This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

Burst Image Restoration and Enhancement

Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, Ming-Husan Yang

(Invited Paper)

Abstract—Burst Image Restoration aims to reconstruct a high-quality image by efficiently combining complementary inter-frame information. However, it is quite challenging since individual burst images often have inter-frame misalignments that usually lead to ghosting and zipper artifacts. To mitigate this, we develop a novel approach for burst image processing named BIPNet that focuses solely on the information exchange between burst frames and filter-out the inherent degradations while preserving and enhancing the actual scene details. Our central idea is to generate a set of pseudo-burst features that combine complementary information from all the burst frames to exchange information seamlessly. However, due to inter-frame misalignment, the information cannot be effectively combined in pseudo-burst. Thus, we initially align the incoming burst features regarding the reference frame using the proposed edge-boosting feature alignment. Lastly, we progressively upscale the pseudo-burst features in multiple stages while adaptively combining the complementary information. Unlike the existing works, that usually deploy single-stage up-sampling with a late fusion scheme, we first deploy a pseudo-burst mechanism followed by the adaptive-progressive feature up-sampling. The proposed BIPNet significantly outperforms the existing methods on burst super-resolution, low-light image enhancement, low-light image super-resolution, and denoising tasks. The pre-trained models and source code are available at https://github.com/akshaydudhane16/BIPNet.

Index Terms—Feature alignment, Feature fusion, Burst processing, Super-resolution, Denoising, Low-light image enhancement

1 Introduction

With the escalating popularity of built-in smartphone cameras, the demand for capturing high-quality images has drawn much attention. However, relative to the larger standalone cameras, e.g., a DSLR, smartphone cameras have several limitations due to the constraints placed on them in order to be integrated into a smartphone's thin profile. The most eminent hardware limitations are the small camera sensor size and the associated lens optics that reduce their spatial resolution and dynamic range [15], thus making noise much more of a problem during smartphone capture. As a remedy for these hardware limitations and to improve the overall image quality on smartphones, image restoration, and enhancement techniques have become indispensable. Image restoration techniques are employed to rectify the deteriorated aspects of an image caused by noise, blurriness, or other artifacts introduced during the image capture process. On the other hand, image enhancement techniques focus on improving the visual appearance of an image such that the viewer deems it pleasant.

In the literature, various approaches for single image restoration and enhancement have been developed to improve image quality. Nevertheless, achieving a truly high-quality output can be challenging due to the limited scene information within a single image. A promising solution gaining traction is the adoption of burst photography, capturing a series of photos in rapid succession rather than relying on a single shot. Burst processing approaches capture multiple shifted images, which are then integrated into a single high-quality output image to retrieve the non-redundant high-frequency details. Three critical factors in designing a novel burst processing approach include feature alignment, fusion, and high-quality image restoration. Generally, the biggest challenge for any burst processing approach is the accuracy of the alignment process, as the scene motion of dynamically moving objects and camera motion results in blurry output. Thus, it is crucial to design a module to facilitate accurate alignment, as the subsequent fusion and reconstruction modules must be robust to misalignment to generate an artifact-free image. We further note that existing burst processing approaches [5], [6] extract and explicitly align the burst features by employing late feature fusion mechanisms, which can hinder flexible information exchange among multiple frames. To address these issues, we present a novel burst image processing approach named BIPNet, which enables inter-frame communication through the proposed pseudo-burst feature fusion mechanism. Specifically, a pseudo-burst is formed by exchanging information within frames, where each feature comprises complementary properties from all burst frames.

The success of the pseudo burst mechanism for inter-frame communication depends upon the alignment among the burst frames. Therefore, it is crucial to accurately align the input burst frames to aggregate the apt pixel-level cues in the later stages before creating pseudo-bursts. We observe that the existing works DBSR [5] and MFIR [6] generally deploy explicit motion estimation techniques (e.g., optical flow) for aligning the burst features, which are typically bulky pre-trained modules (trained on additional data) and cannot be fully blended within an end-to-end learnable pipeline. However, this can result in upstretching of the cascaded errors during the flow estimation stage, and its further propagation to the warping and processing stages

- *Akshay Dudhane, Salman Khan and Fahad Shahbaz Khan are with Mohammed Bin Zayed University of Artificial Intelligence, UAE. E-mail: akshay.dudhane@mbzuai.ac.ae*
- *Syed Waqas Zamir is with the Inception Institute of Artificial Intelligence, UAE.*
- *Ming-Hsuan Yang is with the University of California at Merced, and Google, USA.*

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

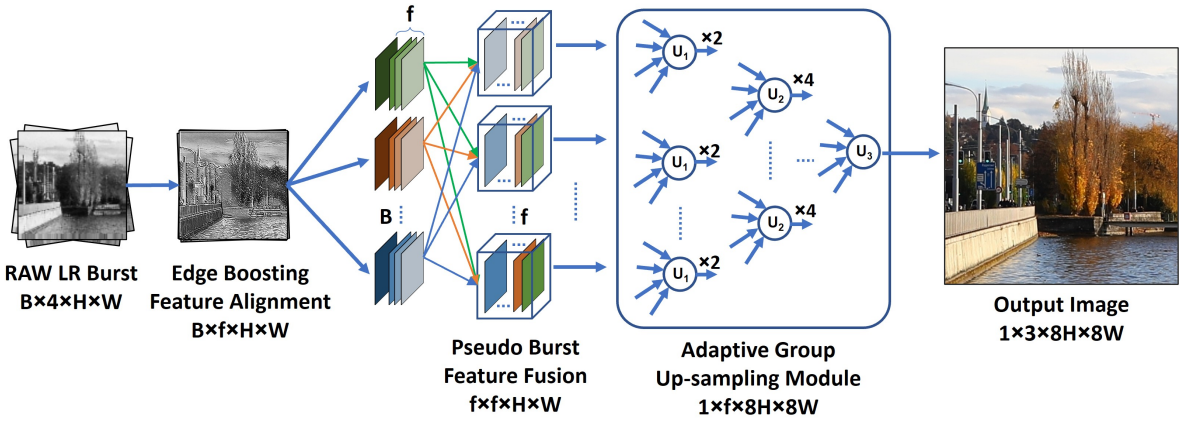


Fig. 1: Comprehensive layout of the proposed burst processing approach. The proposed BIPNet process input RAW burst and reconstruct high-quality RGB image. BIPNet comprises three major stages. (1) Edge boosting feature alignment for tackling noise, inter-frame color, and spatial misalignment issues. (2) Pseudo-burst feature fusion approach for facilitating cross-frame communication and subsequent feature consolidation. (3) Adaptive group upsampling for progressively increasing the spatial resolution in multiple stages while combining the multi-frame information. Though BIPNet is generalized to several other restoration tasks, here we show its application on super-resolution.

negatively affects the generated outputs. In contrast, our proposed BIPNet implicitly learns the frame alignment with deformable convolutions [67] and can adapt to the given problem effectively. Further, we introduce back-projection operation [24] in the proposed feature alignment stage to retain high-frequency information, which helps to align the burst features when burst frames are highly misaligned where alone deformable convolutions may not be sufficient.

In addition, irrespective of the lighting condition, some noise is always inherent in the captured images. Hence, one of our key objectives is to reduce noise [60] in the earlier stage of our network to mitigate the difficulty of the subsequent alignment and fusion stages. Towards this, we embrace residual global context attention in BIPNet for initial feature extraction and subsequent refinement/denoising. With the proposed building blocks, BIPNet can be extrapolated to several burst processing tasks. Our work corroborates its effectiveness on burst super-resolution, burst low-light image super-resolution, burst low-light image enhancement, and burst denoising. For super-resolution (SR), upsampling plays an indispensable role in image reconstruction. The current state-of-the-art burst SR methods [5], [6] initially aggregate the burst features and then utilize pixel-shuffle operation [45] for reconstructing a high-resolution image. Unlike the existing approaches [5], [6], we adaptively utilize the sub-pixel information available in the burst frames and progressively perform feature aggregation and upsampling in an adjustable and effective manner. Particularly, we progressively upscale the burst features through the proposed adaptive group upsampling while merging complimentary features. The schematic of our proposed BIPNet can be seen in Fig. 1.

