MULTI-LEVEL FUSION FOR BURST SUPER-RESOLUTION WITH DEEP PERMUTATION-INVARIANT CONDITIONING

Martina Cilia, Diego Valsesia, Giulia Fracastoro, Enrico Magli

Politecnico di Torino
Department of Electronics and Telecommunications
Turin, Italy

ABSTRACT

Developing deep learning techniques for super-resolving bursts of images acquired by mobile cameras is a topic that has recently gained significant interest. This topic fits the general problem of learning-based multi-image super-resolution (SR), which, contrary to its sibling single-image SR, has so far received little attention despite its potential. In this work, we introduce a neural network architecture for burst SR, called MLB-FuseNet (Multi-Level Burst Fusion Network), that is capable of extracting features in a manner that is invariant to permutations in the burst and to progressively condition features extracted from a reference image. Permutation invariance is desirable as it is known that the order of images in a burst does not matter in this problem, but its study has so far been neglected. Moreover, we also introduce a module exploiting a polyphase decomposition to improve feature extraction from mosaiced raw images. Results show an improvement over the state of the art on the BurstSR dataset – a recent and popular benchmark for this problem.

***Index Terms*—** burst super-resolution, convolutional neural networks, multi-level fusion, self-attention, demosaicing

1. INTRODUCTION

Super-resolution (SR) is a widely studied image processing problem, that aims at reconstructing a high resolution (HR) image from low resolution (LR) ones. Even if the problem is ill-posed, the research community has achieved impressive results using deep convolutional networks [1][2][3] [4]. However, a single image does not contain enough information to estimate high frequency details, and it has been proved that it is beneficial to combine multiple LR observations, conveying complementary information thanks to sub-pixel shifts [5], employing Multi-Image Super Resolution (MISR) techniques. Several model-based approaches [6, 7] have addressed the problem in the past, seeking to carefully model the disparities among the multiple images and define priors for the HR image.

Today, deep-learning methods have revolutionized the field of image super-resolution [8], but deep learning research on MISR is still in its infancy. Existing works have been mostly focused on remote sensing problems [9, 10] and only recently interest in MISR from burst captures of mobile cameras has arisen, also owing to the availabity of a new carefully curated dataset [11]. Indeed, mobile cameras are significantly limited in the capabilities of their optics and sensors, often leading to a lack of sharpness on the captured images. Computational photography solutions processing a burst of images to enhance the spatial resolution are therefore highly desirable. A few learning-based techniques have been proposed for burst SR. Deep Burst Super-Resolution [11] is the first work introducing the MISR setting in the context of computational photography with deep learning. In this work, the authors also introduce a new dataset, called BurstSR, that contains for each set of LR smartphone burst captures a corresponding HR photo captured using a DSLR camera. A few works have recently improved over the approach presented in [11]. In [12] the authors introduce a registration technique in the feature space and a long-range concatenation network to improve the reconstruction. Lecouat et al. [13] present an architecture designed from the unrolling of the iterations of an optimization problem for reconstruction. Finally, Bhat et al. [14] reparametrize the classic *maximum-a-posteriori* reconstruction formulation to model image formation in a latent space.

