Burst Image Restoration and Enhancement

Akshay Dudhane[1]   Syed Waqas Zamir[2]   Salman Khan[1,3]

Fahad Shahbaz Khan[1,4]   Ming-Hsuan Yang[5,6,7]

[1]Mohamed bin Zayed University of AI   [2]Inception Institute of AI   [3]Australian National University
[4]Linköping University   [5]University of California, Merced   [6]Yonsei University   [7]Google Research

Abstract

Modern handheld devices can acquire burst image sequence in a quick succession. However, the individual acquired frames suffer from multiple degradations and are misaligned due to camera shake and object motions. The goal of Burst Image Restoration is to effectively combine complimentary cues across multiple burst frames to generate high-quality outputs. Towards this goal, we develop a novel approach by solely focusing on the effective information exchange between burst frames, such that the degradations get filtered out while the actual scene details are preserved and enhanced. Our central idea is to create a set of pseudo-burst features that combine complimentary information from all the input burst frames to seamlessly exchange information. However, the pseudo-burst cannot be successfully created unless the individual burst frames are properly aligned to discount inter-frame movements. Therefore, our approach initially extracts pre-processed features from each burst frame and matches them using an edge-boosting burst alignment module. The pseudo-burst features are then created and enriched using multi-scale contextual information. Our final step is to adaptively aggregate information from the pseudo-burst features to progressively increase resolution in multiple stages while merging the pseudo-burst features. In comparison to existing works that usually follow a late fusion scheme with single-stage upsampling, our approach performs favorably, delivering state-of-the-art performance on burst super-resolution, burst low-light image enhancement and burst denoising tasks. The source code and pre-trained models are available at https://github.com/akshaydudhane16/BIPNet.

1. Introduction

High-end DSLR cameras can capture images of excellent quality with vivid details. With the growing popularity of smartphones, the main goal of computational photography is to generate DSLR-like images with smartphone cameras [24]. However, the physical constraints of smartphone
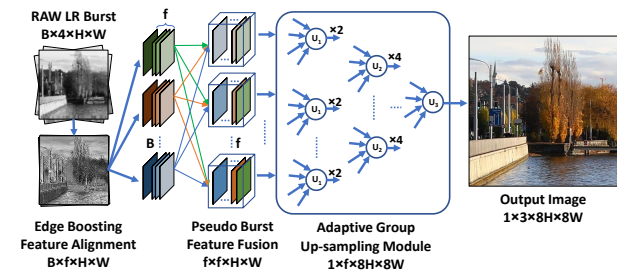

Figure 1. Holistic diagram of our burst image processing approach. Our network BIPNet takes as input a RAW image burst and generates a high-quality RGB image. BIPNet has three key stages. (1) Edge boosting feature alignment to remove noise, and inter-frame spatial and color misalignment. (2) Pseudo-burst feature fusion mechanism to enable inter-frame communication and feature consolidation. (3) Adaptive group upsampling to progressively increase spatial resolution while merging multi-frame information. While BIPNet is generalizable to other restoration tasks, here we show super-resolution application.

cameras hinder the image reconstruction quality. For instance, small sensor size limits spatial resolution and small lens and aperture provides noisy and color distorted images in low-light conditions [12]. Similarly, small pixel cavities accumulate less light therefore yielding low-dynamic range. To alleviate these issues, burst (multi-frame) photography is a natural solution instead of single-frame processing [21].

The goal of burst imaging is to composite a high-quality image by merging desired information from a collection of (degraded) frames of the same scene captured in a rapid succession. However, burst image acquisition presents its own challenges. For example, during image burst capturing, any movement in camera and/or scene objects will cause misalignment issues, thereby leading to ghosting and blurring artifacts in the output image [53]. Therefore, there is a pressing need to develop a multi-frame processing algorithm that is robust to alignment problems and requires no special burst acquisition conditions. We note that existing burst processing techniques [4, 5] extract and align features of burst images separately and usually employ late feature fusion mechanisms, which can hinder flexible infor-

mation exchange among frames. In this paper, we present a burst image processing approach, named BIPNet, which is based on a novel pseudo-burst feature fusion mechanism that enables inter-frame communication and feature consolidation. Specifically, a pseudo-burst is generated by exchanging information across frames such that each feature in the pseudo-burst contains complimentary properties of all input burst frames.

Before synthesizing pseudo-bursts, it is essential to align the input burst frames (having arbitrary displacements) so that the relevant pixel-level cues are aggregated in the later stages. Existing works [4, 5] generally use explicit motion estimation techniques (e.g., optical flow) to align input frames which are typically bulky pretrained modules that cannot be fully integrated within an end-to-end learnable pipeline. This can result in errors caused during the flow estimation stage to be propagated to the warping and image processing stages, thereby negatively affecting the generated outputs. In our case, the proposed BIPNet implicitly learns the frame alignment with deformable convolutions [64] that can effectively adapt to the given problem. Further, we integrate the edge boosting refinement via back-projection operation [19] in the alignment stage to retain high-frequency information. It facilitates sustaining the alignment accuracy in cases where highly complex motions between burst images exist and only the deformable convolutional may not be sufficient for reliable alignment.

