Gated Multi-Resolution Transfer Network for Burst Restoration and Enhancement

Nancy Mehta[1] Akshay Dudhane[2] Subrahmanyam Murala[1] Syed Waqas Zamir[3]
Salman Khan[2,4] Fahad Shahbaz Khan[2,5]
[1]CVPR Lab, Indian Institute of Technology Ropar [2]Mohamed bin Zayed University of AI
[3]Inception Institute of AI [4]Australian National University [5]Linköping University

Abstract

Burst image processing is becoming increasingly popular in recent years. However, it is a challenging task since individual burst images undergo multiple degradations and often have mutual misalignments resulting in ghosting and zipper artifacts. Existing burst restoration methods usually do not consider the mutual correlation and non-local contextual information among burst frames, which tends to limit these approaches in challenging cases. Another key challenge lies in the robust up-sampling of burst frames. The existing up-sampling methods cannot effectively utilize the advantages of single-stage and progressive up-sampling strategies with conventional and/or recent up-samplers at the same time. To address these challenges, we propose a novel ***G****ated* ***M****ulti-Resolution* ***T****ransfer Network (GMTNet) to reconstruct a spatially precise high-quality image from a burst of low-quality raw images. GMTNet consists of three modules optimized for burst processing tasks: Multi-scale Burst Feature Alignment (MBFA) for feature denoising and alignment, Transposed-Attention Feature Merging (TAFM) for multi-frame feature aggregation, and Resolution Transfer Feature Up-sampler (RTFU) to up-scale merged features and construct a high-quality output image. Detailed experimental analysis on five datasets validate our approach and sets a state-of-the-art for burst super-resolution, burst denoising, and low-light burst enhancement. Our codes and models are available at https://github.com/nanmehta/GMTNet.*

1. Introduction

With the soaring popularity of smartphones in day-to-day life, the demand for capturing high-quality images is rapidly increasing. However, the camera in smartphone has several limitations due to the constraints placed on it in order to be integrated into smartphone's thin profile. The most prominent hardware limitations are the small camera sensor size and the associated lens optics that reduce their spatial resolution and dynamic range [14], impeding them in reconstructing DSLR-alike images. To deal with these inherent physical limitations of mobile photography, one emerging solution is to leverage multi-frame (burst) processing instead of single-frame processing. Burst processing techniques primarily focus on extracting high-frequency details by merging non-redundant data from various shifted images to produce a high-quality image.

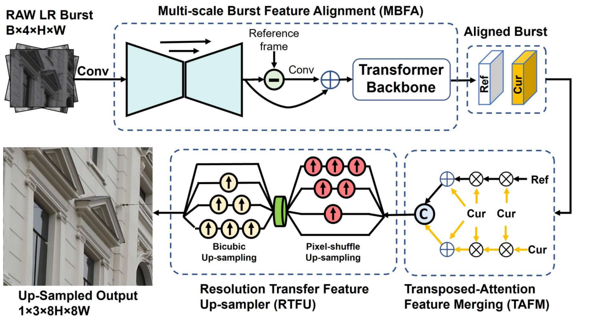


Figure 1. Proposed GMTNet processes RAW burst LR frames and gives a high-quality image through three key stages: (1) Multi-scale Burst Feature Alignment (MBFA), (2) Transposed-Attention Feature Merging (TAFM), and (3) Resolution Transfer Feature Up-sampler (RTFU).

Three critical factors involved in burst processing are feature alignment, fusion, and subsequent reconstruction of the obtained frames. Generally, any burst processing approach is limited by the accuracy of alignment process on account of the camera and scene motion of dynamically moving objects. Therefore, it is crucial to design a module for facilitating accurate alignment, as the subsequent fusion and reconstruction modules must be robust to misalignment for generating an artifact-free image. We further note that the alignment and fusion modules in existing burst processing approaches [3, 17] do not consider the non-local dependencies and mutual correlation among the frames which hinders the flexible inter-frame information exchange. Moreover, the existing burst up-sampling approaches [2, 17] do not take into account the merits of repeatedly transferring the information across several resolutions. To address these issues, we present a novel burst processing framework named Gated Multi-Resolution Transfer network (GMTNet) as illustrated in Figure 1.

In contrast to the previous works [2, 3] which adopt bulky pre-trained modules for alignment, we propose an implicit Multi-scale Burst Feature Alignment (MBFA) to reduce the inter-frame misalignment. Overall, MBFA module implicitly learns feature alignment at multiple scales through the proposed Attention-Guided Deformable Alignment (AGDA) module and obtains an enriched feature representation via Aligned Feature Enrichment (AFE) module. The proposed AFE module is composed of a back-projection mechanism and capable of extracting long-range pixel interactions that ease the feature alignment in complex motions, where simply aligning the frames does not suffice. Additionally, unlike the recent state-of-the-art (SoTA) algorithm, BIPNet [17] that utilizes a computationally intensive pseudo burst mechanism on the aligned burst for inter-frame communication, we propose a simple Transposed-Attention based Feature Merging (TAFM) module that leverages local and non-local correlations to allow an extensive interaction with the reference frame. Finally, our Resolution Transfer Feature Up-sampler (RTFU) combines the complementary features of both single-stage and progressive up-sampling strategies through deployed conventional and recent feature up-samplers. Such a design enables strong feature embedding of LR and HR images that creates a solid foundation for up-sampling in burst SR tasks. In this work, we validate our GMTNet for popular burst processing tasks such as super-resolution, denoising and low-light image enhancement. Overall, the following are our key contributions.