Our main contributions are summarized as:

- We propose an edge-boosting feature alignment module to align burst features with respect to the reference frame. (Sec. 3.1)
- A novel pseudo-burst feature aggregation technique is proposed to enable the interaction within burst frames. (Sec. 3.2)
- To upscale the burst features, we propose an adaptive group upsampling strategy. (Sec. 3.3)

A preliminary version of this work has been published as a conference paper [19], where we validate the proposed BIPNet for burst super-resolution, burst denoising, and burst low-light image enhancement. In this work, we additionally test the proposed BIPNet on a new problem of burst low-light image super-resolution. Furthermore, we validate two lightweight variants of the proposed approach named BIPNet-16 and BIPNet-32 for the burst SR task to reduce the inference time. We investigate more comprehensive ablation studies and add additional visual analysis to emphasize the major determinant factors in BIPNet (Sec. 4, and Sec. 5). The detailed experiments show that the proposed BIPNet outperforms current state-of-the-art methods on real and synthetic datasets for all the discussed applications.

2 Related Work

2.1 Single Image Super-resolution (SISR)

Since the pioneering CNN-based work [17], data-driven approaches have achieved impressive performance gains over the conventional counterparts [21], [58]. The success of CNNs is mainly attributed to their architecture design [2], [62]. Given a low-resolution image (LR), early methods directly learn to generate latent SR image [17], [18]. In contrast, recent approaches learn to produce high-frequency residual to which LR image is added to generate the final SR output [27], [48], [49]. Other notable SISR network designs employ recursive learning [1], [30], progressive reconstruction [32], [56], attention mechanisms [14], [61], [64], [65], and generative adversarial networks [34], [44], [55]. However, the SISR approaches cannot handle multi-degraded frames from an input burst, and our proposed approach belongs to multi-frame SR that assists effective merging of the cross-frame information for a high-quality HR output.

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

2.2 Multi-Frame Super-Resolution (MFSR)

Tsai *et al.* [51] proposed the first frequency domain-based method for the MFSR task. It performs registration and fusion of the aliased LR images to generate an SR image. Since processing multi-frames in the frequency domain generates visual artifacts [51], other works improved results by incorporating image priors in the reconstruction process [46] and making algorithmic choices such as iterative back-projection [28], [43]. Farsui *et al.* [20] design a joint multi-frame demosaicking and SR approach that is robust to noise. MFSR techniques are devised for diverse uses, including handheld devices (Wronski et al., 2019), enhancing facial image spatial resolution (Ustinova et al., 2017), and satellite imagery applications (Deudon et al., 2020; Molini et al., 2019). Lecouat *et al.* [33] 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.* [5] propose a burst SR method that initially aligns the burst image features using an explicit PWCNet [47] and then performs an attention-based fusion mechanism to integrate the features. However, explicit motion estimation and image-warping techniques can pose difficulty in handling scenes with fast object motions. Recent works [50], [54] show that the deformable convolution [67] 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 aggregate the channel-wise information followed by adaptive upsampling, which optimally leverages multi-frame information.

2.3 Low-Light Image Enhancement

Images acquired in low-light conditions are generally darker, noisy, and color distorted. Addressing these issues involves long sensor exposure time, larger aperture lens, camera flash, and exposure bracketing [15], [63]. However, each of these possible solutions comes with its challenges. For instance, long exposure generates images with ghosting artifacts because of camera or object movements. Wide apertures are generally not available on smartphone devices, etc. See-in-the-Dark method [10] is the first attempt to replace the standard camera imaging pipeline with a CNN model. It takes a RAW image captured in extremely low light as input and learns to generate a well-lit sRGB image. Later, this work is further improved by employing a combined pixel-wise and perceptual loss [63] and a new CNN-based architecture [39]. Zaho *et al.* [66] proposes a recurrent convolutional network by using burst imaging to produce a noise-free bright sRGB image from a burst of RAW images. The results are further improved by Karadeniz *et al.* [29] via their two-stage approach: the first sub-network performs denoising, and the second sub-network generates a visually enhanced image. Though these studies exhibit noteworthy progress in low-light image enhancement, they do not effectively consider the inter-frame misalignment and information interaction that we address in this work.

2.4 Low-light Image Super-resolution

Along with the low illumination and noise, distortions in low-light images further increase with physical constraints of the smartphone cameras, such as small sensor size, which limits the spatial resolution of the captured image. Approaches [10], [15], [29], [39], [63], [66] discussed in Sec. 2.3 deals with low-light image enhancement alone, while distortions due to the spatial resolution are not considered. Recently, Han *et al.* [23] have proposed a super-resolution approach for infrared images captured under low-light conditions. Wang *et al.* [52], [53] have proposed a low-light image super-resolution approach for monochromatic low-resolution images. Further, cross-fusion U-Net architecture is proposed in [11] for sRGB low-light image super-resolution. Above discussed approaches jointly deal with image enhancement and super-resolution tasks but operate on a single image captured in low-light conditions. Unlike these approaches, we use multiple low-light images to upscale and enhance the details jointly.

2.5 Multi-Frame Denoising

Earlier works [12], [37], [38] provide extensions on top of the popular image denoising algorithm BM3D [13] to video. Buades *et al.* [9] estimated the noise level from the aligned images followed by the combination of pixel-wise mean and BM3D to perform denoising. A hybrid 2D/3D Wiener filter is used in [25] to denoise and merge burst images for high dynamic range and low-light photography tasks. Godard *et al.* [22] utilize recurrent neural network (RNN) and extend a single image denoising network for multiple frames. Mildenhall *et al.* [42] generate per-pixel kernels through the kernel prediction network (KPN) to merge the input images. In [40], authors extend the KPN approach to predict multiple kernels, while [57] introduces basis prediction networks (BPN) to enable the use of larger kernels. Recently, Bhat *et al.* [6] proposed a deep reparameterization of the maximum a posteriori formulation for the multi-frame SR and denoising.

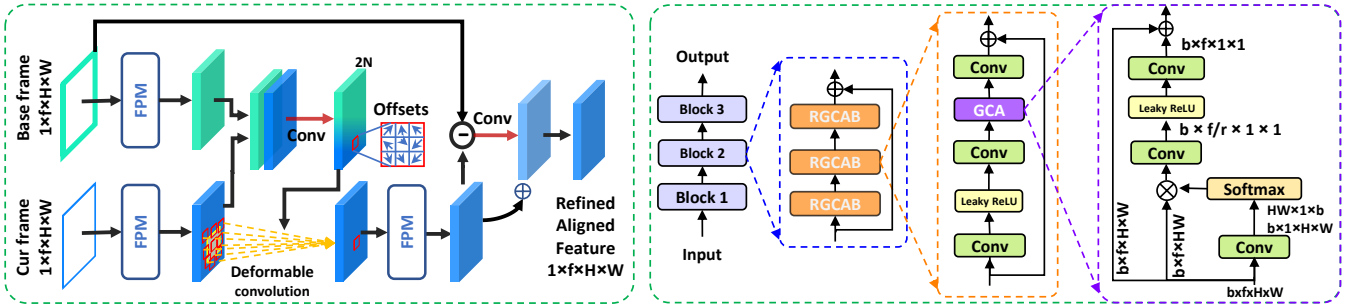
3 Burst Processing Approach

This section describes our burst processing approach, which applies to different image restoration tasks, including burst super-resolution, burst low-light image enhancement, burst low-light image super-resolution, 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 complementary 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

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188



(a) Edge Boosting Feature Alignment Module (EBFA) (b) Feature Processing Module (FPM)

Fig. 2: Edge boosting feature alignment (EBFA) module aligns all other images in the input burst to the base frame. The feature processing module (FPM) is added in EBFA to denoise input frames for facilitating easy alignment. $\otimes$ represents matrix multiplication.