Noise is always present in images irrespective of the lighting condition in which we acquire them. Therefore one of our major goals is to remove noise [57] early in the network to reduce difficulty for the alignment and fusion stages. To this end, we incorporate residual global context attention in BIPNet for feature extraction and refinement/denoising. While the application of BIPNet can be generalized to any burst processing task, we demonstrate its effectiveness on burst super-resolution, burst low-light image enhancement and burst denoising. In super-resolution (SR), upsampling is the key step for image reconstruction. Existing burst SR methods [4, 5] first fuse the multi-frame features, and then use pixel-shuffle operation [43] to obtain the high-resolution image. However, we can leverage the information available in multiple frames to perform merging and upsampling in a flexible and effective manner. As such, we include adaptive group upsampling in our BIPNet that progressively increases the resolution while merging complimentary features. BIPNet schematic is shown in Fig. 1. The main contributions of this work include:

- An edge boosting alignment technique that removes spatial and color misalignment issues among the burst features. (Sec. 3.1)

- A novel pseudo-burst feature fusion mechanism to enable inter-frame communication and feature consolidation. (Sec. 3.2)

- An adaptive group upsampling module for progressive fusion and upscaling. (Sec. 3.3)

Our BIPNet achieves state-of-the-art results on synthetic and real benchmark datasets for the burst super-resolution, low-light image enhancement and burst denoising tasks. We provide comprehensive ablation results and visual examples to highlight the contributing factors in BIPNet (Sec. 4).

2. Related Work

Single Image Super-resolution (SISR). Since the first CNN-based work [14], data-driven approaches have achieved high performance gains over the conventional counterparts [17, 55]. The success of CNNs is mainly attributed to their architecture design [2, 59]. Given a low-resolution image (LR), early methods learn to directly generate latent SR image [14, 15]. In contrast, recent approaches learns to produce high frequency residual to which LR image is added to generate the final SR output [23, 46]. Other notable SISR network designs employ recursive learning [1, 27], progressive reconstruction [29, 52], attention mechanisms [11, 58, 61, 62], and generative adversarial networks [31, 42, 51]. The SISR approaches cannot handle multi-degraded frames from an input burst, while our approach belong to multi-frame SR that allows effectively merging cross-frame information towards a HR output.

Multi-Frame Super-Resolution (MFSR). Tsai *et al.* [48] are the first to deal with the MFSR problem. They propose a frequency domain based method that performs registration and fusion of the multiple aliased LR images to generate a SR image. Since processing multi-frames in the frequency domain leads to visual artifacts [48], several other works aim to improve results by incorporating image priors in HR reconstruction process [44], and making algorithmic choices such as iterative back-projection [25, 41]. Farsui *et al.* [16] design a joint multi-frame demosaicking and SR approach that is robust to noise. MFSR methods are also developed for specific applications, such as for handheld devices [53], to increase spatial resolution of face images [49], and in satellite imagery [13, 40]. Lecouat *et al.* [30] retains the interpretability of conventional approaches for inverse problems by introducing a deep-learning based optimization process that alternates between motion and HR image estimation steps. Recently, Bhat *et al.* [4] propose a multi-frame burst SR method that first aligns burst image features using an explicit PWCNet [45] and then perform feature integration using an attention-based fusion mechanism. However, explicit use of motion estimation and image warping techniques can pose difficulty handling scenes with fast object motions. Recent works [47, 50] show that the deformable convolution [64] effectively handles inter-frame alignment issues due to being implicit and adaptive in

nature. Unlike existing MFSR methods, we implicitly learn the inter-frame alignment and then channel-wise aggregate information followed by adaptive upsampling to effectively utilize multi-frame information.

Low-Light Image Enhancement. Images captured in low-light conditions are usually dark, noisy and color distorted. These problems are somewhat alleviated by using long sensor exposure time, wide aperture, camera flash, and exposure bracketing [12, 60]. However, each of these solutions come with their own challenges. For example, long exposure yields images with ghosting artifacts due to camera or object movements. Wide apertures are not available on smartphone devices, etc. See-in-the-Dark method [8] is the first attempt to replace the standard camera imaging pipeline with a CNN model. It takes as input a RAW input image captured in extreme low-light and learns to generate a well-lit sRGB image. Later this work is further improved with a new CNN-based architecture [36] and by employing a combined pixel-wise and perceptual loss [60]. Zaho *et al*. [63] takes the advantage of burst imaging and propose a recurrent convolutional network that can produce noise-free bright sRGB image from a burst of RAW images. The results are further improved by Karadeniz *et al*. [26] with their two-stage approach: first sub-network performs denoising, and the second sub-network provides visually enhanced image. Although these studies demonstrate significant progress in enhancing low-light images, they do not address inter-frame misalignment and information interaction which we address in this work.

Multi-Frame Denoising. Earlier works [9, 34, 35] extend the popular image denoising algorithm BM3D [10] to video. Buades *et al*. perform denoising by estimating the noise level from the aligned images followed by the combination of pixel-wise mean and BM3D. A hybrid 2D/3D Wiener filter is used in [20] to denoise and merge burst images for high dynamic range and low-light photography tasks. Godard *et al*. [18] utilize recurrent neural network (RNN) and extend a single image denoising network for multiple frames. Mildenhall *et al*. [39] generate per-pixel kernels through the kernel prediction network (KPN) to merge the input images. In [37], authors extend KPN approach to predict multiple kernels, while [54] introduce basis prediction networks (BPN) to enable the use of larger kernels. Recently, Bhat *et al*. [5] propose a deep reparameterization of the maximum a posteriori formulation for the multi-frame SR and denoising.

3. Burst Processing Approach

In this section, we describe our burst processing approach which is applicable to different image restoration tasks, including burst super-resolution, burst low-light image enhancement and burst denoising. The goal is to generate a high-quality image by combining information from multiple degraded images captured in a single burst. Burst images are typically captured with handheld devices, and it is often inevitable to avoid inter-frame spatial and color misalignment issues. Therefore, the main challenge of burst processing is to accurately align the burst frames, followed by combining their complimentary information while preserving and reinforcing the shared attributes. To this end, we propose BIPNet in which different modules operate in synergy to jointly perform denoising, demosaicking, feature fusion, and upsampling tasks in a unified model.