1. A Multi-scale Burst Feature Alignment (MBFA) is proposed which uses both local and non-local features for alignment at multiple scales, resolving the spatial misalignment within burst images (§3.1).
2. A Transposed-Attention Feature Merging (TAFM) is proposed to aggregate the features of the aligned and reference frames (§3.2).
3. A Resolution Transfer Feature Up-sampler (RTFU) is proposed to upscale the merged features. The proposed RTFU integrates the complementary features extracted by single-stage and progressive up-sampling strategies using the conventional and recent up-samplers (§3.3).

Our three-stage design achieves SoTA results on both synthetic as well as real raw datasets for burst super-resolution, denoising and low-light enhancement.

2. Related Work

Multi-Frame Super-Resolution. Compared to the single-image super-resolution (SISR), multi-frame super-resolution (MFSR) encounters new challenges while estimating the offsets among different images caused by camera movement and moving objects. Tsai and Huang [42] were the first to put forward a computationally cheap, frequency domain-based solution for the MFSR problem. Due to significant visual artifacts in frequency domain processing, spatial domain algorithms gained popularity [18, 20]. Following it, Irani and Peleg [23] and Peleg *et al.* [37] proposed an iterative back-projection based approach, and [1] utilized maximum a posteriori (MAP) model to obtain better super-resolved results. But all the above-mentioned approaches were based upon the assumption that motion between input frames, as well as the image formation model can be well estimated. Subsequent works addressed this issue with the joint estimation of the unknown parameters [19, 22].

Recently, a few data-driven approaches have been proposed for different applications, such as satellite imaging [15] and medical images [24]. Bhat *et al.* [2] addressed the problem of MFSR by proposing an explicit feature alignment and attention-based fusion mechanism. However, explicit use of motion estimation and image warping techniques can pose difficulty in handling scenes with fast object motions. Dudhane *et al.* [17, 34] proposed a generalised approach for processing noisy raw bursts through their implicit feature alignment and inter-frame communication strategy. Despite its better accuracy, [17] fails to consider the relevant non-local contextual information at multiple scales while aligning and fusing the features.

Multi-frame Denoising. Existing methods either utilize neural networks that are purely feed-forward [4, 46], recurrent networks [9] or a hybrid of both [10] for multi-frame denoising. Tico *et al.* [41] leveraged a block-based paradigm, and blocks within and across the burst images are used for performing denoising. [12,32] extended the defacto method of single image denoising approaches, BM3D [13] to videos. Liu *et al.* [30] demonstrated superior denoising performance by using a novel homography flow alignment technique via consistent pixel compositing operator. Godard *et al.* [21] proposed a novel multi-frame denoising model by using burst capture strategy and recurrent deep convolutional neural network. Mildenhall *et al.* [36] introduced Kernel Prediction Network (KPN) to generate per-pixel kernels, utilizing information from multiple images to merge input frames. Bhat *et al.* [2] proposed a deep reparametrization of the maximum a posteriori formulation for multi-frame denoising. Dudhane *et al.* [17] proposed a pseudo-burst feature fusion approach for burst frame denoising.

Low-Light Enhancement. Low-light photography in smartphones is limited on account of the small sensor, lens and limited aperture of camera. In [8], authors introduced a dataset of raw short and long-exposure low-light images,

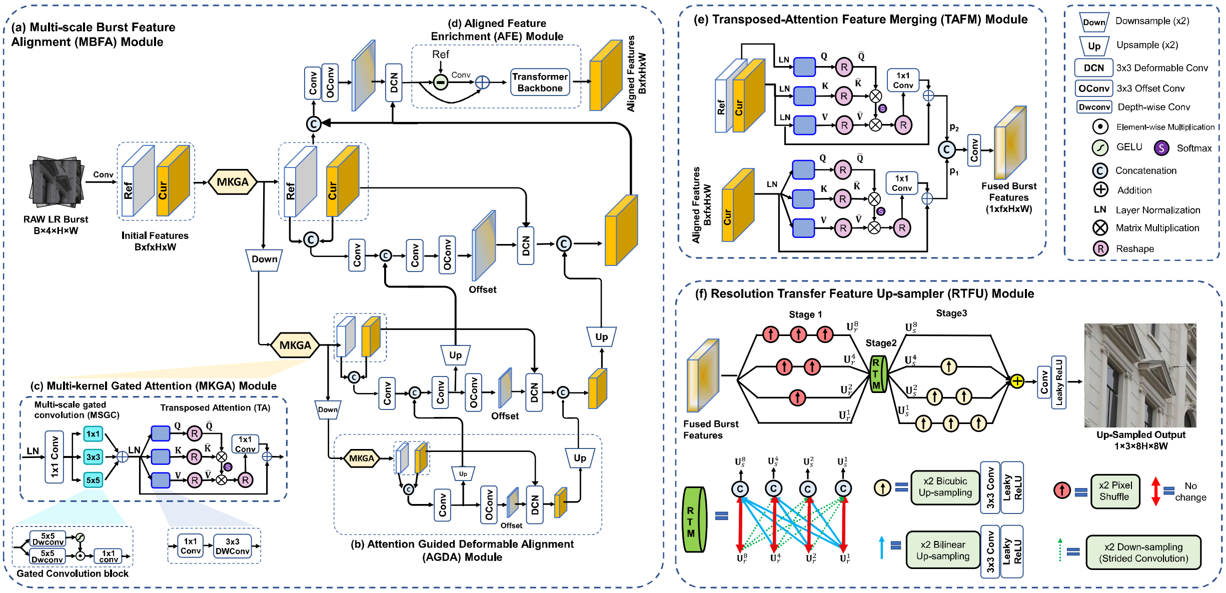


Figure 2. Comprehensive representation of each stage of our proposed GMTNet: (a) The proposed Multi-Scale Burst Feature Alignment (MBFA) module aligns burst features at multiple scales using the proposed (b) Attention-Guided Deformable Alignment (AGDA). The proposed AGDA reduces noise content through our (c) Multi-Kernel Gated Attention (MKGA) module. While, (d) Aligned Feature Enrichment (AFE) boosts high-frequency content through back-projection mechanism and extracts robust features through transformer backbone. (e) Transposed Attention Feature Merging (TAFM) module aggregates the local-non-local pixel interactions within the aligned and reference frames. Lastly, (f) Resolution Transfer Feature Up-sampler (RTFU) up-scales the merged features through single-stage and progressive up-sampling setting using both the conventional and recent up-samplers.