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 [67]. 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)) does 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 [64]. Since capturing long-range pixel dependencies which extract global scene properties is beneficial for a wide range of image restoration tasks [59] (e.g., image/video super-resolution [41] 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 $\{\boldsymbol{x}^b\}_{b\in[1:B]} \in \mathbb{R}^{B\times f\times H\times W}$ be an initial latent representation of the burst having B burst images and f number of feature channels, our residual global context attention block (RGCAB in Fig. 2(b)) is defined as:

$$\boldsymbol{y}^b = \boldsymbol{x}^b + \omega_1\left(\alpha\left(\bar{\boldsymbol{x}}^b\right)\right), \tag{1}$$

where $\bar{\boldsymbol{x}}^b = \omega_3\left(\gamma\left(\omega_3\left(\boldsymbol{x}^b\right)\right)\right)$ and $\alpha\left(\bar{\boldsymbol{x}}^b\right) = \bar{\boldsymbol{x}}^{\boldsymbol{b}} + \omega_1\left(\gamma\left(\omega_1\left(\Psi\left(\omega_1\left(\bar{\boldsymbol{x}}^b\right)\right)\otimes\bar{\boldsymbol{x}}^b\right)\right)\right)$. Here, ω_k represents a convolutional layer with $k\times k$ sized filters and each ω_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:

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

where, ω^d and ω^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:

$$\bar{\boldsymbol{y}}^b_n = \sum_{i=1}^{K} \omega^d_{n_i}\ \boldsymbol{y}^b_{(n+n_i+\Delta n_i)}\cdot\Delta m_{n_i}, \tag{3}$$

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 by the deformable alignment module (DAM) [50] 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 correct the minor alignment errors, we use FPM to obtain refined aligned features (RAF) and the high-frequency residue by taking the difference between the RAF and base frame features and adding it to the RAF. Adding this residue to RAF effectively boosts the edge content within the burst features. The overall process of our EBFA module is summarized as: $\boldsymbol{e}^b = \bar{\boldsymbol{y}}^b + \omega_3\left(\bar{\boldsymbol{y}}^b - \boldsymbol{y}^{b_r}\right)$ where $\boldsymbol{e}^b \in \mathbb{R}^{B\times f\times H\times W}$ represents the aligned burst feature maps, and $\omega_3(\cdot)$ is a 3×3 convolution layer. 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.

1. Here, we consider the first image of a given burst as the base frame.

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

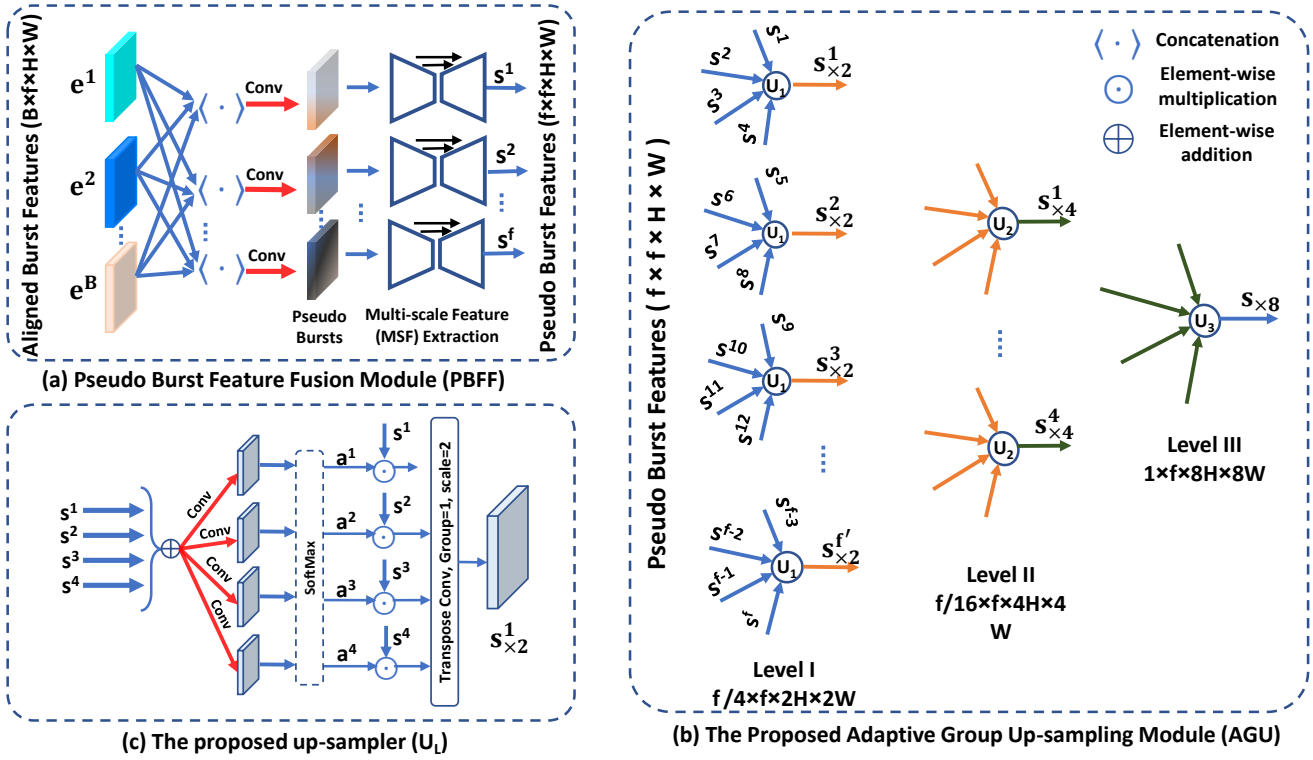


Fig. 3: (a) Pseudo-burst is generated by exchanging information across frames such that each feature tensor in the pseudo-burst contains complementary properties of all frames. Pseudo bursts are processed with (shared) U-Net to extract multi-scale features. (b) AGU module handles pseudo-burst features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

3.2 Pseudo-Burst Feature Fusion Module

Existing burst image processing techniques [5], [6] 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 burst image features. Processing inter-burst feature responses simplify the representation learning task and merge the relevant information by decoupling the burst image feature channels. Given the aligned burst feature set $e = \{\boldsymbol{e}_c^b\}_{c\in[1:f]}^{b\in[1:B]}$ of burst size B and f number of channels, the pseudo-burst is generated by,

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

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

Even after generating pseudo-bursts, obtaining their deep representation is essential. We use a lightweight (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 [5], [6] use pixel-shuffle layer [45] to perform upsampling in one stage. However, in burst image processing, information in multiple frames can be exploited effectively to get into the 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 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 processes the feature maps $\boldsymbol{S} = \{\boldsymbol{S}^c\}_{c\in[1:f]}$ produced by the pseudo-burst fusion module and provides an HR 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 ($\boldsymbol{a}^c$) (attention weights for each pixel). The dense attention maps are element-wise applied to the respective burst features. Finally, the upsampled response for a given group of features $\hat{\boldsymbol{S}}^g = \{\boldsymbol{S}^i : i \in [(g-1)*4+1 : g*4]\}^{g\in[1:f/4]} \subset \boldsymbol{S}$ and associated attention maps $\hat{\boldsymbol{a}}^g$ at the first upsampling level

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

(Level I in Fig. 3(b)) is formulated as:

$$\boldsymbol{S}^{g}_{\times 2} = \omega_T \left(\left\langle \hat{\boldsymbol{S}}^{g} \odot \hat{\boldsymbol{a}}^{g} \right\rangle \right),$$
$$\hat{\boldsymbol{a}}^{g} = \psi \left(\omega_1 \left(\omega_1 \left(\sum_{i=(g-1)*4+1}^{g*4} \boldsymbol{S}^{i} \right) \right) \right), \quad (5)$$

where $\psi(\cdot)$ denotes the softmax activation function, ω_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 $\times 4$, 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. The upsampler at each level is shared among groups to avoid the increase in network parameters.

