MULTI-FRAME SUPER-RESOLUTION WITH RAW IMAGES VIA MODIFIED DEFORMABLE CONVOLUTION

Gongzhe Li, Linwei Qiu*, Haopeng Zhang†, Fengying Xie, Zhiguo Jiang*

Department of Aerospace Information Engineering (Image Processing Center),
School of Astronautics, Beihang University, Beijing 102206, China

ABSTRACT

In this paper we propose a novel model towards multi-frame super-resolution, which leverages multiple RAW images and yields a super-resolved RGB image. To facilitate the pixel misalignment in burst photography, we apply a refined Pyramid Cascading and Deformable Convolution (PCD) feature alignment module. A new 3D deformable convolution fusion module is proposed subsequently to merge the information from all frames adaptively. In addition, we employ an encoder-decoder network to restore color and details in sRGB space after super-resolving images in linear space. Extensive experiments demonstrate the superiority of our architecture and the strength of multi-frame super-resolution with RAW images.

Index Terms— multi-frame super-resolution, RAW image super-resolution, deformable convolution

1. INTRODUCTION

Super-resolution (SR) is a fundamental computer vision problem with numerous applications in mobile photography, remote sensing, medical imaging, etc. Given an image or multiple images of a sense, SR aims to restore a higher-resolution image by adding missing high-frequency details. According to the form of the low-resolution (LR) input, the SR task can be divided into single image super-resolution (SISR) and multi-frame super-resolution (MFSR). There have been many innovations both in model architectures and training strategies [1, 2, 3, 4, 5, 6] for SISR. Nevertheless, SISR methods are fundamentally limited by the available information (*i.e.* single frame), thus only relying on learned image priors to recover missing details. In contrast, MFSR methods aim to restore high-resolution details from multiple spatial-temporal correlation LR images. When multiple-frame images are captured by a non-stationary camera, there is pixel misalignment among multiple-frame images owing to slight camera motion. These different images can be regarded as multiple samples of the same underlying scene [7, 8]. Therefore, MFSR methods can obtain additional information from image misalignment and achieve better performance than SISR task.

Recently, a few deep-learning-based methods have been proposed for MFSR. A MFSR framework termed HighResnet [9] was developed for satellite imagery, which first aligns every input to one reference frame and then merges them by a recursive fusion method. Molini *et al.* [10] assumed only translation motion among frames and introduced 3D convolution to fuse multi-frame features. While these methods focused on remote sensing and have limitation in producing realistic details, Bhat *et al.* [11] proposed a network with raw burst images, which applied cumbersome optical flow [12] to align multiple frames. Different from its complicated designs, we catalyze the potential of deformable convolution [13]: the refined PCD structure is introduced to do feature alignment and a novel 3D deformable convolution is proposed to fuse features. Additionally, they trained the whole model in linear space and simply assumed the color mapping was linear to get a RGB image. A color correction module is developed in our paper to foster color details. On the other hand, previous RAW SR methods [11, 14] usually pack RAW data as input and then apply demosaic operation (it needs extra ×2 super-resolution) to restore the spatial information. Unlike this normal way, our model can implicitly learn demosaicing to improve the performance.

The main contributions of this paper are listed as follows: (i) We propose a novel MFSR model to exploit additional raw information and achieve better SR performance by fusing multiple frames. (ii) We employ the refined PCD structure for feature alignment and develop a novel 3D deformable convolution fusion module for feature fusion to generate high quality HR images. (iii) We utilize a color correction module to transform the feature from linear space to sRGB space and our network can demosaic implicitly while super-resolving.

2. PROPOSED METHOD

For a given RAW sequence of N images, $\mathbf{I}^{\mathrm{LR}}_{\mathrm{RAW}} = \{I^{\mathrm{LR}}_{\mathrm{RAW}_t}\}^{N}_{t=1}$, where $I^{LR}_{\mathrm{RAW}_1}$ is selected as the reference, our goal is to super-resolve $I^{\mathrm{LR}}_{\mathrm{RAW}_1}$ with the help of remaining frames, and yield a

*The first two authors contributed equally. Work supported in part by the National Key Research and Development Program of China (Grant No. 2019YFC1510905) and the Fundamental Research Funds for the Central Universities.

†Corresponding Author.

