Burstormer: Burst Image Restoration and Enhancement Transformer

Akshay Dudhane[1] Syed Waqas Zamir[2] Salman Khan[1,3]
Fahad Shahbaz Khan[1,4] Ming-Hsuan Yang[5,6,7]
[1]Mohamed bin Zayed University of AI [2]Inception Institute of AI [3]Australian National University
[4]Linköping University [5]University of California, Merced [6]Yonsei University [7]Google Research

Abstract

On a shutter press, modern handheld cameras capture multiple images in rapid succession and merge them to generate a single image. However, individual frames in a burst are misaligned due to inevitable motions and contain multiple degradations. The challenge is to properly align the successive image shots and merge their complimentary information to achieve high-quality outputs.

Towards this direction, we propose Burstormer: a novel transformer-based architecture for burst image restoration and enhancement. In comparison to existing works, our approach exploits multi-scale local and non-local features to achieve improved alignment and feature fusion. Our key idea is to enable inter-frame communication in the burst neighborhoods for information aggregation and progressive fusion while modeling the burst-wide context. However, the input burst frames need to be properly aligned before fusing their information. Therefore, we propose an enhanced deformable alignment module for aligning burst features with regards to the reference frame.

Unlike existing methods, the proposed alignment module not only aligns burst features but also exchanges feature information and maintains focused communication with the reference frame through the proposed reference-based feature enrichment mechanism, which facilitates handling complex motions. After multi-level alignment and enrichment, we re-emphasize on inter-frame communication within burst using a cyclic burst sampling module. Finally, the inter-frame information is aggregated using the proposed burst feature fusion module followed by progressive upsampling. Our Burstormer outperforms state-of-the-art methods on burst super-resolution, burst denoising and burst low-light enhancement. Our codes and pre-trained models are available at https://github.com/akshaydudhane16/Burstormer.

1. Introduction

In recent years, smartphone industry has witnessed a rampant growth on account of the fueling demand of smartphones in day-to-day life. While the image quality of smartphone cameras is rapidly improving, there are several barriers that hinder in attaining DSLR-like images. For instance, the physical space available in handheld devices restricts manufacturers from employing high-quality bulky camera modules. Most smartphone cameras use small-sized lens, aperture, and sensor, thereby generating images with limited spatial resolution, low dynamic range, and often with noise and color distortions especially in low-light conditions. These problems have shifted the focus nowadays in developing computational photography (software) solutions for mitigating the hardware limitations and to approach the image quality of DSLRs.

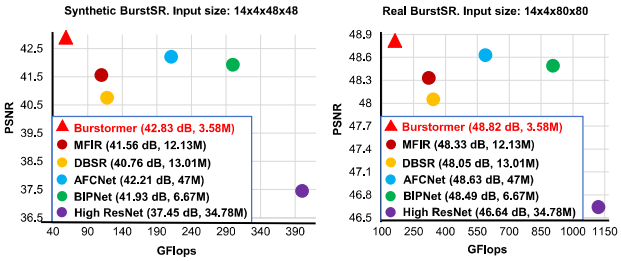

Figure 1. Burst super-resolution results (Tab. 1) vs. efficiency (GFlops). Burstormer advances state-of-the-art, while being compute efficient and light-weight.

One emerging approach to achieve high-quality results from a smartphone camera is to take advantage of burst shots consisting of multiple captures of the same scene. The burst image processing approaches aim to recover the high-quality image by merging the complementary information in multiple frames. Recent works [3, 4, 9] have validated the potential of burst processing techniques in reconstructing rich details that cannot be recovered from a single image. However, these computationally expensive approaches are often unable to effectively deal with the inherent sub-pixel shifts among multiple frames arising due to camera and/or object movements. This sub-pixel misalignment often causes blurring and ghosting artifacts in the final image. To tackle alignment issues, existing methods employ complex explicit feature alignment [3] and deformable con-

volutions [9]. However, these approaches target only the local features at a single level, while the use of global information together with multi-scale features has not been extensively explored. Additionally, while aggregating multi-frame features, existing approaches either employ late fusion strategy [3, 4] or rigid fusion mechanism (in terms of number of frames) [9]. The former one limits the flexible inter-frame communication, while the later one limits the adaptive multi-frame processing.

In this work, we propose Burstormer for burst image processing, which incorporates multi-level local-global burst feature alignment and adaptive burst feature aggregation. In contrast to previous works [3, 4] that employ bulky pre-trained networks for explicit feature alignment, we present a novel enhanced deformable alignment (EDA) module that handles misalignment issues implicitly. Overall, the EDA module reduces noise and extracts both local and non-local features with a transformer-based attention and performs multi-scale burst feature alignment and feature enrichment which is not the case with the recent BIPNet [9].

Unlike existing approaches [3, 4, 9] which allow a one go interaction with the reference frame during alignment process, we add a new reference-based feature enrichment (RBFE) mechanism in EDA to allow a more extensive interaction with the reference frame. This helps in effectively aligning and refining burst features even in complex misalignment cases where the simple alignment approaches would not suffice. In the image reconstruction stage we progressively perform feature consolidation and upsampling, while having access to the multi-frame feature information at all time. This is achieved with our no-reference feature enrichment (NRFE) module. NRFE initially generates burst neighborhoods with the proposed cyclic burst sampling (CBS) mechanism that are then aggregated with our burst feature fusion (BFF) unit. Unlike, the existing pseudo bursts [9], the proposed burst neighborhood mechanism is flexible and enables inter-frame communication with significantly less computational cost.