Overall pipeline. Fig. 1 shows three main stages in the proposed BIPNet. First, the input RAW burst is passed through the edge boosting feature alignment module to extract features, reduce noise, and remove spatial and color misalignment issues among the burst features (Sec. 3.1). Second, a pseudo-burst is generated by exchanging information such that each feature map in the pseudo-burst now contains complimentary properties of all actual burst image features (Sec. 3.2). Finally, the multi-frame pseudo-burst features are processed with the adaptive group upsampling module to produce the final high-quality image (Sec. 3.3).

3.1. Edge Boosting Feature Alignment Module

One major challenge in burst processing is to extract features from multiple degraded images that are often contaminated with noise, unknown spatial displacements, and color shifts. These issues arise due to camera and/or object motion in the scene, and lighting conditions. To align the other images in the burst with the base frame (usually the 1st frame for simplicity) we propose an alignment module based on modulated deformable convolutions [64]. However, existing deformable convolution is not explicitly designed to handle noisy RAW data. Therefore, we propose a feature processing module to reduce noise in the initial burst features. Our edge boosting feature alignment (EBFA) module (Fig. 2(a)) consists of feature processing followed by burst feature alignment.

3.1.1 Feature Processing Module

The proposed feature processing module (FPM), shown in Fig. 2(b), employs residual-in-residual learning that allows abundant low-frequency information to pass easily via skip connections [61]. Since capturing long-range pixel dependencies which extracts global scene properties has been shown to be beneficial for a wide range of image restoration tasks [56] (e.g., image/video super-resolution [38] and extreme low-light image enhancement [3]), we utilize a global context attention (GCA) mechanism to refine the latent representation produced by residual block, as illustrated in Fig. 2(b). Let $\left\{x^b\right\}_{b\in[1:B]}\in\mathbb{R}^{B\times f\times H\times W}$ be an initial latent representation of the burst having B burst images and
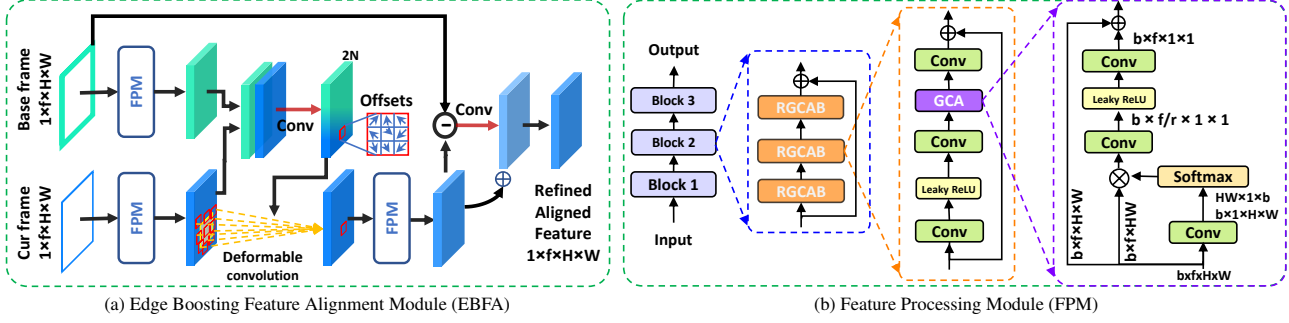
Figure 2. Edge boosting feature alignment (EBFA) module aligns all other images in the input burst to the base frame. Feature processing module (FPM) is added in EBFA to denoise input frames to facilitate the easy alignment. $\otimes$ represents matrix multiplication.

f number of feature channels, our residual global context attention block (RGCAB in Fig. 2(b)) is defined as:

$$\boldsymbol{y}^b = \boldsymbol{x}^b + W_1\left(\alpha\left(\bar{\boldsymbol{x}}^b\right)\right), \qquad (1)$$

where $\bar{\boldsymbol{x}}^b = W_3\left(\gamma\left(W_3\left(\boldsymbol{x}^b\right)\right)\right)$ and $\alpha\left(\bar{\boldsymbol{x}}^b\right) = \bar{\boldsymbol{x}}^b + W_1\left(\gamma\left(W_1\left(\Psi\left(W_1\left(\bar{\boldsymbol{x}}^b\right)\right)\otimes\bar{\boldsymbol{x}}^b\right)\right)\right)$. Here, W_k represents a convolutional layer with $k \times k$ sized filters and each W_k corresponds to a separate layer with distinct parameters, γ denotes leaky ReLU activation, Ψ is softmax activation, $\otimes$ represents matrix multiplication, and $\alpha(\cdot)$ is the global context attention.

3.1.2 Burst Feature Alignment Module

To effectively fuse information from multiple frames, these frame-level features need to be aligned first. We align the features of the current frame $\boldsymbol{y}^b$ with the base frame[1] $\boldsymbol{y}^{b_r}$. EBFA processes $\boldsymbol{y}^b$ and $\boldsymbol{y}^{b_r}$ through an offset convolution layer and predicts the offset Δn and modulation scalar Δm values for $\boldsymbol{y}^b$. The aligned features $\bar{\boldsymbol{y}}^b$ computed as:

$$\bar{\boldsymbol{y}}^b = W^d\left(\boldsymbol{y}^b,\ \Delta n,\ \Delta m\right),\ \ \Delta m = W^o\left(\boldsymbol{y}^b,\ \boldsymbol{y}^{b_r}\right), \qquad (2)$$

where, W^d and W^o represent the deformable and offset convolutions, respectively. More specifically, each position n on the aligned feature map $\bar{\boldsymbol{y}}^b$ is obtained as:

$$\bar{\boldsymbol{y}}_n^b = \sum_{i=1}^{K} W_{n_i}^d\ \boldsymbol{y}_{(n+n_i+\Delta n_i)}^b \cdot \Delta m_{n_i}, \qquad (3)$$

where, K=9, Δm lies in the range [0, 1] for each $n_i \in \{(-1,1),(-1,0),...,(1,1)\}$ is a regular grid of 3×3 kernel.

