DDet: Dual-Path Dynamic Enhancement Network for Real-World Image Super-Resolution

Yukai Shi, Haoyu Zhong, Zhijing Yang, *Member, IEEE*, Xiaojun Yang, and Liang Lin, *Senior Member, IEEE*

***Abstract*—Different from traditional image super-resolution task, real image super-resolution(Real-SR) focus on the relationship between real-world high-resolution(HR) and low-resolution(LR) image. Most of the traditional image SR obtains the LR sample by applying a fixed down-sampling operator. Real-SR obtains the LR and HR image pair by incorporating different quality optical sensors. Generally, Real-SR has more challenges as well as broader application scenarios. Previous image SR methods fail to exhibit similar performance on Real-SR as the image data is not aligned inherently. In this article, we propose a Dual-path Dynamic Enhancement Network(DDet) for Real-SR, which addresses the cross-camera image mapping by realizing a dual-way dynamic sub-pixel weighted aggregation and refinement. Unlike conventional methods which stack up massive convolutional blocks for feature representation, we introduce a content-aware framework to study non-inherently aligned image pair in image SR issue. First, we use a content-adaptive component to exhibit the Multi-scale Dynamic Attention(MDA). Second, we incorporate a long-term skip connection with a Coupled Detail Manipulation(CDM) to perform collaborative compensation and manipulation. The above dual-path model is joint into a unified model and works collaboratively. Extensive experiments on the challenging benchmarks demonstrate the superiority of our model.**

***Index Terms*—Real Image Super-resolution, Dual-path, Neural Network.**

I. Introduction

OWING to the rapid development of artificial intelligence technology, emerging applications, such as Alexa, Woogie, and Prisma, grow more prevalent than ever to change human lives. As a classical problem in computer vision, image super-resolution technology [1]–[7] also achieves tremendous progress and be widely used in many mobile devices, such as mobile phones, for photo enhancement [8]. With such a light-weight algorithm, mobile devices are capable of providing a high-quality photograph and free from purchase expensive sensors. However, due to the existing image SR algorithms are restricted to handle simple down-scaling kernels(e.g., Bicubic, Bilinear), the robustness of traditional image SR models is limited when dealing with intricate data.

Example- based approaches [9] open a new solution for image SR by a data-driven fashion and address the relationship between internal- and external- sample [10]–[12]. Yang *et al* [13] uses spare representation to build a dictionary learning algorithm between LR and HR image. Deep convolutional neural networks have also generally applied in image SR. For instance, SRCNN [14] uses a three-layer convolution network to simulate the spare coding process. VDSR [15] first adopts a long-term skip connection to enforce neural network approximate residual map(i.e., the difference between LR image and HR image), rather than image content in the value domain. Similarly, the long-term skip connection technology is widely used in the latter image SR frameworks. EDSR [16] employs an efficient convolutional block and long-term skip connection to realize high-quality image restoration. RCAN [17] also employs feature squeeze and extraction strategy and multiple lengths skip connection for advanced image SR. These methods aim at restoring the corrupted image with fixed degeneration kernel and the real-world degeneration means may complicated and diverse, which limit the restoration quality when meeting real-world photo. Also, the fixed degeneration kernel inherently makes the input and the ground-truth image has a stable relationship and therefore turns the long-term residual connection to become practical. Nevertheless, the real-world LR and HR image pair are unable to have a strict aligned relationship as the filming conditions change and absolute alignment without information loss is impractical. To this end, these methods have two drawbacks: 1) Most of the image SR methods obtain the LR input by applying fixed degradation metrics on the ground-truth image, which limited the generalization ability toward real-world cases. 2) It is potentially less optimal that forward the original corrupt input when dealing with real-world image restoration, regardless of their pixel displacement.