Most of the ideas in state-of-the-art works on burst SR originate from video applications. Despite the similarities between videos and bursts, the former typically displays a coherent temporal evolution, making frame ordering important. However, burst captures do not typically possess this property, as noted in earlier works on burst captures [8], and ordering within the burst should not matter in producing the SR image. However, this issue is currently neglected by methods proposed for burst SR. This means that those models need to learn this important property from the data, resulting in a suboptimal use of the (limited) training examples and ultimately lower performance. In this paper, we conceive a neural network for the burst SR problem, called MLB-FuseNet, that is, by construction, able to estimate features that are invariant to a temporal permutation. These features are derived from $T-1$ images in an T-length burst and serve to augment the features extracted from an image that is considered as reference for the SR process. Designing the network layers to be mathematically invariant to temporal permutation of the $T-1$ images is the key feature that allows to improve the network performance. This is especially true for real bursts that are not synthetically generated from one image, where the network should efficiently use the often limited data. These permutation-invariant features are then progressively fused with those of the reference image in a slow, multi-level process. Moreover, an added benefit of our design with respect of existing methods is that bursts of arbitrary length can be processed by the same model. Finally, similarly to other state-of-the-art works, MLB-FuseNet works on raw images and integrates demoisacing in the network functionality. However, we improve upon existing techniques by designing a feature extraction block better suited for mosaiced data, called Mosaiced Convolution Feature Extractor (MCFE). The MCFE design is based on a polyphase decomposition to allow easier learning of suitable convolutional kernels, as each kernel only processes a specific Bayer pattern instead of mixed ones as it strides. These contributions allow MLB-FuseNet to reach state-of-the-art performance on the BurstSR

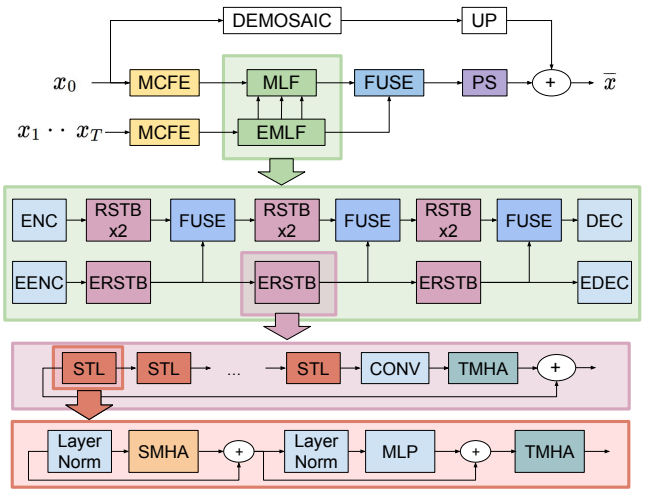

Fig. 1. MLB-FuseNet overview.

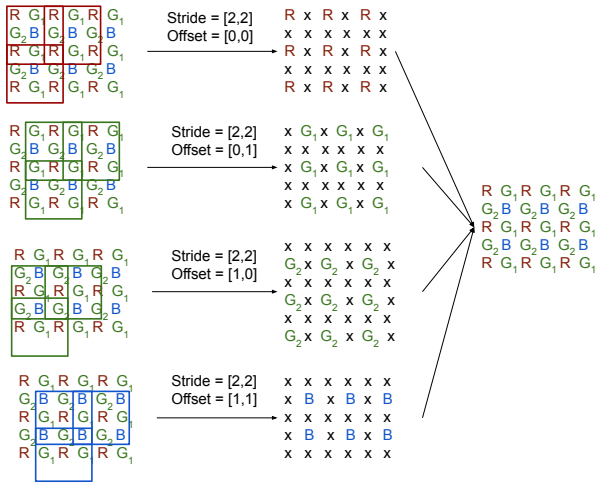

Fig. 2. MFCE principle. Separate convolutional kernels process the 4 phases with stride of 2 so that the pattern observed by the kernel is consistent throughout all spatial locations.

dataset, providing comparable quality to models with significantly more parameters in the synthetic burst setting and superior quality in the real burst setting.