The convolution operation will be performed on the non-uniform positions $(n_i + \Delta n_i)$, where n_i can be fractional. To avoid fractional values, the operation is implemented using bilinear interpolation.

The proposed EBFA module is inspired from the deformable alignment module (DAM) [47] with the following

differences. Our approach does not provide explicit ground-truth supervision to the alignment module, instead it learns to perform implicit alignment. Furthermore, to strengthen the feature alignment and to correct the minor alignment errors, using FPM, we obtain refined aligned features (RAF) followed by computing the high-frequency residue by taking the difference between the RAF and base frame features and add it to the RAF. The overall process of our EBFA module is summarized as: $\boldsymbol{e}^b = \bar{\boldsymbol{y}}^b + W_3\left(\bar{\boldsymbol{y}}^b - \boldsymbol{y}^{b_r}\right)$ where $\boldsymbol{e}^b \in \mathbb{R}^{B \times f \times H \times W}$ represents the aligned burst feature maps, and $W_3(\cdot)$ is the convolution. Although the deformable convolution is shown only once in Fig. 2(a) for brevity, we sequentially apply three such layers to improve the transformation capability of our EBFA module.

3.2. Pseudo-Burst Feature Fusion Module

Existing burst image processing techniques [4, 5] separately extract and align features of burst images and usually employ late feature fusion mechanisms, which can hinder flexible information exchange between frames. We instead propose a pseudo-burst feature fusion (PBFF) mechanism (see Fig. 3 (a)). This PBFF module generates feature tensors by concatenating the corresponding channel-wise features from all burst feature maps. Consequently, each feature tensor in the pseudo-burst contains complimentary properties of all actual burst image features. Processing inter-burst feature responses simplifies the representation learning task and merges the relevant information by decoupling the burst image feature channels. Given the aligned burst feature set $e = \left\{\boldsymbol{e}_c^b\right\}_{c\in[1:f]}^{b\in[1:B]}$ of burst size B and f number of channels, the pseudo-burst is generated by,

$$\boldsymbol{S}^c = W^\rho\left(\left\langle\boldsymbol{e}_c^1,\ \boldsymbol{e}_c^2,\ \cdots,\ \boldsymbol{e}_c^B\right\rangle\right), \quad s.t. \quad c \in [1:f], \ (4)$$

where, $\langle\cdot\rangle$ represents concatenation, $\boldsymbol{e}_c^1$ is the cth feature map of 1st aligned burst feature set $\boldsymbol{e}^1$, W^ρ is the convolution layer with f output channel, and $\boldsymbol{S} = \{\boldsymbol{S}^c\}_{c\in[1:f]}$ represents the pseudo-burst of size $f \times f \times H \times W$. In this paper, we use $f = 64$.

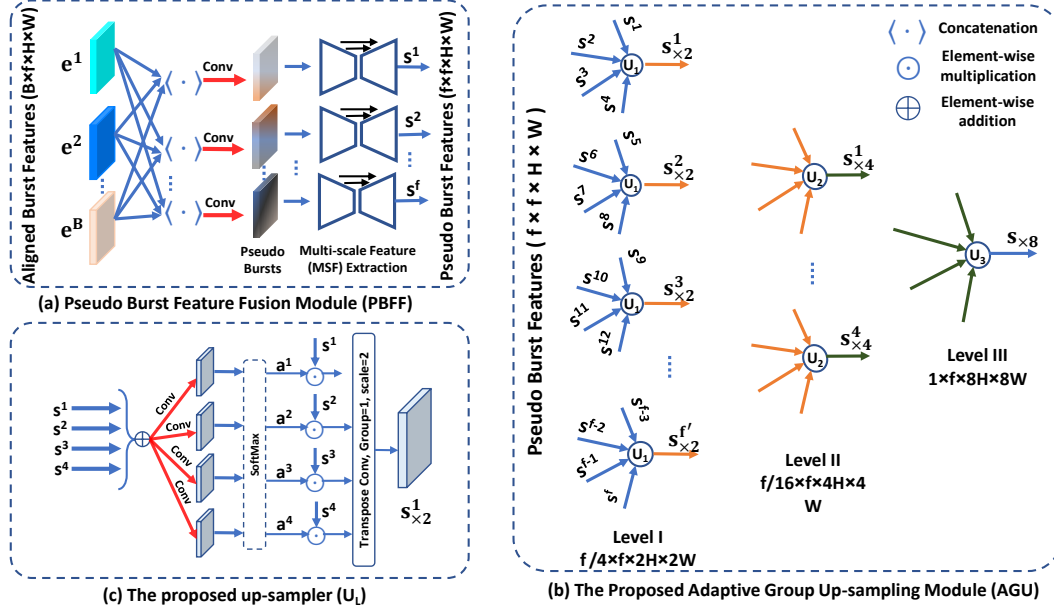[1]In this work, we consider first input burst image as the base frame.

Figure 3. (a) Pseudo-burst is generated by exchanging information across frames such that each feature tensor in the pseudo-burst contains complimentary properties of all frames. Pseudo bursts are processed with (shared) U-Net to extract multi-scale features. (b) AGU module handles pseudo-bursts features in groups and progressively performs upscaling. (c) Schematic of dense-attention based upsampler.

Even after generating pseudo-bursts, obtaining their deep representation is essential. We use a light-weight (3-level) U-Net to extract multi-scale features (MSF) from pseudo-bursts. We use shared weights in the U-Net, and also employ our FPM instead of regular convolutions.

3.3. Adaptive Group Upsampling Module

Upsampling is the final key step to generate the super-resolved image from LR feature maps. Existing burst SR methods [4, 5] use pixel-shuffle layer [43] to perform up-sampling in one-stage. However, in burst image processing, information available in multiple frames can be exploited effectively to get into HR space. To this end, we propose to *adaptively* and *progressively* merge multiple LR features in the upsampling stage. For instance, on the one hand it is beneficial to have uniform fusion weights for texture-less regions in order to perform denoising among the frames. On the other hand, to prevent ghosting artifacts, it is desirable to have low fusion weights for any misaligned frame.