The key highlights of our work are outlined below:

- Our Burstormer is a novel Transformer based design for burst-image restoration and enhancement that leverages multi-scale local and non-local features for improved alignment and feature fusion. Its flexible design allows processing bursts of variable sizes.
- We propose an enhanced deformable alignment module which is based on multi-scale hierarchical design to effectively denoise and align burst features. Apart from aligning burst features it also refines and consolidates the complimentary burst features with the proposed reference-based feature enrichment module.
- We propose no-reference feature enrichment module to progressively aggregate and upsample the burst features with less computational overhead. To enable inter-frame interactions, it generates burst neighborhoods through the proposed cyclic burst sampling mechanism followed by the burst feature fusion.

Our Burstormer sets new state-of-the-art on several real and synthetic benchmark datasets for the task of burst super-resolution, burst low-light enhancement, and burst denoising. Compared to existing approaches, Burstormer is more accurate, light-weight and faster; see Fig. 1. Further, we provide detailed ablation studies to demonstrate the effectiveness of our design choices.

2. Related Work

Multi-Frame Super-Resolution. Unlike single image super-resolution, multi-frame super-resolution (MFSR) approaches are required to additionally deal with the sub-pixel misalignments among burst frames caused by camera and object motions. While being computationally efficient, the pioneering MFSR algorithm [36] processes burst frames in frequency domain, often producing images with noticeable artifacts. To obtain better super-resolved results, other methods operate in the spatial domain [10, 17], exploit image priors [33], use iterative back-projection [31], or maximum a posteriori framework [1]. However, all these approaches assume that the image formation model, and motion among input frames can be estimated reliably. Successive works addressed this constraint with the joint estimation of the unknown parameters [11, 15]. To deal with noise and complex motion, the MSFR algorithm of [38] employs non-parametric kernel regression and locally adaptive detail enhancement model.

The DBSR algorithm [3] addresses the MFSR problem by applying the explicit feature alignment and attention-centric fusion mechanisms. However, their image warping technique and explicit motion estimation may find difficult in handling scenes with fast moving objects. The EBSR [25] builds on prior PCD alignment techniques [37] by aligning enhanced features specifically for the burst SR task. In addition, the BSRT [24] employs a combination of optical flow and deformable convolution for feature alignment and utilizes a Swin Transformer [21] for feature extraction. More recently, BIPNet [9] was introduced to process noisy raw bursts using implicit feature alignment and pseudo-burst generation. Building on BIPNet, AFCNet [29] incorporates existing Restormer [41] to improve feature extraction for burst SR tasks. Despite having an effective inter-frame communication, their approach is rigid to using certain number of burst frames during alignment and fusion.

Multi-Frame Denoising. Aside from aforementioned MFSR approaches, several multi-frame methods have been developed to perform denoising [7, 14, 26, 27]. The algorithm of [35] leverages visually similar image block within and across frames to obtain denoised results. Other works

[7, 27] extend the state-of-the-art single image denoising technique BM3D [7] to videos. The method of [22] yields favorable denoising results by employing a novel homography flow alignment technique with consistent pixel compositing operator. In the work of [12], the authors extend single-image denoising network to multi-frame task via recurrent deep convolutional neural network. The kernel prediction network [30] generates per-pixel kernels for fusing multiple-frames. RViDeNet [40] uses deformable convolutions to perform explicit frame alignment in order to provide improved denoising results. The re-parametrization approach of MFIR [4] learns image formation model in deep feature space for the multi-frame denoising. BIPNet [9] presents a novel pseudo-burst feature fusion approach to perform denoising on burst frames.

Multi-Frame Low-light Image Enhancement. In low-light conditions, smartphone cameras often yield noisy and color-distorted images due to their small aperture and sensor pixel cavities. [6] collect a multi-frame dataset for low-light image enhancement, and present a data-driven approach to learn camera imaging pipeline in order to map under-exposed RAW images directly to well-lit sRGB images. The quality of output image is further improved with the perceptual loss presented by [43]. The works of [28] and [45], respectively, use residual learning framework and recurrent convolution network to obtain enhanced images from multiple degraded low-lit input frames. The two-stage approach of [18] employs one subnet for explicitly denoising multiple frames followed by the second subnet to provide us with the enhanced image. Along with super-resolution and denoising, BIPNet [9] is also capable of performing multi-frame low-light image enhancement. Unlike the existing multi-frame approaches, our Burstormer aligns burst features at multiple-scales and enables flexible inter-frame communication without much computational overhead. It also incorporates progressive feature merging to obtain high-quality images.

3. Proposed Burst Image Processing Pipeline

Burst sequences are usually acquired with handheld devices. The spatial and color misalignments among burst frames are unavoidable due to hand-tremor and camera/object motions. These issues negatively affect the overall performance of the burst processing approaches. In this work, our goal is to effectively *align* and progressively *merge* the desired information from multiple degraded frames to reconstruct a high-quality composite image. To this end, we propose Burstormer, a novel unified model for multi-frame processing where different modules jointly operate to perform feature denoising, alignment, fusion, and upsampling tasks. Here, we describe our method for the task of burst super-resolution, nevertheless, it is applicable to different burst restoration tasks such as burst denoising and burst enhancement (see experiments Sec. 4).

Overall Pipeline. Fig. 2 shows the overall pipeline of the proposed Burstormer. *First,* the RAW input burst is passed through the proposed enhanced deformable alignment (EDA) module which extract noise-free deep features that are aligned and refined with respect to the reference frame features. *Second,* an image reconstruction module is employed that takes as input the burst of aligned features and progressively merges them using the proposed no reference feature enrichment (NRFE) module. To obtain the super-resolved image, the upsampling operation is immediately applied after each NRFE module in the reconstruction stage. Next, we describe each stage of our approach.