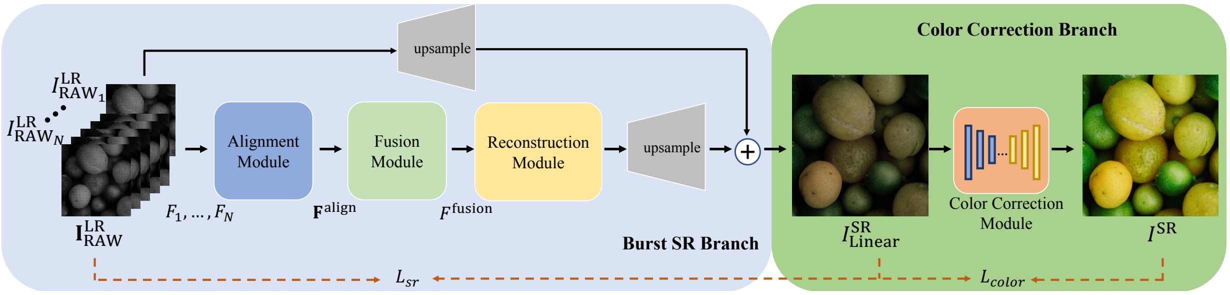

Fig. 1. An overview architecture of our method. The proposed method consists of the burst SR branch and the color correction branch. The burst SR branch generates HR details in linear space from the RAW data. Subsequently the color correction branch supplements the color information for the super-revolve results. The two branches are trained separately.

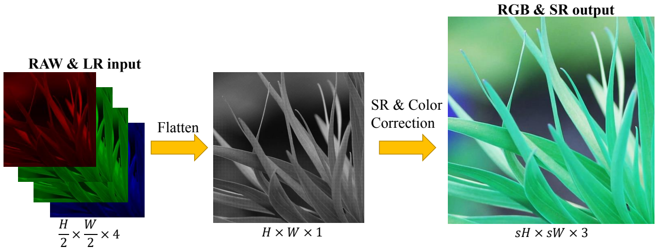

Fig. 2. RAW data pipeline of our method. Our approach can demosaic implicitly without extra ×2 super-resolution.

high-resolution color image I^{SR}.

2.1. Network Architecture

An overview of our architecture is shown in Fig. 1. Our network can be divided into two cascading branches, i.e. the burst SR branch and the color correction branch. As depicted in Fig. 2 , the RAW data are flattened to obtain a higher resolution and then super-resolution and color correction are conducted through our method.

2.2. Alignment Module

One challenge in MFSR is the input sequence encompasses shifted RAW images with unknown displacements which stem from both global camera motion and scene variations. The optical flow networks [12] are widely used to align the target frame to the reference frame. Nevertheless, the optical flow method is tricky to deal with the optical parallax, which makes it unsuitable for burst images. An effective method is to explicitly align the individual image to a common reference image. To tackle this problem, we adopt the modulated deformable modules [13] to perform feature level alignment between images inspired by Pyramid Cascading and Deformable convolution (PCD) module [15, 16]. The PCD feature alignment module applies the concatenate on the feature of reference frame (denoted as F_r) and the features of remaining frames (denoted as F_t) to get the offsets for deformable convolution layer (DCN). Then DCN generates the aligned features by these features and offsets. This operation will be applied on the 3-level pyramid to fuse them to get the aligned features, *i.e.*

$$F_t^{\text{align}} = \begin{cases} \textbf{PCD}(F_r, F_t), t \neq r \\ F_t, t = r \end{cases}, t \in [1, N] \tag{1}$$

$\mathbf{F}^{\text{align}} = \{F_t^{\text{align}}\}_{t=1}^{N}$ denotes the set of aligned features from reference and remaining frames.

2.3. Fusion Module

For a image sequence $\mathbf{I}^{\text{LR}}_{\text{RAW}}$, it is very easy to concatenate the input aligned features and apply multi-channel convolution along the channel dimension. However, this operation treats each frame and each pixel equally, and cannot make full use of the information. To remedy this information loss, deformation convolution [13] is introduced to perform feature fusion. The plain 3D deformable convolution uses the offsets both on spatial and temporal dimension, which can easily extract deeper features. When used for MFSR task, the temporal offsets usually focus on the reference frame and ignore the other frames. To solve this issue, we design a novel 3D deformable convolution by applying deformable convolution on spatial dimension and paying more attention to frames temporally closer to the reference frame. Given a deformable convolution kernel of K sampling locations [13], $\mathbf{p}_k$ is the pre-specified offsets for the k-th location (K is the total sampling location). The proposed 3D deformable convolution fusion module can be modified from Conv3d to augment the regular sampling offsets with extra learnable offsets $\Delta\mathbf{p}_{[t,\mathbf{p}],k}$, *i.e.*