Fig. 3(b) shows the proposed adaptive group upsampling (AGU) module that takes as input the feature maps $S = \{S^c\}_{c \in [1:f]}$ produced by the pseudo-burst fusion module and provides a super-resolved output via three-level progressive upsampling. In AGU, we sequentially divide the pseudo-burst features into groups of 4, instead of following any complex selection mechanism. These groups of features are upsampled with the architecture depicted in Fig. 3(c) that first computes a dense attention map (a^c), carrying attention weights for each pixel lo-

cation. The dense attention maps are element-wise applied to the respective burst features. Finally, the up-sampled response for a given group of features $\hat{S}^g = \left\{ S^i : i \in [(g-1) * 4 + 1 : g * 4] \right\}^{g \in [1:f/4]} \subset S$ and associated attention maps $\hat{a}^g$ at the first upsampling level (Level I in Fig. 3(b)) is formulated as:

$$S_{\times 2}^g = W_T \left(\left\langle \hat{S}^g \odot \hat{a}^g \right\rangle \right),$$

$$\hat{a}^g = \psi \left(W_1 \left(W_1 \left(\sum_{i=(g-1)*4+1}^{g*4} S^i \right) \right) \right), \quad (5)$$

where $\psi (\cdot)$ denotes the softmax activation function, W_T is the 3×3 Transposed convolution layer, and $\hat{a}^g \in \mathbb{R}^{4 \times f \times H \times W}$ represents the dense attention map for g^{th} burst feature response group ($\hat{S}^g$).

To perform burst SR of scale factor $\times 4$, we need in fact $\times 8$ upsampling (additional $\times 2$ is due to the mosaicked RAW LR frames). Thus, in AGU we employ three levels of $\times 2$ upsampling. As, our BIPNet generates 64 pseudo bursts, this naturally forms 16, 4 and 1 feature groups at levels I, II, and III, respectively. Upsampler at each level is shared among groups to avoid the increase in network parameters.

4. Experiments

We evaluate the proposed BIPNet and other state-of-the-art approaches on real and synthetic datasets for **(a)** burst

super-resolution, **(b)** burst low-light image enhancement, and **(c)** burst denoising.

Implementation Details. Our BIPNet is end-to-end trainable and needs no pretraining of any module. For network parameter efficiency, all burst frames are processed with shared BIPNet modules (FPM, EBFA, PBFF and AGU). Overall, the proposed network contains 6.67M parameters. We train separate model for burst SR, burst low-light image enhancement and burst denoising using L_1 loss only. While for SR on real data, we fine-tune our BIPNet with pre-trained weights on SyntheticBurst dataset using aligned L_1 loss [4]. The models are trained with Adam optimizer. Cosine annealing strategy [33] is employed to steadily decrease the learning rate from 10^{-4} to 10^{-6} during training. We use horizontal and vertical flips for data augmentation. Additional network details and visual results are provided in the supplementary material.

4.1. Burst Super-resolution

We perform SR experiments for scale factor $\times 4$ on the SyntheticBurst and (real-world) BurstSR datasets [4]

Datasets. (1) SyntheticBurst dataset consists of 46,839 RAW bursts for training and 300 for validation. Each burst contains 14 LR RAW images (each of size 48×48 pixels) that are synthetically generated from a single sRGB image. Each sRGB image is first converted to the RAW space using the inverse camera pipeline [6]. Next, the burst is generated with random rotations and translations. Finally, the LR burst is obtained by applying the bilinear downsampling followed by Bayer mosaicking, sampling and random noise addition operations. **(2) BurstSR** dataset consists of 200 RAW bursts, each containing 14 images. To gather these burst sequences, the LR images and the corresponding (ground-truth) HR images are captured with a smartphone camera and a DSLR camera, respectively. From 200 bursts, 5,405 patches are cropped for training and 882 for validation. Each input crop is of size 80×80 pixels.

SR results on synthetic data. The proposed BIPNet is trained for 300 epochs on training set while evaluated on validation set of SyntheticBurst dataset [4]. We compare our BIPNet with the several burst SR method such as High-ResNet [13], DBSR [4], LKR [30], and MFIR [5] for $\times 4$ upsampling. Table 1 shows that our method performs favorably well. Specifically, our BIPNet achieves PSNR gain of 0.37 dB over the previous best method MFIR [5] and 0.48 dB over the second best approach [30].

Visual results provided in Fig. 4 show that the SR images produced by BIPNet are more sharper and faithful than those of the other algorithms. Our BIPNet is capable of reconstructing structural content and fine textures, without introducing artifacts and color distortions. Whereas, the results of DBSR, LKR and MFIR contain splotchy textures and compromise image details.

Table 1. Performance evaluation on synthetic and real burst validation sets [4] for $\times 4$ burst super-resolution.

Methods	SyntheticBurst		(Real) BurstSR	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Single Image	36.17	0.909	46.29	0.982
HighRes-net [13]	37.45	0.92	46.64	0.980
DBSR [4]	40.76	0.96	48.05	0.984
LKR [30]	41.45	0.95	-	-
MFIR [5]	41.56	0.96	48.33	0.985
BIPNet (Ours)	**41.93**	**0.96**	**48.49**	**0.985**

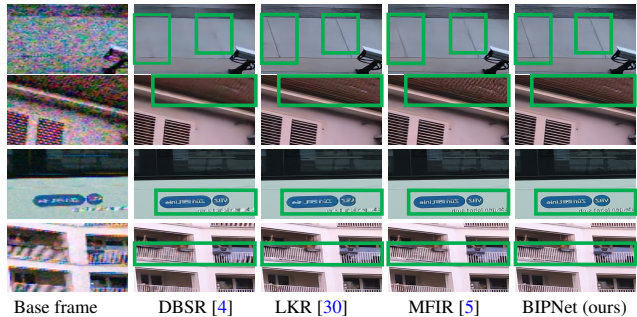Base frame   DBSR [4]   LKR [30]   MFIR [5]   BIPNet (ours)

Figure 4. Comparisons for $\times 4$ burst SR on SyntheticBurst [4]. Our BIPNet produces more sharper and clean results than other competing approaches.