4 Experiments

We evaluate the proposed BIPNet and other approaches on real and synthetic datasets for **(a)** burst super-resolution, **(b)** burst low-light image enhancement, **(c)** burst low-light image super-resolution, and **(d)** burst denoising.

4.1 Implementation Details.

Our BIPNet is end-to-end trainable and needs no pre-training 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 models for burst SR, burst low-light image enhancement, burst low-light image SR, and burst denoising using L_1 loss only. While for burst SR on real data, we fine-tune our BIPNet with pre-trained weights on the SyntheticBurst dataset using aligned L_1 loss [5]. The models are trained with an Adam optimizer. Cosine annealing strategy [36] 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.

4.2 Burst Super-resolution

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

4.2.1 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 [7]. 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.

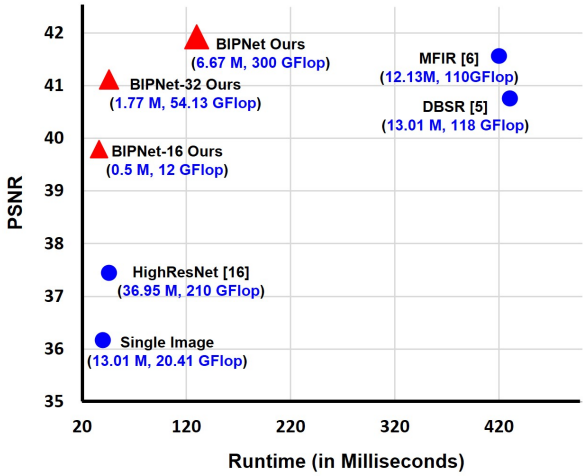


Fig. 4: Burst SR results (Table 1) vs inference time. The proposed BIPNet-32 achieves 41.12 dB PSNR and outperforms the existing DBSR [5] approach while reducing 89%↓ inference time, 86%↓ parameters and 54%↓ GFlops.

Methods	SyntheticBurst		(Real) BurstSR	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Single Image	36.17	0.909	46.29	0.982
HighRes-net [16]	37.45	0.92	46.64	0.980
DBSR [5]	40.76	0.96	48.05	0.984
LKR [33]	41.45	0.95	-	-
MFIR [6]	41.56	0.96	48.33	0.985
BIPNet (Ours)	**41.93**	**0.96**	**48.49**	**0.985**

TABLE 1: Performance evaluation of the proposed BIPNet and other existing methods on synthetic and real burst validation sets [5] for $\times 4$ burst super-resolution task.

4.2.2 SR results on synthetic data

The proposed BIPNet is trained for 300 epochs on the training set while evaluated on a validation set of SyntheticBurst dataset [4]. We compare our BIPNet with the several burst SR methods such as HighResNet [16], DBSR [5], LKR [33], and MFIR [6] 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 MFIR [6] and 0.48 dB over the second best approach [33].

Visual results provided in Fig. 5 show that the SR images produced by BIPNet are sharper and more 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 DBSR, LKR, and MFIR results contain splotchy textures and compromise image details.

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

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

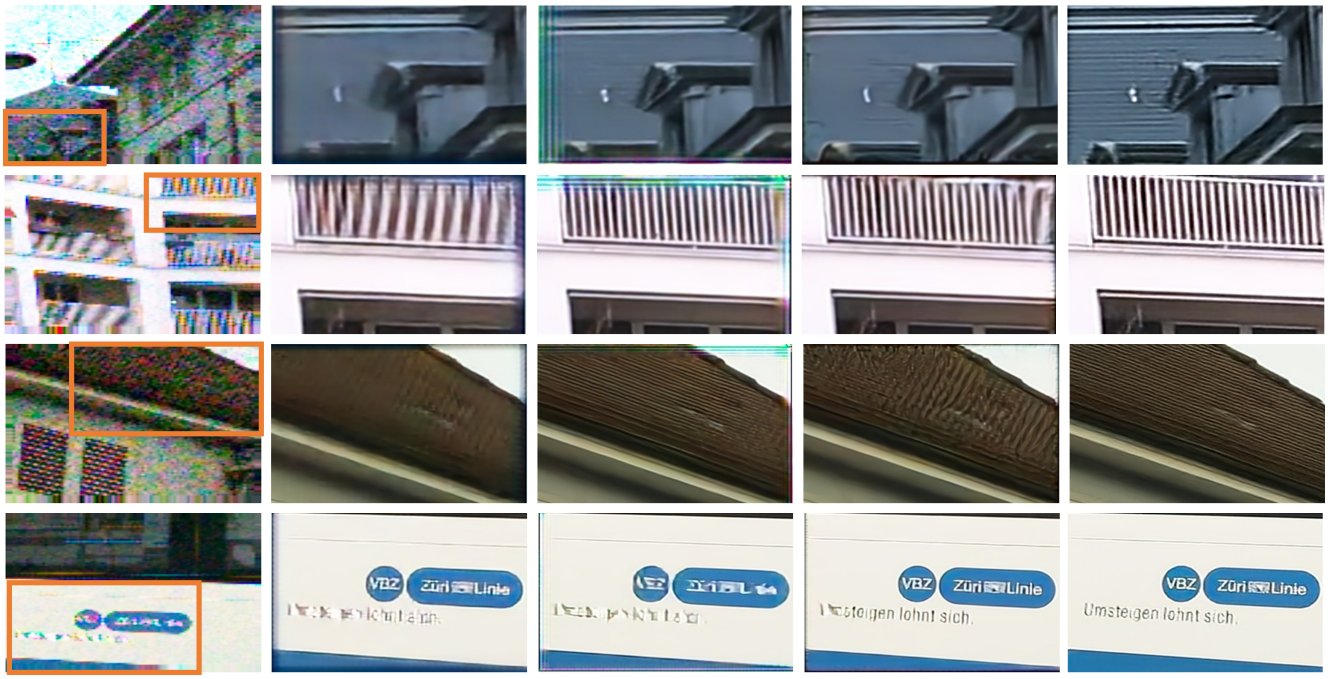


Fig. 5: Comparisons for $\times 4$ burst SR on validation set of the SyntheticBurst dataset [4]. Our BIPNet produces more sharper and clean results than other competing approaches.

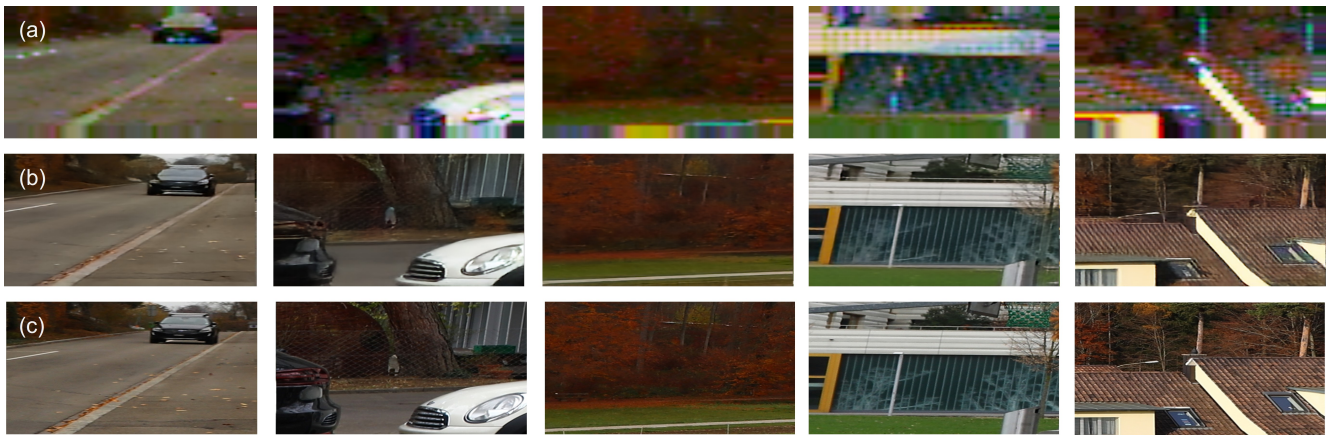


Fig. 6: Results for $\times 8$ burst SR on SyntheticBurst dataset [5]. (a) Base frame, (b) BIPNet (Ours), (c) Ground truth. Our method effectively recovers image details in extremely challenging cases.