2. PROPOSED METHOD

We propose a network, called MLB-FuseNet, that takes as input a burst of T raw LR images, registered at a precision of one LR pixel, and generates a corresponding HR image. As illustrated in Figure 1, the proposed architecture is divided into two branches: the one on top processes one of the T images, taken as reference frame $\mathbf{x}_0$, through a single-image SR network; the one at the bottom processes the remaining images all at once in a permutation-invariant fashion with the goal of extracting features that are used to augment the features of the top branch. This allows to draw features from the burst that are referenced to $\mathbf{x}_0$, and particularly to extract features related to subpixel misalignments between the reference and the rest of the burst. At the same time, the network treats the remaining $T-1$ images in a permutation-invariant way, without enforcing an arbitrary temporal order that does not carry meaningful information. Both branches exhibit as first block a Mosaiced Convolution Feature Extractor to improve feature extraction from the raw mosaiced data. Then, the features estimated by the two branches are fused at multiple levels of the processing chain. Finally, the Pixel Shuffle (PS) [2] block generates the super-resolved image. A skip connection provides a basic demosaicing operation [15] and bilinear upsampling so that the network only learns a residual correction.

2.1. Mosaiced Convolution Feature Extractor

We introduce a novel convolutional block, namely MCFE, in order to improve feature extraction from mosaiced sensor data. The input to the network is a raw burst of images at low resolution $\mathbf{x}_{\text{LR}} \in \mathbb{R}^{1\times 2H\times 2W}$, where pixels are acquired according to the Bayer pattern of the camera. The idea of the MCFE is to extract high level features without disrupting the Bayer color arrangement. In particular, when a convolutional kernel slides over the mosaiced image with a stride equal to 1, it observes different Bayer patterns at every stride. This is undesirable because the learned weights in the kernel would converge to a value that is a compromise for all the different combinations. We instead propose to use a polyphase decomposition where different kernels work with a stride of 2 with 4 different offset arrangements. This ensures that each kernel only sees a consistent pattern. For example, following the depiction in Figure 2 for an RGGB Bayer pattern and 3×3 kernels, the first phase kernel will always see the RGRGBGRGR pattern. To accomplish this, we apply a 3×3 filter with stride equal to 2, and a different offset of either 0 or 1 pixels in the vertical and horizontal direction, resulting in 4 phase branches. As shown in the central part of Fig. 2, each branch generates a set of F channels with halved spatial dimensions. Finally, we combine the features generated by each kernel, considering the correct spatial position in the original Bayer grid. When multiple images have to be processed, as in the permutation-invariant pipeline working on $T-1$ images, the same kernels are shared across the time dimension. Multiple MFCE layers can be used in sequence and in our MLB-FuseNet we use two of them.

2.2. Multi-Level Fusion block

The main part of the network is composed of two modules: the Multi-Level-Fusion (MLF) block in the top branch, and the Equivariant Multi-Level Fusion (EMLF) block in the bottom branch. These blocks are inspired by the SwinIR network [16] in the use of windowed transformer operations for spatial attention.

In the lower branch, the EMLF module extends the SwinIR Deep Feature Extraction block to a multi-image setting. The main idea of EMLF is to extract features that are equivariant to temporal permutation, i.e., they stay the same but permuted in ordering when the input is permuted. These features are then used to derive a permutation-invariant representation via averaging in temporal dimension. In order to make the EMLF temporally equivariant, we use a temporally-equivariant encoder (EENC), followed by a sequence of K Equivariant Residual Swin Transformer blocks (ERSTBs) and a temporally-equivariant decoder (EDEC). In each ERSTB, we first extract the spatial features of each image separately using a sequence of Swin Transformer Layers (STLs) that exploit the Spatial Multi Head Attention (SMHA) in its local-windowed version [16]. Then, we combine the spatial features of the different images using Multi Head

Table 1. Performance comparison on synthetic data

Method	#params	Synthetic data	
		PSNR (dB) ↑	SSIM ↑
DBSR [11]	13M	39.17	0.946
DeepRep [14]	12M	41.55	0.964
EBSR [12]	26M	42.98	0.972
MLB-FuseNet (ours)	9M	42.34	0.969

Table 2. Performance comparison on real BurstSR

Method	#params	BurstSR data	
		PSNR (dB) ↑	SSIM ↑
DBSR [11]	13M	47.70	0.984
DeepRep [14]	12M	48.33	0.985
EBSR [12]	26M	48.23	0.985
MLB-FuseNet (ours)	9M	48.66	0.986

Attention along the temporal dimension (TMHA) [17][18]. This operation allows to mix the features of the different images along the temporal dimension, in a mathematically equivariant way, i.e., obtaining the same output, albeit permuted, from a permuted input. The EENC is defined in a similar way: a sequence of 2D convolutional layers is applied to each frame separately in order to mix up the spatial features obtained by the MFCE and combine the pixels of the Bayer mosaic, then a TMHA combines the features of the different images. Similarly, the EDEC is formed by a sequence of 2D convolutions and a TMHA.