Although conventional convolutional networks can extract diverse features with various inputs, the parameter is fixed during the inference process. DFN [18] opens up a new solution to turn convolutional features into dynamic filters for feature aggregation. Similar idea [19], [20] is addressed in the image restoration task yet. KPN [20] proposes a dynamic prediction mechanism by employing the high-dimensional feature into convolutional

Manuscript received January 21, 2020; revised February 26, 2020; accepted February 29, 2020. Date of publication March 4, 2020; date of current version March 24, 2020. This work was supported in part by the National Nature Science Foundation of China under Grant U1701266, and in part by Guangdong Provincial Key Laboratory of Intellectual Property and Big Data under Grant 2018B030322016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ananda Shankar Chowdhury. *(Corresponding author: Zhijing Yang.)*

Yukai Shi, Haoyu Zhong, Zhijing Yang, and Xiaojun Yang are with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: ykshi@gdut.edu.cn; 2111903112@mail2.gdut.edu.cn; yzhj@gdut.edu.cn; yangxj18@gdut.edu.cn).

Liang Lin is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: linliang@ieee.org).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/LSP.2020.2978410

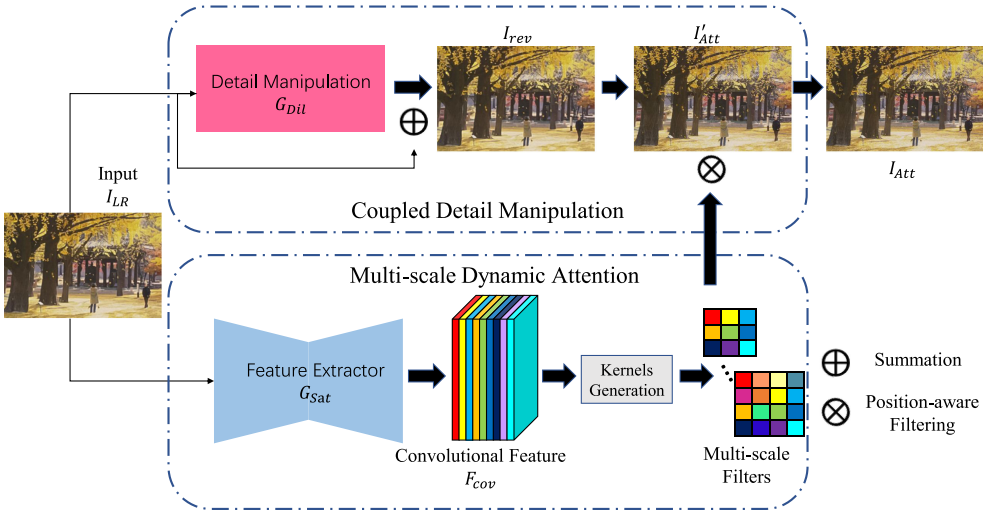


Fig. 1. Illustration of our dual-path dynamic enhancement network.

kernels. LP-KPN [20] further conducts the dynamic kernels on multi-scale feature for image restoration. As the multi-scale strategy [21]–[24] prefers to conduct kind of down-samplings to obtain large receptive field, the relative local context is sufficient for restoration. Therefore, the usage of applying various size kernels is implied to aggregate local context effectively and avoid information lose.

This paper addresses the aforementioned issues by presenting a dual-path dynamic enhancement network. Our contribution is to conduct a multi-scale dynamic attention module(MDA). Compared with the traditional convolutional kernel, our spatial attention kernel is learned in a content-adaptive fashion, which is more flexible. Compared with dynamic filter pipelines [20], [25], our model uses multiple dynamic kernels from the different receptive field without information loss, which make our method enjoys rich diversity. We also conduct a lightweight coupled detail manipulation(CDM) to relax pixel displacement in real image SR. Compared with other long-term skip connection pipeline [15], our lightweight model exhibits a supporting pixel manipulation and correction. The above two components are joint together by a dual-way fashion to demonstrate charming results on real image SR benchmark.

II. Methodology

A. Overview

By convention, the image super-resolution aims at restoring the high-quality image I_{HR} from poor-quality input I_{LR}. In classical image SR issue, the I_{LR} is obtained from I_{HR} with a down-sampling operation:

$$I_{\mathrm{LR}} = \partial \left(I_{\mathrm{HR}} \otimes K_{\mathrm{Gauss}} \right), \tag{1}$$

where K_{Gauss} is a Gaussian blur operator and ∂ denotes the degradation process. ∂ usually adopts a Bicubic down-scale algorithm in traditional image SR. Real-SR dataset [20] is proposed to address real-world degeneration metrics by capturing I_{HR} and I_{LR} with different quality optical sensors and resolution settings. The data collection manner inherently make the image pair has different resolution property.