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 the supplementary material.

4.2.3 SR results on real data

The LR input bursts and the corresponding HR ground truth in the BurstSR dataset suffer from minor misalignment as they are captured with different cameras. To mitigate this issue, we used aligned L1 loss for training and aligned PSNR/SSIM for evaluating our model, as in previous works [5], [6]. We fine-tuned the pre-trained BIPNet for 15 epochs on the training set while evaluating on the validation set of the BurstSR dataset. The image quality scores are reported in Table 1. Compared to the previous best approach MFIR [6], our BIPNet provides a performance gain of 0.16 dB. The visual comparisons in Fig. 7 show that our BIPNet is more effective in recovering fine details in the reproduced images than other competing approaches.

4.3 Light-weight BIPNet for Burst SR

The proposed BIPNet is designed with 64 filters in each convolution layer. It has 6.67M parameters and 300 GFlops. To reduce the GFlops and increase the burst processing speed, we obtain two lightweight versions, BIPNet-32 and BIPNet-16, by reducing the convolution filters from 64 to 32 and 64 to 16, respectively. Compared to BIPNet (6.67M, 300 GFlops), BIPNet-32 (1.8 M, 54.3 GFlops) has 73% fewer parameters and 81% fewer GFlops. While BIPNet-16 (0.5 M, 12 GFlops) has 93% fewer parameters and 96% fewer

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

Fig. 7: Comparisons for $\times 4$ burst super-resolution on Real BurstSR dataset [5]. Our BIPNet produces more sharper and clean results than other competing approaches.

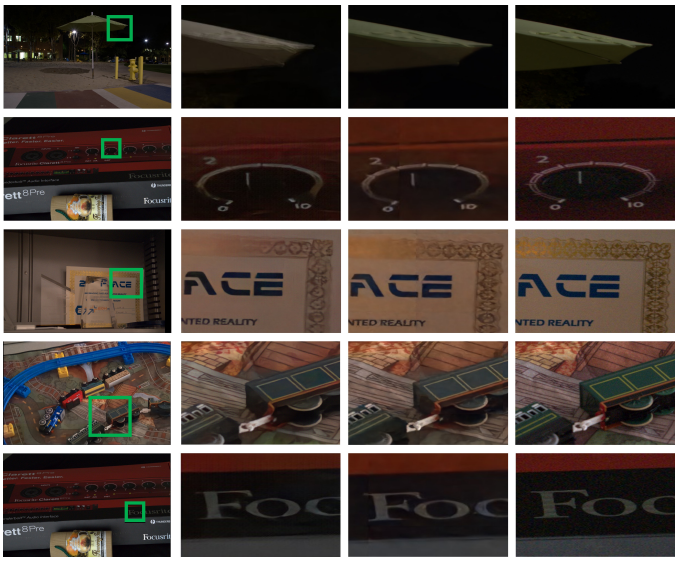


Fig. 8: Burst low-light image enhancement on Sony subset [10]. BIPNet better preserves color and structural details.

GFlops. In Fig. 4, we compare PSNR, inference time (in Milliseconds), and GFlops of the proposed light-weight BIPNet versions with the existing networks on the SyntheticBurst [4] dataset for burst SR task. As shown in Fig. 4, the proposed BIPNet-32 has an inference time of 45.85 ms, 54.3 GFlops, and achieves 41.12 dB PSNR which is better than the recent DBSR [5] approach (431 ms, 118 GFlops, 40.76 dB).

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
SID [10]	29.38	0.892	0.484
ELID [39]	29.57	0.891	0.484
LDCP [63]	29.13	0.881	0.462
RFCN [66]	29.49	0.895	0.455
LEED [29]	30.04	0.890	0.308
BIPNet (Ours)	**32.87**	**0.936**	**0.305**

TABLE 2: Burst low-light image enhancement methods evaluated on the SID dataset [10]. Our BIPNet advances state-of-the-art by 2.83 dB.

While BIPNet-16 is comparatively less accurate (achieves 39.8 dB PSNR), it is extremely efficient with only 36.16 ms inference time, 503K parameters, and 12 GFlops, reducing 91%↓ inference time, 96%↓ parameters and 89%↓ GFlops compared to the existing baseline DBSR [5] approach.

4.4 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.

4.4.1 Dataset

SID dataset [10] consists of short-exposure burst raw images taken under extremely dark indoor (0.2-5 lux) or outdoor (0.03-0.3 lux) scenes and their corresponding ground truth sRGB images. Burst RAW images are acquired with three

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

Fig. 9: Comparisons for $\times 4$ burst low-light image super-resolution on SID-SR [10] dataset. Our BIPNet produces sharper and enhanced results compared to the other approaches.

different exposure times of 1/10, 1/25, and 1/30 sec, where the corresponding ground truth images are obtained with 10 seconds or 30 seconds exposures depending on the scene. For each burst low-light image, the amplification ratio (either of $\times 100$, $\times 250$, $\times 300$) is provided. The amplification ratio is measured as the ratio between the exposure times of the dark input image and the long-exposure ground truth. 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 eight from the training set of the Sony subset of SID to train the network for 50 epochs. We use the same pre-processing steps as in SID [10] paper.

4.4.2 Enhancement results

In Table 2, we report the results of several low-light enhancement methods. Our BIPNet yields a significant performance gain of 2.83 dB over the existing best method [29]. Similarly, the visual examples provided in Fig. 8 also corroborate the effectiveness of our approach.

4.5 Burst Low-Light Image Super-resolution

Existing joint low-light image enhancement and super-resolution approaches operate on a single image captured in low-light conditions. They are not benefited by additional information through the multiple frames. Conversely, existing works [4], [5], [6] perform joint denoising and super-resolution while operating on LR burst captured in normal-light/day-time. Unlike these approaches, we use burst low-light image and denoise, enhance and upscale its details jointly. Thus, in this work, we further extend the proposed BIPNet for burst low-light image super-resolution (LSR) task. We perform an LSR experiment for scale factor $\times 4$ on SID dataset [10].

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
LDCP [63]	26.43	0.62	0.58
DBSR [5]	26.71	0.74	0.51
MFIR [6]	27.61	0.76	0.48
LEED [29]	27.30	0.76	0.51
BIPNet (Ours)	**29.16**	**0.81**	**0.43**

TABLE 3: Burst low-light image super-resolution methods evaluated on the SID-SR dataset [10].