Figure 5. Results for $\times 8$ burst SR on SyntheticBurst dataset [4]. (a) Base frame (b) BIPNet (Ours) (c) Ground truth. Our method effectively recovers image details in extremely challenging cases.

To show the effectiveness of our method BIPNet on large scale factor, we perform experiments for the $\times 8$ burst SR. We synthetically generate LR-HR pairs following the same procedure as we described above for the SyntheticBurst dataset. Visual results in Fig. 5 show that our BIPNet is capable of recovering rich details for such large scale factors as well, without any artifacts. Additional examples can be found in supplementary material.

SR results on real data. The LR input bursts and the corresponding HR ground-truth in BurstSR dataset suffer with minor misalignment as they are captured with different cameras. To mitigate this issue, we use aligned L1 loss for training and aligned PSNR/SSIM for evaluating our model, as in previous works [4, 5]. We fine-tuned the pre-trained BIPNet for 15 epochs on training set while evaluated on validation set of BurstSR dataset. The image quality scores

Table 2. Importance of BIPNet modules evaluated on SyntheticBurst validation set for ×4 burst SR.

Modules	A1	A2	A3	A4	A5	A6	A7	A8
Baseline	✓	✓	✓	✓	✓	✓	✓	✓
FPM (§3.1.1)		✓	✓	✓	✓	✓	✓	✓
DAM (§3.1.2)			✓	✓	✓	✓	✓	✓
RAF (§3.1.2)				✓	✓	✓	✓	✓
PBFF (§3.2)					✓	✓	✓	✓
MSF (§3.2)						✓	✓	✓
AGU (§3.3)							✓	✓
EBFA (§3.1)								✓
PSNR	36.38	36.54	38.39	39.10	39.64	40.35	41.25	41.55

Table 3. Importance of the proposed alignment and fusion module evaluated on SyntheticBurstSR dataset for ×4 SR.

	Methods	PSNR ↑	SSIM ↑
(a) Alignment	Explicit [4]	39.26	0.944
	TDAN [47]	40.19	0.957
	EDVR [50]	40.46	0.958
(b) Fusion	Addition	39.18	0.943
	Concat	40.13	0.956
	DBSR [4]	40.16	0.957
(c)	**BIPNet (Ours)**	**41.55**	**0.96**

Table 4. Burst low-light image enhancement methods evaluated on the SID dataset [8]. Our BIPNet advances state-of-the-art by 3.07 dB.

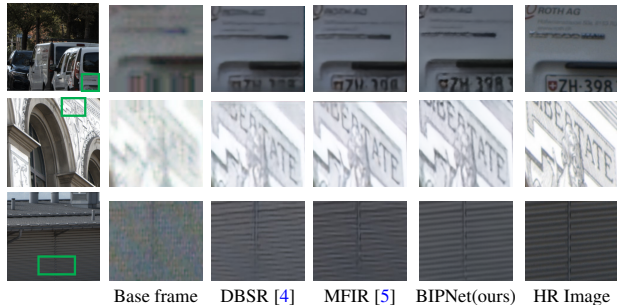| Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Chen *et al.* [8]	29.38	0.892	0.484
Maharjan *et al.* [36]	29.57	0.891	0.484
Zamir *et al.* [60]	29.13	0.881	0.462
Zhao *et al.* [63]	29.49	0.895	0.455
Karadeniz *et al.* [26]	29.80	0.891	0.306
BIPNet (Ours)	**32.87**	**0.936**	**0.305**

| Base frame | DBSR [4] | MFIR [5] | BIPNet(ours) | HR Image |

Figure 6. Comparisons for ×4 burst super-resolution on Real BurstSR dataset [4]. Our BIPNet produces more sharper and clean results than other competing approaches.

are reported in Table 1. Compared to the previous best approach MFIR [5], our BIPNet provides performance gain of 0.16 dB. The visual comparisons in Fig. 6 show that our BIPNet is more effective in recovering fine details in the reproduced images than other competing approaches.

Ablation Study. Here we present ablation experiments to demonstrate the impact of each individual component of our approach. All ablation models are trained for 100 epochs on SyntheticBurst dataset [5] for SR scale factor ×4. Results are reported in Table 2. For the baseline model, we employ Resblocks [32] for feature extraction, simple concatenation operation for fusion, and transpose convolution for upsampling. The baseline network achieves 36.38 dB PSNR. When we add the proposed modules to the baseline, the results improve significantly and consistently. For example, we obtain performance boost of 1.85 dB when we consider the deformable alignment module DAM. Similarly, RAF contributes 0.71 dB improvement towards the model. With our PBFF mechanism, the network achieves significant gain of 1.25 dB. AGU brings 1 dB increment in the upsampling stage. Finally, EBFA demonstrates its effectiveness in correcting alignment errors by providing 0.3 dB improvement in PSNR. Overall, our BIPNet obtains a compelling gain of 5.17 dB over the baseline method.

Finally, we perform ablation experiments to demonstrate the importance of the proposed EBFA and PBFF modules by replacing them with existing alignment and fusion modules. Table 3(a) shows that replacing our EBFA with other alignment modules have negative impact (PSNR drops at least over 1 dB). Similar trend can be observed when using fusion strategies other than our PBFF, see Table 3(b).