DDet consists of two components (i.e., MDA and CDM). Given a low-resolution image I_{LR}, DDet send I_{LR} into the dual-path model parallelly:

$$I_{\mathrm{Rev}} = \mathbf{G}_{\mathrm{Dil}} \left(I_{\mathrm{LR}} | \Theta_{\mathrm{Dil}} \right) + I_{\mathrm{LR}}, \tag{2}$$

where the Θ_{Dil} indicates the parameter of CDM. Since the overall reveal process based on a residual fashion, we use the I_{LR} as the reference for the output of CDM. After obtaining the I_{Rev}, we employ a dynamic kernel to exhibit a local spatial aggregation. Then, we extract the I_{LR} into high-dimensional feature and represent it into dynamic kernel for spatial attention:

$$K_{Dyn} = \mathbf{G}_{Sat} \left(I_{\mathrm{LR}} | \Theta_{Sat} \right), \tag{3}$$

where K_{Dyn} is learned content-adaptive kernel. As the kernel is flexible and aware to image content, we conduct a content-adaptive spatial attention by:

$$I_{Att} = \mathbf{F}_{\mathbf{PR}} \left(\mathbf{g} \left(I_{\mathrm{Rev}} | K_{Dyn} \right) \right), \tag{4}$$

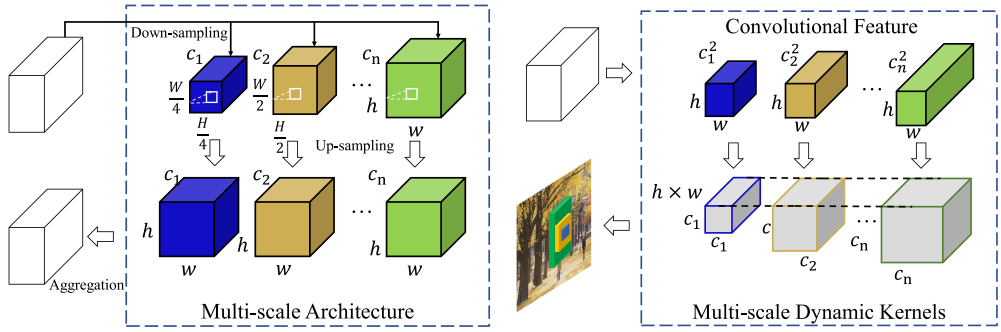
where $\mathbf{g}$ indicates a convolution operator, K_{Dyn} is corresponding convolutional kernels and $\mathbf{F}_{\mathbf{PR}}$ is a post-refinement operation. The above two branch work complementary for real image SR.

B. Multi-Scale Dynamic Attention

The MDA is used for extract valuable local information and enhance the original low-quality input into high-quality image. As Fig. 1 show, MDA consists of a feature extractor G_{Sat} and the kernel generation process. Meanwhile, G_{Sat} inherits the structure of auto-encoder. We first down-sampling I_{LR} with scale $4\times$:

$$y_{\mathrm{down}} = \Downarrow C(I_{\mathrm{LR}}), \tag{5}$$

where $\Downarrow$ indicates the down-sampling and C means the convolutional feature extraction. We use a set of residual block for the feature representation of y_{down}. Each residual block consists of two convolutional layer with size of $64 \times 3 \times 3$. We use 16 residual block in G_{Sat} to obtain deeper representation y'_{down}.



(a) Laplacian Pyramid [20] (b) Our

Fig. 2. Illustration of multi-scale dynamic attention mechanism. Typical multi-scale methods (e.g., (a)) tend to compress the feature into small size for the large receptive field, and then conduct convolutional feature extraction and re-scaling. Since relative local information is enough for the image SR, the proposed multi-scale dynamic attention module applies the various size of dynamic convolutional filters, which not only demonstrate a multi-scale enhancement but also free from re-scaling.

