BURST IMAGE RESTORATION AND ENHANCEMENT

Anonymous authors
Paper under double-blind review

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. The pseudo-burst representations encode channel-wise features from the original burst images, thus making it easier for the model to learn distinctive information offered by multiple burst frames. 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 preprocessed 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 and low-light image enhancement tasks. Our codes and models will be publicly released.

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 (Ignatov et al., 2017). However, the physical constraints of smartphone 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 (Delbracio et al., 2021). Similarly, small pixel cavities accumulate less light therefore yielding low-dynamic range images. To alleviate these issues, one natural solution is to use burst (multi-frame) photography instead of single-frame processing (Hasinoff et al., 2016).

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 (almost always the case in handheld devices) will cause misalignment issues, thereby leading to ghosting and blurring artifacts in the output image (Wronski et al., 2019). 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 (Bhat et al., 2021a;b) extract and align features of burst images separately and usually employ late feature fusion mechanisms, which can hinder flexible information 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 tensor in the pseudo-burst contains complimentary properties of all input frames in the burst sequence.

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 (Bhat et al., 2021a;b) 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 (Zhu et al., 2019) that can effectively adapt to the given problem. Further, we integrate the edge boosting refinement via back-projection operation (Haris et al., 2018) 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 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 and burst low-light image enhancement. In super-resolution (SR), upsampling is the key step for image reconstruction. Existing burst SR methods (Bhat et al., 2021a;b) first fuse the multi-frame features, and then use pixel-shuffle operation (Shi et al., 2016) 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.

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 and low-light image enhancement tasks. We provide visual examples and comprehensive ablation experiments to highlight the main contributing factors in proposed solution (Sec. 4).

2 Related Work

Single Image Super-resolution (SISR). Since the first CNN-based work (Dong et al., 2014), data-driven approaches have achieved tremendous performance gains over the conventional counterparts (Yang et al., 2010; Freeman et al., 2002). The success of CNNs is mainly attributed to their architecture design. Given a low-resolution image (LR), early methods learn to directly generate latent super-resolved image (Dong et al., 2014; 2015). In contrast, recent approaches learns to produce high frequency residual to which LR image is added to generate the final SR output (Tai et al., 2017; Hui et al., 2018). Other notable SISR network designs employ recursive learning (Kim et al., 2016; Ahn et al., 2018), progressive reconstruction (Wang et al., 2015; Lai et al., 2017), attention mechanisms (Zhang et al., 2018a; Dai et al., 2019; Zhang et al., 2020), and generative adversarial networks (Wang et al., 2018; Sajjadi et al., 2017; Ledig et al., 2017). The SISR approaches cannot handle multi-degraded frames from an input burst, while our approach belong to multi-frame SR family that allows effectively merging cross-frame information towards a high-resolution output.

Multi-Frame Super-Resolution (MFSR). Tsai & Huang (1984) 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 (Tsai & Huang, 1984), several other works aim to improve results by incorporating image priors in HR reconstruction process (Stark & Oskoui, 1989), and making algorithmic choices such as iterative back-projection (Peleg et al., 1987; Irani & Peleg, 1991). Farsiu et al. (2004) designed 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 (Wronski et al., 2019), to increase spatial resolution of face images (Ustinova & Lempitsky, 2017), and in satellite imagery (Deudon et al., 2020; Molini et al., 2019). Lecouat et al. (2021) 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. (2021a) propose a multi-frame burst SR method that first aligns burst image features using an explicit PWCNet (Sun et al., 2018) 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 (Tian et al., 2020; Wang et al., 2019) show that the deformable convolution (Zhu et al., 2019) 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 (Delbracio et al., 2021; Zamir et al., 2021). 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 (Chen et al., 2018) 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 (Maharjan et al., 2019) and by employing a combined pixel-wise and perceptual loss (Zamir et al., 2021). Zhao et al. (2019) 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. (2020) 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 inter-frame information interaction which we address in this work.

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, and burst low-light image enhancement. 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 a new architecture 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 burst image processing framework. 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 1^{st} frame for simplicity) we propose an alignment module based on modulated deformable convolutions (Zhu et al., 2019). However, existing deformable convolution is not explicitly designed to handle noisy RAW data. Therefore, we propose a feature

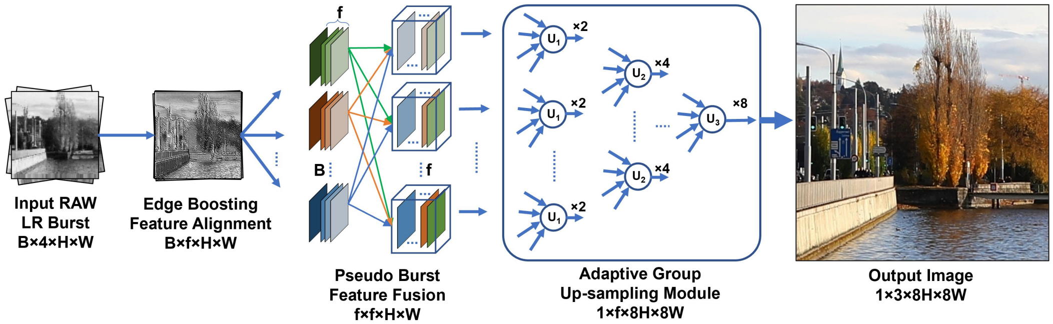


FIGURE 1: Holistic diagram of our burst image processing approach. Our network BIPNet takes as input a RAW image burst and generates a high-quality RGB image. BIPNet has three key stages. (1) Edge boosting feature alignment to remove noise, and inter-frame spatial and color misalignment. (2) Pseudo-burst 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.