and proposed a learning based pipeline for mapping the degraded low-lit input frames to well-lit sRGB images. Zamir *et al.* [48] proposed a data-driven method for mapping underexposed RAW images to a well-exposed sRGB image. Jung *et al.* [25] leveraged a novel cycle adversarial network for generating frames in low lighting conditions. Liu *et al.* [29] used synthetic events from multiple frames for guiding the enhancement and restoration of low-light frames. Maharjan *et al.* [33] and Zhao *et al.* [51], respectively leveraged a residual learning-based and recurrent convolution network based framework to process burst photos acquired under extremely low-light conditions. Besides super-resolution and denoising, BIPNet [17] is also adept at performing multi-frame low-light image enhancement.

3. Methodology

We present the overall pipeline of our burst processing approach in Figure 1. Given a raw burst image, the goal of our GMTNet is to reconstruct a clean, high-quality image by exploiting the shifted complementary information from the noisy LR image burst. As shown in Figure 1, the input RAW LR burst features are aligned to the reference frame through our proposed **M**ulti-scale **B**urst **F**eature **A**lignment (MBFA) module. Further, aligned burst features are aggregated using the **T**ransposed-**A**ttention **F**eature **M**erging (TAFM) module. Lastly, our **R**esolution **T**ransfer **F**eature **U**p-sampler (RTFU) up-scales the merged features to reconstruct a high-quality image.

3.1. Multi-scale Burst Feature Alignment

Generating an artifact-free, high-quality image through burst processing is highly reliant upon the alignment of the mismatched burst frames. However, proper alignment is quite challenging, specifically in low-light and low-resolution images, where noise excessively contaminates the input burst frames. Previous burst restoration and video SR methods [2, 3, 5, 17, 38, 44] often seek to alleviate these issues by following alignment on locally extracted features. However, they do not explicitly consider the long-range dependencies which are crucial for restoration tasks. Consequently, the generated feature maps have limited receptive field making it difficult to align the burst features in case of complex motions. We develop the Multi-scale Burst Feature Alignment (MBFA) module to address the mentioned challenges, streamlining burst feature alignment across various scales and facilitating long-range pixel interactions for improved alignment. As seen in Figure 2(a), MBFA works in two phases: first, it aligns burst features at multiple scales with the Attention-Guided Deformable Alignment (AGDA) module; second, it refines aligned features via the Aligned Feature Enrichment (AFE) module.

3.1.1 Attention-Guided Deformable Alignment

As discussed in [7], noise disturbs the prediction of dense correspondences among multiple frames which is the key concern of several alignment methods. However, we find that a well-designed module can easily tackle noisy raw data. Therefore, in order to reduce the noise content in the initial burst features and eventually ease the alignment process, we propose an Attention-Guided Deformable Alignment (AGDA) module that operates at multiple scales to align the burst features as shown in Figure 2(b). The proposed AGDA module is inspired from the deformable alignment proposed in TDAN [40] and EDVR [44]. But, their alignment approaches [40, 44] directly apply deformable convolution on the input features, *making them prone to miss the detailed information in case of noisy RAW burst features.* Additionally, they also lack at extracting long-range pixel interactions which are useful in complex motions. Our AGDA block addresses these issues by performing implicit feature denoising using MKGA prior to burst feature alignment, instead of directly applying deformable convolution to incoming features. Further, the denoised burst features are aligned through the modulated deformable convolution (DCN) as shown in Figure 2(b).

Multi-Kernel Gated Attention. The proposed MKGA block offers dynamic adjustment of its receptive field to learn multi-scale *local context* through our Multi-Scale Gated Convolution (MSGC) sub-module and *non-local context* with the Transposed attention (TA) sub-module as demonstrated in Figure 2(c). This adaptability of transitioning between small (local) and large receptive fields is useful for dealing with various types of image degradation. Given an input tensor $\mathbf{Y} \in \mathbb{R}^{C\times H\times W}$, the overall operation of MSGC, outputting $\hat{\mathbf{Y}}$ is formulated as:

$$\hat{\mathbf{Y}} = W_1 * (G_1(\mathbf{Y})) + W_1 * (G_3(\mathbf{Y})) + W_1 * (G_5(\mathbf{Y})) \quad (1)$$

Here, W_1 denotes a convolution filter with size 1×1, and $*$ is a convolution operation. $G_k(\mathbf{Y})$ represents the output of the Gated Convolution block *(See Figure 2 (c))*, that is mapped out as the element-wise product of two parallel paths for depth-wise convolution layers with filter size *k* and formulated as $G_k(\mathbf{Y}) = \lambda(W_k^{dep}) \odot W_k^{dep}$. Here, W_k^{dep} denotes a depth-wise convolution layer, λ and $\odot$ represents the GELU non-linearity, and element-wise multiplication.

