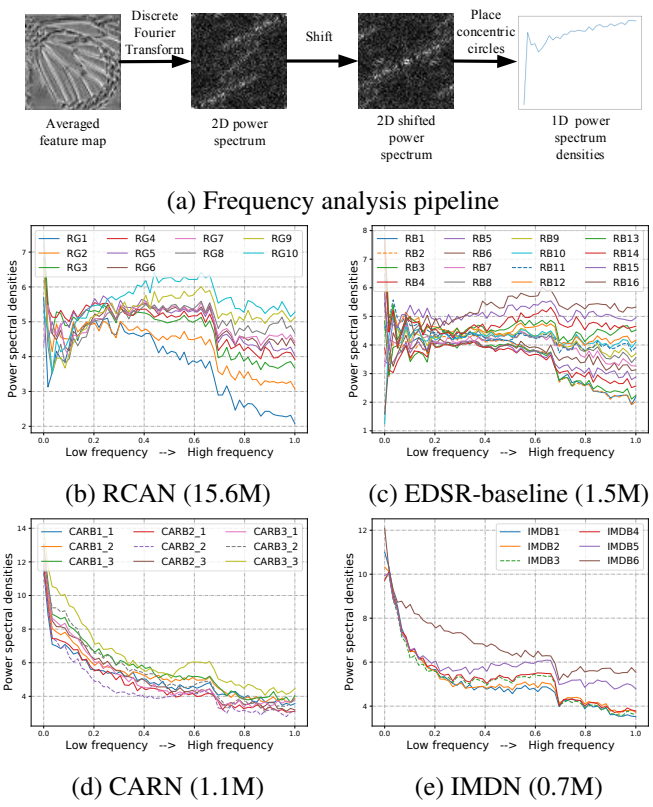
CNN-based methods is gaining more and more popularity due to its supremacy and some creative ideas on lightweight design [9, 6, 10, 8] have captivated the SR community. However, we observe two phenomena in current lightweight networks, which restricts the powerful learning ability of CNNs. First, it is commonly known that CNNs super-resolve a LR image by restoring its high-frequency component (edges, textures) layer by layer, but we find it is not always true especially for small models. To better understand this, we choose RCAN [5] as big baseline model, and EDSR-baseline [4], single-scale CARN [6], IMDN [7] as small baseline models. We analyze the power spectrum densities (PSD) of each residual group (RG) output in RCAN, residual block (RB) output in EDSR-baseline, cascading residual block (CARB) output in CARN and information multi distillation block (IMDB) output in IMDN. This is illustrated in Figure 1, where RCAN shows excellent progressive learning ability compared with small models. As the depth grows, current RG can restore more details than its predecessor. However, in EDSR-baseline, RB1 outperforms RB2 and RB10 outperforms RB11 in high-frequency domain (see dash line in Fig. 1(c)), which is contradictory to what we expect. The same thing also happens to the other two small models. Second, residual learning (RL) is widely adopted in ISR [4, 5, 11, 8, 6] to ease the training difficulty, and one of the key insights is to learn small residual rather than large value mapping so the training process should be more stable. Unfortunately, we observe lightweight architectures can not meet this condition well compared with over-parameterized model. We depict residual values of different models in Fig. 2. It can be seen that RCAN is able to learn really small residual values, while some layers in small models are in the absence of such potentiality.

To alleviate above problems, we put forward two strategies, global-guided attention strategy (GGAS) and channel wise scaling strategy (CWSS). GGAS generates an attention map which approximately describes the importance of every location. In this way, flat areas are suppressed and the network is qualified for concentrating on renovating high-frequency details. As for learning smaller residuals, CWSS adopts trainable parameters to re-scale the inputs, with the goal of reducing data variance.

The contributions are summarized as follows:

- We observe current lightweight super-resolution models suffer from discontinuous high-frequency learning and large residual problems.
- We propose global-guided attention strategy (GGAS) and channel-wise scaling strategy (CWSS) to solve the found issues.
- The proposed strategies can be applied to current state-of-the-art models and significantly improve their performance with negligible overheads.

