Learnable Global Spatio-Temporal Adaptive Aggregation for Bracketing Image Restoration and Enhancement

Xinwei Dai[1], Yuanbo Zhou[1], Xintao Qiu[1], Hui Tang[1], Wei Deng[2],
Qingquan Gao[1,2], Tong Tong[1,2,*]
[1]Fuzhou University [2]Imperial Vision Technology
xinweidai@163.com, ttraveltong@gmail.com

Abstract

*Employing specific networks to address different types of degradation often proved to be complex and time-consuming in practical applications. The **B**racket **I**mage **R**estoration and **E**nhancement (BIRE) aimed to address various image restoration tasks in a unified manner by restoring clear single-frame images from multiple-frame shots, including denoising, deblurring, enhancing high dynamic range (HDR), and achieving super-resolution under various degradation conditions. In this paper, we propose LGSTANet, an efficient aggregation restoration network for BIRE. Specifically, inspired by video restoration methods, we adopt an efficient architecture comprising alignment, aggregation, and reconstruction. Additionally, we introduce a **L**earnable **G**lobal **S**patio-**T**emporal **A**daptive (LGSTA) aggregation module to effectively aggregate inter-frame complementary information. Furthermore, we propose an adaptive restoration modulator to address specific degradation disturbances of various types, thereby achieving high-quality restoration outcomes. Extensive experiments demonstrate the effectiveness of our method. LGSTANet outperforms other state-of-the-art methods in Bracket Image Restoration and Enhancement and achieves competitive results in the NTIRE2024 BIRE challenge.*

1. Introduction

Smartphone and camera manufacturers have always strived to capture clear photos under low-light conditions. While long exposures could enhance brightness in photography, they might also result in motion blur or overexposure due to camera shake or subject motion. Conversely, short exposures may result in the camera capturing a limited amount of photons, causing noise in the images and rendering dark areas invisible. Moreover, the use of lower-quality photography equipment often results in issues such as low resolution and loss of detail. To enhance photography quality, there is a desire to obtain high dynamic range (HDR) images. Despite extensive research on single-image restoration methods, such as deblurring [18, 35], denoising [53, 54], super-resolution [9, 39], and HDR reconstruction [24, 46] for the aforementioned issues, these methods are limited when faced with the simultaneous existence of multiple degradation types in real-world scenarios, as illustrated in Fig. 1(a).

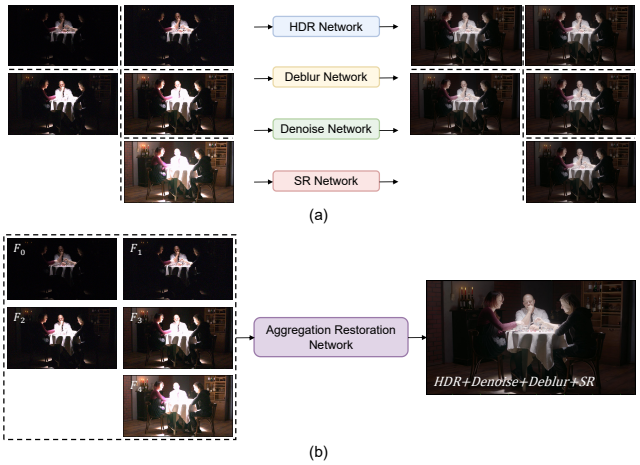

Figure 1. (a) indicated the utilization of a dedicated network to process a single image with a specific type of degradation. (b) indicated the usage of an aggregation restoration network to restore and enhance multiple degraded multi-frame shots.

Recently, there has been significant attention on restoring and enhancing multi-frame images [1, 2, 10, 11, 23, 29, 36, 46]. By using images of the same scene captured with varying exposure times, denoising and super-resolution were conducted, or the complementary benefits of long and short exposures were combined to improve deblurring effects. Additionally, HDR reconstruction typically necessitates multiple exposure support. Inspired by these processing paradigms, the task of restoring and enhancing bracket images was proposed, aiming to address practical issues such as motion blur, noise, low resolution,

*Corresponding author

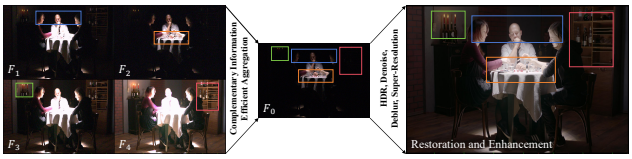

Figure 2. An example of a multi-frame burst image and the restored image. In the multiple frames, corresponding regions contained complementary information to aid in the restoration and enhancement of the reference frame.