Transposed Attention. The extracted multi-kernel features from the MSGC module are passed through the transposed attention (TA) sub-module *(see Figure 2(c))* for capturing their long-range pixel interactions. From a layer normalized tensor $\hat{\mathbf{Y}}$, our TA sub-module first generates query **(Q)**, key **(K)**, and value **(V)** projections by applying 1×1 convolutions followed by 3×3 depth-wise convolutions for encoding the non-local and channel-wise spatial context. Thereafter, we reshape (**Q**,**K**,**V**) into $\hat{\mathbf{Q}}$, $\hat{\mathbf{K}}$ and $\hat{\mathbf{V}}$ projections such that the subsequent dot-product interactions between query and key generate a transposed-attention map of size $\mathbb{R}^{C\times C}$ [47], instead of the huge regular attention map of size $\mathbb{R}^{HW\times HW}$ [43]. And, the overall TA process, outputting $\tilde{\mathbf{Y}}$ is defined as:

$$\begin{aligned} \tilde{\mathbf{Y}} &= LN(\hat{\mathbf{Y}}) + W_1 * (TA(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}})); \\ TA(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) &= \hat{\mathbf{V}} \otimes S(\hat{\mathbf{K}} \otimes \hat{\mathbf{Q}}) \end{aligned} \quad (2)$$

Here, $\hat{\mathbf{Y}}$ is the feature map obtained from the MSGC module, LN denotes the layer normalization; TA and S denotes the operation of the TA sub-module and Softmax, respectively, $\hat{\mathbf{Q}} \in \mathbb{R}^{HW\times C}$, $\hat{\mathbf{K}} \in \mathbb{R}^{C\times HW}$, and $\hat{\mathbf{V}} \in \mathbb{R}^{HW\times C}$ matrices are obtained after reshaping the tensors from the original size, $\mathbb{R}^{C\times H\times W}$, and $\otimes$ denotes matrix multiplication. Altogether, the employed MKGA module at each scale allows each pyramidal level to focus on fine details, generating contextualized features that reduce noise and thus ease the subsequent alignment mechanism.

Modulated Deformable Convolution. After extracting the features from the MKGA module, we implicitly align the current frame features, $\boldsymbol{f^b}$ with the reference frame features (*we considered the first frame as reference*), $\boldsymbol{f^{b^r}}$ via modulated deformable convolution [40, 52] (learnable offsets for deformable convolution layer are obtained through a 3×3 offset convolution layer) as shown in Figure 2(b). To ensure better learning, the predicted offsets and aligned burst features are shared from the lower-scale to upper-scale in a bottom-up fashion to ensure semantically stronger and cleaner aligned features.

3.1.2 Aligned Feature Enrichment

To fix the remaining minor alignment and noise issues, we embed a novel Aligned Feature Enrichment (AFE) module on the obtained aligned features. The proposed AFE module differs from conventional high-frequency enhancement methods as it extracts local & non-local features through a transformer backbone, in addition to a back-projection mechanism. This results in a more effective approach for high-frequency enhancement. During the back-projection process, we simply compute the high-frequency residue between the aligned burst features and reference frame as shown in Figure 2(d). Thereafter, the local-non-local pixel interactions are enabled by processing the aligned edge boosted burst features through the existing transformer backbone [47]. In a nutshell, besides capturing multi-scale local-global representation among the bursts, the AFE module also bridges the gap between the relevant and irrelevant features of the aligned frames.

3.2. Transposed-Attention Feature Merging

In burst processing, temporal relation among the multiple frames plays an indispensable role in feature fusion on account of blurry frames from camera perturbations. Considering the fact, that incoming multiple frames have quite a few similar patterns at the feature level, it is infeasible to directly concatenate or add them as it will naively introduce a large amount of redundancy into the network. Existing DBSR [2] proposed an attention-based fusion approach but it is limited in exploiting the complementary (global and local) relations that can hinder the information exchange among multiple frames. Further, the recently proposed BIPNet [17] tries to merge the relevant information by concatenating channel-wise features from all burst feature maps. Though it is effective in extracting complementary information, it is computationally extensive.

Unlike the aforementioned fusion techniques, we propose a Transposed-Attention Feature Merging (TAFM) to efficiently encode *inter-frame and intra-frame correlations before merging the frames.* As shown in Figure 2(e), TAFM takes queries **(Q)** and a set of key-value (**K**,**V**) pairs as input and outputs the linear combination of values that are determined by correlations between the queries and corresponding keys [49]. The proposed TAFM module has been designed with two parallel blocks (*see Figure 2(e)*), where the lower block (outputting p_1) performs the query-key interactions across channels of the aligned neighboring frames to encode the channel-wise local context. While the upper block (outputting p_2) enhances the feature representations of the reference and current frames by bridging their global correlations. This design allows TAFM to effectively reduce feature redundancy and extract complementary information from multiple frames. After encoding the feature correlations globally and locally for a given aligned frame, $\bar{\boldsymbol{f}}^{\boldsymbol{b}}$ with b number of burst frames, the overall merged features of TAFM, $\boldsymbol{F_m} \in \mathbb{R}^{1\times C\times H\times W}$ is obtained as follows:

$$\boldsymbol{F_m} = W_3 * (p_1 \text{Ⓒ} p_2) \tag{3}$$

where, W_3 is a convolution layer with filter size 3×3, and Ⓒ refers to the concatenation.