3.1. Enhanced Deformable Alignment

In burst processing, effective alignment of mismatched frames is essential as in any error arising in this stage will propagate to later stages, subsequently making the reconstruction task difficult. Existing methods perform image alignment either explicitly [3, 4], or implicitly [9]. While, these techniques are suitable to correcting mild pixel displacements among frames, they might not adequately handle fast moving objects. In Burstormer, we propose enhanced deformable alignment (EDA) which employs a multi-scale design as shown in Fig. 2(a). Since sub-pixel shifts among frames are naturally reduced at low-spatial resolution, using the multi-level hierarchical architecture provides us with more robust alignment. Therefore, EDA starts feature alignment from the lowest level (3^{rd} in this paper) and progressively passes offsets to upper high-resolution levels to help with the alignment process. Furthermore, at each level, the aligned features are passed through the proposed reference-based feature enrichment (RBFE) module to fix remaining misalignment issues in burst frames by interacting them again with the reference frame. EDA has two key components: **(1)** Feature alignment, and **(2)** Reference-based feature enrichment.

Feature alignment. Burst images are often contaminated with random noise that impedes in finding the dense correspondences among frames. Therefore, before performing alignment operation, we extract noise-free burst features by using burst feature attention (BFA) module which is built upon the existing transformer block [41]. Unlike in other approaches [3, 4, 9], the BFA module computes encodes local and non-local context using MDTA block [41] and controls feature transformation through the GDFN [41] block. Furthermore, BFA module also efficient enough to be applied to high-resolution images. The denoised features from BFA are passed further for alignment. Figure 2(b) shows the feature alignment (FA) module that utilizes a modulated deformable convolution [46] to align features

