Super-Resolution Imaging Using a Focus Pixel Sensor

Sung-Min Woo*, Jeong-Won Ha† and Jong-Ok Kim‡
* Korea University of Technology and Education, Korea
E-mail: innosm@koreatech.ac.kr
† Korea University, Korea
E-mail: jwon9339@korea.ac.kr
‡ Korea University, Korea
E-mail: jokim@korea.ac.kr

***Abstract*—Modern camera sensors are equipped with a focus pixel, a special type of pixel that can collect separate light rays from the left (L) and right (R) directions. The phase difference between the corresponding L/R pixels is utilized to facilitate quick auto-focusing. In this study, we expand the usability of the special pixels to super-resolve an image. We design a neural net to best fuse multiple low-resolution focus pixel images with a normal image based on repetitive channel and spatial attention layer structures. Empirical results show that focus pixel images contribute to the creation of fine details by providing additional information to super-resolve an image, especially for textured areas and that the proposed neural net-based method enhances the state-of-the-art super-resolution methods that do not use focus pixels in quantitative and qualitative measures.**

I. Introduction

Super-resolution (SR) methods are required to generate a high resolution images in many smartphone applications. For example, smart phone cameras typically do not have optical zoom lenses due to space and cost constraints. When zooming is needed, a common method is to crop the required portion of the entire sensor area and then up-sample or super-resolve the cropped image to restore the original image size. However, blind super resolution methods inherently lack of high-frequency spatial information. Various methods have been proposed over the years to tackle this problem [1], [2], [3], [4], [5]. In the meantime, deep-learning techniques have been successfully applied and have broken records in all literature. SR has also been greatly improved by deep learning techniques [6], [7], [8], [9]. Nevertheless, obstacles remain in accurately recovering the missing content. In general, the performance of SR increases as the depth of the network increases, which cannot be implemented in smartphone cameras for real-world use cases.

Instead of designing a heavy and complicated network structure, we leverage focus pixels to enhance SR by providing the network with more convincing information hidden in the focus pixels. Focus pixels in modern camera consist of pairs of left (L) and right (R) pixels that separate light from the L/R directions. The L/R pixel groups are created as L and R images with spatial phase difference. The two images finally construct a single disparity map for fast and smooth auto-focus, reducing unnecessary lens searches. Some studies have demonstrated that utilizing the focus pixels is highly beneficial in areas such as depth estimation [10], reflection removals [11], de-blurring [12], and HDR [13].

In this study, we proposed a new deep-learning method for super-resolving a single image from a focus pixel sensor. Even though the eight L/R images from the focus pixels in the study have 1/8 resolution to the corresponding low resolution image at the input, they are jointly fused and extracted into useful features through the novel dual attention structure, and they are concatenated with the feature of the original low resolution image. The combined features are up-scaled with pixel shuffling step by step to construct a high-quality high-resolution image at output. To the best of our knowledge, this is the first attempt exploiting focus pixels at SR imaging.

Section II is a review of the related works and Subsection III-A briefly explains the focus pixel sensor used in the study. The proposed deep neural network architecture is introduced in the rest of Section III. The experimental results and ablation studies are presented in Section IV. Finally, the conclusions are discussed in Section V.

II. Related Work

Single image SR (SISR) methods are mainly divided into several branches. Interpolation-based SR methods such as bilinear and bicubic interpolation [14] are the simplest and most straightforward techniques to obtain a high resolution image. These methods are widely used in many computer vision system due to their high processing speed, but suffer from the blurring of the resulting image. Reconstruction-based SR methods have been proposed to overcome this problem [15], [16], [17]. They often generate sharp details by regularizing output pixels to lie in a predefined space using image priors. Learning-based (or example-based) SR methods [18], [19] learn the statistical relationships between the low resolution (LR) and the corresponding high resolution (HR) images exploiting external datasets. These methods utilize LR-HR feature matching to produce visually pleasing image textures.

Multiple image SR (MISR) methods take the advantage of variations in sample data at slightly different locations and time intervals [1], [4]. Unlike SISR, these methods have

demonstrated reliable performances on real-world stationary scenes due to the additional pixel information. Deep-learning based methods have shown outstanding performances on both SISR [6], [20], [7] and MISR [21]. However, these methods are cumbersome to apply to lightweight mobile processing engines because they rely on an explosive number of parameters and large computations to generate plausible and sharp images.