processing module to reduce noise in the initial burst features. Our edge boosting feature alignment (EBFA) module (Fig. 2(b)) 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(a), employs residual-in-residual learning that allows abundant low-frequency information to pass easily via skip connections (Zhang et al., 2018b). 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 (e.g., image/video super-resolution (Mei et al., 2020) and extreme low-light image enhancement (Arora et al., 2021)), we utilize a global context attention (GCA) mechanism to refine the latent representation produced by residual block, as illustrated in Fig. 2(a). Let $\left\{\boldsymbol{x}^b\right\}_{b\in[1:B]} \in \mathbb{R}^{B\times f\times H\times W}$ be an initial latent representation of the burst having B number of burst images and f number of feature channels, our residual global context attention block (RGCAB in Fig. 2(a)) is defined as:

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

where $\bar{\boldsymbol{x}}^b = W_3\left(\gamma\left(W_3\left(\boldsymbol{x}^b\right)\right)\right)$ and $\alpha\left(\bar{\boldsymbol{x}}^b\right) = \bar{\boldsymbol{x}}^{\boldsymbol{b}} + W_1\left(\gamma\left(W_1\left(\Psi\left(W_1\left(\bar{\boldsymbol{x}}^b\right)\right)\odot\bar{\boldsymbol{x}}^b\right)\right)\right)$. 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, $\odot$ represents element-wise 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 features of the base $\boldsymbol{y}^{b_r}$ frame[1]. It processes $\boldsymbol{y}^b$ and $\boldsymbol{y}^{b_r}$ through an offset convolution layer (W^o) and predicts the offset Θ and modulation scalar Δm values for $\boldsymbol{y}^b$. With Θ, Δm and $\boldsymbol{y}^b$, the aligned features $\bar{\boldsymbol{y}}^b$ can be computed by the deformable convolution:

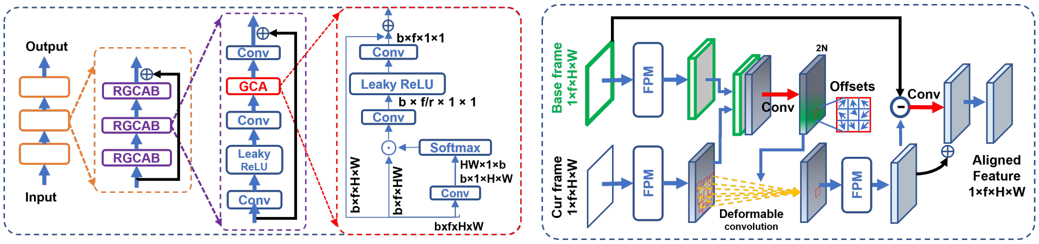
$$\bar{\boldsymbol{y}}^b = W^d\left(\boldsymbol{y}^b,\ \Theta,\ \Delta m\right), \text{ and } \Delta m = W^o\left(\boldsymbol{y}^b,\ \boldsymbol{y}^{b_r}\right), \tag{2}$$

where, W^d and W^o represent the deformable and offset convolutions, respectively. The set $\Theta = \{\Delta n_i \mid i = 1, \cdots, |\Re|\}$ denotes offsets where $\Re$ =(-1, 1), (-1, 0), ..., (1,1) is a regular grid of 3×3 kernel. While, Δm lies in the range [0, 1] for each n_i. More specifically, each position n on the aligned feature map $\bar{\boldsymbol{y}}^b$ is obtained as:

$$\bar{\boldsymbol{y}}^b_n = \sum_{n_i\in\Re} W^d_{n_i}\ \ \boldsymbol{y}^b_{(n+n_i+\Delta n_i)}\cdot\Delta m_{n_i} \tag{3}$$

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

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



(a) Feature Processing Module (FPM)

(b) Edge Boosting Feature Alignment Module (EBFA)

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. $\odot$ represents element-wise multiplication.