Corresponding author

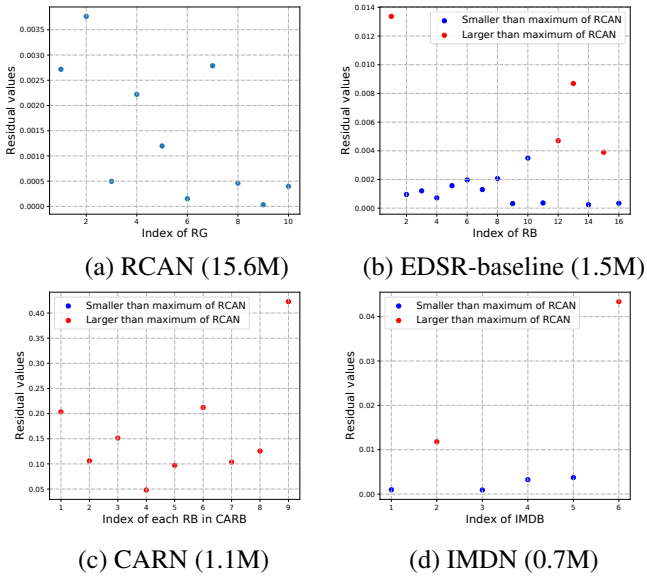


(a) Frequency analysis pipeline

(b) RCAN (15.6M) (c) EDSR-baseline (1.5M)

(d) CARN (1.1M) (e) IMDN (0.7M)

Fig. 1. 1D PSD of each block output in the big baseline model and three small baseline models. We test the same baby picture in Set5 using official pretrained models for x4 scale. The dash line denotes current block output has less high-frequency component than that of the last block output.



(a) RCAN (15.6M) (b) EDSR-baseline (1.5M)

(c) CARN (1.1M) (d) IMDN (0.7M)

Fig. 2. Averaged residual values of each RG, RB, RB in CARB and IMDB, respectively. We test the same baby picture in Set5 using official pretrained models for x4 scale. The red point denotes its value is larger than the maximum residual value of RG in RCAN.

2. PROPOSED METHOD

2.1. General architecture

When we apply GGAS to a specific architecture, it becomes a global attention block (GAB). We show general architecture equipped with GAB in Fig. 3(a). Let's denote I_{LR} and I_{SR} as the input and output of the general network. We obtain shallow feature F_0 by:

$$F_0 = H_{SFE}(I_{LR}) \tag{1}$$

where $H_{SFE}(\cdot)$ represents the mapping function of shallow feature extraction. Then a global attention map GA is calculated through:

$$GA = H_{GAB}(F_0) \tag{2}$$

where $H_{GAB}(\cdot)$ means the mapping function of GAB. Then, the output of i-th block F_i can be expressed as:

$$F_i = H_{B_i}(F_{i-1}) * GA, i = 1, \ldots, N, \tag{3}$$

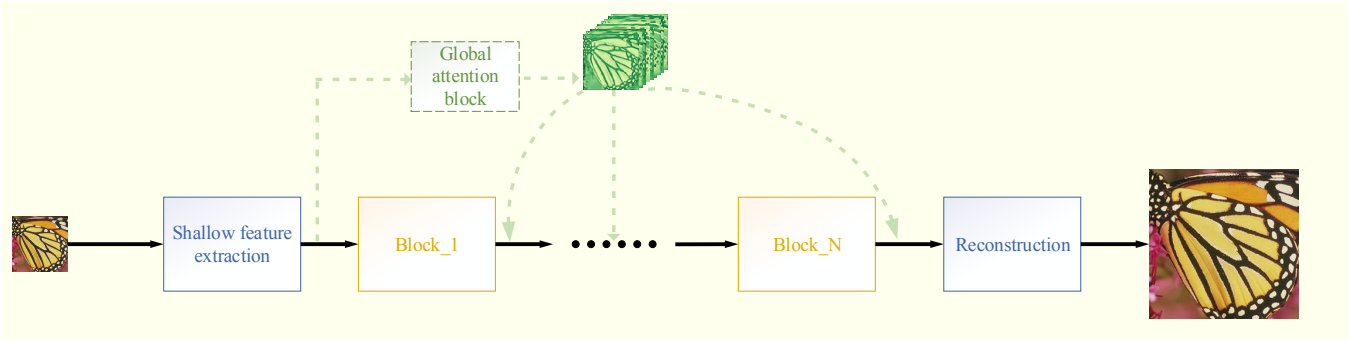
where $H_{B_i}(\cdot)$ indicates the mapping function of the i-th block, N is the total number of blocks, and $*$ denotes element-wise multiplication. Lastly, the restored image is obtained:

$$I_{SR} = H_{Re}(F_N) \tag{4}$$

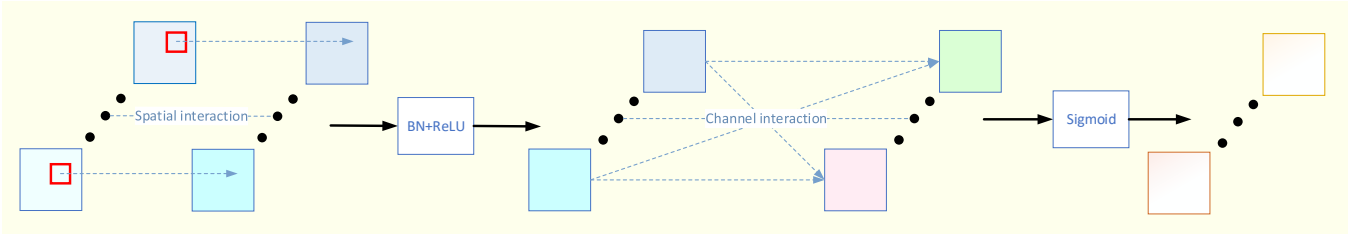
where $H_{Re}(\cdot)$ is the reconstruction function.

2.2. Global-guided attention strategy

Three aspects should be carefully taken into consideration when converting GGAS to a particular block. Most importantly, it should bring negligible overheads including number of parameters, number of FLOPs and inference speed cost. Second, it should contain trainable parameters which can be automatically adjusted according to different data distribution. Besides, the importance of every location should be determined by its neighbors, which means spatial and channel information should be utilized together. We employ depth-wise separable convolution to achieve this goal. A depthwise convolution is first used to model the spatial dependencies, then batch normalization (BN) and ReLU is performed to normalize the output and introduce non-linearity. Finally, a pointwise convolution is used to model the channel dependencies, followed by a sigmoid function to restrict the importance values to $[0, 1]$. Suppose the channel number of F_0 is C, and the kernel size of depthwise convolution is K, then the total



(a) General architecture with proposed GGAS, denoted as green dash line.



(b) Detailed structure of GAB.

Fig. 3. Illustration of how to apply GGAS to general architectures.

number of parameters during training in GAB is:

$$T_{params} = \underbrace{(K^2+1)C}_{\substack{depthwise\\convolution}} + \underbrace{2C}_{BN} + \underbrace{C^2+C}_{\substack{pointwise\\convolution}} \tag{5}$$

In the inference stage, the BN layer can be integrated into previous convolution layer, so the number of parameters is:

$$I_{params} = C^2 + (K^2+2)C \tag{6}$$

If we choose $K = 7$ and $C = 64$, the number of parameters in GAB is roughly 7.3K.