3.3. Resolution Transfer Feature Up-sampler

The popular up-sampling techniques deployed in SoTA burst SR methods DBSR [2], DRSR [3] perform direct one-stage up-sampling without leveraging the benefits of information exchange between the HR features and their corresponding LR counterparts. Considering the fact that HR features contain abundant global information and LR features are rich in edge information [35,50], we design a Resolution Transfer Feature Up-sampler (RTFU) module that is the first upsampler to extract *unique features* of *different resolution spaces.* The proposed RTFU module stems from the observation that the transfer of LR and HR features through a multi-resolution framework can be propitious in adaptively recovering the textural information from the fused frames as shown in the ablation study. In RTFU, we target at exploiting the dual benefits of both direct [16] and progressive up-sampling [27] strategies using the conventional [6] and recent learnable up-sampling layers [39] to adequately get into the HR space. As shown in Figure 2(f), RTFU achieves its desired HR feature space via a three-stage design: two sets of four parallel progressive multi-resolution streams (Stage1 and Stage3) and a Resolution-Transfer Merging (RTM) module (Stage2).

We first apply progressive up-sampling strategy with pixel-shuffle [39] (*extreme left of Figure 2(f)*) parallelly in Stage1 for generating ($\times 1$, $\times 2$, $\times 4$, and $\times 8$) multi-resolution SR feature responses, which are then forwarded to the RTM module (Stage2). RTM module consists of four input representations: U_r^i (output of Stage1), i = 1, 2, 4, and 8 with i being the input resolution index, and the associated output representations are given by U_s^o, o = 1, 2, 4, and 8 with o being the output resolution index. Each output representation (U_s^o) is the concatenation of the transformed representations of the corresponding four inputs *(as shown in the middle of Figure 2(f)).* Thus, the overall operation of Stage2 (RTM module) can be formulated as follows:

$$U_s^o = \left[\left[f(U_r^i) \right]\right]_{i=1,2,4,8};\; f = \begin{cases} 1 \;\; \forall \;\; i = o \\ \frac{o}{i} \uparrow \;\; \forall \;\; o > i \\ \frac{i}{o} \downarrow \;\; \forall \;\; o < i \end{cases} \tag{4}$$

Here, $o \in \{1, 2, 4, 8\}$, and the mathematical definition of the symbol used in Eq. 4 is given as $\left[\left[A^j \right]\right]_{j=1,2,...n} = A^1 \text{Ⓒ} A^2 ..... \text{Ⓒ} A^n$, where Ⓒ denotes the concatenation operation among the inputs, and f represents the corresponding transformation operation (upsample or downsample) applied to the input feature U_r and is dependent upon the input resolution index (i), and the output resolution index (o). For instance, as shown above, if $o > i$, then the corresponding input representation U_r^i is up-sampled ($\uparrow$) by a factor of *o*/*i*. In Stage2 (RTM), we deploy bilinear interpolation and strided convolution for feature up-sampling and down-sampling, respectively. Thereafter, the resulting features from each branch of Stage3 are again up-sampled progressively using bicubic interpolation to generate an up-sampled feature map of the size $\mathbb{R}^{1\times C\times 8H\times 8W}$. Finally, we add the individual branch output of Stage3 to generate the final high-quality image. Thus, for each pixel location, RTFU can leverage the underlying content information from input frames at multiple-scales and utilize it to get better performance than the mainstream up-sampling operations, pixel-shuffle or interpolations.

4. Experimental Analysis

We validate the proposed GMTNet on real and synthetic datasets for **(a)** Burst Super-resolution, **(b)** Burst denoising, and **(c)** Burst low-light image enhancement tasks.

Implementation Details. We train separate models for all

Table 1. **Burst super-resolution** results for $\times 4$ factor.

Methods	SyntheticBurst [2]				BurstSR [2]	
	GFlops	Time (s)	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SingleImage	20.41	0.04	36.86	0.919	46.60	0.979
WMKPN [11]	-	-	36.56	0.912	41.87	0.958
HighResNet [15]	400	0.05	37.45	0.924	46.64	0.980
DBSR [21]	118	0.43	40.76	0.959	48.05	0.984
MFIR [3]	110	0.42	41.56	**0.964**	48.33	0.985
BIPNet [17]	300	0.13	41.93	0.960	48.49	0.985
Ours	157	0.20	**42.36**	0.961	**48.95**	**0.986**

Base frame DBSR [2] LKR [28] MFIR [3] BIPNet [17] Ours

Figure 3. Visual results on SyntheticBurst [2] for $\times 4$ burst SR.
Figure 4. Visual results on SyntheticBurst [2] for $\times 4$ burst SR.

the considered tasks in an end-to-end manner. For better parameter efficiency, we shared each GMTNet module for all burst frames. Our GMTNet has 12.7M parameters with 157 GFLOPs for the burst of size 14×4×48×48 with a running time of 24 fps. To train GMTNet with 4 V100 GPUs, it takes 29 hours for real SR, 97 hours for synthetic SR, 72 hours for grayscale/color denoising, and 38 hrs for burst enhancement. All the models are trained with Adam optimizer with L_1 loss function. We employ cosine annealing strategy [31] to decrease the learning rate from 10^{-4} to 10^{-6} during training. For real-world SR, we fine-tune our GMTNet (*with pre-trained weights on SyntheticBurst dataset*) using aligned L_1 loss [2]. We provide the task-specific experimental details in the corresponding sections. *Additional experimental details and visual results are provided in the supplementary material.*

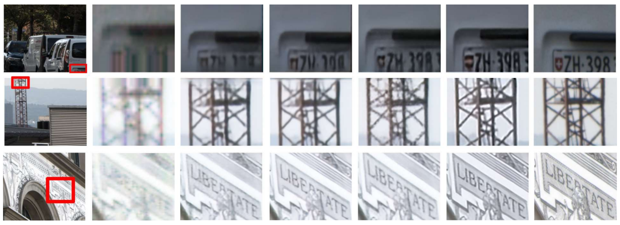
4.1. Burst Super-Resolution

We evaluate our proposed GMTNet on synthetic [2] and real-world datasets [2] for scale factor $\times 4$. Following the settings in [2], we utilized **SyntheticBurst** dataset (46,839 and 300 RAW burst sequences for training and validation respectively, where each burst sequence consists of 14 images), and **BurstSR** dataset consisting of 200 RAW burst sequences (5,405 and 882 patches of size 80×80 for training and validation, respectively).