and low dynamic range (LDR) simultaneously in a unified manner. High-quality images were restored by eliminating various types of degradation from multiple frames. Compared to the previous task, this one was more closely aligned with real-world applications, but it was also more challenging due to the requirement to handle multiple types of degradation. Fortunately, having more input frames provided a wealth of information for restoration. As shown in Fig. 2, we observed complementary information among the input multiple frames. For example, images captured with short exposure were clear, assisting in restoring blurry regions in long-exposure images. Moreover, the content in long-exposure shots was clearer compared to the darker areas in short-exposure images. Additionally, combining multiple exposure images facilitated HDR reconstruction, enabling better preservation of details compared to single-image processing. Therefore, the key to the BIRE task lies in effectively and efficiently utilizing the complementary information present in these inter-frame corresponding regions to achieve high-quality image restoration.

To tackle the aforementioned issues, we introduced LGSTANet. Inspired by tasks like video super-resolution [4, 5] and video deblurring [57], which aim to utilize inter-frame information for video restoration and enhancement. Our method adopted a similar architecture, incorporating alignment, aggregation, and reconstruction. Regarding alignment, considering the varying exposure levels across multiple frames resulting in variations in color and brightness, we initially performed color alignment. Acknowledging potential motion misalignment between frames, especially in long exposures, we also conducted alignment in the feature space, using a flow-guided deformable convolution paradigm. Regarding aggregation, through a recurrent structure, we propagated information from the current frame to the next for fusion, achieving implicit feature aggregation. In this recurrent structure, we utilized both weight-sharing and non-weight-sharing residual blocks to mitigate degradation across multiple types of frames.

Considering the BIRE task entails multiple-frame inputs and single-frame outputs, effectively aggregating multiple frames is crucial. Notably, not all regions in non-reference frames contribute to the reference frame, and merely concatenating them may introduce additional degradation noise. Hence, we introduced a trainable global spatio-temporal adaptive aggregation module, utilizing global pooling and nonlinear activation functions to prioritize valuable regions for reference frame aggregation. Moreover, considering the diverse degradation types present in input frames (blur, noise, low light, overexposure), and the fact that different degradation types entail different degradation disturbances [42] that affect the quality of recovery. Therefore, following each layer's aggregation with the reference frame, we introduce a non-weight-shared learnable tensor to adapt to these variations, thereby achieving more detailed texture restoration.

Based on the above design, our method utilized a pure CNN architecture. Compared to methods [11, 36] utilizing complex structures or self-attention and cross-attention mechanisms for spatio-temporal interaction, our approach exhibited greater model efficiency. This makes our method more suitable for deployment on edge devices in real-world scenarios.

The main contributions of this work are summarized below:

- We proposed LGSTANet, an aggregated restoration network for aggregating inter-frame complementary information, which proved to be efficient and effective.
- We introduced a Learnable Global Spatio-Temporal Adaptive (LGSTA) aggregation module, utilized to efficiently extract inter-frame complementary regions and aggregate them.
- We proposed an ultra-lightweight adaptive restoration modulator capable of adapting to various types of degradation disturbances, significantly enhancing the recovery quality with its straightforward design.
- Extensive experiments demonstrated the effectiveness of LGSTANet. Our algorithm ranked 5th and 3rd in the NTIRE 2024 Bracket Image Restoration and Enhancement Challenge's two tracks, respectively. In track 2, our model's inference speed reached the **SOTA** level.

2. Related Work

2.1. Single Image Restoration and Enhancement

Image restoration and enhancement are among the classical tasks in computer vision, aimed at recovering clear images from degraded ones, such as denoise [53, 54], deblur [18, 35], derain [15, 52], dehaze [12, 33], desnow [7, 22], and super-resolution [9, 39]. Traditional methods typically depended on hand-crafted features, assumptions, or statistical priors to constrain the solution space [14]. However, while these methods excelled on specific datasets, they demonstrated limited generalization and generalizability. In recent years, the rapid development of deep learning has led to the emergence of numerous convolutional neural network-based methods for single-image restoration

and enhancement. These methods have comprehensively outperformed traditional algorithms [6, 50] across multiple tasks, achieving remarkable results. Additionally, the utilization of Transformer models for global modeling in low-level vision tasks [27, 34] demonstrates impressive capabilities, leveraging the self-attention mechanism to attain global dependencies, thereby outperforming convolutional neural network-based approaches. To solve the quadratic computational complexity problem of self-attention, some methods optimized the computation of self-attention and improved the computational efficiency [19, 42, 51]. However, relying solely on information from a single image may not fully harness the potential of restoration methods.