4.2. Burst Low-Light Image Enhancement

To further demonstrate the effectiveness of BIPNet, we perform experiments for burst low-light image enhancement. Given a low-light RAW burst, our goal is to generate a well-lit sRGB image. Since the input is mosaicked RAW burst, we use one level AGU to obtain the output.
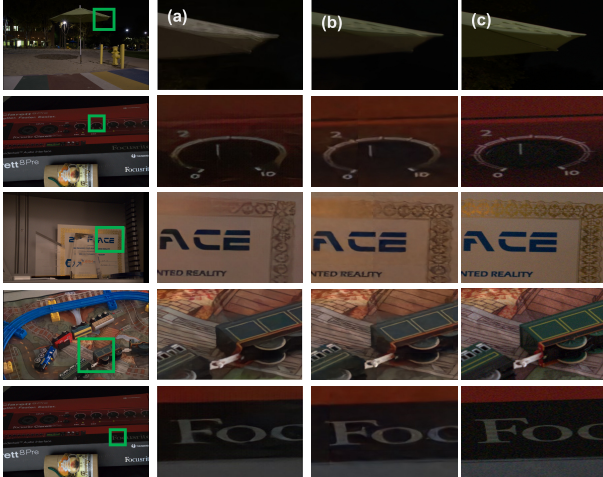
Dataset. SID dataset [8] consists of input RAW burst images captured with short-camera exposure in low-light conditions, and their corresponding ground-truth sRGB images. The Sony subset contains 161, 20 and 50 distinct burst sequences for training, validation and testing, respectively. We prepare 28k patches of spatial size 128×128 with burst size 8 from the training set of Sony subset of SID to train the network for 50 epochs.

Enhancement results. In Table 4, we report results of several low-light enhancement methods. Our BIPNet yields significant performance gain of 3.07 dB over the existing best method [26]. Similarly, the visual examples provided in Fig. 7 also corroborates the effectiveness of our approach.

4.3. Burst Denoising

Here, we demonstrate the effectiveness of the proposed BIPNet on the burst denoising task. BIPNet processes the input noisy sRGB burst and obtains a noise-free image. Since, there is no need to up-sample the extracted features, transpose convolution in the proposed AGU is replaced by a simple group convolution while rest of the network architecture is kept unmodified.

Dataset. We evaluate our approach on the grayscale and color burst denoising datasets introduced in [39] and [54]. These datasets contain 73 and 100 burst images respectively. In both datasets, a burst generated synthetically by applying random translations to the base image. The shifted images are then corrupted by adding heteroscedastic Gaussian noise [22] with variance $\sigma_r^2 + \sigma_s x$. The networks are then evaluated on 4 different noise gains $(1, 2, 4, 8)$, corresponding to noise parameters $(\log(\sigma_r), \log(\sigma_s)) \rightarrow$

Figure 7. Burst low-light image enhancement on Sony subset [8]. BIPNet better preserves color and structural details.

	Gain ∝ 1	Gain ∝ 2	Gain ∝ 4	Gain ∝ 8	Average
HDR+ [20]	31.96	28.25	24.25	20.05	26.13
BM3D [10]	33.89	31.17	28.53	25.92	29.88
NLM [7]	33.23	30.46	27.43	23.86	28.75
VBM4D [35]	34.60	31.89	29.20	26.52	30.55
KPN [39]	36.47	33.93	31.19	27.97	32.39
MKPN [37]	36.88	34.22	31.45	28.52	32.77
BPN [54]	38.18	35.42	32.54	29.45	33.90
MFIR [5]	39.37	36.51	33.38	29.69	34.74
BIPNet (Ours)	**41.26**	**38.74**	**35.91**	**31.35**	**36.81**

Table 5. Comparison of our method with prior approaches on the grayscale burst denoising set [39] in terms of PSNR. The results for existing methods are from [5].

	Gain ∝ 1	Gain ∝ 2	Gain ∝ 4	Gain ∝ 8	Average
KPN [39]	38.86	35.97	32.79	30.01	34.41
BPN [54]	40.16	37.08	33.81	31.19	35.56
MFIR [5]	42.21	39.13	35.75	32.52	37.40
BIPNet (Ours)	**42.28**	**40.20**	**37.85**	**34.64**	**38.74**

Table 6. Comparison with previous methods on the color burst denoising set [54] in terms of PSNR. The results for existing methods are from [5]. Our approach outperforms BPN on all four noise levels with average margin of 1.34dB.
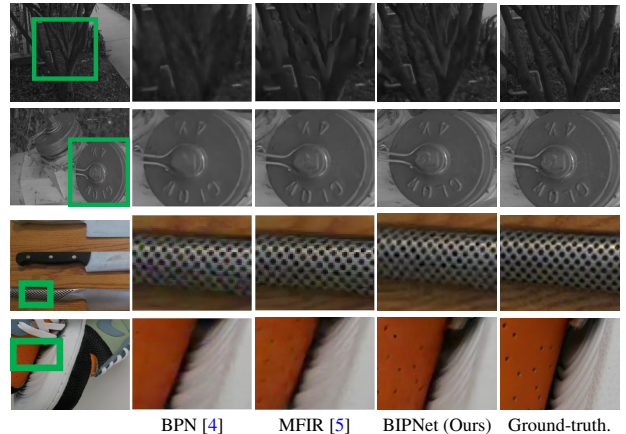

Figure 8. Comparisons for burst denoising on gray-scale [39] and color datasets [54]. Our BIPNet produces more sharper and clean results than other competing approaches. Many more examples are provided in the supplementary material.

(-2.2, -2.6), (-1.8, -2.2), (-1.4, -1.8), and (-1.1, -1.5), respectively. Note that the noise parameters for the highest noise gain (Gain ∝ 8) are unseen during training. Thus, performance on this noise level indicates the generalization of the network to unseen noise. Following [5], we utilized 20k samples from the Open Images [28] training set to generate the synthetic noisy bursts of burst-size 8 and spatial size 128×128. Our BIPNet is trained for 50 epochs both for the grayscale and color burst denoising tasks and evaluated on the benchmark datasets [39] and [54] respectively.