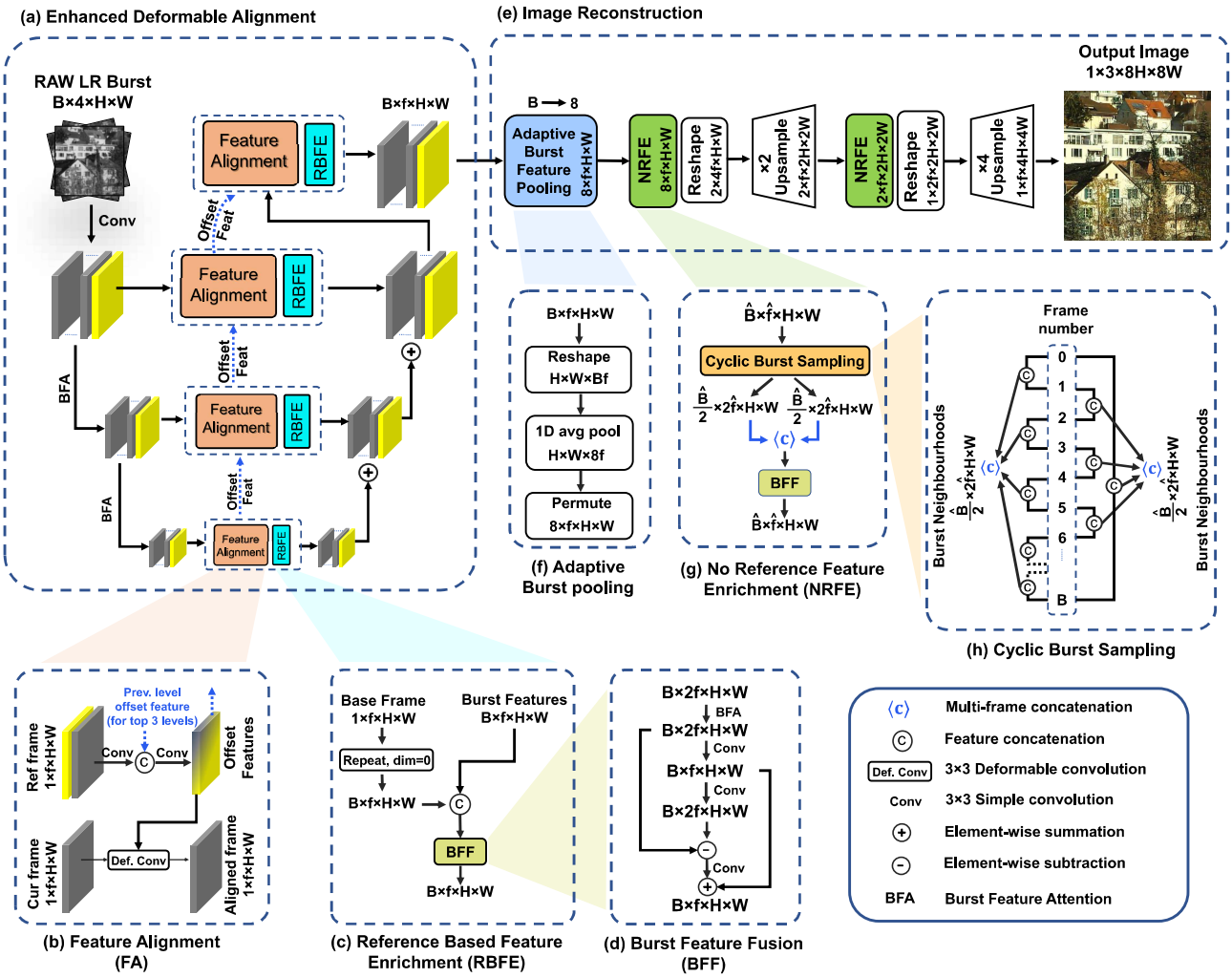

Figure 2. Overall pipeline of the proposed Burstormer for burst image processing. Burstormer takes as input a RAW burst of degraded images and outputs a clean high-quality sRGB image. It has two main parts: enhanced deformable alignment (EDA), and the image reconstruction. EDA, labeled as (a), is a multi-scale hierarchical module that, at each level, first extracts noise-free local and non-local features with the burst feature attention (BFA), performs feature alignment (b), and finally refines and consolidates features with an additional interaction with the base frame via (c) the proposed reference-based feature enrichment (RBFE) module. RBFE further employs (d) the burst feature fusion (BFF) unit for merging features by using the back-projection and squeeze-excitation mechanisms. The aligned burst of features are then passed to the image reconstruction stage (e). Here (f) the adaptive burst pooling module transforms the input burst size (B frames) to constant 8 frames through an average pooling operator. Finally, (g) the no-reference feature enrichment (NRFE) module progressively aggregates, and upsamples the burst features to generate the final HR image.

of each burst frame to those of the reference frame. Let, $\{\boldsymbol{g}^b : b \in [1, \ldots, B]\} \in \mathbb{R}^{B\times f\times H\times W}$ denotes the burst features obtained after BFA module, where B denotes number of burst frames, f is the number of feature channels, and $H\times W$ is the spatial size. We align the features of the current frame $\boldsymbol{g}^b$ with the reference frame* $\boldsymbol{g}^{b_r}$. Feature alignment module processes $\boldsymbol{g}^b$ and $\boldsymbol{g}^{b_r}$ via an offset convolution layer and outputs the offset Δn and modulation scalar Δa values for $\boldsymbol{g}^b$. In Fig. 2(a), for simplicity, only offset Δn is shown. The aligned features $\bar{\boldsymbol{g}}^b$ are computed as:

$$\bar{\boldsymbol{g}}^b = W_{\text{def}}\left(\boldsymbol{g}^b,\ \{\Delta n,\ \Delta a\}\right), \{\Delta n, \Delta a\} = W_{\text{off}}\left(\boldsymbol{g}^b,\ \boldsymbol{g}^{b_r}\right), \tag{1}$$

where, $W_{\text{def}}(\cdot)$ and $W_{\text{off}}(\cdot)$ represent the deformable and offset convolutions, respectively. Specifically, every position n on the aligned feature map $\bar{\boldsymbol{g}}^b$ is calculated as:

$$\bar{\boldsymbol{g}}^b_n = \sum_{i=1}^{K} W^d_{n_i}\ \boldsymbol{y}^b_{(n+n_i+\Delta n_i)} \cdot \Delta a_{n_i}, \tag{2}$$

where, K=9, Δa in range $[0, 1]$ for each $n_i \in$

*We consider the first burst image to be the reference frame.

$\{(-1,1),(-1,0),...,(1,1)\}$ is a regular 3×3 kernel grid. The convolution is performed on non-uniform positions $(n_i+\Delta n_i)$, where n_i may be fractional. To tackle the fractional values, this operation is implemented with the bilinear interpolation.

Reference-Based Feature Enrichment. In the presence of complex pixel displacements among frames, simple alignment techniques [3, 4, 9] may not able to align burst features completely. Thus, to fix the remaining minor misalignment issues, we propose the reference-based feature enrichment (RBFE) module, shown in Fig. 2(c). RBFE enables additional interaction of aligned frames features $\bar{\boldsymbol{g}}^b$ with the reference frame features $\boldsymbol{g}^{b_r}$ to generate consolidated and refined representations. This interactive feature merging is performed via our burst feature fusion (BFF) unit as illustrated in Fig. 2(d). The BFF mechanism is built upon the principles of feature back projection [13] and squeeze-excitation techniques [16]. Given the concatenated feature maps of the current frame and the reference frame $[\bar{\boldsymbol{g}}^b, \boldsymbol{g}^{b_r}] \in \mathbb{R}^{1\times 2*f\times H\times W}$, BFF applies BFA to generate representations $\boldsymbol{g}_{\boldsymbol{a}}^{\boldsymbol{b}}$ encoding the local non-local context. Overall, BFF yields fused features $\boldsymbol{g}_{\boldsymbol{f}}^{\boldsymbol{b}} \in \mathbb{R}^{1\times f\times H\times W}$:

$$\boldsymbol{g}_{\boldsymbol{f}}^{\boldsymbol{b}} = \boldsymbol{g}_{\boldsymbol{s}}^{\boldsymbol{b}} + W\left(\boldsymbol{g}_{\boldsymbol{a}}^{\boldsymbol{b}} - \boldsymbol{g}_{\boldsymbol{e}}^{\boldsymbol{b}}\right), \tag{3}$$

where $\boldsymbol{g}_{\boldsymbol{s}}^{\boldsymbol{b}}{=}W_s\boldsymbol{g}_{\boldsymbol{a}}^{\boldsymbol{b}}$ represents squeezed features and $\boldsymbol{g}_{\boldsymbol{e}}^{\boldsymbol{b}}{=}W_eW_s\boldsymbol{g}_{\boldsymbol{a}}^{\boldsymbol{b}}$ are the expanded features. W_s and W_e denote simple convolutions to squeeze and expand feature channels. The squeezed features $\boldsymbol{g}_{\boldsymbol{s}}^{\boldsymbol{b}}$ poses complementary properties of multiple input features. While, $\boldsymbol{g}_{\boldsymbol{e}}^{\boldsymbol{b}}$ is used to compute the high-frequency residue with the attentive features $\boldsymbol{g}_{\boldsymbol{a}}^{\boldsymbol{b}}$. The aggregation of this high-frequency residual with the squeezed features $\boldsymbol{g}_{\boldsymbol{s}}^{\boldsymbol{b}}$ helps to learn the feature fusion process implicitly and provides the capability to extract high-frequency complementary information from multiple inputs. While illustrated for fusing features of two frames in Fig. 2(d), the proposed BFF can be flexibly adapted to any number of inputs.