2.2. Multi-Frame Image Reconstruction

Unlike single-image processing, multi-frame image restoration and enhancement require consideration of misalignment and motion blur caused by camera and object movement, as well as noise introduced by differences in device parameter settings. The advantage was that inter-frame information could be utilized to achieve better restoration than single-image processing. MFSR [40] pioneered the processing of burst frames in the frequency domain, which produced significant artifacts despite its computational efficiency. DBSR [1] addressed the MFSR problem by employing attention-centered explicit feature alignment. Some methods achieved better restoration by exploring the complementary information of short exposure noise and long exposure blur. AHDRNet [47] proposed an attention-guided end-to-end deep neural network, to produce high-quality ghost-free HDR images. Kalantari *et al*. [16] explored frame alignment and merging of images with optical flow. Wu *et al*. [44] proposed free optical flow HDR image reconstruction of large motion scenes. HDR-Transformer [24] and SCTNet [37] utilized self-attention and cross-attention for context and inter-frame interaction. Besides, a few methods [1, 10, 13, 28, 49] take noise into account. Although the above method achieved decent results, it only considered one or two types of degradation, and employing self-attention mechanisms often imposed inappropriate computational burdens for practical application scenarios.

2.3. Video Image Restoration and Enhancement

The video restoration and enhancement task was also about dealing with multi-frame degraded image inputs, with the key being to effectively utilize inter-frame information to recover a clear image. Deep learning-based video restoration methods mainly included time-sliding window-based and recurrent-based methods [3, 20, 21, 32, 41]. In the temporal sliding window method, given an LR video sequence, reference frames and neighboring frames were aligned to estimate a single HR output. The alignment module played a key role in this process. Previous methods [8, 38, 41] used deformable convolution and optical flow for alignment, while more recent methods [45] proposed implicit alignment modules to perform alignment in high-dimensional feature space. Some methods [48, 58] also address image alignment by designing complex network structures. On the other hand, methods based on recurrent structures propagated features from the current frame to the next, achieving the effect of implicit aggregation. BasicVSR [4] and BasicVSR++ [5] proposed a VSR method that combined bidirectional propagation of past and future frames into the features of the current frame, achieving a significant improvement. ESTRNN [57] combined a dense residual block with a recurrent structure and efficiently fused past and future frames to achieve an efficient video deblurring method. The aforementioned methods for video restoration and enhancement have inspired the BIRE task, which also involves processing multiple frames of images.

3. Method

In this section, we detailed our method. We will first provide an overview of our method in Sec. 3.1. Then, we will introduce the processing method for alignment in Sec. 3.2, and our proposed learnable global spatio-temporal adaptive (LGSTA) aggregation module in Sec. 3.3. **The main idea of this paper** is to effectively aggregate burst frames with complementary information to achieve high-quality bracket image restoration and enhancement.

3.1. Overall Framework

We proposed LGSTANet, depicted in Fig. 3, primarily comprising alignment, aggregation, and reconstruction components. The alignment part included color alignment and feature space alignment. The aggregation part consisted of a recurrent network and a learnable aggregation module. In track 2, the reconstruction part involved $4\times$ super-resolution. Initially, given multiple degraded frames $F_i \in \mathbb{R}^{H\times W\times 4}$, F_0, F_1, F_2, F_3, and F_4 underwent color alignment to obtain normalized features $\widetilde{F}_{ci} \in \mathbb{R}^{H\times W\times 8}$. Subsequently, shallow features were extracted using shared-weight 3×3 convolutions, followed by alignment in feature space to obtain aligned features $\widetilde{F}_1$, $\widetilde{F}_2$, $\widetilde{F}_3$, $\widetilde{F}_4$ relative to F_b. Subsequently, $\widetilde{F}_i \in \mathbb{R}^{H\times W\times C}$ was fused with F_b and fed into the recurrent network, which produced outputs F_{T0}, F_{T1}, F_{T2}, F_{T3}, F_{T4} at each temporal layer. These outputs were then fused into F_R using the learnable aggregation module. Finally, they enter the reconstruction module stacked with residual blocks, where the output $F_o \in \mathbb{R}^{H\times W\times C}$ is added to F_b connected via skip connections. Subsequently, channel reduction is achieved using a 1×1 convolution, followed by upsampling, and then added to the result of the same upsampling operation on F_0. For track 1, the reconstruction module maintained the resolu-