4.5.1 Dataset

Here, we discuss the SID dataset for a low-light super-resolution task called as SID-SR dataset. As discussed in Sec. 4.4, SID dataset [10] consists of RAW bursts captured with short-camera exposure in low-light conditions with respective long exposure sRGB images. We prepare 6440, 800, and 1880 patches from training (161), validation (20), and testing (50) splits, respectively, of the SID dataset. We follow the same pre-processing steps for raw data as described in SID [10]. First, the raw array is converted into channels, subtracting the black level and using the given amplification

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

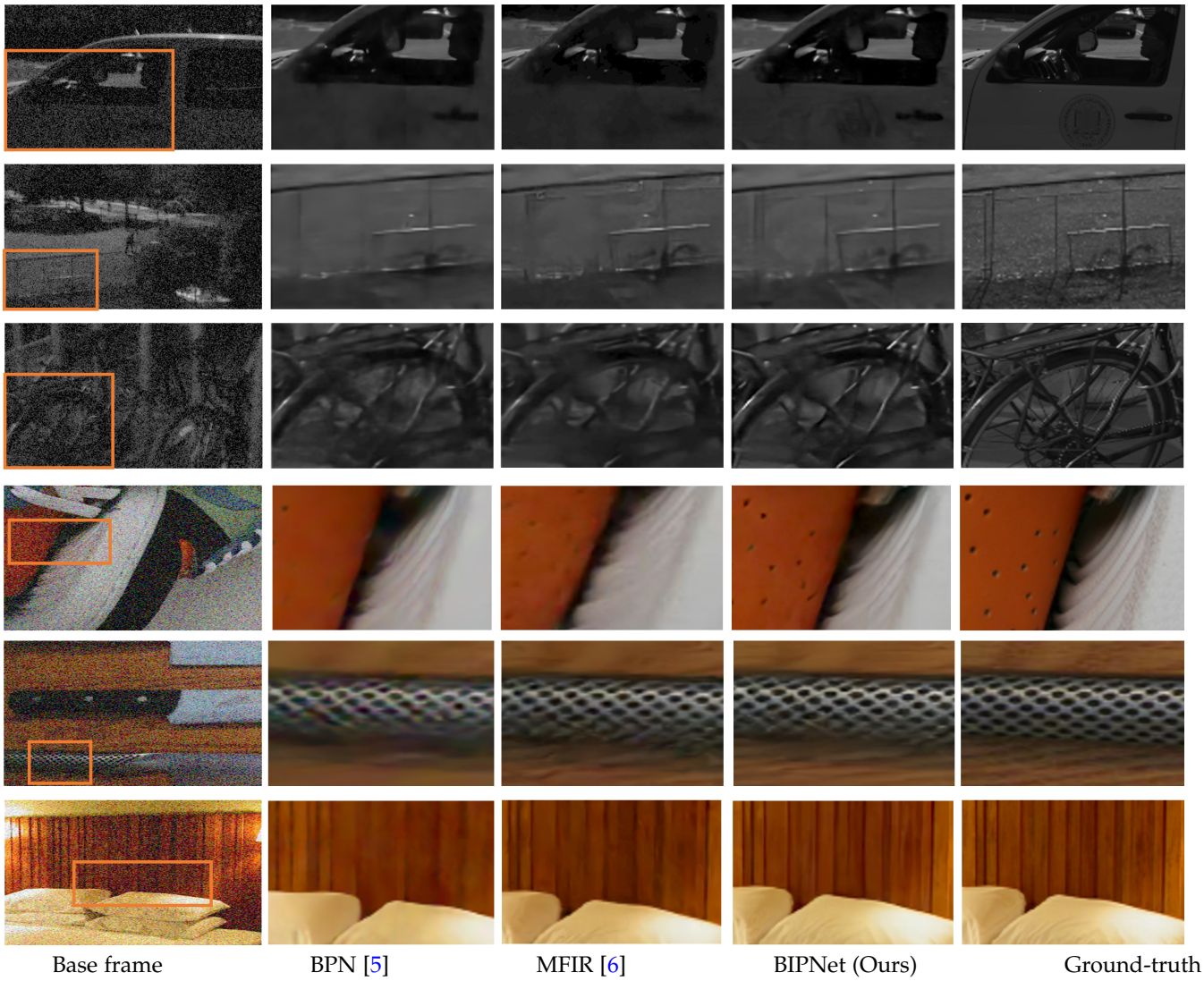


Fig. 10: Comparisons for burst denoising on gray-scale [42] and color datasets [57]. Our BIPNet produces more sharper and clean results than other competing approaches. Many more examples are provided in the supplementary material.

ratio to align the exposure. After pre-processing, each burst patch is of size 256×256×4×B while ground truth sRGB patch is of size 512×512×3, where B denotes the number of burst images ranging from 2 to 8. Further, to mold the SID dataset for the low-light super-resolution (LSR) task, we apply bilinear downsampling by a factor ×4 on the pre-processed burst to get the LR burst of size 64×64×4×B.

4.5.2 LSR results

We compare the proposed BIPNet with existing base methods from burst super-resolution: DBSR [5], MFIR [6] and burst low-light image enhancement: LDCP [61], LEED [29] for ×4 LSR task. For LDCP [61], and LEED [29] methods, we have deployed a pixel-shuffle layer to upscale the burst features. We train the proposed and existing methods for 100 epochs on a training set of the SID-SR dataset. Table 3 shows that the proposed BIPNet outperforms the other methods by a large margin. Visual results given in Fig. 9 show that the proposed BIPNet produces more enhanced results when compared with the existing methods.

4.6 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 the rest of the network architecture is kept unmodified.

4.6.1 Dataset

We evaluate our approach on the grayscale and color burst denoising datasets introduced in [42] and [57]. These datasets contain 73 and 100 burst images, respectively. In both datasets, a burst is generated synthetically by applying random translations to the base image. The shifted images are then corrupted by adding heteroscedastic Gaussian noise [26] 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 $(\log(\sigma_r), \log(\sigma_s)) \rightarrow$ (-2.2, -2.6), (-1.8, -2.2), (-1.4, -1.8), and (-1.1, -1.5), respectively. Note

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

	Gain $\propto 1$	Gain $\propto 2$	Gain $\propto 4$	Gain $\propto 8$
HDR+ [25]	31.96	28.25	24.25	20.05
BM3D [13]	33.89	31.17	28.53	25.92
NLM [8]	33.23	30.46	27.43	23.86
VBM4D [38]	34.60	31.89	29.20	26.52
KPN [42]	36.47	33.93	31.19	27.97
MKPN [40]	36.88	34.22	31.45	28.52
BPN [57]	38.18	35.42	32.54	29.45
MFIR [6]	39.10	36.14	32.89	28.98
BIPNet (Ours)[2]	**38.53**	**35.94**	**33.04**	**29.89**

TABLE 4: Comparison of our method with prior approaches on the grayscale burst denoising set [42] in terms of PSNR. The results for existing methods are from [6].

	Gain $\propto 1$	Gain $\propto 2$	Gain $\propto 4$	Gain $\propto 8$
KPN [42]	38.86	35.97	32.79	30.01
BPN [57]	40.16	37.08	33.81	31.19
MFIR [6]	41.90	38.85	35.48	32.29
BIPNet (Ours)[2]	**40.58**	**38.13**	**35.30**	**32.87**

TABLE 5: Comparison with previous methods on the color burst denoising set [57] in terms of PSNR. The results for existing methods are from [6]. Our approach outperforms BPN on the highest noise level by 0.58 dB.

that the noise parameters for the highest noise gain (Gain $\propto 8$) are unseen during training. Thus, performance on this noise level indicates the generalization of the network to unseen noise. Following [6], we utilized 20k samples from the Open Images [31] training set to generate the synthetic noisy bursts of burst-size eight and spatial size 128×128. Our BIPNet is trained for 50 epochs for the grayscale and color burst denoising tasks and evaluated on the benchmark datasets [42] and [57] respectively.