The MLF module in the upper branch of the network is composed of an Encoder (ENC), $2K$ Residual Swin Transformer blocks (RSTBs) interlayed with fusion modules (FUSE) and a final Decoder (DEC). The encoder and decoder are composed by traditional convolutional layers, while the RSTBs consist of a sequence of STLs. After each RSTB block there is a fusion module that combines the features of the reference frame $\mathbf{x}_0$ obtained in the upper branch with the ones obtained in the lower branch from the other frames of the burst. The fusion is performed by averaging the features of the lower branch along the temporal dimension and merging, through a 1D convolution after channel concatenation. These fusion modules allow the upper branch to produce a super resolved image guided by the information extracted from the multiple images in the lower branch. We use multiple stages of fusion module to slowly incorporate the features of the burst at various levels of abstraction.

After the MLF module, there is a last fusion block that merges the final outputs of the MLF and EMLF blocks, and finally a series of 2D convolutions and pixel shuffling [2] to generate the super-resolved image.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed super-resolution method on both synthetically-generated data and real-world data using the recent BurstSR dataset introduced in [11]. We first train the proposed network on synthetic data. Then, we fine-tune our model on the real-world BurstSR dataset. In both settings, the model is composed of $K = 3$ ERSTBs, 1 convolutional layer in all the encoders and decoders, $F = 120$ features in the MCFE and the MLF, 90 feature channels in the EMLF and 128 feature channels in the final FUSE block. We compare the proposed method with DBSR [11] as baseline and with two additional recent state-of-the-art methods, DeepRep [14] and EBSR [12], using for both training and evaluation a burst of size $T = 14$. Training on the synthetic dataset requires approximately 200 epochs on 4 Nvidia RTX A6000, while fine-tuning on the real data is performed for 80 epochs on the same GPUs.

3.1. Synthetic data

The synthetic raw bursts used for training and testing are generated from the Zurich RAW to RGB dataset [19], respectively from the training and test split, using the inverse camera pipeline described in [20]. Each burst comprises a reference image $\mathbf{x}_0$ and its variations, created applying random translation and rotations. To train our network, we extract crops of dimension 48×48 and we use an already trained PWC-Net [21] to align the burst to the reference frame, as a preprocessing operation. Notice that, each raw image is represented as a 4-channel image because of the Bayer CFA pattern in this stage. After the alignment, we flatten the color channels, to obtain a burst of dimensions $T \times 96 \times 96$, and we crop these images by 8 pixels per side, to remove border effects created by the pixel translation. Thus, the final dimension of the input image is $T \times 80 \times 80$. At this point, the neural network model will take care of the demosaicing process and target a SR factor equal to $\times 4$. It is important to notice that state-of-the-art techniques adopt a trainable alignment block inside the neural network. Instead, we use preprocessing to align the multiple frames at a precision of a single LR pixel. We consider this preferable for two main reasons: i) reduced computational complexity of the network; ii) improved robustness to geometric perturbations outside the trained range. In the training phase we use the entire training split from the Zurich dataset. We employ Adam optimizer with learning rate 10^{-4}, a batch size of 56 and an L_1 loss function ignoring the boundary pixels to prevent learning boundary artifacts. During the testing phase, we repeat all the preprocessing steps used in training and we further improve the performance using an ensembling technique, i.e., we average the SR prediction with an additional one, obtained transposing the images in the burst, as also done by the other methods.