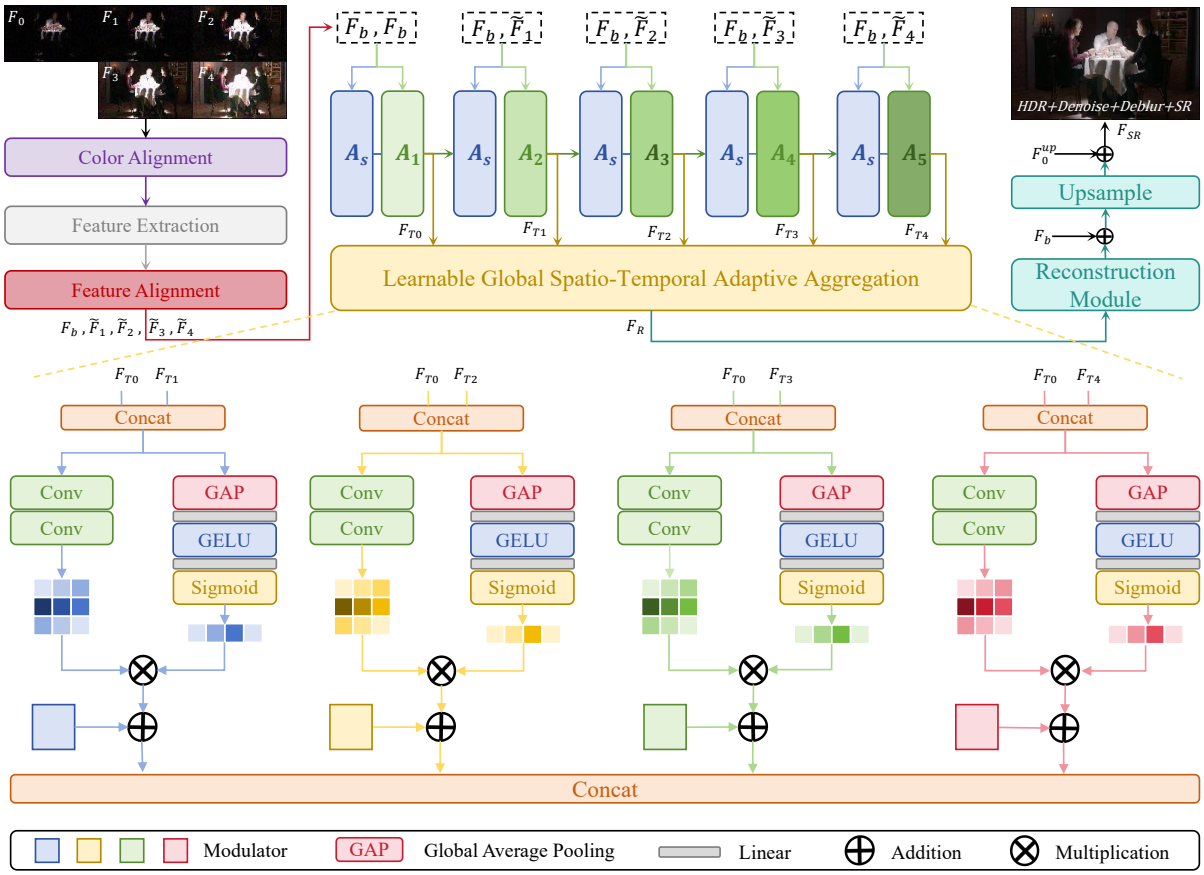

Figure 3. Architecture of LGSTANet for Bracketing Image Restoration and Enhancement. Our LGSTANet consists of an alignment module, an implicitly aggregated part based on recurrent structures, an explicitly aggregated part with a Learnable Global Spatio-Temporal Adaptive (LGSTA) aggregation module, and a restoration and super-resolution reconstruction part.

tion $F_{SR} \in \mathbb{R}^{H\times W\times 4}$, whereas for track 2, it employs PixelShuffle for $4\times$ super-resolution reconstruction, resulting in $F_{SR} \in \mathbb{R}^{4H\times 4W\times 4}$.