4.6.2 *Burst Denoising results*

We compare the proposed BIPNet[2] with the several approaches (KPN [42], MKPN [40], BPN [57] and MFIR [6]) both for grayscale and color burst denoising tasks. Since the proposed BIPNet is trained without any extra data or supervision, we consider the results of the MFIR [6] variant that uses a custom optical flow sub-network (without pre-training it on extra data). Table 4 shows that our BIPNet significantly advances state-of-the-art on grayscale burst denoising dataset [42]. Specifically, the BIPNet outperforms the previous best method MFIR [6] by 0.91 dB on the highest noise level (Gain $\propto 8$), which is unseen during training levels. A similar performance trend can be observed in Table 5 for color denoising on color burst dataset [57]. Particularly, our BIPNet provides a PSNR boost of 0.58 dB over the previous best method MFIR [6] for the highest noise level (Gain $\propto 8$). In Figure 10, BIPNet's reproduced images appear cleaner and sharper than other methods.

4.7 Ablation Study

Here we present ablation experiments to demonstrate the impact of each individual component of our approach. All

2. In conference version [19] of this work, we mistakenly calculated the PSNR before post-processing [6]. This paper rectifies the error, and the corrected PSNR scores can be found in Table 4 and 5.

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)								✓
Parameters (M)	5.27	5.64	6.23	6.27	6.35	6.44	6.57	6.67
PSNR	36.38	36.54	38.39	39.10	39.64	40.35	41.25	41.55

TABLE 6: Importance of BIPNet modules evaluated on SyntheticBurst validation set for $\times 4$ burst SR.

	Methods	**PSNR ↑**	**SSIM ↑**
(a) Alignment	Explicit [5]	39.26	0.944
	TDAN [50]	40.19	0.957
	EDVR [54]	40.46	0.958
(b) Fusion	Addition	39.18	0.943
	Concat	40.13	0.956
	DBSR [5]	40.16	0.957
(c) Up-sampling	Pixel-shuffle [45]	40.35	0.951
(d)	**BIPNet (Ours)**	**41.55**	**0.960**

TABLE 7: Importance of the proposed alignment, fusion, and up-sampling modules evaluated on SyntheticBurstSR dataset [4] for $\times 4$ burst SR.

ablation models are trained for 100 epochs on SyntheticBurst dataset [4] for SR scale factor $\times 4$. Results are reported in Table 6. For the baseline model, we employ Resblocks [35] 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 a performance boost of 1.85 dB when considering the deformable alignment module DAM. Similarly, RAF contributes 0.71 dB improvement toward the model. With our PBFF mechanism, the network achieves a significant gain of 1.25 dB. AGU brings a 1 dB increment in the upsampling stage. Finally, EBFA demonstrates its effectiveness in correcting alignment errors by providing a 0.30 dB improvement in PSNR. Overall, our BIPNet obtains a gain of 5.17 dB over the baseline.

Finally, we carry ablation experiments to show the importance of the proposed EBFA and PBFF modules by replacing them with existing alignment and fusion modules. Table 7(a) shows that replacing our EBFA with other alignment modules has a negative impact (PSNR drops at least over 1 dB). A similar trend can be observed with fusion strategies other than our PBFF and AGU; see Table 7(b) and (c).

Visual analysis: In addition to conducting a quantitative ablation study, we analyze restored results to validate the efficacy of the proposed Edge Boosting Feature Alignment (EBFA) module. We use checkerboard image for ease of understanding. We obtain a burst of sub-pixel shifted checkerboard images with the process described in Sec. 4.2.1 (1). Finally, the synthetically generated checkerboard burst is processed through the proposed EBFA module, which aligns all the neighboring frames with respect to the base frame

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

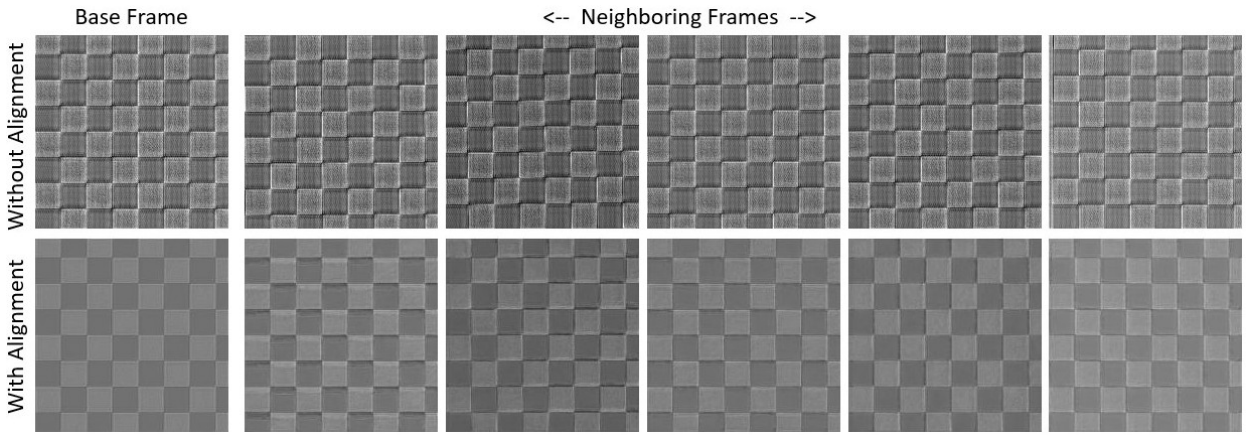


Fig. 11: Illustration of the feature maps with and without the proposed Edge Boosting Feature Alignment (EBFA) module. The upper row shows the unaligned burst features, whereas the lower row displays the corresponding aligned burst features. The EBFA module significantly reduces the noise and aligns the neighboring frames with the base frame.

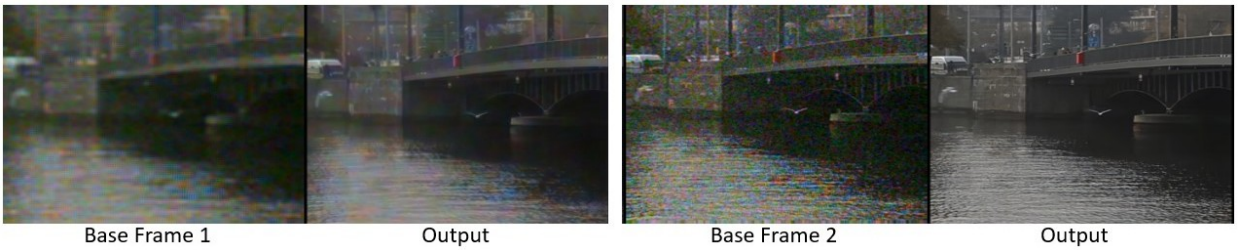


Fig. 12: Visualization of the BIPNet results utilizing various base frames. The left side showcases a heavily distorted base frame alongside its restored image, whereas the right side displays a moderately distorted base frame and its restored image. This analysis shows that the final result gets influenced with respect to the base frame distortions.

(first frame). Fig. 11 shows the feature representations of the base and neighboring frames with and without the inclusion of the EBFA module. These results facilitate us to gain deeper insights into the functionality of the EBFA module. Notably, in the absence of feature alignment, the neighboring frames exhibit noticeable sub-pixel shifts compared to the base frame. Conversely, employing the EBFA module for feature alignment results in minimal sub-pixel shifts for the neighboring frame feature and a notable reduction in noise compared to the base frame.

5 Limitations and Future Scope

An inherent limitation of the proposed BIPNet lies in its assumption that the first frame serves as the base frame, guiding the alignment of subsequent frames. As a result, if the base frame contains significant distortions, it will significantly influence the final result. To illustrate, we have shown an example of this limitation in Fig. 12. We tested this with a heavily distorted base frame and a moderately distorted one. The findings of this analysis demonstrate that the performance of the burst processing technique may degrade if significant distortion is present within the chosen base frame, which subsequently reflects in the final result. Therefore, an essential future direction involves the development of an adaptive reference frame selector that tailors the choice of reference frame for each burst, potentially enhancing the algorithm's performance in diverse scenarios. Moreover, the proposed network's modules could extend to the applications where challenges are in feature alignment, fusion, and reconstruction, testing the modules' robustness and adaptability.