SR results on SyntheticBurst dataset for $\times 4$ and $\times 8$. The proposed GMTNet is trained for 300 epochs on the training split of SyntheticBurst dataset for both $\times 4$, and $\times 8$ upsampling tasks and evaluated on the validation set of SyntheticBurst dataset [2]. We compared our proposed GMTNet with several SoTA approaches for $\times 4$ as shown in Table 1. Particularly, our GMTNet obtains a PSNR gain of about 0.43 dB over the previously best-performing BIPNet [17] and 0.80 dB over the second-best approach [3]. To further prove the potency of our proposed GMTNet on large scale factors, we conduct an experiment for $\times 8$ burst SR. The LR-HR pairs are synthetically generated using the same procedure described for SyntheticBurst dataset [2]. Visual results shown for a few challenging images in Figure 4 ($\times 4$) and Figure 5 ($\times 8$) clearly prove that results obtained by GMTNet are sharper and it efficiently reconstructs the structural content and fine textures, without compromising details. In Table 1, we also compare the computational complexity of several state-of-the-art burst SR methods.

Base frame BIPNet [17] Ours Ground Truth

Figure 5. Visual results on SyntheticBurst [2] for $\times 8$ burst SR.



HR Image Base frame DBSR [2] MFIR [3] BIPNet [17] Ours GT-Image

Figure 6. Results on real BurstSR dataset [2] for $\times 4$ burst SR.

SR results on BurstSR dataset. Since, the LR-HR pairs for BurstSR dataset are captured using different cameras, they suffer from minor misalignment. Thus we follow the previous work [2] and use aligned L_1 loss for fine-tuning the GMTNet for 25 epochs and evaluate our model by using aligned PSNR/SSIM. Table 1 shows that our proposed GMTNet obtain conducive results, outperforming SoTA BIPNet [17] by a substantial gain of 0.46 dB. Visual comparisons in Figure 6 depict that unlike other compared methods, the proposed GMTNet is more effective for generating minute details in the reconstructed images, with better color and structure preservation.

4.2. Burst Denoising Results

This section presents the results of burst denoising on color (test split: 100 bursts) [45] as well as gray-scale (test split: 73 bursts) [36] datasets. Both these datasets have four 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. We train grayscale and color burst denoising models for 200

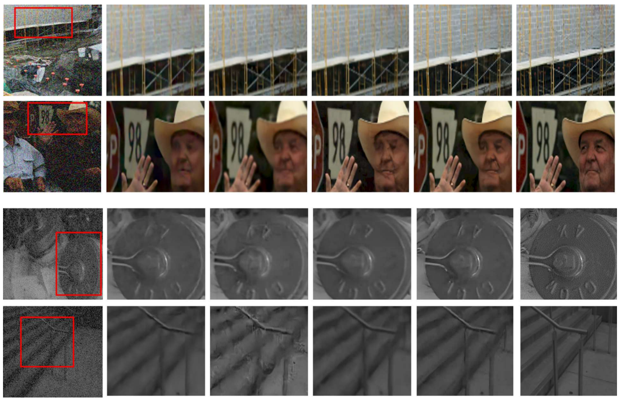


Figure 7. Visual results on color datasets [45] (first two rows) and gray-scale [36] (last two rows) for burst denoising.

epochs on 20k synthetic noisy samples (generated as in [3]).

Table 2. **Gray-scale burst denoising** [36] results with PSNR.

Methods	Gain $\propto$ 1	Gain $\propto$ 2	Gain $\propto$ 4	Gain $\propto$ 8	Average
KPN [36]	36.47	33.93	31.19	27.97	32.19
BPN [45]	38.18	35.42	32.54	29.45	33.90
BIPNet [17]	38.53	35.94	33.08	29.89	34.36
MFIR [2]	**39.37**	**36.51**	33.38	29.69	34.74
Ours	39.07	36.46	**33.52**	**30.46**	**34.87**

Denoising results. Table 2 shows the results on the gray-scale burst denoising dataset against SoTA methods. Our GMTNet outperforms the recent BIPNet[1] [17] by about 0.57 dB for the highest noise gain (Gain $\propto$ 8). Similarly, for color denoising, our approach outperforms existing MFIR [2] on all four noise levels (except the lowest noise gain) with an average margin of 0.25 dB as shown in Table 3. Qualitative comparison in Figure 7 clearly proves the efficacy of our approach in recovering the required subtle contextual details, thus generating cleaner denoised outputs.

4.3. Low-Light Enhancement Results

Following other existing works [17, 26], we test the performance of our GMTNet on the SONY-subset from the SID dataset [8]. It contains 161 input RAW burst sequences for training, 36 for validation, and 93 for testing. We train the proposed GMTNet with L_1 loss for 200 epochs on 5000 cropped patches of size 256×256 from the training set of SONY-subset. Table 4 gives the image quality scores for several competing approaches. The proposed GMTNet provides 0.26 dB improvement over the existing best BIPNet [17]. Visual comparisons in Figure 8 show that the enhanced images are relatively cleaner, sharper and preserves more structural content than other compared approaches.