The proposed EBFA module is inspired from the deformable alignment module (DAM) (Tian et al., 2020) with the following difference. 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:

$$\boldsymbol{e}^b = \bar{\boldsymbol{y}}^b + W_3\left(\bar{\boldsymbol{y}}^b - \boldsymbol{y}^{b_r}\right), \tag{4}$$

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(b) 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 (Bhat et al., 2021a;b) 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 $e = \left\{\boldsymbol{e}^b_c\right\}^{b\in[1:B]}_{c\in[1:f]}$ of burst size B and f number of channels, the pseudo-burst is generated by,

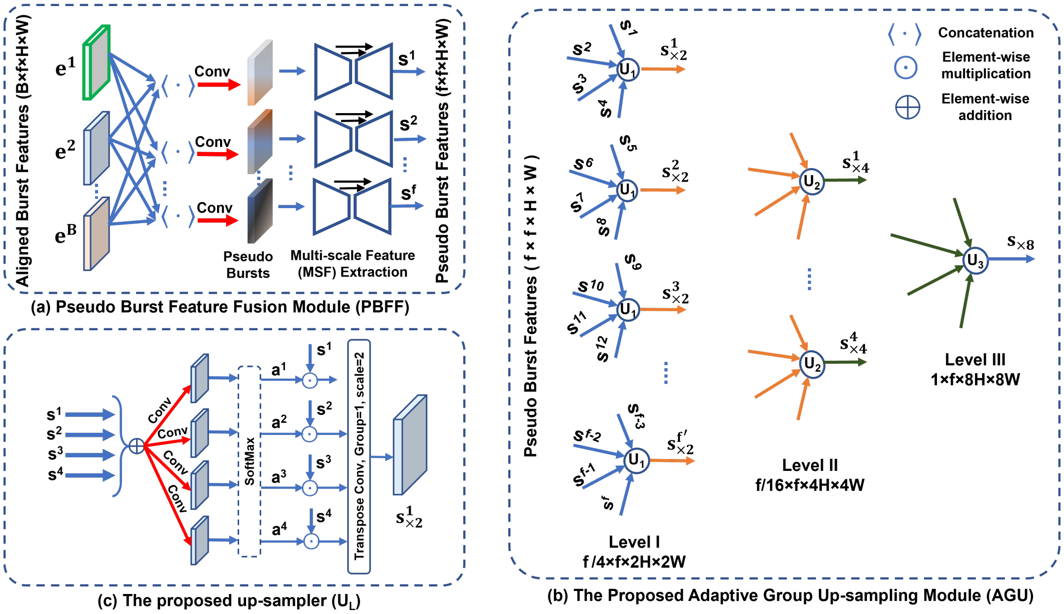
$$\boldsymbol{S}^c = W^\rho\left(\left\langle \boldsymbol{e}^1_c,\ \boldsymbol{e}^2_c,\ \cdots,\ \boldsymbol{e}^B_c\right\rangle\right), \quad s.t. \quad c \in [1:f], \tag{5}$$

where, $\langle\cdot\rangle$ represents concatenation, $\boldsymbol{e}^1_c$ is the c^{th} feature map of 1^{st} 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 $f = 64$.

Even after generating pseudo-bursts, obtaining its deep representation is essential. For this we use a light-weight (3-level) UNet to extract multi-scale features (MSF) from pseudo-bursts. We use shared weights in the UNet, and also employ our FPM (Sec. 3.1.1) 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 (Bhat et al., 2021a;b) use pixel-shuffle layer (Shi et al., 2016) to perform upsampling 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.



(a) Pseudo Burst Feature Fusion Module (PBFF)

(c) The proposed up-sampler (U_L)

(b) The Proposed Adaptive Group Up-sampling Module (AGU)

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) UNet 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.

Fig. 3(b) shows the proposed adaptive group upsampling (AGU) module that takes as input the feature maps $\boldsymbol{S} = \{\boldsymbol{S}^c\}_{c\in[1:f]}$ 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 ($\boldsymbol{a}^c$), carrying attention weights for each pixel location. The dense attention maps are element-wise applied to the respective burst features. Finally, the upsampled response for a given group of features $\hat{\boldsymbol{S}}^g = \left\{\boldsymbol{S}^i : i \in [(g-1)*4+1 : g*4]\right\}^{g\in[1:f/4]} \subset \boldsymbol{S}$ and associated attention maps $\hat{\boldsymbol{a}}^g$ at the first upsampling level (Level I in Fig. 3(b)) is formulated as:

$$\boldsymbol{S}^g_{\times 2} = W_T\left(\left\langle \hat{\boldsymbol{S}}^g \odot \hat{\boldsymbol{a}}^g \right\rangle\right), \text{ and } \hat{\boldsymbol{a}}^g = \psi\left(W_1\left(W_1\left(\sum_{i=(g-1)*4+1}^{g*4} \boldsymbol{S}^i\right)\right)\right), \qquad (6)$$

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

To perform burst SR of scale factor ×4, we need in fact ×8 upsampling[2]. In AGU, we employ three levels of progressive upsampling due the dimensionality of the pseudo-bursts ($\boldsymbol{S}^c \in \mathbb{R}^{64\times 64\times H\times W}$). We form 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 BIPNet and SOTA approaches on real and synthetic datasets for **(a)** burst super-resolution, and **(b)** burst low-light image enhancement. The source code and trained models will be made available to the public.