2.3. Channel-wise scaling strategy

Suppose the input of RL is x, and H_{RL} is residual learning function, then the original output y can be expressed as:

$$y = x + H_{RL}(x) \tag{7}$$

The insight of CWSS is to learn smaller residuals, thus we can utilize the good property of RL such as training with larger learning rate to achieve better performance. To this end, we first re-scale the input of RL channel by channel and denote the re-scaled input as $\hat{x}$. The i-th output channel can be expressed as:

$$y_i = \alpha_i x_i + H_{RL}(\hat{x})_i \tag{8}$$

Feature variance can be reduced by setting α less than one. This is also displayed in Fig. 4. The number of introduced parameters is the same with the number of input channels C.

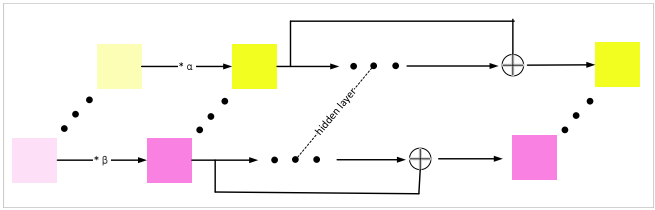


Fig. 4. Residual learning structure equipped with CWSS.

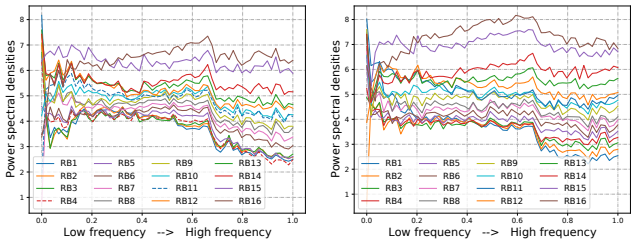
3. EXPERIMENTS

3.1. Datasets and metrics.

We adopt a 2K resolution high-quality dataset DIV2K [12] as our training and validating set. We test on five standard benchmark datasets: Set5 [13], Set14 [14], B100 [15], Urban100 [16], and Manga109 [17]. The SR results are evaluated using two common metrics, peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM).

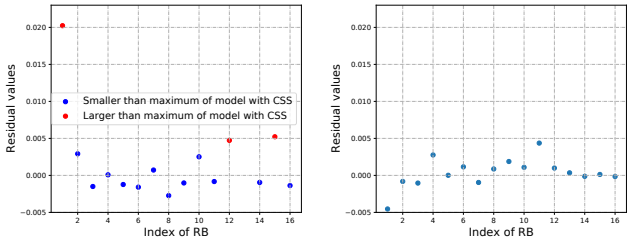
3.2. Implementation details.

We apply our strategies to FSRCNN [18], MemNet [19], CARN [6], IDN [10], EDSR-baseline [4], IMDN [7] and RCAN [5], using the official codes. We adopt two times larger learning rate than that of the original method. In GGAS, the kernel size of depthwise covolution is set to 7. In CWSS, the initial α value is experimently set to 0.9.



(a) Without GGAS (b) With GGAS

Fig. 5. Comparision of model without GGAS and model with GGAS in frequency domain. We test both on butterfly picture in Set5 with EDSR-baseline model for x4 scale.



(a) Without CWSS (b) With CWSS

Fig. 6. Residual values of EDSR-baseline x4 model with or without CWSS. CWSS is beneficial for learning smaller residuals, thus it's possible to train with larger learning rate.

3.3. Effectiveness of GGAS and CWSS

To better understand the functionality of GGAS, we send the same image to the model without GGAS and model with GGAS, then we calculate the PSD of intermediate features. This is illustrated in Fig. 5, where the model trained with GGAS is able to consistently restore high-frequency component. For verifying CWSS, we also send the same image to two models, and estimate their average residual values. Fig. 6 exhibits that the model with CWSS tends to learn smaller residuals compared with the model without CWSS.

3.4. Comparison with other attention mechanisms

In this subsection, we compare recent attention mechanisms with our scheme. We construct different models based on shallow EDSR (8 RBs). More concretely, we investigate average channel attention (CA) [20], contrast-aware channel attention (CCA) [7, 8], spatial attention (SA) [20] and enhanced spatial attention (ESA) [21]. Table. 1 shows that the performance of baseline can be improved by embedding current attention blocks, but the gain is lower than that brought by our scheme (+0.13dB on DIV2K set, +0.17dB on Set5).