$$F^{\text{fusion}} = \sum_{t=1}^{N} \sum_{k=1}^{K^2} \omega_{t,k} \cdot F_t^{\text{align}}(\mathbf{p} + \mathbf{p}_k + \Delta\mathbf{p}_{[t,\mathbf{p}],k}) \tag{2}$$

where $\omega_{t,k}$ represents the weight of convolution kernel. As illustrated in Fig. 3, the offset prediction network generates the offsets from the aligned features of each frame. The learnable offsets $\Delta\mathbf{p}_{[t,\mathbf{p}],k}$ are position-specific, which will be assigned for each convolution layer at spatial and temporal position $\Delta\mathbf{p}_{[t,\mathbf{p}]}$ to guide this module to fuse features. The offsets will be applied for each frame, leading to modeling the spatial deformations and temporal features.

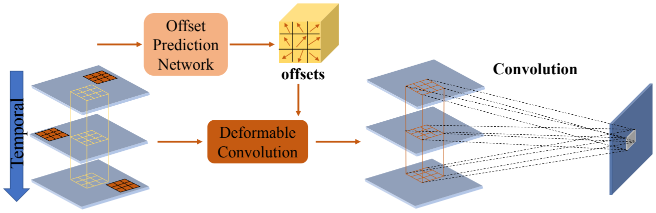

Fig. 3. An example of our 3D deformable convolution. The offsets are generated by an offset prediction network (simply 3×3×3 Conv layer), which represent the deformation position on spatial dimension of 3D deformable convolution.

2.4. Reconstruction Module

The reconstruction module generates the output high-resolution image from the fused feature maps. We use 16 residual blocks as the reconstruction module. The upsample layer is comprised of Pixel-Shuffle [2] and a 3×3 convolution layer. Moreover, we introduce an extra residual pathway structure as the low-frequency information to reconstruct the final SR image, as shown in Fig. 1.

2.5. Color Correction Module

$I^{\text{SR}}_{\text{Linear}}$ is obtained by the burst SR branch, which lacks diversified color and bright information compared with the I^{SR} in sRGB space. Therefore, we design the second branch for color correction, for which many traditional methods use post-processing *e.g.* color correction, gamma compression or tone-mapping to restore the RGB image for a better perceptual quality. DBSR [11] used I^{SR} as a reference to estimate the 3×3 color matrix between I^{SR} and $I^{\text{SR}}_{\text{Linear}}$. However, this global correction strategy dose not work well when the color transformation involves spatially-variant operations. To solve this problem, we adopt a encode-decoder structure to learn the pixel-wise transformation. Thus, we can generate the final results as $I^{\text{SR}} = M_c(I^{\text{SR}}_{\text{Linear}})$, where $M_c(\cdot)$ stands for the color correction branch.

2.6. Loss Function

SR Loss: Our burst SR branch is similar to the most SR methods in a fully supervised manner by minimizing the L1 loss between the branch prediction and the ground truth image computed in the linear sensor space, *i.e.*

$$L_{sr} = \left\| I^{\text{HR}}_{\text{Linear}} - I^{\text{SR}}_{\text{Linear}} \right\|_1 \tag{3}$$

Color Loss: To minimize the color difference between the linear space and the sRGB space, we propose a mixed color loss for color correction branch to restore details and textures. It linearly combines the L1 loss L_1 and Laplacian loss L_{lap} between I^{HR} and I^{SR} as

$$L_{color} = \lambda_1 \cdot L_1 + \lambda_2 \cdot L_{lap} \tag{4}$$

where the coefficients of loss functions in this paper are set as $\lambda_1 = 1, \lambda_2 = 0.5$. L_1 compares the difference between two images pixel by pixel as

$$L_1 = \left\| I^{\text{HR}} - I^{\text{SR}} \right\|_1 \tag{5}$$

L_{lap} uses the Laplacian pyramid [17] to represent color information and calculate 5 levels of L1 loss as

$$L_{lap} = \sum_{j=1}^{5} 2^{2j} \left\| \phi^j(I^{\text{HR}}) - \phi^j(I^{\text{SR}}) \right\|_1 \tag{6}$$

where ϕ^j represents the j-th level of the Laplacian pyramid representation. Laplacian loss blurs the details of an image thus focusing on the color of this image.