3.2. Image Reconstruction

Figure 2(e) illustrates the overall image reconstruction stage. To operate on bursts of arbitrary sizes, we propose an adaptive burst feature pooling (ABFP) mechanism that returns the constant burst-size features. As shown in Fig. 2(f), the burst features $(B*f)$ are concatenated along channel dimension followed by 1D average pooling operation which adaptively pools out the burst features $(8*f)$ as per the requirement. Next, the pooled burst feature maps pass through the no-reference feature enrichment (NRFE) module, shown in Fig. 2(g). The key idea of the proposed NRFE module is to pair immediate neighborhood frames along feature dimension and fuse them using the BFF module. However, doing this would limit the inter-frame communication to successive frames only. Therefore, we propose cyclic burst sampling (CBS) that gathers the neighborhood frames in zigzag manner (called here as burst neighborhoods) such that reference frame could interact with the last frame as well via intermediate frames; see Fig. 2(h). This cyclic scheme of sampling the burst frames helps in long-range communication without increasing the computational overhead unlike the existing pseudo burst technique [9]. Next, the sampled neighborhood features are combined along burst dimension and processed with BFF to integrate the useful information available in multiple frames of the burst sequence.

To upscale the burst features, we adapt pixel-shuffle [32] such that the information available in burst frames is shuffled to increase the spatial resolution. This helps in reducing the compute cost and the overall network parameters.

4. Experiments and Analysis

We evaluate the performance of the proposed Burstormer on three different burst image processing tasks: **(a)** super-resolution (on synthetic and real burst images), **(b)** low-light image enhancement, and **(c)** denoising (on grayscale and color data). Additional visual results, ablation experiments, and more details on the network and training settings are provided in the supplementary material.

Implementation Details. We train separate models for different tasks in an end-to-end manner without pre-training any module. We pack the input mosaicked raw burst into 4-channel RGGB format. All burst frames are handled with shared Burstormer modules (FA, RBFE, BFF, NRFE) for better parameter efficiency. The following training settings are common to all tasks, whereas task-specific experimental details are provided in their corresponding sections. The EDA module of Burstormer is a 3-level encoder-decoder, where each level employs 1 FA (containing single deformable conv. layer) and 1 RBFE module. The BFF unit both in RBFE and NRFE consists of 1 BFA module. Each BFA module consists of 1 multi-dconv head transposed attention (MDTA) and 1 gated-Dconv feed-forward network (GDFN) [41]. In the image reconstruction stage, we use 2 NRFE modules. We train models with L_1 loss and Adam optimizer with the initial learning rate $1e^{-4}$ that is gradually reduced to $1e^{-6}$ with the cosine annealing scheduler [23] on four RTX6000 GPUs. Random horizontal and vertical image flipping is used for data augmentation.

4.1. Burst Super-resolution

We evaluate the proposed Burstormer on synthetic as well as on real-world datasets [2, 3] for the SR scale factor $\times4$. For comparisons, we consider several burst SR approaches such as DBSR [3], LKR [19], HighResNet [8], MFIR [4] and BIPNet [9].

Table 1. **Burst super-resolution results** on synthetic and real datasets [3] for factor $4\times$.

Methods	SyntheticBurst			(Real) BurstSR	
	PSNR ↑	SSIM ↑	Time (ms)	PSNR ↑	SSIM ↑
Single Image	36.17	0.91	40.0	46.29	0.982
HighRes-net [8]	37.45	0.92	46.3	46.64	0.980
DBSR [3]	40.76	0.96	431	48.05	0.984
LKR [19]	41.45	0.95	-	-	-
MFIR [4]	41.56	0.96	420	48.33	0.985
BIPNet [9]	41.93	0.96	130	48.49	0.985
AFCNet [29]	42.21	0.96	140	48.63	0.986
Burstormer (Ours)	**42.83**	**0.97**	**55.0**	**48.82**	**0.986**

Datasets. **(1)** SyntheticBurst dataset [3] contains 46,839 RAW burst sequences for training and 300 for validation. Each sequence consists of 14 LR RAW images (with spatial resolution of 48×48 pixels) that are synthetically generated from a single sRGB image as follows. The given sRGB image is first transformed to RAW space with the inverse camera pipeline [5]. Next, random rotations and translations are applied to this RAW image to generate the HR burst sequence. The HR burst is finally converted to LR RAW burst sequence by applying the downsampling, Bayer mosaicking, sampling and random noise addition operations.
(2) BurstSR dataset [3] has 200 RAW burst sequences, each containing 14 images. The LR images of these sequences are captured with a smartphone camera, whereas their corresponding HR (ground-truth) images are taken with a DSLR camera. From 200 full-resolution sequences, the original authors extract 5,405 patches of size 80×80 for training and 882 patches for validation.

SR results on synthetic dataset. We train Burstormer with batch size 4 for 300 epochs on SyntheticBurst dataset [3]. Table 1 shows that our approach significantly advances the state of the art. When compared to the previous best BIPNet [9], our Burstormer yields performance gain of 0.9 dB, while having 47% fewer parameters, 80% less FLOPs, and runs $2\times$ faster. Fig. 3 shows that Burstormer-restored images are visually superior with enhanced structural and textural details compared to competing methods. Specifically, the reproductions of DBSR [3], LKR [19], and MFIR [4] contain blotchy textures and color artifacts.