Table 1. Investigation of different attention mechanisms. All models are validated on DIV2K validation set with scaling factor $\times 4$ in 300 epochs. We also report the PSNR and average inference time on Set5 $\times 4$ on Titan Xp GPU. The last column shows time increase rate compared with baseline.

	Params	DIV2K	Set5	Inference time	Increase rate
Baseline	620K	30.75	31.52	4.7ms	0.0%
Baseline+CA	620K	30.80	31.59	9.1ms	93.6%
Baseline+CCA	621K	30.81	31.60	6.7ms	42.6%
Baseline+SA	620K	30.77	31.54	5.8ms	23.4%
Baseline+ESA	622K	30.82	31.63	7.6ms	61.7%
Baseline+proposed	621K	**30.88**	**30.69**	4.9ms	4.3%

3.5. Integrated into other small models

We apply our strategies to FSRCNN [18], MemNet [19], CARN [6], IDN [10], EDSR-baseline [4] and IMDN [7]. All models are training from scratch using the official codes with scaling factor $\times 4$. It is worthy mentioning that the original FSRCNN is trained only with Y channel, we train it in RGB space to keep consistence with other methods. Table. 2 display the results, from which we can see performance can be further improved, demonstrating the generalization ability of our strategies.

Table 2. Quantitative results of applying our strategies to other lightweight architectures.

Model	Params	Set5	Set14	BSD100	Urban100	Manga109
		PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM
FSRCNN	33K	31.08 / 0.8696	27.88 / 0.7609	27.03 / 0.7166	24.73 / 0.7288	28.12 / 0.8655
+ GGAS	34K	+0.08 / 0.8733	+0.07 / 0.7650	+0.06 / 0.7188	+0.10 / 0.7296	+0.11 / 0.8689
MemNet	714K	31.96 / 0.8923	28.44 / 0.7803	27.60 / 0.7321	25.68 / 0.7664	28.12 / 0.8655
+ GGAS	721K	+0.07 / 0.8946	+0.06 / 0.7833	+0.09 / 0.7342	+0.08 / 0.7688	+0.12 / 0.8676
+ CWSS	714K	+0.06 / 0.8948	+0.05 / 0.7841	+0.09 / 0.7339	+0.12 / 0.7672	+0.13 / 0.8679
+ both	721K	+0.09 / 0.8952	+0.08 / 0.7847	+0.12 / 0.7345	+0.15 / 0.7689	+0.17 / 0.8683
CARN	1112K	32.01 / 0.8932	28.44 / 0.7796	27.51 / 0.7350	25.94 / 0.7842	30.48 / 0.9084
+ GGAS	1119K	+0.07 / 0.8948	+0.06 / 0.7821	+0.03 / 0.7379	+0.08 / 0.7861	+0.13 / 0.9089
+ CWSS	1112K	+0.00 / 0.8950	+0.04 / 0.7811	+0.06 / 0.7369	+0.06 / 0.7862	+0.09 / 0.9088
+ both	1119K	+0.05 / 0.8955	+0.09 / 0.7830	+0.08 / 0.7382	+0.11 / 0.7878	+0.17 / 0.9091
IDN	600K	31.76 / 0.8901	28.44 / 0.7732	27.36 / 0.7287	25.60 / 0.7634	29.38 / 0.8933
+ GGAS	607K	+0.06 / 0.8903	+0.05 / 0.7783	+0.04 / 0.7299	+0.04 / 0.7652	+0.08 / 0.8942
+ CWSS	600K	+0.05 / 0.8923	+0.06 / 0.7750	+0.06 / 0.7310	+0.08 / 0.7662	+0.12 / 0.8960
+ both	607K	+0.09 / 0.8929	+0.08 / 0.7790	+0.08 / 0.7315	+0.12 / 0.7671	+0.12 / 0.8970
EDSR-baseline	1518K	32.09 / 0.8938	28.58 / 0.7813	27.57 / 0.7357	26.04 / 0.7849	30.35 / 0.9067
+ GGAS	1525K	+0.12 / 0.8959	+0.08 / 0.7834	+0.11 /0.7371	+0.15 / 0.7869	+0.18 / 0.9088
+ CWSS	1518K	+0.08 / 0.8969	+0.07 / 0.7824	+0.08 / 0.7361	+0.10 / 0.7858	+0.13 / 0.9084
+ both	1525K	+0.15 / 0.8973	+0.12 / 0.7837	+0.14 / 0.7374	+0.22 / 0.7872	+0.16 / 0.9092
IMDN	715K	32.21 / 0.8948	28.56 / 0.7809	27.55 / 0.7350	26.02 / 0.7831	30.45 / 0.9076
+ GGAS	722K	+0.04 / 0.8952	+0.05 / 0.7814	+0.03 / 0.7356	+0.07 / 0.7839	+0.08 / 0.9082
+ CWSS	715K	+0.03 / 0.8951	+0.04 / 0.7812	+0.03 / 0.7355	+0.05 / 0.7840	+0.05 / 0.9079
+ both	722K	+0.07 / 0.8955	+0.09 / 0.7821	+0.05 / 0.7360	+0.11 / 0.7844	+0.16 / 0.9088