Burst Denoising results. We compare the proposed BIP-Net with the several approaches (KPN [39], MKPN [37], BPN [54] and MFIR [5]) both for grayscale and color burst denoising tasks. Table 5 shows that our BIPNet significantly advances state-of-the-art on grayscale burst denoising dataset [39]. Specifically, the BIPNet outperforms the previous best method MFIR [5] on all four noise levels. On average, BIPNet achieves 2.07 dB improvement over MFIR [5]. Similar performance trend can be observed in Table 6 for color denoising on color burst dataset [54]. Particularly, our BIPNet provides PSNR boost of 1.34 dB over previous best method MFIR [5]. Figure 8 shows that the images reproduced by BIPNet are more cleaner and sharper than those of the other approaches.

5. Conclusion

We present a burst image restoration and enhancement framework which is developed to effectively fuse complimentary information from multiple burst frames. Instead of late information fusion approaches that merge cross-frame information towards late in the pipeline, we propose the idea of pseudo-burst sequence that is created by combining the channel-wise features from individual burst frames. To avoid mismatch between pseudo-burst features, we pro-

pose an edge-boosting burst alignment module that is robust to camera-scene movements. The pseudo-burst features are enriched using multi-scale information and later progressively fused to create upsampled outputs. Our state-of-the-art results on three burst image restoration and enhancement tasks (super-resolution, low-light enhancement, denoising) corroborate the generality and effectiveness of BIPNet.

Acknowledgement

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018. 2

[2] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 2020. 2

[3] Aditya Arora, Muhammad Haris, Syed Waqas Zamir, Munawar Hayat, Fahad Shahbaz Khan, Ling Shao, and Ming-Hsuan Yang. Low light image enhancement via global and local context modeling. *arXiv:2101.00850*, 2021. 3

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021. 1, 2, 4, 5, 6, 7, 8

[5] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *ICCV*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[6] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019. 6

[7] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. In *CVPR*, 2005. 8

[8] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 3, 7, 8

[9] Kostadin Dabov, A. Foi, and K. Egiazarian. Video denoising by sparse 3d transform-domain collaborative filtering. *2007 15th European Signal Processing Conference*, pages 145–149, 2007. 3

[10] Kostadin Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *TIP*, 2007. 3, 8

[11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 2

[12] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *arXiv:2102.09000*, 2021. 1, 3

[13] 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:2002.06460*, 2020. 2, 6

[14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2015. 2

[16] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multi-frame demosaicing and super-resolution from undersampled color images. In *Computational Imaging II*, 2004. 2

[17] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *CG&A*, 2002. 2

[18] C. Godard, K. Matzen, and Matthew Uyttendaele. Deep burst denoising. In *ECCV*, 2018. 3

[19] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 2

[20] S. W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, J. Barron, F. Kainz, Jiawen Chen, and M. Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *TOG*, 2016. 3, 8

[21] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *TOG*, 2016. 1

[22] G. Healey and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *TPAMI*, 1994. 7

[23] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, 2018. 2

[24] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017. 1

[25] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP*, 1991. 2

[26] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *arXiv:2006.09845*, 2020. 3, 7, 8

[27] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 2

[28] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available at https://github.com/openimages*, 2017. 8

[29] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *CVPR*, 2017. 2

[30] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *ICCV*, 2021. 2, 6

[31] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[32] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 7

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv:1608.03983*, 2016. 6

[34] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian. Video denoising using separable 4d nonlocal spatiotemporal transforms. In *Electronic Imaging*, 2011. 3

[35] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *TIP*, 2012. 3, 8

[36] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *ICME*, 2019. 3, 7

[37] Talmaj Marinc, V. Srinivasan, S. Gül, C. Hellge, and W. Samek. Multi-kernel prediction networks for denoising of burst images. In *ICIP*, 2019. 3, 8

[38] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 3

[39] Ben Mildenhall, J. Barron, Jiawen Chen, Dillon Sharlet, R. Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, 2018. 3, 7, 8

[40] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *TGRS*, 2019. 2

[41] Shmuel Peleg, Danny Keren, and Limor Schweitzer. Improving image resolution using subpixel motion. *PRL*, 1987. 2

[42] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017. 2

[43] 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 *CVPR*, 2016. 2, 5

[44] Henry Stark and Peyma Oskoui. High-resolution image recovery from image-plane arrays, using convex projections. *JOSA A*, 1989. 2

[45] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 2

[46] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017. 2

[47] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 2, 4, 7

[48] Roger Y. Tsai and Thomas S. Huang. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1984. 2

[49] Evgeniya Ustinova and Victor Lempitsky. Deep multi-frame face super-resolution. *arXiv:1709.03196*, 2017. 2

[50] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 2, 7

[51] 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 *ECCVW*, 2018. 2

[52] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *ICCV*, 2015. 2

[53] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *TOG*, 2019. 1, 2

[54] Zhihao Xia, Federico Perazzi, M. Gharbi, Kalyan Sunkavalli, and A. Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *CVPR*, 2020. 3, 7, 8

[55] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *TIP*, 2010. 2

[56] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 3

[57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: real image restoration via improved data synthesis. In *CVPR*, 2020. 2

[58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020. 2

[59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 2

[60] Syed Waqas Zamir, Aditya Arora, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Learning digital camera pipeline for extreme low-light imaging. *Neurocomputing*, 2021. 3, 7

[61] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2, 3

[62] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *TPAMI*, 2020. 2

[63] Di Zhao, Lan Ma, Songnan Li, and Dahai Yu. End-to-end denoising of dark burst images using recurrent fully convolutional networks. *arXiv:1904.07483*, 2019. 3, 7

[64] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 2, 3