Focus pixel sensors were first introduced to the mobile market in the early 2010s. Researchers have recently become interested in the focus pixel sensor because its use cases were limited in auto-focusing in its infancy. The sensor has L/R subpixels in a single micro-lens, and they create parallax to measure the distance to an object in front, just like the human eyes do. Various types of focus pixel sensors have been developed. Some dual pixel sensors have a pair of L/R subpixels for every single pixel on the sensor. For high-resolution imaging, only a limited percentage of the pixels are used for focus pixels.

A few researchers have conducted on studies on applying the focus pixel images to computational photography such as depth extraction [10] and reflection removal [11]. Instead of using a stereo camera or depth sensor, the L/R subpixels in the focus pixel sensor played a key role in estimating depth information in these methods, proposing that a focus pixel sensor is an alternative to a distance-measuring device. Interestingly, focus pixels were also used in recent deblurring [12] or HDR [13] studies that seemed unrelated to depth extraction, but they also showed a huge improvement over the previous methods that did not use focus pixels in those studies. Motivated by the fact that the L/R pixels in focus pixel sensors can be a useful resource to train a deep neural-net in a more robust manner, we propose a novel deep learning-based SR method that best utilizes the focus pixels by combining them with the conventional RGB images. The proposed method has a relatively simple neural-net structure, but outperforms the existing state-of-the-art SISR methods that do not use focus pixels.

III. Proposed Method

A. Focus pixel sensor

The focus pixel sensor in this study is a type of a sensor that performs L/R beam separation in the on-chip lens (OCL) as depicted in Fig. 1(a), and the corresponding sensor array pattern is illustrated in Fig. 1(b). The total number of focus pixel marked by yellow in Fig. 1(b) is approximately 1/8 of the total number of pixels in the sensor. Unlike the conventional RGGB or RGBG Bayer pattern, this type of sensor has a 2×2 array of pixels with the same color filter. In the sum-binning operating mode, the photons of the 2×2 pixel array of the same color are merged and converted into a single pixel value for better sensitivity. The focus pixels are usually darker than the merged RGB pixels for this reason.

From the raw sensor structure in Fig. 1(b), a full-resolution RGB and eight focus pixel images are obtained by the process depicted in Fig. 2. The merged Bayer raw is demosaiced, white-balanced, color corrected, and converted into a sRGB image sequentially. The resulting RGB images are used as high-resolution images for groundtruth in this study. Focus pixels are grouped in the form of an image depending on the same location, and the resulting eight images are directly converted to sRGBs and stacked channel-wise because they are achromatic. Please refer to [13] for a more detailed explanation.

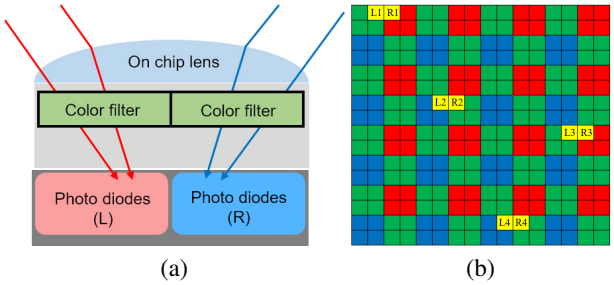

Fig. 1. (a) Separation of light on the focus pixel. (b) schematic diagram of on-chip lens focus pixel sensor. The focus pixels are marked as yellow in (a) [13].

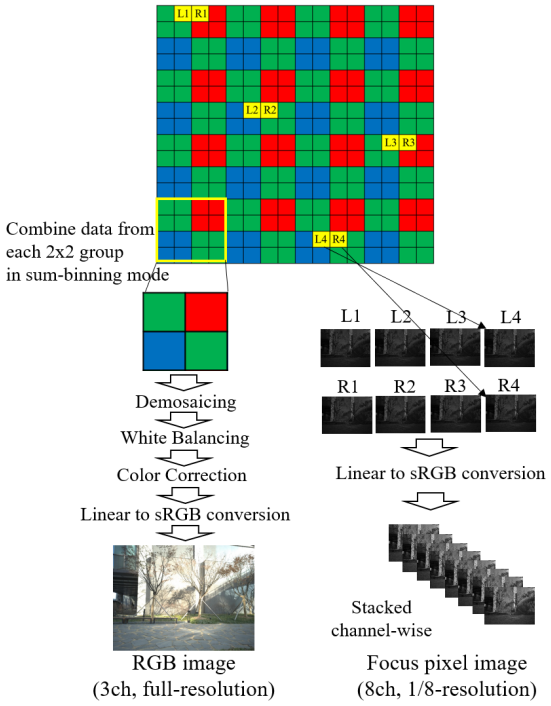

Fig. 2. Decomposed RGB and focus pixel images from the raw image of the focus pixel sensor for inputs [13].