4. CONCLUSION

In this paper, we analyze the problems existing in current lightweight super-resolution models, one is the lack of consistent learning ability of high-frequency components, the other is large residual problem in residual learning. And we put forward two relative strategies, global-guided attention strategy (GGAS) and channel-wise scaling strategy (CWSS) to solve the found problems. Our methods can improve the performance of current lightweight architectures with negligible overheads, achieving more accurate reconstructions at the same time.

5. REFERENCES

[1] Carlos Fernandezgranda and Emmanuel Candès, "Super-resolution via transform-invariant group-sparse regularization," 2013.

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

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV (4)*. 2014, vol. 8692 of *Lecture Notes in Computer Science*, pp. 184–199, Springer.

[4] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," 2017.

[5] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," 2018.

[6] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *ECCV (10)*. 2018, vol. 11214 of *Lecture Notes in Computer Science*, pp. 256–272, Springer.

[7] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang, "Lightweight image super-resolution with information multi-distillation network," in *ACM Multimedia*. 2019, pp. 2024–2032, ACM.

[8] Luo Xiaotong, Xie Yuan, Zhang Yulun, Qu Yanyun, Li Cuihua, and Fu Yun, "Latticenet: Towards lightweight image super-resolution with lattice block," 2020.

[9] Ying Tai, Jian Yang, and Xiaoming Liu, "Image super-resolution via deep recursive residual network," in *CVPR*. 2017, pp. 2790–2798, IEEE Computer Society.

[10] Zheng Hui, Xiumei Wang, and Xinbo Gao, "Fast and accurate single image super-resolution via information distillation network," in *CVPR*. 2018, pp. 723–731, IEEE Computer Society.

[11] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," 2018.

[12] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming Hsuan Yang, and Qi Guo, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[13] Marco Bevilacqua, A. Roumy, Christine Guillemot, and Marie-Line Alberi-Morel, "Low-complexity single image super-resolution based on nonnegative neighbor embedding," 09 2012.

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

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE International Conference on Computer Vision*, 2002.

[16] 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*, 2015.

[17] 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, no. 20, pp. 21811–21838, 2017.

[18] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV (2)*. 2016, vol. 9906 of *Lecture Notes in Computer Science*, pp. 391–407, Springer.

[19] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, "Memnet: A persistent memory network for image restoration," in *ICCV*. 2017, pp. 4549–4557, IEEE Computer Society.

[20] Sanghyun Woo, Jongchan Park, Joon Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," 2018.

[21] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu, "Residual feature aggregation network for image super-resolution," 06 2020, pp. 2356–2365.