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

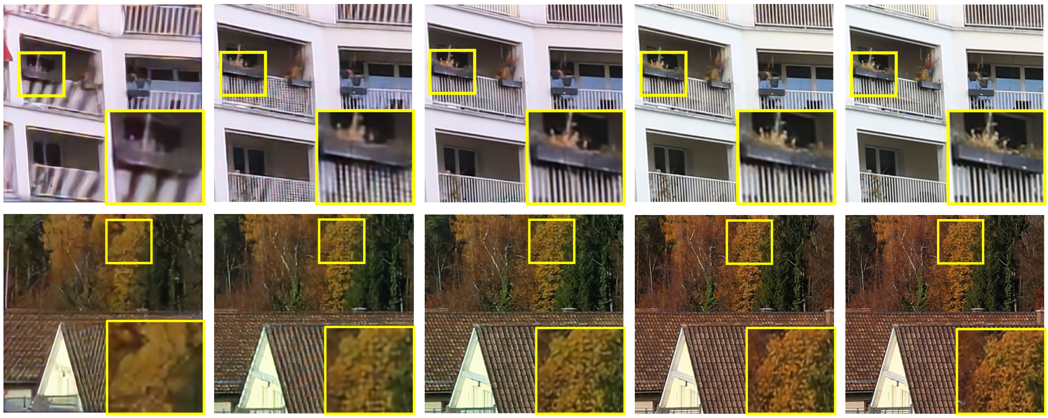
[2] The actual task is to upsample by ×4, additional ×2 is due to the mosaicked RAW LR frames.

TABLE 1: Performance evaluation on synthetic and real burst validation sets (Bhat et al., 2021a) for ×4 burst super-resolution.

Methods	SyntheticBurst		(Real) BurstSR	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Single Image	36.17	0.909	46.29	0.982
Deudon et al. (2020)	37.45	0.92	46.64	0.980
DBSR (Bhat et al., 2021a)	40.76	0.96	48.05	0.984
LKR (Lecouat et al., 2021)	41.45	0.95	-	-
MFIR (Bhat et al., 2021b)	41.56	0.96	48.33	0.985
BIPNet (Ours)	**41.93**	**0.96**	**48.49**	**0.985**

TABLE 2: Importance of BIPNet modules evaluated on synthetic burst 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



Demosaick + SISR

DBSR (Bhat et al., 2021a)

LKR (Lecouat et al., 2021)

BIPNet (Ours)

HR Ground-truth.

FIGURE 4: Visual results for ×4 burst SR on SyntheticBurst dataset (Bhat et al., 2021a). Compared to other approaches, our BIPNet yields images that are more vivid and visually closer to the ground-truth.

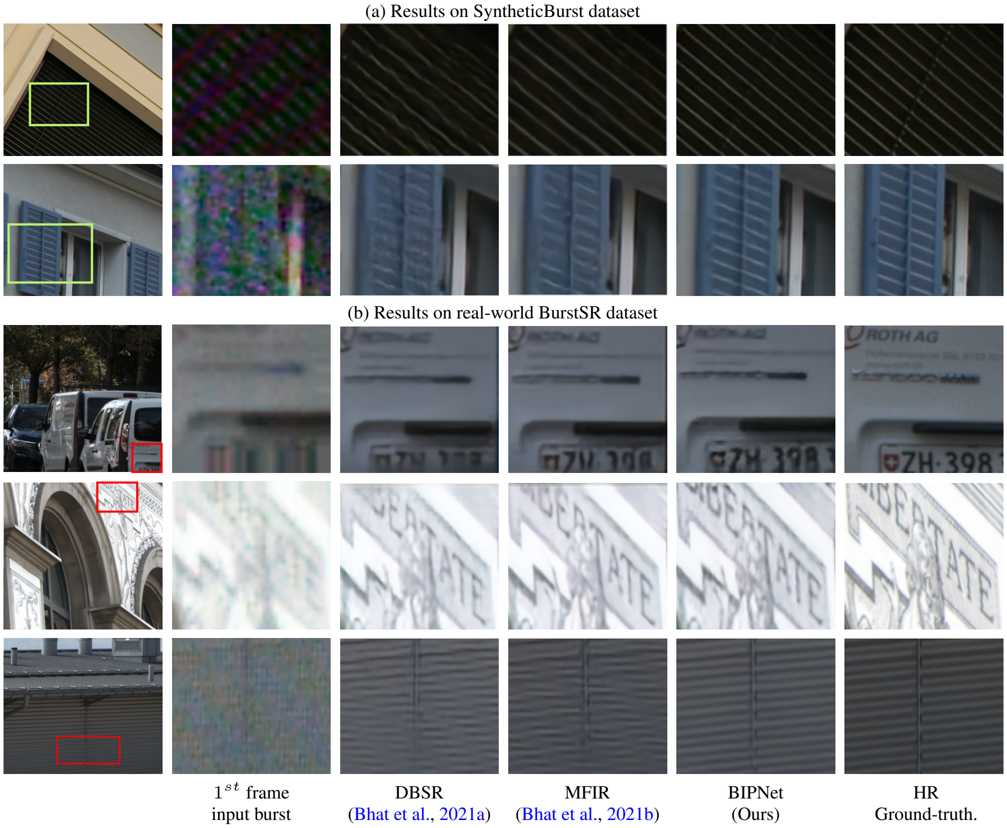
train two separate models: (1) SR on synthetic data, and (2) image enhancement. The models are trained with Adam optimizer for 300 epochs for synthetic SR and 100 epochs for image enhancement. While for SR on real data, we fine-tuned our BIPNet for 15 epochs with pre-trained weight on SyntheticBurst dataset. Cosine annealing strategy (Loshchilov & Hutter, 2016) 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 are given in Appendix B.