B. Proposed Network Architecture

The proposed network uses a low resolution (LR) image and focus pixel images as input and estimates a high resolution image. The LR image and the corresponding focus pixel images are made by bicubic downsampling of the original RGB image and focus pixel images as described in Section

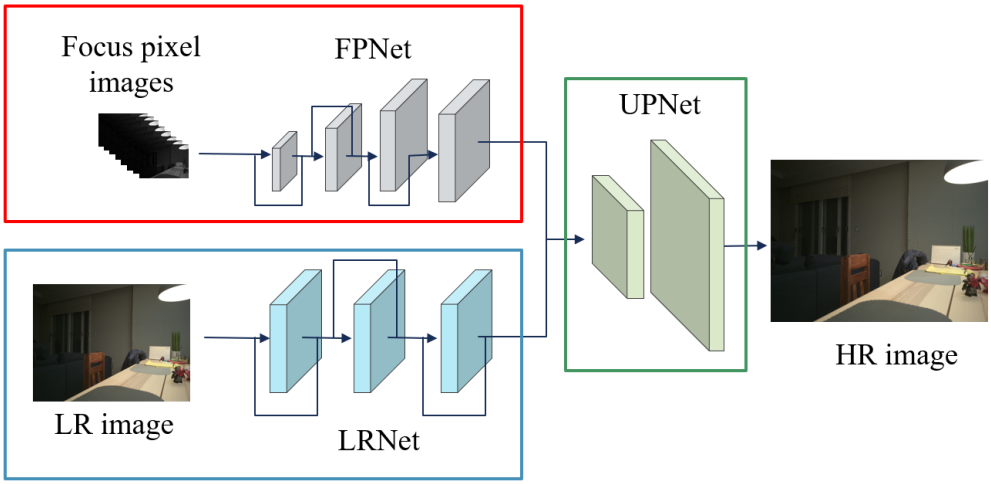

Fig. 3. The proposed network. The network use a LR image and the corresponding focus pixel images as input and a generates HR image.

III.A. Fig. 3 shows the network architecture of the proposed method. The proposed network includes two sub-branch nets, LRNet and FPNet, to extract the features of the LR image and focus pixel images independently. The features of FPNet and LRNet are concatenated and upscaled in UPNet. Since the resolutions of the focus pixel and LR image are different, the process matching the resolution of two features is required for fusion. The resolution of the LR image is set to the reference and the features of focus pixel images are upsampled to the reference resolution through FPNet.

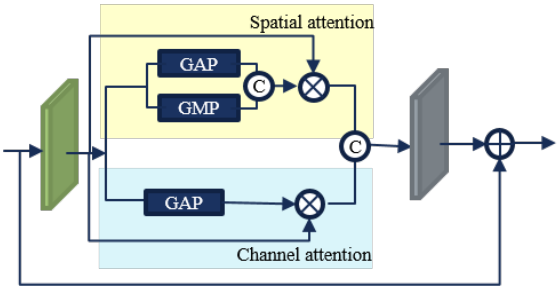

Fig. 4. Structure of Dual Attention Block. It consists of two sub-branch layers performing spatial and channel attentions.

1) Network structure: The LR image and the corresponding focus pixel images (L1-R4) are fed into LRNet and FPNet, respectively. LRNet is constructed with Residual convolution block, and the feature size in LRNet does not change. FPNet is composed of Dual Attention Block (DAB) [22], [23] and pixel-shuffling layers [24]. FPNet performs upscaling by extracting the features of the focus pixels. To effectively infer the lost spatial detail from the focus pixel images which is 1/8 of the LR image, upscaling is done incrementally in three steps with a pixel-shuffling operation.

Fig. 4 illustrates the structure of DAB used in FPNet. DAB conducts channel and spatial attentions in parallel. For the spatial attention process, Global Average Pooling (GAP) and Global Max Pooling (GMP) are operated along channel axis, and the resulting output features are concatenated thereafter. A convolution layer and sigmoid activation are followed. The channel attention branch applies GAP, and a convolution layer and sigmoid activation are followed for obtaining the attention map. The resolution of the focus pixels is not equivalent to the LR image, and thus the additional information provided by the focus pixel is limited. However, DAB learns the location and channel which are more important to find missing details among the eight focus pixel images, through the channel and spatial attention branches.

UPNet uses the concatenated features of FPNet and LRNet, and generates an image with desired resolution. UPNet is composed of residual convolution blocks and pixel-shuffling layers.