3.2. Alignment Module

Effective alignment is crucial for aggregating multiple frames effectively. Various alignment methods have been widely used in video restoration [5, 20, 41]. Considering the brightness and contrast changes caused by different exposure times, the potential color variations between frames might affect the subsequent aggregation and restoration results. Therefore, color alignment was conducted as the initial step. Specifically, we reshape the input tensor F_i into a four-dimensional tensor F_{ci}, with a shape of $(n \times t, 2c, h, w)$. Then, we create a zero tensor $\widetilde{F}_{ci}$ with the same data type as F_{ci}, with a shape of $(n \times t, 2c, h, w)$. Next, we fill the even channel positions of $\widetilde{F}_{ci}$ with the corresponding channel values from F_{ci}:

$$\widetilde{F}_{ci}\left[:, 0::2, :, :\right] = F_{ci}. \tag{1}$$

Finally, we perform clamp and gamma correction on F_{ci}, and fill the results into the odd channel positions of $\widetilde{F}_{ci}$:

$$\widetilde{F}_{ci}\left[:, 1::2, :, :\right] = \left(clamp\left(F_{ci}, min = 0\right)\right)^{1/\gamma}, \tag{2}$$

where γ represents the gamma correction parameter and is generally set to 2.2. Thus, we have completed the transformation from the input tensor $F_i \in \mathbb{R}^{H\times W\times 4}$ to the output tensor $\widetilde{F}_{ci} \in \mathbb{R}^{H\times W\times 8}$.

The feature extraction module, composed of stacked residual blocks with shared weights, mapped the image to the feature space. Further inspired by the classic video super-resolution algorithm BasicVSR++ [5], we also adopted the approach of feature space alignment through flow-guided deformable convolutions. Specifically, by using SPyNet [31] to compute the optical flow s_i between the reference frame F_0 and the frames to be aligned $[F_{s1}, F_{s2}, F_{s3}, F_{s4}]$, we warp s_i with the frames to be aligned:

$$\widetilde{F}_{Pi} = \mathcal{W}\left(F_{si}, s_i\right), \tag{3}$$

where $\mathcal{W}$ denotes the spatial warping operation. We compute the residue to the optical flow. The pre-aligned fea-

Table 1. Quantitative comparison with state-of-the-art methods on the synthetic of BracketIRE and BracketIRE+ tasks, respectively.

	Method	BracketIRE			BracketIRE+		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Burst Processing Networks	DBSR [1]	34.05	0.9026	0.187	28.22	0.8359	0.336
	MFIR [2]	33.92	0.9026	0.196	28.33	0.8381	0.330
	BIPNet [10]	35.97	0.9314	0.145	28.44	0.8453	0.311
	Burstormer [11]	37.01	0.9454	0.127	28.59	0.8516	0.292
	RBSR [43]	37.88	0.9453	0.119	28.70	0.8549	0.279
HDR Reconstruction Networks	AHDRNet [46]	36.32	0.9273	0.154	28.17	0.8422	0.309
	HDRGAN [29]	35.07	0.9157	0.177	27.80	0.8359	0.342
	HDR-Tran. [23]	36.54	0.9341	0.127	28.18	0.8483	0.282
	SCTNet [36]	36.90	0.9437	0.120	28.28	0.8461	0.282
	Kim *et al*. [17]	37.93	0.9452	0.115	28.33	0.8494	0.270
Ours	TMRNet [55]	38.19	0.9488	0.112	28.91	0.8572	0.273
	LGSTANet	38.46	0.9527	0.105	29.82	0.8537	0.282

tures $\widetilde{F}_{Pi}$ are then used to compute the DCN [8] offsets o_i and modulation masks m_i. A DCN is then applied to the unwarped feature $\widetilde{F}_i$:

$$\widetilde{F}_i = \mathcal{D}\left(F_{s_i}; o_i, m_i\right), \tag{4}$$

where $\mathcal{D}$ denotes a deformable convolution. It is worth noting that, unlike BasicVSR++ [5], we aligned only the residual frames with the reference frame, without performing inter-frame alignment, thereby improving computational efficiency.