4.1 BURST SUPER-RESOLUTION

We perform SR experiments for scale factor ×4 on the SyntheticBurst and BurstSR (real-world) datasets, recently presented in (Bhat et al., 2021a).

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 (Brooks et al., 2019). 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. We evaluate our BIPNet with the several burst SR method such as HighResNet (Deudon et al., 2020), DBSR (Bhat et al., 2021a), LKR (Lecouat et al., 2021), and MFIR (Bhat et al., 2021b) for ×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 (Bhat et al., 2021b) and 0.48 dB over the second best approach (Lecouat et al., 2021).



(a) Results on SyntheticBurst dataset

(b) Results on real-world BurstSR dataset

FIGURE 5: Comparisons for ×4 burst super-resolution on SyntheticBurst and BurstSR datasets (Bhat et al., 2021a). Our BIPNet produces more sharper and clean results than other competing approaches. Many more examples are provided in Appendix C.

Fig. 4 shows that the reproductions of the competing algorithms contain color shifts (top row), and less vivid than those produced by our BIPNet (bottom row). Similarly, visual results provided in Fig. 5(a) show that the super-resolved images produced by our method are more sharper and faithful to the ground-truth 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 reproductions of DBSR, and MFIR contain splotchy textures.

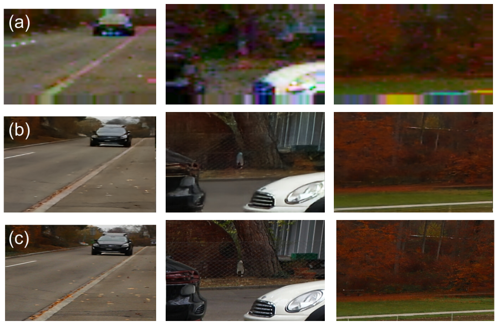


FIGURE 6: Results for ×8 burst SR on SyntheticBurst dataset (Bhat et al., 2021a). (a) 1^{st} burst frame. (b) Our BIPNet. (c) Ground truth. Our method effectively recovers image details in extremely challenging cases.

To show the effectiveness of our method on large scale factor, we perform experiments for the ×8 burst SR. We synthetically generate LR-HR pairs following the same procedure as we described above for SyntheticBurst dataset. Visual results in Fig. 6 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 Appendix C.

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 (Bhat et al., 2021a;b). To perform training on real BurstSR dataset for ×4 upsampling, we initialize our BIPNet with the pre-trained weights on SyntheticBurst dataset. The image quality scores are reported in Table 1. Compared to the previous best approach MFIR (Bhat et al., 2021b), our BIPNet provides performance gain of 0.16 dB. The visual comparisons in Fig. 5(b)

TABLE 3: Burst low-light image enhancement methods evaluated on the SID dataset (Chen et al., 2018). Our BIPNet provides 3.07 dB improvement over the previous best algorithm in literature.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Chen et al. (2018)	29.38	0.892	0.484
Maharjan et al. (2019)	29.57	0.891	0.484
Zamir et al. (2021)	29.13	0.881	0.462
Zhao et al. (2019)	29.49	0.895	0.455
Karadeniz et al. (2020)	29.80	0.891	0.306
BIPNet (Ours)	**32.87**	**0.9365**	**0.305**

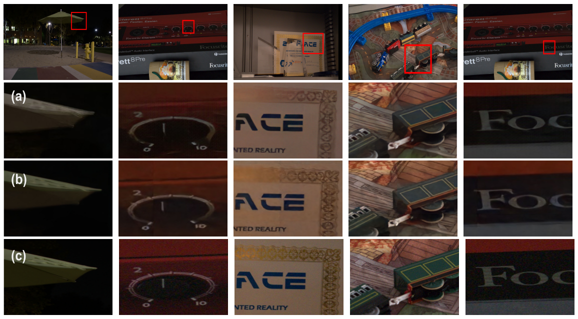


FIGURE 7: Burst low-light image enhancement on Sony subset (Chen et al., 2018). (a) Karadeniz et al. (2020). (b) BIPNet (Ours). (c) Ground truth. Our BIPNet better preserves color and structural detail in the enhanced images.

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 (Bhat et al., 2021b) for SR scale factor $\times 4$. Results are reported in Table 2. For the baseline model, we employ Resblocks (Lim et al., 2017) 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 demonstrate 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.

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 (Chen et al., 2018) consists of input RAW burst images captured with short-camera exposure in low-light conditions, and their corresponding ground-truth sRGB images. Following Karadeniz et al. (2020), we use the Sony subset of SID to train the network. The Sony subset contains 161, 20 and 50 distinct burst sequences for training, validation and testing, respectively.

Burst low-light image enhancement results. In Table 3, we report results of several low-light enhancement methods. Our BIPNet yields significant performance gain of 3.07 dB over the existing best method (Karadeniz et al., 2020). Similarly, the visual example provided in Fig. 7 also corroborates the effectiveness of our approach.