2) Training details: For training, the L1 reconstruction loss between the generated SR image and the groundtruth HR image is applied, and is expressed as follows:

$$L_{recon} = \frac{1}{N} \sum_{i \in \hat{I}} | \hat{I}(i) - I(i) |, \tag{1}$$

where N is the number of pixels in an image patch and i is the pixel index of the estimated SR image $\hat{I}$ and the ground truth image HR image I.

The dataset in this study was created for the purpose of studying HDR imaging [13]. Many of the RGB and focus pixel image pairs contain flat regions with no edges, which is not appropriate for this study. In order for the proposed network to learn the details more efficiently, we used a patch with a mean of the sobel-filtered values greater than the threshold for training. For optimization, Adam is used with the batch size of 4 and the learning rate of 1×10^{-4}. The resolution of the original HR image and focus pixel image is 4624 $\times$ 3456 and

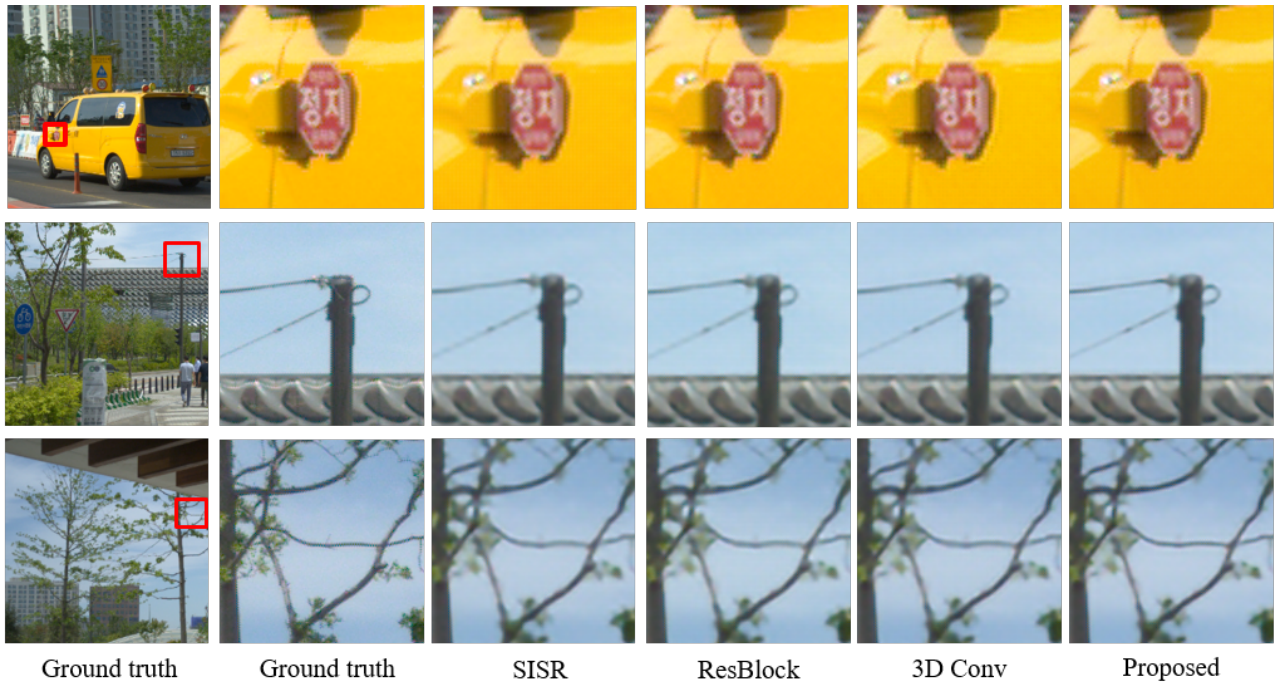

Fig. 5. Visual comparisons. The first row is the resulting images of an upscaling factor of 2, and the second and third rows are the resulting images of an upscaling factor of 4. Note that the proposed method creates more sharp and clear details due to the fusion of the focus pixels through DAB.

TABLE I
PSNR COMPARISONS FOR REAL-WORLD DATASET

	Bilinear	Bicubic	SISR	The proposed (DAB)	ResBlock	3D Conv
$\times 2$	35.803	36.562	39.827	**40.667**	40.335	40.199
$\times 4$	32.710	33.215	34.586	**34.764**	34.474	34.683

578 $\times$ 432, respectively.

IV. Experimental Result and Ablation Study

For training and evaluating the proposed method, the dataset that contains the LR-HR image pairs and the corresponding focus pixel images is required. We utilized the focus pixel dataset performed in the previous HDR study [13]. For training and test, 640 and 107 scenes are used, respectively. Evaluations were performed on upscaling factors 2 ($\times 2$) and 4 ($\times 4$).