SR results on real dataset. In BurstSR dataset [3], the LR and HR bursts are slightly misaligned as they are captured with different cameras. We address this by training Burstormer using aligned L_1 loss and evaluating with aligned PSNR/SSIM, as in prior works [3, 4, 9]. Instead of training from scratch, we fine-tune the pre-trained model (of SyntheticBurst dataset) for 100 epochs on the BurstSR dataset. Table 1 shows that our Burstormer performs favorably well by providing PSNR gain of 0.33 dB over the previous best method BIPNet [9]. We present visual comparisons in Fig. 4. Burstormer-generated images exhibit higher detail, sharpness, and visual accuracy.

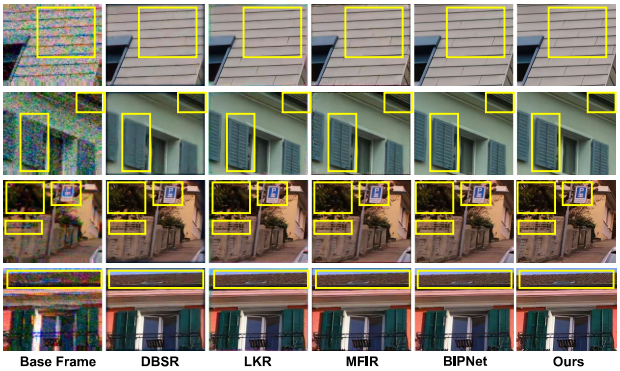

Figure 3. **Burst super-resolution** ($\times4$) results on SyntheticBurst dataset [3]. The SR images by our Burstormer retain more texture and structural content than the other approaches.

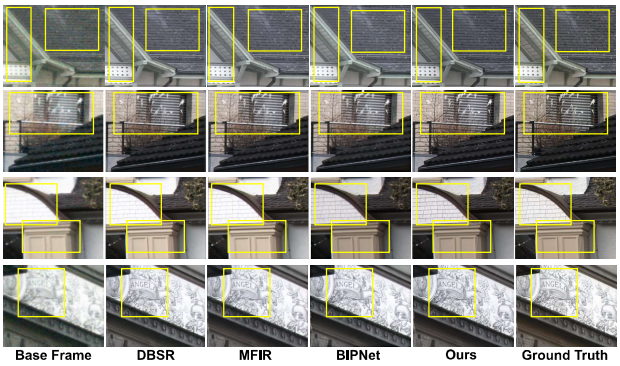

Figure 4. **Burst super-resolution** ($\times4$) results on BurstSR dataset [3]. Our results recover better visual details.

Ablation experiments. To study the impact of different modules of the proposed architecture on the final performance, we train several ablation models on the SyntheticBurst dataset [3] for 100 epochs. Results are provided in Fig. 5. In the baseline model, we use Resblocks [20] for feature extraction, simple concatenation-based fusion, and the pixel-shuffle operation for upsampling. It can be seen that inclusion of the proposed RBFE in feature alignment stage leads to substantial PSNR boost of 1.02 dB. This performance gain is further increased by 1.49 dB when we add the proposed burst fusion (NRFE) and upsampling modules. Overall, when deployed all our modules, we achieve 5.67 dB increment over the baseline. Further, Table 2 shows that replacing the proposed alignment and fusion methods with other existing techniques causes significant performance drop, *i.e.,* 0.43 dB and 0.34 dB, respectively. Specifically, our Burstormer lead to 0.79 dB boost when compared with existing multi-level PCD alignment [37]. The proposed RBFE module with local-non-local feature extraction ability is a key difference between the existing PCD and our enhanced deformable alignment. Further, we observe 0.34 dB drop in PSNR when we replace the proposed NRFE (fusion module) with existing compute extensive PBFF [9]. Ablation experiments show that with compute efficient in nature

Table 2. Comparison of **alignment and fusion** techniques. PSNR is computed on SyntheticBurst [3] for $4\times$ SR.

Task	Methods	PSNR ↑
Alignment	Explicit [3]	39.84
	TDAN [34]	40.58
	PCD [37]	41.26
	EBFA [9]	41.62
	Burstormer (Ours)	**42.05**
Burst Fusion	Addition	40.20
	Concat	40.65
	DBSR [3]	41.08
	PBFF [9]	41.71
	Burstormer (Ours)	**42.05**

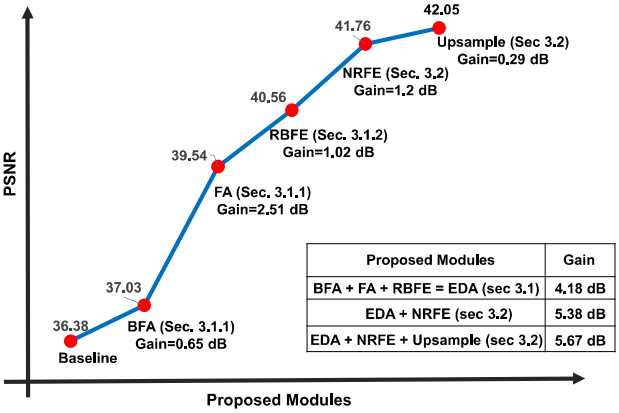

Figure 5. **Ablation experiments** for Burstormer contributions. PSNR is reported on SyntheticBurst dataset [3] for $4\times$ SR. All our major components contribute significantly. As given in Table, our Burstormer achieves 5.67 dB gain over the baseline approach.

our modules outperform other existing modules in all manner without any compromise in performance.