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. In order to avoid mismatch between pseudo-burst features, we propose 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 two image restoration and enhancement applications corroborate the generality and effectiveness of BIPNet.

References

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

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. 4

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021a. 1, 2, 3, 5, 7, 8, 14, 15, 16

Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *ICCV*, 2021b. 1, 2, 5, 7, 8, 9, 14, 15

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019. 7

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 3, 9

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

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

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. 3, 7

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

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

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

William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 2002. 2

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

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

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

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

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

Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *arXiv:2006.09845*, 2020. 3, 9

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

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

Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *ICCV*, 2021. 3, 7

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

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 9

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv:1608.03983*, 2016. 7

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, 9

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. 4

Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *TGRS*, 2019. 3

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

Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017. 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 *CVPR*, 2016. 2, 5

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

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. 3

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

Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 3, 5

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

Evgeniya Ustinova and Victor Lempitsky. Deep multi-frame face super-resolution. *arXiv:1709.03196*, 2017. 3

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

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 3

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

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 TOG*, 2019. 1, 3

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

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, 9

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*, 2018a. 2

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*, 2018b. 4

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

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, 9

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

A Appendix

Here we describe the architectural details of the proposed BIPNet, and present additional visual comparisons with existing state-of-the-art approaches for burst SR.

B Network Architecture Details

B.1 Edge Boostting Feature Alignment (EBFA)

The proposed feature processing module (FPM) consists of three residual-in-residual (RiR) groups. Each RiR is made up of three RGCAB and each RGCAB contains basic residual block followed by global context attention as shown in Fig.2(a). Although, the deformable convolution layer is shown only once in the Fig.2(b) for simplicity, we apply three such layers to improve the feature alignment ability of the proposed EBFA module.

B.2 Pseudo Burst Feature Fusion (PBFF)

The proposed PBFF is as shown in Fig.3(a). It consists of multi-scale feature (MSF) extraction module which is made up of light-weight 3-level UNet. We employed one FPM (with 2 RiR and 2 RGCAB in each RiR) after each downsample and upsample convolution layer. Number of convolution filters are increased by a factor of 1.5 at each downsampling and decreased by the rate of 1.5 after each upsampling operation. We simply add features extracted at each level to the upsampled features via skip connections.

B.3 Adaptive Group Up-sampling (AGU)

Our AGU module is shown in Fig.3(c). It aggregates the input group of pseudo bursts and pass them through a bottleneck convolution layer of kernel size 1×1 followed by a set of four parallel convolution layers, each with kernel size of 1×1 and 64 filters. Further, the outputs from previous step are passed through the softmax activation to obtain the dense attention maps.

C Additional Visual Results for Burst SR

The results provided in Fig. C.1 and Fig. C.2 show that our method performs favorably on both real and synthetic images for the scale factor . The true potential of the proposed approach is demonstrated in C.3, where it successfully recover the fine-grained details from extremely challenging LR burst images (that are downscaled by a factor of ×8).