3.3. Aggregatiton Structure

Following multiple-frame alignment, the recurrent aggregation method [4, 5, 57] has been extensively utilized in video restoration. This method recursively propagated features extracted from the current frame to the next frame, thereby achieving implicit feature fusion. We adopted a similar recurrent structure for aggregation and continued the design approach from previous work [55]. We alternately stacked weight-shared and non-weight-shared residual blocks into the recurrent architecture to address multiple-frame degradation of the same type and different types of degradation:

$$\begin{aligned} F_{l_s} &= \mathcal{A}_s\left(concat\left(F_0, \widetilde{F}_i\right), F_{T_i}\right), \\ F_{T_i} &= \mathcal{A}_i\left(concat\left(F_0, \widetilde{F}_i\right), F_{l_s}\right), \end{aligned} \tag{5}$$

where $\mathcal{A}_s$ denotes weight-shared, $\mathcal{A}_i$ denotes non-weight-shared. F_{l_s} represents the features processed by convolutions with shared weights, while F_{T_i} represents the features processed by convolutions with unshared weights. For the first F_{T_0}, we set it to zero.

In previous work [4, 5], only the final output of the recurrent structure was used for reconstruction. Due to the presence of various degradation types in multi-frame images, and not all regions containing complementary information—sometimes even noise, directly aggregating may result in performance degradation. Therefore, we proposed a Learnable Global Spatio-Temporal Adaptive (LGSTA) aggregation module, which aims to explicitly aggregate complementary information from adjacent frames to fully leverage it for high-quality bracket image restoration and enhancement. Specifically, before propagating features to the next frame in the recurrent structure, we stored the current feature in the aggregate list awaiting processing. After the recurrent structure propagation was completed, each frame in the aggregate list underwent effective alignment and preliminary restoration. Take the first frame in the list as the reference frame, and then aggregate the other four frames as supplementary frames with the reference frame. First, use the concat operation to fuse the reference frame with each supplementary frame separately. On one hand, extract features from the fusion of the two using convolution. On the other hand, utilize global average pooling and activation functions to filter out hierarchical features from complementary regions. Then, multiply the two to obtain efficient aggregated features from effective complementary regions:

$$F_{c_i} = concat\left(F_{T_0}, F_{T_i}\right), \tag{6}$$

$$\widetilde{F}_{c_i} = \mathcal{L}\left(GAP\left(F_{c_i}\right)\right) \otimes Conv\left(F_{c_i}\right), \tag{7}$$

where $\mathcal{L}$ denotes a series of linear transformations with activation function and Sigmoid function for channel weight generation. After the above processing, four frames have been effectively aggregated.

Due to the different types of degradation present in the input frames (such as blur, noise, low light, etc.), each type of image degradation has its distinctive perturbation pattern [42] that needs to be addressed or restored. To further enhance the ability of our method to handle various disturbances, we proposed a lightweight restoration modulator to

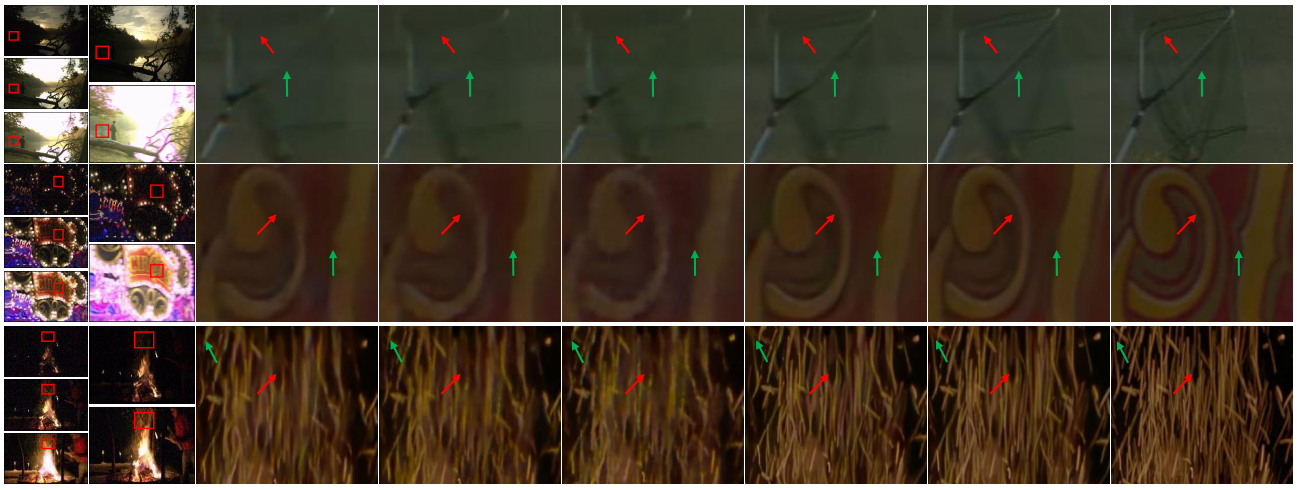

Figure 4. Visual comparison on the synthetic dataset of BracketIRE task. Please zoom in for more details.