4.2. Burst Low-Light Image Enhancement

We test the performance of our Burstormer on the Sony subset from the SID dataset, as in other existing works [9, 18, 42, 45]. In addition to L_1 loss, we use the perceptual loss [44] for network optimization.

Dataset. SID [6] contains input RAW burst sequences captured with short-camera exposure in extreme low ambient light, and their corresponding well-exposed sRGB ground-truth images. The dataset consists of 161 burst sequences for training, 36 for validation, and 93 for testing. We crop 6,500 patches of size 256×256 with burst size varying from 4 to 8 and train the network for 200 epochs. Since the input RAW burst is mosaicked, we use single $2\times$ upsampler in our Burstormer to obtain the final image.

Enhancement results. The image quality scores for competing approaches are summarized in Table 3. Our Burstormer achieves PSNR gains of 0.47 dB over the previous best method BIPNet [9] and 3.54 dB over the second best algorithm LEED [18]. Figure 6 shows enhanced

Table 3. **Burst low-light image enhancement** evaluation on the SID dataset [6]. Burstormer performs well across three metrics.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
Chen *et al.* [6]	29.38	0.892	0.484
Maharjan *et al.* [28]	29.57	0.891	0.484
Zamir *et al.* [43]	29.13	0.881	0.462
Zhao *et al.* [45]	29.49	0.895	0.455
LEED [18]	29.80	0.891	0.306
BIPNet [9]	32.87	0.936	0.305
Ours	**33.34**	**0.941**	**0.285**

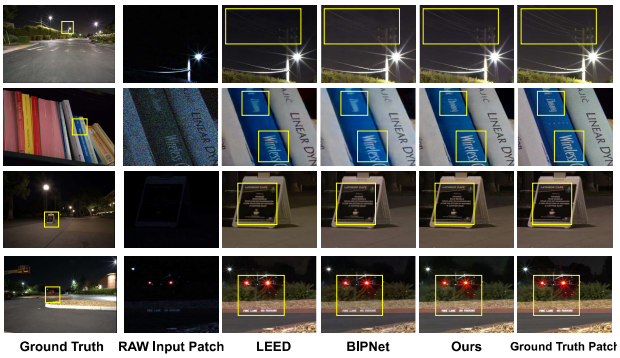

Figure 6. **Burst low-light image enhancement** comparisons on the Sony subset of SID dataset [6]. Our Burstormer retains color and structural details faithfully relative to the ground-truth.

images produced by different approaches. Our Burstormer yields images with more faithful color and structural content than the other compared approaches.

4.3. Burst Denoising

This section presents the results of burst denoising on grayscale data [30] as well as on color data [30]. As there is no need to upscale the burst features, we replace the upsampler in Burstormer with a simple convolution to generate the output image.

Datasets. Following the experimental protocols of [30] and [39], we prepare training datasets for grayscale denoising and color denoising, respectively. We train separate denoising models for 300 epochs on 20K synthetic burst patches. Each burst contains 8 frames of 128×128 spatial resolution. Testing is performed on 73 grayscale bursts and 100 color bursts. Both of these test sets contain 4 variants with 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.

Denoising results. We compare various existing methods such as KPN [30], MKPN, BPN [39], MFIR [4], and BIPNet [9]. Since the proposed Burstormer trained without any extra data or supervision, we consider results of the MFIR [4] variant that uses a custom optical flow subnetwork (without pre-training it on extra data). Table 4