6 Conclusion

We have proposed a novel burst restoration and enhancement approach for effectively fusing complementary information from multiple burst frame features. Unlike the existing late feature fusion methods, which combine the multi-frame feature information in the late part of the pipeline, we present a novel concept of pseudo-burst arrangement by individually integrating the channel-wise attentive features from each burst frame. To avert any sort of mismatch among the generated pseudo-burst features, we design an edge-boosting burst alignment module to implicitly align the frames by being robust to the camera-scene motions. Subsequently, the generated pseudo-burst features are refined by utilizing multi-scale information and later progressively fused for generating the upsampled reconstructed output. Extensive experiments on four burst restoration and enhancement tasks (super-resolution, low-light enhancement, low-light image super-resolution, and denoising) validate the authenticity and potency of BIPNet.

Acknowledgments

M.-H. Yang is supported in part by the NSF CAREER Grant 1149783. The authors would like to thank Martin Danelljan,

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

Goutam Bhat (ETH Zurich) and Bruno Lecouat (Inria and DIENS) for their useful feedback and for providing burst super-resolution results.

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. 4

[4] Goutam Bhat, Martin Danelljan, and Radu Timofte. Ntire 2021 challenge on burst super-resolution: Methods and results. In *CVPR*, 2021. 6, 7, 8, 9, 11

[5] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7, 8, 9, 10, 11

[6] 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, 5, 6, 7, 8, 9, 10, 11

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

[8] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. In *CVPR*, 2005. 11

[9] Toni Buades, Yifei Lou, J. M. Morel, and Z. Tang. A note on multi-image denoising. *2009 International Workshop on Local and Non-Local Approximation in Image Processing*, 2009. 3

[10] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 3, 8, 9

[11] Deqiang Cheng, Liangliang Chen, Chen Lv, Lin Guo, and Qiqi Kou. Light-guided and cross-fusion u-net for anti-illumination image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 3

[12] 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

[13] Kostadin Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *TIP*, 2007. 3, 11

[14] 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

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

[16] 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. 6

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

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

[19] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022. 2, 11

[20] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multiframe demosaicing and super-resolution from undersampled color images. In *Computational Imaging II*, 2004. 3

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

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

[23] Tae Young Han, Dae Ha Kim, Seung Hyun Lee, and Byung Cheol Song. Infrared image super-resolution using auxiliary convolutional neural network and visible image under low-light conditions. *Journal of Visual Communication and Image Representation*, 51:191–200, 2018. 3

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

[25] 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, 11

[26] G. Healey and R. Kondepudy. Radiometric ccd camera calibration and noise estimation. *TPAMI*, 1994. 10

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

[28] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP*, 1991. 3

[29] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *arXiv:2006.09845*, 2020. 3, 8, 9, 10

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

[31] 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. 11

[32] 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

[33] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *ICCV*, 2021. 3, 6, 7

[34] 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

[35] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 11

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

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

[38] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *TIP*, 2012. 3, 11

[39] 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, 8

[40] Talmaj Marinc, V. Srinivasan, S. Gül, C. Hellge, and W. Samek. Multi-kernel prediction networks for denoising of burst images. In *ICIP*, 2019. 3, 11

[41] 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

[42] Ben Mildenhall, J. Barron, Jiawen Chen, Dillon Sharlet, R. Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, 2018. 3, 10, 11

[43] Shmuel Peleg, Danny Keren, and Limor Schweitzer. Improving image resolution using subpixel motion. *PRL*, 1987. 3

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

[45] 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, 11

[46] Henry Stark and Peyma Oskoui. High-resolution image recovery from image-plane arrays, using convex projections. *JOSA A*, 1989. 3

[47] 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

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

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3356188

[49] Omkar Thawakar, Prashant W. Patil, Akshay Dudhane, Subrahmanyam Murala, and Uday Kulkarni. Image and video super-resolution using recurrent generative adversarial network. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019. 2
[50] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 3, 4, 11
[51] Roger Y. Tsai and Thomas S. Huang. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1984. 3
[52] Bowen Wang, Yan Zou, Linfei Zhang, Yan Hu, Hao Yan, Chao Zuo, and Qian Chen. Low-light-level image super-resolution reconstruction based on a multi-scale features extraction network. In *Photonics*, volume 8, page 321. MDPI, 2021. 3
[53] Bowen Wang, Yan Zou, Linfei Zhang, Le Li, and Chao Zuo. Super resolution reconstruction of low light level image based on the feature extraction convolution neural network. In *Computational Imaging VI*, volume 11731, pages 74–79. SPIE, 2021. 3
[54] 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, 11
[55] 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
[56] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *ICCV*, 2015. 2
[57] 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, 10, 11
[58] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *TIP*, 2010. 2
[59] 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. 4
[60] 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
[61] 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, 10
[62] 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
[63] 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, 8, 9
[64] 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, 4
[65] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *TPAMI*, 2020. 2
[66] 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, 8
[67] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 2, 3, 4

Akshay Dudhane is a Research Scientist at Mohommad Bin Zayed University of Artificial Intelligence in UAE. He received the Ph.D. degree from the Indian Institute of Technology Ropar, India, in 2021. His research interests include image and video processing, medical image analysis, image restoration, and enhancement.

Sayed Waqas Zamir received the Ph.D. degree from University Pompeu Fabra, Spain, in 2017. He is a Research Scientist at Inception Institute of Artificial Intelligence in UAE. His research interests include low-level computer vision, computational imaging, image and video processing, color vision and image restoration and enhancement.

Salman Khan is an Associate Professor at MBZ University of Artificial Intelligence. He has been an Adjunct faculty member at Australian National University since 2016. He has been awarded the outstanding reviewer award at CVPR multiple times, won the best paper award at 9th ICPRAM 2020, and 2nd prize in the NTIRE Image Enhancement Competition at CVPR 2019. He served as a program committee member for several premier conferences including CVPR, ICCV, ICLR, ECCV and NeurIPS. He received his Ph.D. degree from the University of Western Australia in 2016. His thesis received an honorable mention on the Dean's List Award. His research interests include computer vision and machine learning.

Fahad Shahbaz Khan is currently a Full Professor and Deputy Department Chair of Computer Vision at the MBZUAI, Abu Dhabi, United Arab Emirates. He also holds a faculty position (Universitetslektor + Docent) at Computer Vision Laboratory, Linköping University, Sweden. From 2018 to 2020 he worked as a Lead Scientist at the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. He received the M.Sc. degree in Intelligent Systems Design from Chalmers University of Technology, Sweden and a Ph.D. degree in Computer Vision from Autonomous University of Barcelona, Spain. He has achieved top ranks on various international challenges (Visual Object Tracking VOT: 1st 2014 and 2018, 2nd 2015, 1st 2016; VOT-TIR: 1st 2015 and 2016; OpenCV Tracking: 1st 2015; 1st PASCAL VOC 2010). His research interests include a wide range of topics within computer vision and machine learning, such as object recognition, object detection, action recognition and visual tracking. He has published articles in high impact computer vision journals and conferences in these areas. He serves as a regular program committee member for leading computer vision conferences such as CVPR, ICCV, and ECCV.

Ming-Hsuan Yang is affiliated with Google, UC Merced, and Yonsei University. Yang serves as a program co-chair of IEEE International Conference on Computer Vision (ICCV) in 2019, program co-chair of the Asian Conference on Computer Vision (ACCV) in 2014, and general co-chair of ACCV 2016. Yang served as an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and is an associate editor of the International Journal of Computer Vision, Image and Vision Computing and Journal of Artificial Intelligence Research. He received the NSF CAREER award and Google Faculty Award. He is a Fellow of the IEEE.

Authorized licensed use limited to: Univ of Calif Merced. Downloaded on February 02,2025 at 19:07:43 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.