Finally, we upscale y'_{down} as:

$$\Upsilon_{Dyn} = C \Uparrow C(y'_{\text{down}}), \tag{6}$$

where $\Uparrow$ is upscale operation and Υ_{Dyn} is final output of G_{Sat}.

Dynamic filter formulation: We then use Υ_{Dyn} to formulate our feature-based dynamic convolutional filters. we reformulate the feature into kernel shape across the channel. For instance, we have a feature with a size of $N \times C \times W \times H$, which indicates batch size, channel number, feature width, and height, respectively. Meanwhile, the feature width and height own the same shape with I_{HR}. Hence, a single dynamic convolutional filter can be obtained with:

$$K_{Dyn}^{(w,h)} = \Upsilon_{Dyn}(w,h), \tag{7}$$

w and h indicate the location. Besides, $1 \leq w \leq W$ and $1 \leq h \leq H$. Since we adopt a point-aware filtering strategy, each pixel in I_{HR} own a individual dynamic filter. Suppose the $I_{\text{HR}}(w,h)$ is a pixel in the I_{HR}, we need to collect spatial information with pure local perspective by a dynamic filter $K_{Dyn}^{(w,h)}$. Therefore, $K_{Dyn}^{(w,h)}$ has corresponding location with $\Upsilon_{Dyn}(w,h)$. We obtain each element in $K_{Dyn}^{(w,h)}$ alone the channel of $\Upsilon_{Dyn}(w,h)$ as:

$$\begin{cases} K_{Dyn}^{(w,h)}(i_1,j_1) = \Upsilon_{Dyn}(w,h,c_n), n=1 \\ K_{Dyn}^{(w,h)}(i_1,j_2) = \Upsilon_{Dyn}(w,h,c_n), n=2 \\ \cdots\cdots \\ K_{Dyn}^{(w,h)}(i_n,j_n) = \Upsilon_{Dyn}(w,h,c_{n^2}), n^2=C. \end{cases} \tag{8}$$

Then, as illustrated in Eq. 4, the generated K_{Dyn} is directly used on I_{HR} to demonstrate a pure local attention.

Kernels generation: We conduct a multi-scale dynamic attention strategy, we generate multiple dynamic spatial kernels as illustrated in Eq. 7 and 8. To fully capture multi-scale information, we adopt three dynamic spatial kernels with a size of 3×3, 5×5 and 7×7. We conduct the dynamic convolution filters work on I_{Rev} directly:

$$I'_{Att} = \sum_{n=1}^{N} \left(\mathbf{g_n} \left(I_{\text{Rev}} | K_{Dyn}^{n} \right) \right), \tag{9}$$

(a) I_{LR} (b) I_{rev}

Fig. 3. Demonstration of coupled detail manipulation.

N denotes the kernel number, we use 3 dynamic kernels for multi-scale dynamic attention. As shown in Fig. 2, our model adopt a multi-scale attention based on various size of dynamic convolutional filters$\{K_{Dyn}^{1}, \ldots, K_{Dyn}^{n}\}$. Specifically, we incorporate several dynamic kernels on I_{Rev} in parallel. The parallel results are aggregated together by summation to obtain the I'_{Att}. Since the kernel size is decided by channel number of Υ_{Dyn}, we adopt 9, 25 and 49 channel numbers of Υ_{Dyn} to formulate the corresponding dynamic kernels.

C. Coupled Detail Manipulation

Suppose we have the I'_{Att}, a critical problem is the input image may have pixel displacement and disparity. Therefore, a complementary component G_{Dil} is incorporated for compensation and enhancement. G_{Dil} adopts a residual learning strategy for the detail revealed. Supposed we have I_{LR}, the residual block is incorporated for feature representation:

$$I_{rev} = C_{Res}(I_{\text{LR}}) + I_{\text{LR}}, \tag{10}$$

C_{Res} means residual block. The incorporated residual block has three convolutional layers with 3×3 convolutional kernel, the channel number is 64, 64 and 1, respectively. Once we obtain the I_{Rev}, as illustrated in Eq. 9, we employ the learned K_{Dyn} work on enhanced image I_{Rev} for a more accurate spatial aggregation. As shown in Fig. 3, compared with a single branch baseline, the proposed dual-way mechanism is able to realize a compensation and correction on original input, which exhibits an accurate image enhancement. Finally, we introduce a post-refinement operation by employing a convolution kernels with size of 3×3