Several ablation studies are conducted to confirm the effectiveness of the proposed method. 1) The FPNet is removed to evaluate the effect of focus pixel images, which is equivalent to the single image SR method denoted as SISR in Table 1. 2) DAB is replaced by Residual convolution block and 3D convolution block. In this case, unlike DAB in the proposed network, Residual convolution block does not learn the channel-wise relationship, which resulted slight less quantitative scores both in $\times 2$ and $\times 4$ denoted as ResBlock in Table 1. 3) 3D convolution has been proposed for video SR to exploit temporal features between multiple frames [25]. 3D convolution extracts the features in spatial and channel domains simultaneously, whereas DAB does it separately.

Table 1 summarizes the quantitative measure for average PSNR scores for upscaling factors of $\times 2$ and $\times 4$ for the comparison methods. The proposed network achieved the highest PSNR scores in both x2 and x4 cases among all comparison methods. In common, the methods using focus pixels, the proposed, ResBlock and 3D Conv, performed better than SISR without focus pixels, which proves that deep-learning based methods fuses the focus pixel information to effectively reconstruct fine details. Also, the proposed network (using DAB) outperformed ResBlock or 3D Block in both x2 and x4 cases, demonstrating that exploiting channel-wise relationships with DAB finds more useful features for missing details.

Fig. 5 illustrates the visual comparisons of the proposed and the ablation studies for real-world images. The first row is the resulting images of an upscaling factor of 2, and the second and third lows are the resulting images of an upscaling factor of 4. The red box regions of the first column images are zoomed in for better visibility. The proposed method consistently generates sharp images with less blurred, compared with SISR in both x2 and x4 cases. In the case of x4, the proposed method looks slightly compelling images

than the ResBlock and 3D conv.

V. Conclusion

In this study, we proposed a new method to exploit a focus pixel image in SR imaging. Focus pixels provide a left and right view image by splitting light ray bidirectionally. The parallax between the left and right pixels helps assisting auto-focusing. Inspired by the fact that the half-divided raw pixels are still a resource to recover high frequency details in SR imaging, we demonstrated that these pixels are successfully fused to an original LR image to enhance the performance SR with the novel proposed network architecture. The quantitative measure and visual comparison showed that the proposed method using DAB makes the best use of focus pixels among ablation studies.

Acknowledgment

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1005834) and the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2020-0-01749) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation)

References

[1] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C.-K. Liang, M. Levoy, and P. Milanfar, "Handheld multi-frame super-resolution," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–18, 2019.

[2] S. P. Belekos, N. P. Galatsanos, and A. K. Katsaggelos, "Maximum a posteriori video super-resolution using a new multichannel image prior," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1451–1464, 2010.

[3] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE transactions on image processing*, vol. 6, no. 12, pp. 1646–1658, 1997.

[4] S. Farsiu, M. Elad, and P. Milanfar, "Multiframe demosaicing and super-resolution of color images," *IEEE transactions on image processing*, vol. 15, no. 1, pp. 141–159, 2005.

[5] C. Liu and D. Sun, "A bayesian approach to adaptive video super resolution," in *CVPR 2011*. IEEE, 2011, pp. 209–216.

[6] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[7] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[8] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[9] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 252–268.

[10] A. Punnappurath and M. S. Brown, "Reflection removal using a dual-pixel sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1556–1565.

[11] R. Garg, N. Wadhwa, S. Ansari, and J. T. Barron, "Learning single camera depth estimation using dual-pixels," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7628–7637.

[12] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *European Conference on Computer Vision*. Springer, 2020, pp. 111–126.

[13] S.-M. Woo, J.-H. Ryu, and J.-O. Kim, "Ghost-free deep high-dynamic-range imaging using focus pixels for complex motion scenes," *IEEE Transactions on Image Processing*, vol. 30, pp. 5001–5016, 2021.

[14] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE transactions on acoustics, speech, and signal processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[15] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, and A. K. Katsaggelos, "Softcuts: a soft edge smoothness prior for color image super-resolution," *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 969–981, 2009.

[16] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[17] Q. Yan, Y. Xu, X. Yang, and T. Q. Nguyen, "Single image superresolution based on gradient profile sharpness," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3187–3202, 2015.

[18] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[19] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1920–1927.

[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, Cham, 2014, pp. 184–199.

[21] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3224–3232.

[22] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Cycleisp: Real image restoration via improved data synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2696–2705.

[23] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[24] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. 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.

[25] S. Li, F. He, B. Du, L. Zhang, Y. Xu, and D. Tao, "Fast spatio-temporal residual network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 522–10 531.