[1]Existing BIPNet results are collected from their official GitHub repository.

Table 3. **Color burst denoising** [45] results with PSNR.

Methods	Gain $\propto$ 1	Gain $\propto$ 2	Gain $\propto$ 4	Gain $\propto$ 8	Average
KPN [36]	38.86	35.97	32.79	30.01	34.40
BPN [45]	40.16	37.08	33.81	31.19	35.56
BIPNet [17]	40.58	38.13	35.30	32.87	36.72
MFIR [2]	**41.90**	38.85	35.48	32.29	37.13
Ours	41.74	**38.91**	**35.74**	**33.09**	**37.38**

Figure 8. Visual results on SONY-subset of SID dataset [8] for burst low-light image enhancement.

Table 4. **Burst low-light enhancement** on Sony-subset [8].

Methods	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓
Chen *et al.* [8]	29.38	0.89	0.48
Maharjan *et al.* [33]	29.57	0.89	0.48
Zamir *et al.* [48]	29.13	0.88	0.46
Zhao *et al.* [51]	29.49	0.89	0.45
Karadeniz *et al.* [26]	29.80	0.89	0.30
BIPNet [17]	32.87	0.93	**0.30**
Ours	**33.13**	**0.94**	0.31

Table 5. Ablation study for GMTNet contributions. PSNR is reported on SyntheticBurst dataset [2] for ×4 burst SR task.

Task	**Modules**								
	Baseline	✓	✓	✓	✓	✓	✓	✓	✓
Align ment (§3.1)	w/O MKGA		✓						
	with MKGA			✓	✓	✓	✓	✓	✓
	AFE				✓	✓	✓	✓	✓
Fusion (§3.2)	with p_1					✓			
	with p_2						✓		
	with p_1+p_2							✓	✓
Upsample (§3.3)									✓
	PSNR	36.38	38.02	39.12	39.40	39.84	40.23	40.74	**41.82**

5. Ablation Study

Here we analyze the influence of every key component and design choice in our formulation. All models are trained for 100 epochs on SyntheticBurst dataset [2] for ×4 burst SR task. As reported in Table 5, the baseline model achieves a PSNR of 36.38 dB. For the baseline model we

Table 6. Impact of the proposed modules in terms of PSNR/SSIM on SyntheticBurst SR dataset for ×4 burst SR task.

Task	Methods	PSNR↑	SSIM↑
(a) Alignment	GMTNet + PCD [44]	40.99	0.953
	GMTNet + Explicit [2]	39.26	0.944
	GMTNet + EBFA [17]	41.10	0.958
	GMTNet + MBFA	**41.82**	**0.960**
(b) Burst Fusion	GMTNet + TSA [44]	39.97	0.947
	GMTNet + DBSR [21]	40.32	0.950
	GMTNet + PBFF [17]	41.60	0.954
	GMTNet + TAFM	**41.82**	**0.960**
(c) Upsampler	GMTNet + Bil	40.22	0.940
	GMTNet + PS [21]	40.41	0.943
	GMTNet + AGU [17]	41.30	0.951
	GMTNet + RTFU	**41.82**	**0.960**

deploy addition operation for fusion and pixel-shuffle for up-sampling. After adding the proposed modules to the baseline network, the results improve persistently and notably. For instance, we attain a performance gain of 3.02 dB when we incorporate our alignment module into the baseline model. The insertion of the proposed fusion and up-sampling modules in our network further improves the PSNR of the overall network by about 1.34 dB and 1.08 dB, respectively. Overall, GMTNet obtains a compelling gain of 5.44 dB over the baseline model.

Effectiveness of MBFA module. As reported in Table 5, the inclusion of MKGA and AFE modules into our alignment (MBFA) module provides a performance boost of around 1.10 dB and 0.28 dB, respectively which supports the effectiveness of the proposed modules in capturing motion cues. Further, we compare the GMTNet results in Table 6 (a) by replacing MBFA with other popular explicit and implicit alignment approaches *(Keeping the rest of the modules same)*. We observe that the MBFA module obtains a performance gain of about 0.83 dB over PCD module proposed in EDVR [44]. To further highlight the ability of MBFA module in aligning burst features, we visualize the features (of few frames) before and after applying it as shown in Figure 9. It clearly reveals our MBFA works well without any dedicated supervision.

How to design TAFM module? A trivial design of our TAFM module is to use a single stream for extracting the information and then concatenating the features. However, from Table 5, it is clear that utilizing both the p_1 and p_2 outputs for subsequent merging results in a performance boost of around 0.90 dB. It clearly signifies that two-stream TAFM performs better than any single-stream.

Impact of TAFM module. The results in Table 6 for burst fusion tasks further show that replacing our TAFM module with other popular fusion modules have a detrimental influence on the overall performance of our model, with PSNR drop of around 0.22 dB when utilizing the recently proposed PBFF [17] module in our network.

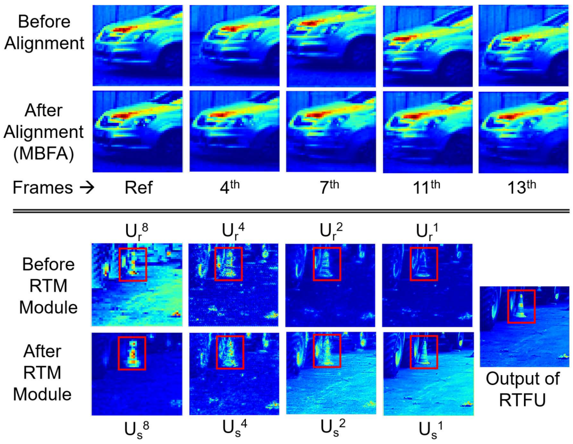


Figure 9. Feature map visualizations before and after applying proposed MBFA (Figure 2(b)) and RTM (middle of Figure 2(f)) modules into our GMTNet.