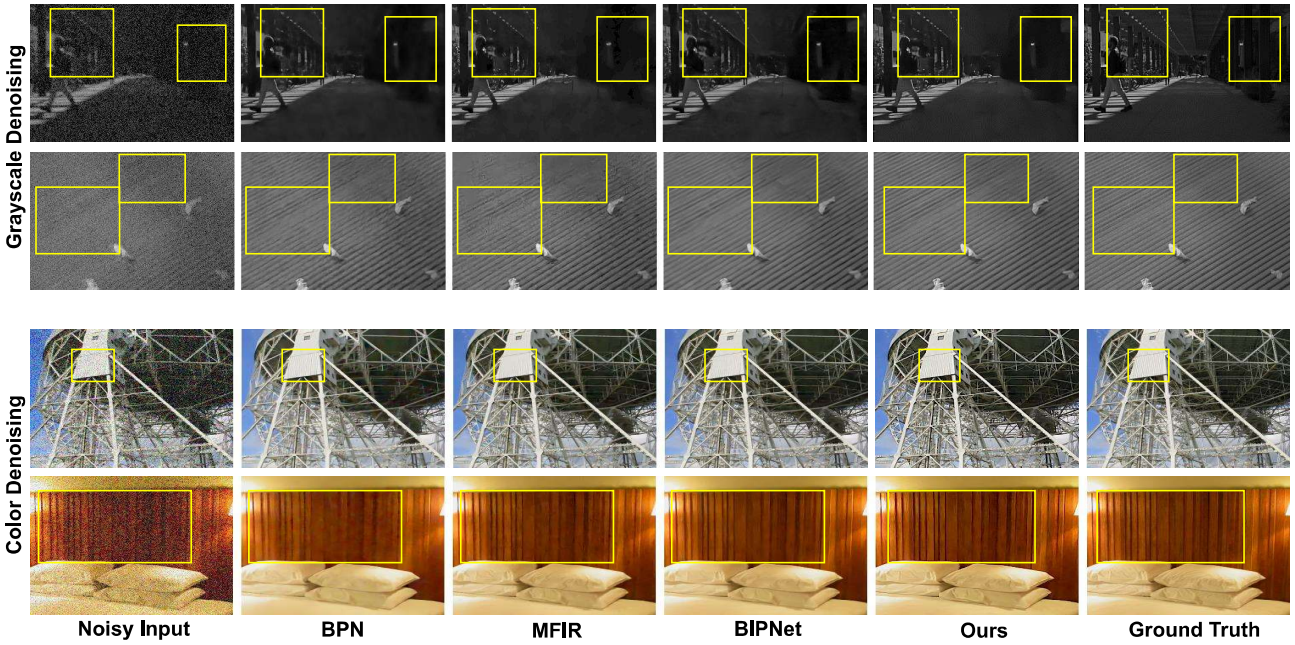

Figure 7. **Burst denoising** results on burst images from the grayscale [30] and color datasets [39]. Our Burstormer produces more sharp and clean results than other competing approaches. More examples are provided in the supplementary material.

Table 4. **Grayscale burst denoising** PSNR on the dataset by [30].

	Gain $\propto$ 1	Gain $\propto$ 2	Gain $\propto$ 4	Gain $\propto$ 8	**Average**
KPN [30]	36.47	33.93	31.19	27.97	32.39
BPN [39]	38.18	35.42	32.54	29.45	33.90
MFIR [4]	39.10	36.14	32.89	28.98	34.28
BIPNet [9] †	38.53	35.94	33.08	29.89	34.36
Burstormer (Ours)	**39.49**	**36.70**	**33.71**	**30.55**	**35.11**

Table 5. **Color burst denoising** PSNR on the dataset by [39].

	Gain $\propto$ 1	Gain $\propto$ 2	Gain $\propto$ 4	Gain $\propto$ 8	**Average**
KPN [30]	38.86	35.97	32.79	30.01	34.41
BPN [39]	40.16	37.08	33.81	31.19	35.56
BIPNet [9] †	40.58	38.13	35.30	32.87	36.72
MFIR [4]	**41.90**	38.85	35.48	32.29	37.13
Burstormer (Ours)	41.70	**39.15**	**36.09**	**33.44**	**37.59**

reports grayscale denoising results where our Burstormer consistently performs well. When averaged across all noise levels, our method provides 0.75 dB PSNR improvement over the state-of-the-art BIPNet [9]†. Table 5 shows that the performance trend of Burstormer is similar on color denoising as well. For instance, on high noise level bursts (Gain $\propto$ 8), Burstormer provides PSNR boost of 0.57 dB over BIPNet [9]. Visual comparisons in Fig. 7 show that Burstormer's denoised outputs are relatively cleaner, sharper and preserve subtle textures. Additional qualitative results are provided in supplementary material.

†We use BIPNet results from the official Github repository.

5. Conclusion

We present a transformer-based framework for burst image processing. The proposed Burstormer is capable of generating a single high-quality image from a given burst of noisy images having pixel misalignments among them. Burstormer employs a multi-scale hierarchical module EDA that, at each scale, first generates denoised features encoding local and non-local context, and then aligns each burst frame with the reference frame. To fix any remaining minor alignment issues, we incorporate a reference-based feature enrichment RBFE module in EDA that enables additional interaction of the features of each frame with the base frame features. Overall, EDA improves model robustness by yielding a burst of features that are well denoised, aligned, consolidated and refined. In the image reconstruction stage, we repeatedly apply the no-reference feature enrichment NRFE and upsampling modules in tandem until the final image is obtained. NRFE progressively and adaptively fuses each pair of frame features that are obtained with the proposed cyclic burst sampling. Experiments performed on three representative burst processing tasks (super-resolution, denoising, low-light image enhancement) demonstrate that our Burstormer provides state-of-the-art results and generalizes well compared to recent burst processing approaches.

References

[1] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *ECCV*, 1996.

[2] Goutam Bhat, Martin Danelljan, and Radu Timofte. Ntire 2021 challenge on burst super-resolution: Methods and results. In *CVPR*, 2021.

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021.

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

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

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

[7] Kostadin Dabov, Alessandro Foi, and Karen. Egiazarian. Video denoising by sparse 3d transform-domain collaborative filtering. *ESPC*, 2007.

[8] 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.

[9] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022.

[10] Michael Elad and Yacov Hel-Or. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *TIP*, 2001.

[11] Esmaeil Faramarzi, Dinesh Rajan, and Marc Christensen. Unified blind method for multi-image super-resolution and single/multi-image blur deconvolution. *TIP*, 2013.

[12] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *ECCV*, 2018.

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

[14] 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. *ACM ToG*, 2016.

[15] Yu He, Kim-Hui Yap, Li Chen, and Lap-Pui Chau. A nonlinear least square technique for simultaneous image registration and super-resolution. *TIP*, 2007.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[17] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 1991.

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

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

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

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[22] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM ToG*, 2014.

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

[24] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Bsrt: Improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 998–1008, 2022.

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

[26] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising using separable 4d nonlocal spatiotemporal transforms. In *Electronic Imaging*, 2011.

[27] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *TIP*, 2012.

[28] 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.

[29] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Adaptive feature consolidation network for burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1279–1286, 2022.

[30] Ben Mildenhall, Jonathan Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, 2018.

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

[32] 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.

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

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

[35] Marius Tico. Multi-frame image denoising and stabilization. *ESPC*, 2008.

[36] Roger Tsai. Multiframe image restoration and registration. *ACVIP*, 1984.

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

[38] 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.

[39] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *CVPR*, 2020.

[40] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *CVPR*, 2020.

[41] 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.

[42] 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.

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

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[45] 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.

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