3. EXPERIMENTS

3.1. Dataset

Following the previous methods [11], we perform extensive experiments on our synthetic dataset which is produced using publicly available codes [1]. We first synthesize the ground truth of linear color space $I^{\text{HR}}_{\text{Linear}}$ by applying inverse ISP pipeline [18] for I^{HR} so that each pixel can have its own red, green and blue values. Then, the translation and rotation are randomly generated using Euclidean motion, and each frame is downsampled by bilinear interpolation to simulate an LR frame containing aliasing. Specially, we sample the translation and rotation independently from the range [-24,24] pixels and [-1,1] degrees respectively. Two color channels of per pixel are discarded obeying the Bayer pattern to obtain the mosaicked RAW bursts. To cover the image as much as possible, we crop the image to a fixed size. Synthetic data used for training and testing models are generated through DIV2K train dataset and valid dataset,respectively.

3.2. Implementation Details

We use Adam optimizer with momentum 0.9. For training burst SR branch, the learning rate is set to 2×10^{-4} and then reduced to half every 200 epochs. For color correction branch, the learning rate is 1×10^{-4}. Then we train two branches jointly for a fine-tune with L1 loss. We implement the proposed method using PyTorch framework with 2 NVIDIA 2080Ti GPUs. We report popular PSNR and SSIM to evaluate SR performance. We test 10 times to calculate the mean and standard deviation for each model configure to reduce the impact of random crop. Unless mentioned, all the models are evaluated with a burst size of 8.

3.3. Analysis of Experimental Results

Comparison with existing approaches. We compare our burst SR branch with several representative methods on our synthetic dataset in Table 1. Notice that these methods are hardly able to obtain RGB images directly when applied to MFSR with RAW data. Color correction branch is all employed to help transform $I^{\text{SR}}_{\text{Linear}}$ to I^{SR} for fairness. As tabulated in Table 1, our model outperforms existing methods

[1] https://github.com/goutamgmb/NTIRE21_BURSTSR

Table 1. Quantitative results in comparison with the existing methods. Average and standard deviation of PSNR/SSIM for scale×2,×3,×4 on BSD100, Urban100 and Manga109 datasets. We crop the image for 240 × 240 on BSD100, and 512 × 512 on Urban100 and Manga109. The best performance is **highlighted**.

Method	Scale	BSD100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
HighResNet [9]	x2	29.914±0.108	0.871±0.002	27.833±0.120	0.880±0.002	29.498±0.168	0.941±0.001
DeepJoint [19]+RRDB [4]	x2	30.829±0.030	0.882±0.001	30.382±0.067	0.900±0.002	33.167±0.371	0.958±0.001
EDSR [3]	x2	30.097±0.064	0.872±0.001	28.468±0.112	0.886±0.002	30.078±0.138	0.945±0.001
EDVR [15]	x2	34.123±0.042	0.949±0.001	32.147±0.136	0.945±0.002	34.431±0.260	0.973±0.001
DBSR [11]	x2	30.378±0.190	0.879±0.002	29.978±0.113	0.901±0.001	32.699±0.144	0.955±0.001
Ours	x2	**35.509±0.103**	**0.963±0.001**	**33.944±0.101**	**0.962±0.001**	**35.449±0.273**	**0.978±0.001**
HighResNet [9]	x3	25.956±0.056	0.734±0.002	23.900±0.125	0.750±0.003	25.925±0.210	0.869±0.002
DeepJoint [19]+RRDB [4]	x3	26.369±0.108	0.746±0.004	25.337±0.114	0.797±0.003	27.947±0.124	0.894±0.001
EDSR [3]	x3	26.282±0.048	0.741±0.002	24.899±0.129	0.781±0.006	27.309±0.162	0.886±0.002
EDVR [15]	x3	30.870±0.074	0.898±0.002	29.209±0.059	0.901±0.002	**32.561±0.170**	0.953±0.001
DBSR [11]	x3	30.535±0.203	0.884±0.005	28.474±0.087	0.886±0.002	32.219±0.123	0.954±0.001
Ours	x3	**30.955±0.066**	**0.898±0.001**	**29.628±0.082**	**0.901±0.002**	32.522±0.109	**0.957±0.001**
HighResNet [9]	x4	25.563±0.091	0.691±0.002	23.393±0.136	0.708±0.005	25.104±0.146	0.831±0.003
DeepJoint [19]+RRDB [4]	x4	25.780±0.057	0.697±0.003	24.544±0.064	0.745±0.004	26.440±0.126	0.856±0.003
EDSR [3]	x4	25.731±0.087	0.695±0.003	23.918±0.127	0.729±0.006	25.897±0.119	0.849±0.002
EDVR [15]	x4	27.059±0.088	0.847±0.002	26.951±0.125	0.845±0.004	30.564±0.190	0.934±0.001
DBSR [11]	x4	29.286±0.100	0.848±0.002	27.130±0.081	0.851±0.002	30.518±0.170	0.933±0.001
Ours	x4	**29.868±0.061**	**0.864±0.011**	**27.512±0.043**	**0.862±0.001**	**30.792±0.150**	**0.938±0.001**