Effectiveness of the proposed RTFU. To validate the effectiveness of our RTFU, we replace it with the conventional and recent, bilinear interpolation (Bil) and pixel-shuffle (PS), AGU respectively. The accuracy scores in Table 6, clearly demonstrate its ability to reconstruct a high-quality image.

How important is the proposed RTM module in RTFU? To prove the imperativeness of the RTM module, in Figure 9 we visualize the feature maps before and after embedding it in RTFU. It clearly proves that our model benefits from the efficient use of both LR and HR information to complete the restoration of sharp regions.

6. Conclusion

We present a generalised network for burst processing to reconstruct a single high-quality image from a given RAW burst of low-quality noisy images. In the proposed approach, our Multi-scale Burst Feature Alignment (MBFA) module aligns the noisy burst features at multiple scales using the proposed Attention-Guided Deformable Alignment (AGDA). The inclusion of Aligned Feature Enrichment (AFE) module improves the aligned features by fixing any minor misalignment issue, thus yielding well-refined, denoised and aligned features. To further improve model robustness, Transposed Attention Feature Merging (TAFM) module manifests efficient fusion performance by analyzing the global and local correlations among the incoming frames. Finally, the proposed Resolution Transfer Feature Up-sampler (RTFU) up-scales the merged features by consolidating information from both LR and HR feature spaces to reconstruct a high-quality image. Consistent achievement of SoTA results for burst super-resolution, denoising and low-light enhancement on synthetic and real datasets corroborates the robustness and potency of our approach.

References

[1] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*, pages 571–582. Springer, 1996.

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021.

[3] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021.

[4] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *2012 IEEE conference on computer vision and pattern recognition*, pages 2392–2399. IEEE, 2012.

[5] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[6] Ralph E Carlson and Frederick N Fritsch. Monotone piecewise bicubic interpolation. *SIAM journal on numerical analysis*, 22(2):386–400, 1985.

[7] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3185–3194, 2019.

[8] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018.

[9] Xinyuan Chen, Li Song, and Xiaokang Yang. Deep rnns for video denoising. In *Applications of digital image processing XXXIX*, volume 9971, pages 573–582. SPIE, 2016.

[10] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016.

[11] Wooyeong Cho, Sanghyeok Son, and Dae-Shik Kim. Weighted multi-kernel prediction network for burst image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 404–413, 2021.

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

[13] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

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

[15] 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 preprint arXiv:2002.06460*, 2020.

[16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[17] 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.

[18] Michael Elad and Yacov Hel-Or. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Transactions on image Processing*, 10(8):1187–1193, 2001.

[19] Esmaeil Faramarzi, Dinesh Rajan, and Marc P Christensen. Unified blind method for multi-image super-resolution and single/multi-image blur deconvolution. *IEEE Transactions on Image Processing*, 22(6):2101–2114, 2013.

[20] Sina Farsiu, Michael Elad, and Peyman Milanfar. Multi-frame demosaicing and super-resolution from undersampled color images. In *Computational Imaging II*, volume 5299, pages 222–233. SPIE, 2004.

[21] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *Proceedings of the European conference on computer vision (ECCV)*, pages 538–554, 2018.

[22] Yu He, Kim-Hui Yap, Li Chen, and Lap-Pui Chau. A nonlinear least square technique for simultaneous image registration and super-resolution. *IEEE Transactions on Image Processing*, 16(11):2830–2841, 2007.

[23] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991.

[24] Amod Jog, Aaron Carass, and Jerry L Prince. Self super-resolution for magnetic resonance images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–560. Springer, 2016.

[25] Eunah Jung, Nan Yang, and Daniel Cremers. Multi-frame gan: Image enhancement for stereo visual odometry in low light. In *Conference on Robot Learning*, pages 651–660. PMLR, 2020.

[26] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *IEEE Transactions on Image Processing*, 30:9372–9385, 2021.

[27] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

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

[29] Lin Liu, Junfeng An, Jianzhuang Liu, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, Yanfeng Wang,

and Qi Tian. Low-light video enhancement with synthetic event guidance. *arXiv preprint arXiv:2208.11014*, 2022.

[30] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM Transactions on Graphics (TOG)*, 33(6):1–9, 2014.

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

[32] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *IEEE Transactions on image processing*, 21(9):3952–3966, 2012.

[33] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 916–921. IEEE, 2019.

[34] 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.

[35] Nancy Mehta and Subrahmanyam Murala. Con-net: A consolidated light-weight image restoration network. *IEEE Transactions on Broadcasting*, 68(4):862–875, 2022.

[36] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018.

[37] Shmuel Peleg, Danny Keren, and Limor Schweitzer. Improving image resolution using subpixel motion. *Pattern recognition letters*, 5(3):223–226, 1987.

[38] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. *arXiv preprint arXiv:2207.08494*, 2022.

[39] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[40] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020.

[41] Marius Tico. Multi-frame image denoising and stabilization. In *2008 16th European Signal Processing Conference*, pages 1–4. IEEE, 2008.

[42] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[45] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11844–11853, 2020.

[46] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25, 2012.

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021.

[48] Syed Waqas Zamir, Aditya Arora, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Learning digital camera pipeline for extreme low-light imaging. *Neurocomputing*, 452:37–47, 2021.

[49] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019.

[50] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–284, 2018.

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

[52] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.