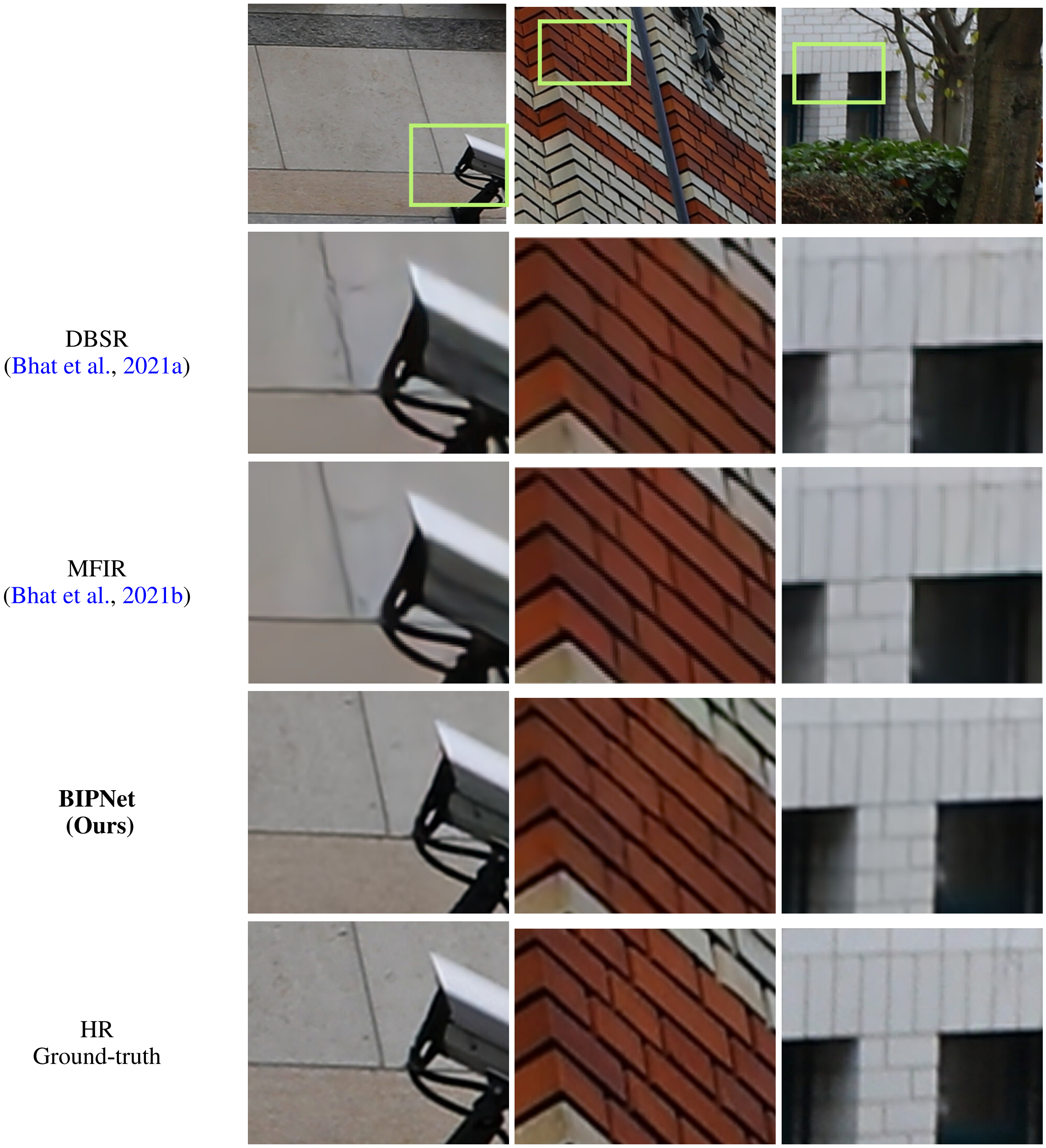


FIGURE C.1: Comparison for ×4 burst SR on SyntheticBurst dataset (Bhat et al., 2021a).