The results on the synthetic dataset are shown in Table 1. As we can see, our method significantly outperforms both DBSR and DeepRep, which have a comparable number of parameters, and it shows competitive performance with respect to EBSR which, on the contrary, is a significantly larger model with almost three times the number of trainable parameters.

3.2. Real data

In this section, we analyze the performance of our model on the real BurstSR data. This dataset contains a collection of 200 LR photo bursts, and a corresponding HR photo for each burst. The LR images are acquired using a handheld smartphone camera, while the HR one is captured using a DSLR camera. Given the different acquisition device, the main challenge of the BurstSR dataset is the lack of alignment between the LR photos and the HR ground truth. In order to deal with the misalignment, we employ the same approach as proposed in [11]. That is, the network output is registered with the ground truth via optical flow methods before the calculation of the pixel-wise loss function or quality metrics. Since the alignment may not be perfect, only a valid subset of the pixels is actually used for the calculation. This procedure is applied both during the training phase on the loss function and in the testing phase, for all methods considered in the experiments.

As in the synthetic framework, we register the T frames in advance and we crop the images, to remove the undesired borders.

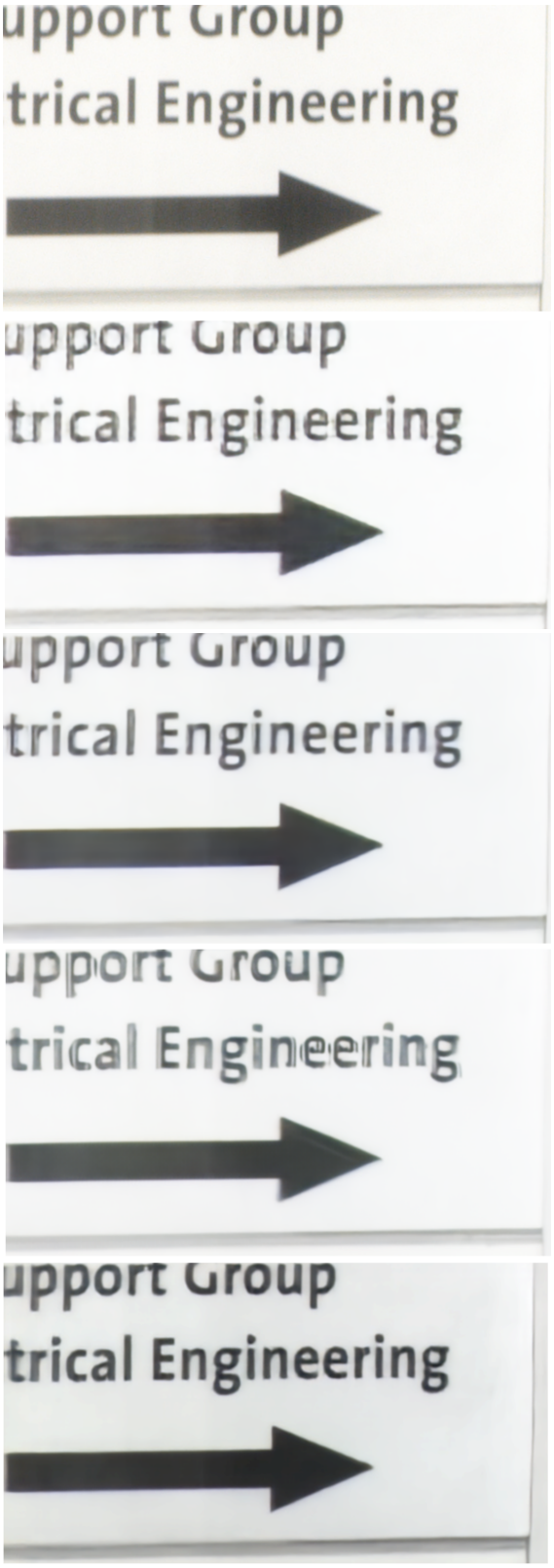

Fig. 3. Qualitative comparison on the BurstSR test set. Top to bottom: ground truth, DBSR [11], DeepRep [14], EBSR [12], MLB-FuseNet (ours).