TABLE I
COMPARISON BETWEEN OUR MODELS AND OTHER METHODS ON THE PSNR/SSIM INDEXES, IN WHICH THE **RED** INDICATES THE FIRST PLACE. WE USE THE MEGA BYTE(MB) TO REFLECT PARAMETER NUMBER

Methods		×2		×3		×4		Parameter	Time
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
Traditional SR	SRResNet	33.79	0.918	30.52	0.858	29.03	0.821	4.9	2.1
	EDSR	33.57	0.916	30.13	0.847	28.77	0.812	153.5	8.7
	RCAN	33.76	0.917	30.36	0.861	28.89	0.820	54.8	3.61
Real SR	EDRN	33.72	0.920	30.21	0.852	28.79	0.809	131.9	0.29
	KPN(K=5)	33.41	0.913	30.47	0.860	28.80	0.826	5.1	0.042
	KPN(K=7)	33.42	0.913	30.49	0.861	28.84	0.826	5.2	0.047
	KPN(K=13)	33.44	0.913	30.52	0.863	28.92	0.829	5.5	0.067
	KPN(K=19)	33.45	0.914	30.57	0.864	28.99	0.832	5.9	0.087
	LP-KPN(K=5)	33.49	0.917	30.60	0.865	29.05	0.834	5.7	0.138
Our		**34.37**	**0.932**	**31.08**	**0.880**	**29.23**	**0.837**	5.7	0.123

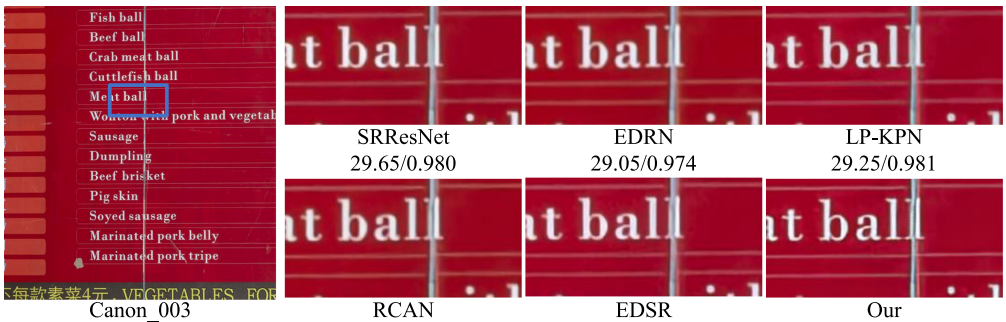


Fig. 4. Visual comparison for 2 × SR. The best results in term of PSNR/SSIM are highlighted.

on I'_{Att} for additional enhancement and restoration:

$$I_{Att} = \mathbf{F_{PR}}\left(I'_{Att}, \Theta_{PR}\right) \tag{11}$$

III. EXPERIMENT

Dataset and Baselines: We employ real-world single image super-resolution dataset(RealSR) [20] to evaluate the proposed method. RealSR dataset has 3,147 images, which are collected from 559 scenes. The images are mainly captured from Canon 5D3 and Nikon D810 devices. The cross-sensor image pair are aligned with iterative pixel-wise registration and alignment. RealSR dataset contains three scale(i.e., 2× , 3× and 4×). Each scale has 400 image pairs for training and 100 image pair for testing. Meanwhile, the training samples are collected from 459 scenes and testing images are collected from 100 scenes. To justify the effectiveness of our model, we conduct traditional image SR approaches, including SRResNet [26], RCAN [17] and EDSR [27] for comparison. Furthermore, we also employ real image SR methods, including KPN [20], EDRN [16], for comparison.

Comparison: As Table I shown, our model achieves promising results on the real image SR benchmark. Compared with LP-KPN, our model achieves 0.88 dB promotion. Compared with RCAN [17], our method exhibit a 0.61 dB gain. It notes that EDSR gets into overfitting as the training sample number is limited. As shown in Fig. 4, our model restore a clear structure with less artifact, which justifies that the restoration quality improvement of our model is significant.