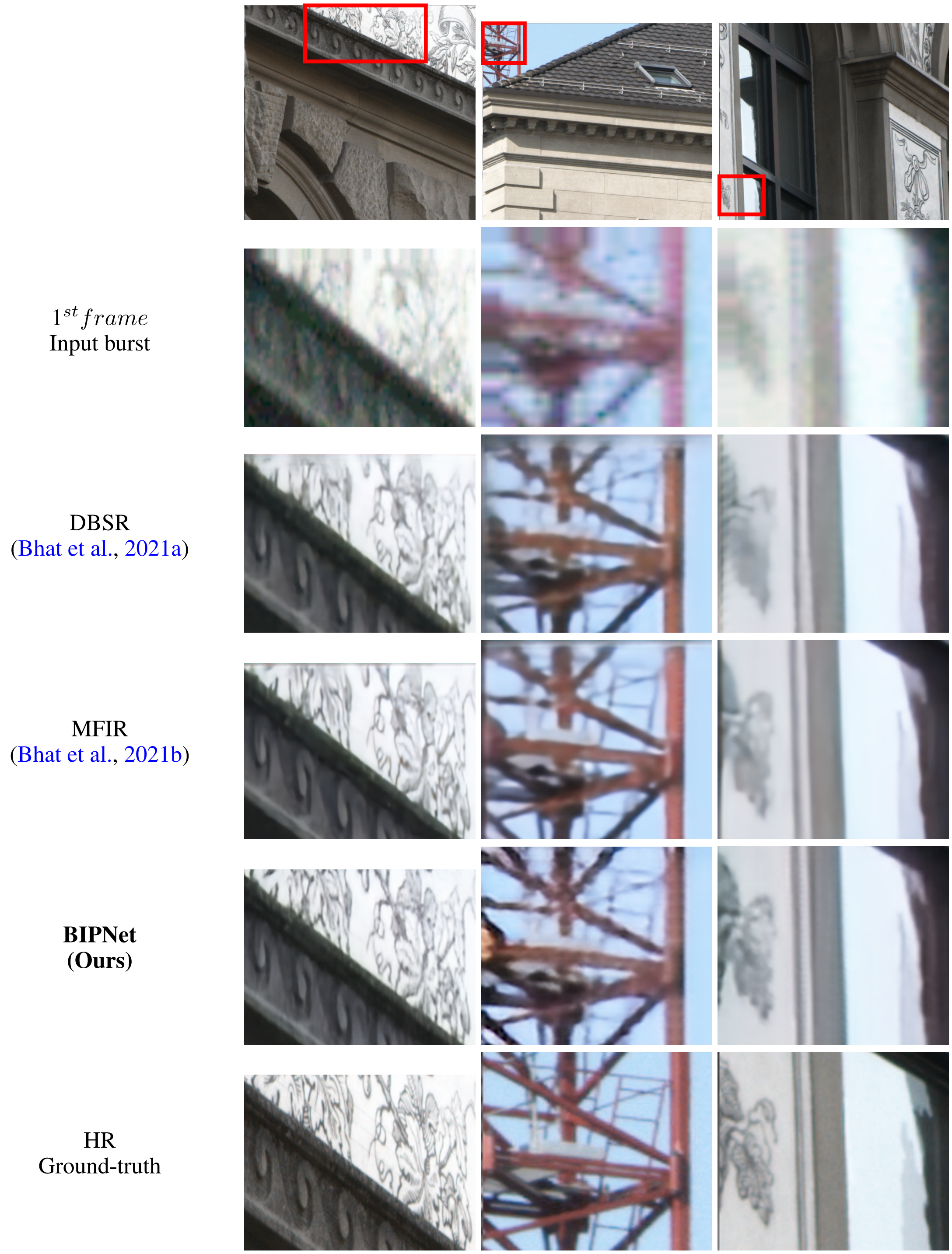
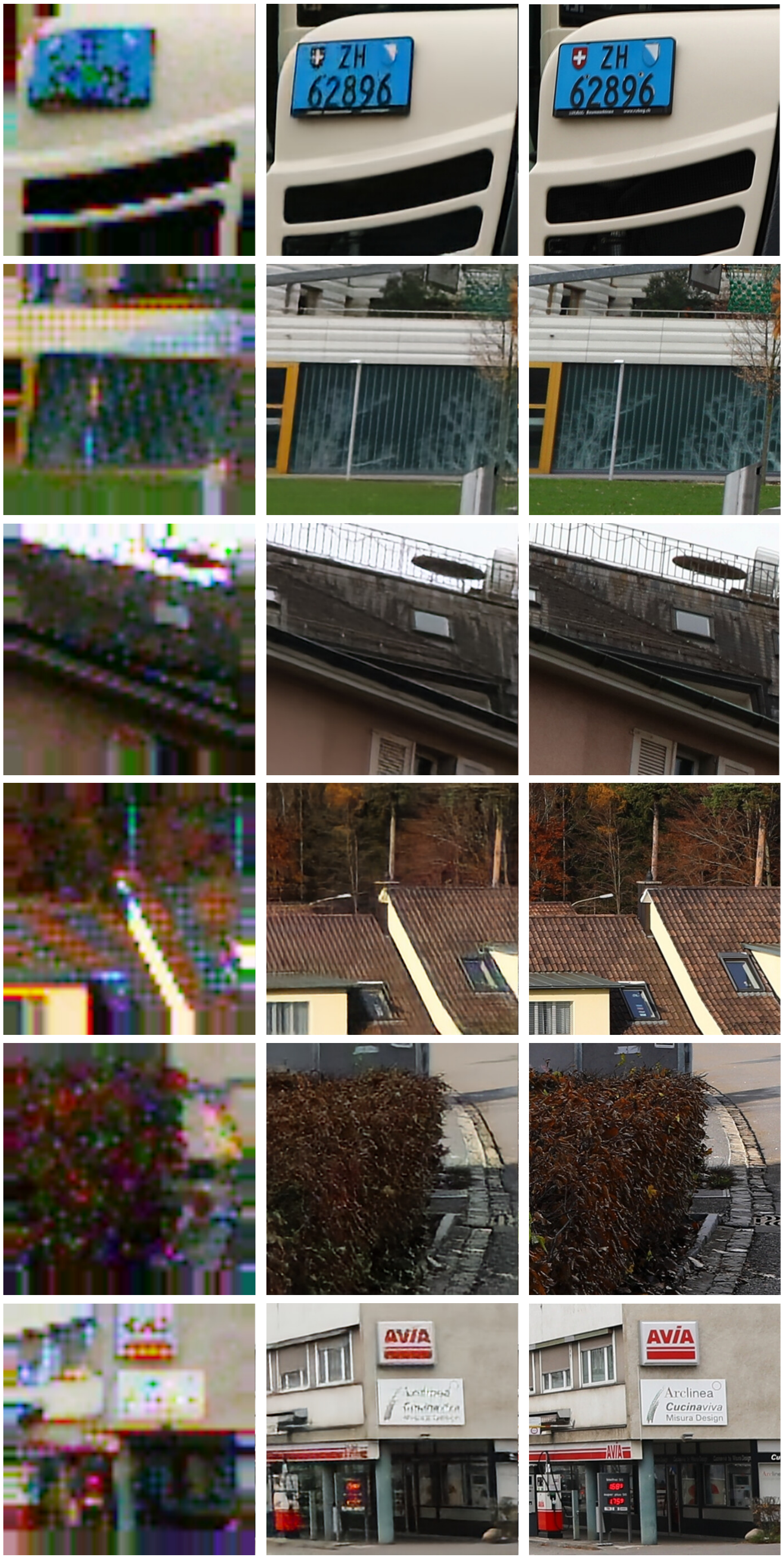


FIGURE C.2: Comparison for ×4 burst SR on real BurstSR dataset (Bhat et al., 2021a). The reproductions of our BIPNet are perceptually more faithful to the ground-truth than those of other methods.



1st frame of input burst **BIPNet (Ours)** Ground-truth

FIGURE C.3: Results for ×8 SR on images from SyntheticBurst dataset (Bhat et al., 2021a). Our method effectively recovers image details in extremely challenging cases.