Figure 5. Compared the input and output results of our method, as well as the ground truth. Please zoom in for more details.

adaptively process the previously obtained frames:

$$F_{A_i} = M_{T_i} \oplus \widetilde{F}_{c_i}, \tag{8}$$

where M_{T_i} denotes learnable tensor. By adapting to different degradation disturbances, we aim to restore more details. We deployed four modulators for these four frames with minimal additional parameters. The results show that our adaptive restoration modulator indeed contributes to recovering details with minimal computational cost.

4. Experiments

4.1. Datasets

We used the dataset provided by the NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge organizers: BracketIRE [55] and BracketIRE+ [55]. The dataset includes 1,335 data pairs in 35 scenes. 1,045 pairs from 31 scenes were used for training, and the remaining 290 pairs from the other 4 scenes were used for testing. Setting the exposure time ratio S to 4 and the frame number T to 5 covers most of the dynamic range with fewer images. For Track 1, both LR and GT have a resolution of 1920×1080. For Track 2, which includes $\times 2$ and $\times 4$ super-resolution tasks, LR resolutions are 960×540 and 480×270 respectively, while GT resolutions remain at 1920×1080. The proposed BracketIRE involves denoising, deblurring, and HDR reconstruction, while BracketIRE+ adds support for SR task.

4.2. Implemental Details

We randomly crop patches and augment them with flips and rotations. The batch size is set to 8. We adopt AdamW [26] optimizer with $\beta_1 = 0.9$ and $\beta_1 = 0.999$. The initial learning rate was set to 1e-4. Cosine annealing strategy [25] is employed to decrease the learning rates to 1e-6. All experiments are conducted with PyTorch [30] on a single Nvidia RTX 3090 GPU.

Track 1, we used the progressive training strategy [51] to increase patch size and reduce batch size. The patch size of the training includes [128, 160, 192, 256, 320, 384], and a total of 800 epochs were trained, and only L1 loss was used for supervision and optimization in the whole training process.

Track 2, We used the best weights from Track1 and the $2\times$ resolution data provided in BracketIRE+ for training $2\times$

Table 2. Results on **track 1** of on Bracketing Image Restoration and Enhancement Challenge [56].

Rank	Team	Full Images PSNR ↑ / SSIM↑ / LPIPS↓	Cropped Images PSNR ↑ / SSIM↑ / LPIPS↓	#Params (M)	#FLOPs (T)	Time (s)	Memory (GB)
1	SRC-B	40.54 / 0.9637 / 0.077	41.77 / 0.9633 / 0.076	94.34	48.238	3.102	20
2	MegIRE	39.78 / 0.9556 / 0.102	39.82 / 0.9550 / 0.105	19.75	30.751	$2.383_{(3)}$	$16_{(3)}$
3	UPN1	39.03 / 0.9500 / 0.117	39.02 / 0.9493 / 0.120	$13.32_{(2)}$	$10.409_{(1)}$	$1.090_{(1)}$	$6_{(1)}$
4	CVG	38.78 / 0.9543 / 0.102	39.89 / 0.9557 / 0.104	$13.29_{(1)}$	$21.340_{(2)}$	7.518	$11_{(2)}$
5	Ours	38.46 / 0.9527 / 0.105	39.61 / 0.9540 / 0.107	$14.04_{(3)}$	$22.283_{(3)}$	$1.829_{(2)}$	$16_{(3)}$

Table 3. Results on **track 2** of on Bracketing Image Restoration and Enhancement Challenge [56].

Rank	Team	Full Images PSNR ↑ / SSIM↑ / LPIPS↓	Cropped Images PSNR ↑ / SSIM↑ / LPIPS↓	#Params (M)	#FLOPs (T)	Time (s)	Memory (GB)
1	SRC-B	34.26 / 0.8913 / 0.206	34.80 / 0.8913 / 0.208	95.00	5.285	0.813	$5_{(3)}$
2	NWPU	30.59 / 0.8728 / 0.268	29.93 / 0.8633 / 0.274	$13.37_{(1)}$	$1.426_{(1)}$	0.887	$3_{(1)}$
3	Ours	29.82 / 0.8537 / 0.282	31.35 / 0.8660 / 0.277	$14.34_{(2)}$	$1.500_{(2)}$	$0.493_{(1)}$	$3_{(1)}$
4	CYD	29.66 / 0.8598 / 0.284	30.27 / 0.8632 / 0.285	$17.60_{(3)}$	$1.560_{(3)}$	$0.751_{(3)}$	$4_{(2)}$
5	CVG	29.25 / 0.8521 / 0.278	30.63 / 0.8645 / 0.275	71.82	6.898	$0.679_{(2)}$	$4_{(2)}$