TABLE II
ABLATION STUDY ON REALSR DATASET WITH SCALE × 2

Algorithm	Plain	w/ PR	w/ CDM	w/ MDA
PSNR	33.43	33.80	34.28	34.37

Ablation study: In the Table II, we make an ablation study to justify the effectiveness of each component. We use the structure of the KPN(K = 7) as the 'plain' model. We append post-refinement on it(i.e., 'w/PR') and obtain 0.37 dB gain. Then, we analyze the effectiveness of CDM and MDA by incorporating 'w/CDM' and 'w/MDA'. As Table II show, the CDM bring 0.48 dB promotion and the MDA give 0.09 gain. Since the 'w/CDM' achieves competitive result yet, the MDA component still brings significant improvement.

IV. CONCLUSION

In this article, we propose a dual-path dynamic enhancement network for real image SR. To correct and relax pixel displacement, we incorporate a lightweight model for detail manipulation. To capture multi-scale information, we use multiple dynamic kernels with various sizes for information aggregation. Experimental results well justify the effectiveness of the proposed dual-path model.

REFERENCES

[1] Y. Shi, L. Guanbin, Q. Cao, K. Wang, and L. Lin, "Face hallucination by attentive sequence optimization with reinforcement learning," *IEEE Trans. Pattern Anal. Mach. Intell.* to be published, doi: 10.1109/TPAMI.2019.2915301.

[2] C. Ren, X. He, and Y. Pu, "Nonlocal similarity modeling and deep CNN gradient prior for super resolution," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 916–920, Jul. 2018.

[3] D. L. Cosmo and E. O. T. Salles, "Multiple sequential regularized extreme learning machines for single image super resolution," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 440–444, Mar. 2019.

[4] J. Chu, J. Zhang, W. Lu, and X. Huang, "A novel multiconnected convolutional network for super-resolution," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 946–950, Jul. 2018.

[5] Z. Lu, X. Jiang, and A. Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 526–530, Apr. 2018.

[6] Y. Shi, K. Wang, C. Chen, L. Xu, and L. Lin, "Structure-preserving image super-resolution via contextualized multitask learning," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2804–2815, Dec. 2017.

[7] Y. Wang, L. Wang, H. Wang, and P. Li, "Information-compensated downsampling for image super-resolution," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 685–689, May 2018.

[8] A. Ignatov *et al.*, "Pirm challenge on perceptual image enhancement on smartphones: Report," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 315–333.

[9] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, 2002.

[10] T. Tirer and R. Giryes, "Super-resolution via image-adapted denoising CNNs: Incorporating external and internal learning," *IEEE Signal Process. Lett.*, vol. 26, no. 7, pp. 1080–1084, Jul. 2019.

[11] J. Y. Cheong and I. K. Park, "Deep CNN-based super-resolution using external and internal examples," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1252–1256, Aug. 2017.

[12] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 690–698.

[13] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[14] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[15] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1646–1654.

[16] G. Cheng, A. Matsune, Q. Li, L. Zhu, H. Zang, and S. Zhan, "Encoder-decoder residual network for real super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019.

[17] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 294–310.

[18] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 667–675.

[19] A. Bhowmik, S. Shit, and C. S. Seelamantula, "Training-free, single-image super-resolution using a dynamic convolutional network," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 85–89, Jan. 2018.

[20] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3086–3095.

[21] Y. Yang, D. Zhang, S. Huang, and J. Wu, "Multilevel and multiscale network for single image super-resolution," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1877–1881, Dec. 2019.

[22] L. Huang, J. Zhang, Y. Zuo, and Q. Wu, "Pyramid-structured depth map super-resolution based on deep dense-residual network," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1723–1727, Dec. 2019.

[23] W. Yang, W. Wang, X. Zhang, S. Sun, and Q. Liao, "Lightweight feature fusion network for single image super-resolution," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 538–542, Apr. 2019.

[24] X. Liang *et al.*, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1386–1394.

[25] B. De Brabandere, X. Jia, T. Tuytelaars, and L. Van Gool, "Dynamic filter networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 667–675.

[26] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[27] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1132–1140.