Table 2. Evaluation results of different number of input frames with comparison with SISR network.

	Urban100 ×4	
	PSNR	SSIM
Single Image	24.464±0.128	0.743±0.004
Burst-2	25.224±0.081	0.783±0.004
Burst-4	26.259±0.156	0.823±0.003
Burst-8	27.512±0.043	0.862±0.001
Burst-14	28.255±0.118	0.882±0.002

on BSD100 and Urban100 datasets for all three scales. Fig. 4 demonstrates that our model can recover more details with less blurring.

Parameter efficiency.The DBSR [11] applies a pre-trained optical flow network as the alignment module, the parameter of which is about 37M. It is nearly close to our whole model (38M), demonstrating that the deformation-based convolution alignment module used in our method is lightweight and efficient.

Impact of using multiple frames. We investigate the impact of using multiple frames for SR by comparing our MFSR approach with a SISR baseline. The comparison results are shown in Table 2. It encourages that our approach can effectively utilize the information from multiple frames to improve SR performance.

Impact of fusion module. We evaluate the effectiveness of Laplacian loss (Lap) and 3D deformable fusion module (Fusion) via ablation experiments. The results are summarized in Table 3. When Lap or Fusion is equipped up individually, an improvement about 0.1∼0.2dB in PSNR can be obtained. Moreover, a collaboration of the both can bring nearly 0.3dB increase in PSNR, on account of the full use of the inter-frame and intra-frame information within burst images.

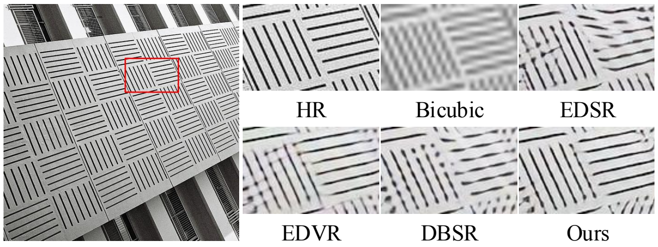

Fig. 4. An illustration of result (×4) on "image_092" from Urban100 dataset.

Table 3. Ablation study on the proposed method. All the models are evaluated for scale×4 on Urban100.

Models	Model1	Model2	Model3	Model4
Lap			✓	✓
Fusion		✓		✓
PSNR	27.213±0.103	27.317±0.130	27.385±0.043	27.512±0.043
SSIM	0.852±0.004	0.857±0.004	0.860±0.001	0.860±0.001

4. CONCLUSIONS

We put forward a novel two-branch network towards MFSR. Our model can align features by a refined PCD feature alignment module while the information from multiple input images is adaptively combinee using a 3D deformable convolution fusion module. We also utilized color correction branch and a designed color loss to restore color details and improve image quality. Experimental results show that our approach can obtain promising SR results from multi-frame RAW data.

5. REFERENCES

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *Proceedings of the European conference on computer vision*. Springer, 2014, pp. 184–199.

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung 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.

[3] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung 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.

[4] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *The European Conference on Computer Vision Workshops*, 2018, pp. 63–79.

[5] 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*, 2018, pp. 286–301.

[6] 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.

[7] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar, "Handheld multi-frame super-resolution," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–18, 2019.

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

[9] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio, "Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery," *arXiv preprint arXiv:2002.06460*, 2020.

[10] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli, "Deepsum: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, 2019.

[11] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte, "Deep burst super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9209–9218.

[12] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[14] Xiangyu Xu, Yongrui Ma, and Wenxiu Sun, "Towards real scene super-resolution with raw images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1723–1731.

[15] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1954–1963.

[16] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3360–3369.

[17] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam, "Optimizing the latent space of generative networks," *arXiv preprint arXiv:1707.05776*, 2017.

[18] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron, "Unprocessing images for learned raw denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11036–11045.

[19] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 191:1–191:12, 2016.