Table 3. Impact of MCFE: test on synthetic data

	Synthetic data	
	PSNR (dB) ↑	SSIM ↑
MLB-FuseNet	42.34	0.969
No MCFE	39.98	0.955

However, during our experiments we observed that the results on the real data benefit from bigger training patches. For this reason, the input to our network during training is an image of dimension 112×112 after alignment and cropping. Instead, for testing, we utilize the entire available LR image having dimension 144×144. To train the network we use Adam optimizer with learning rate 10^{-5}, a batch size of 28 and the aligned L1 loss from [11] described above.

The results on the BurstSR real dataset are shown in Table 2. We can observe that on the real data the proposed network outperforms all the other methods considered in the experimental evaluation. This result validates that our permutation-invariant approach is indeed appropriate for real burst captures, when the dataset itself does not exhibit an explicit reference as it happens in the synthetic framework. For a qualitative comparison, Fig. 3 shows a detail of one of the images in the BurstSR test set. We can see how our proposed method exhibits sharper details and less artifacts in the super-resolved text (best seen on a computer screen).

3.3. Ablation: effect of the MCFE

We study the impact of the newly conceived MCFE module, by evaluating the performance of the network when the mosaic convolution is not used. We replace all the MCFE blocks with usual 2D convolutions, which mix up all the $2H \times 2W$ pixels disrupting the Bayer pattern. Then, we re-train the network following the exact steps described in Sec. 3.1 and we evaluate this architecture on the synthetic test set. The results of this experiment are illustrated in Table 3. We can observe that the network without MCFE exhibits poorer performance, confirming that the MCFE plays an important role in the proposed design.

4. CONCLUSIONS

In this paper, we introduce a new neural network, namely MLB-FuseNet, for burst super-resolution. The network exploits the inherent temporal permutation invariance of the problem, a multi-stage fusion and a polyphase approach to process mosaiced data. Our experiments show that the proposed solution achieves competitive performance on the synthetic test set and it outperforms the state-of-the-art methods on the BurstSR dataset with real burst captures. In future works, we will test the stability of the network to perturbations and we will improve the performance by building a model that does not need a reference,.

5. REFERENCES

[1] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1874–1883.

[3] Kai Zhang, Luc Van Gool, and Radu Timofte, "Deep unfolding network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[5] 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*, vol. 38, no. 4, 2019.

[6] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 10, pp. 1327–1344, Oct 2004.

[7] Kato Toshiyuki, Hino Hideitsu, and Murata Noboru, "Multi-frame image super resolution based on sparse coding," *Neural Networks*, vol. 66, pp. 64 – 78, 2015.

[8] Miika Aittala and Frédo Durand, "Burst image deblurring using permutation invariant convolutional neural networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 731–747.

[9] 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.

[10] Francesco Salvetti, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sensing*, vol. 12, no. 14, 2020.

[11] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte, "Deep burst super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.

[12] Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu, "Ebsr: Feature enhanced burst super-resolution with deformable alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2021, pp. 471–478.

[13] Bruno Lecouat, Jean Ponce, and Julien Mairal, "Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[14] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte, "Deep reparametrization of multi-frame super-resolution and denoising," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 2460–2470.

[15] "Convolutional pytorch debayering," https://github.com/cheind/pytorch-debayer.

[16] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, vol. 30.

[18] Diego Valsesia and Enrico Magli, "Permutation invariance and uncertainty in multitemporal image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[19] Andrey Ignatov, Luc Van Gool, and Radu Timofte, "Replacing mobile camera isp with a single deep learning model," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition Workshops (CVPRW)*, 2020, pp. 2275–2285.

[20] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.