super-resolution. We trained for 400 epochs and a crop size of 64×64. The pre-trained weights were further used for $4\times$ super-resolution. For Track2, we trained for a total of 801 epochs and only used L1 loss for supervision and optimization throughout the training process. Progressive training strategies and any ensemble operations were not used.

4.3. Result

Evaluation metrics. We utilize PSNR, SSIM and LPIPS to quantitatively evaluate the BIRE performance.

Quantitative Results. Tab. 1 shows the quantitative results of our method and other methods, which include Burst Processing methods DBSR [1], MFIR [2], BIPNet [10], Burstotmer [11] and RBSR [43], HDR Reconstruction methods AHDRNet [46], HDRGAN [29], HDRTran. [23], SCTNet [36] and Kim *et al*. [17], BracketIRE method TMRNet [55]. As shown in Tab. 1, our proposed LGSTANet achieves SOTA results in terms of PSNR on the BracketIRE and BracketIRE+ datasets. Notably, on the BracketIRE+ dataset ($4\times$ super-resolution), our method improves PSNR by 0.91dB compared to the suboptimal method. On the BracketIRE dataset, our method comprehensively surpasses previous state-of-the-art methods in PSNR, SSIM, and LPIPS metrics.

Visual Comparison. Fig. 4 shows that our model has better visual image quality than the other different methods, including the AHDRNet [46], HDRformer [23] and SCTNet [36]. The method we proposed achieves better results in deblurring and denoising. In the third group of images, our results outperform other methods in fine texture restoration. This validates the superiority and effectiveness of our proposed LGSTANet method. Fig. 5 shows the output of our model compared to the input of multi-frame burst images, achieving significant improvements in denoising and debluring, as well as enhancing low-light regions. It can be observed that our results closely approximated the ground truth, further validating the effectiveness of our method.

4.4. NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge

The top 5 results of the NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge Track1 and Track 2 selected by the NTIRE 2024 committee [56] are presented in Tab. 2 and Tab. 3.

Our method participated in two tracks and ranked 5th and 3rd under the PSNR evaluation metric. We didn't use any ensemble operations during the testing period. Our method achieved competitive results in both tracks. Although the final scores are ranked only according to the PSNR metrics, Tab. 2 and Tab. 3 show more metrics on model efficiency, including Params, Inference time, Memory usage, and FLOPs, in which these metrics are commonly meaningful for real-world applications. For track 1, our approach achieves the top three results in each of the metrics for evaluating model efficiency. For track 2, with a more competitive PSNR, our method achieves even better results in terms of model efficiency compared to track 1. In particular, we take the state-of-the-art performance in terms of inference speed. This demonstrates the better applicability of our method in edge devices.

4.5. Ablation Study

To validate the effectiveness of our LGSTANet, we conducted a series of experiments. To enhance training efficiency, we performed ablation experiments using a crop size of 64×64 and trained for 400 epochs on the BracketIRE dataset.

Table 4. Ablation studies of different components in LGSTANet.

Method	PSNR↑	SSIM↑	LPIPS↓	Params
w/o Color Alignment	36.44	0.9337	0.136	14.041M
w/o Feature Alignment	37.75	0.9452	0.120	12.322M
w/o Recurrent Network	36.47	0.9334	0.146	3.332M
w/o LGSTA	37.71	0.9455	0.118	13.286M
Ours	37.95	0.9469	0.117	14.041M

Effectiveness of LGSTANet. To further validate the effectiveness of the architecture proposed, we conducted a series of ablation experiments, the results of which are shown in Tab. 4. We conducted ablation studies on feature alignment and recurrent aggregation components, respectively. The results revealed that when feature alignment or recurrent aggregation was abandoned, the PSNR decreased by 0.2dB and 1.48dB, respectively. This indicates that the recurrent architecture in the video restoration domain was equally significant in the BIRE task. In terms of alignment, the alignment approaches from the perspectives of color space alignment and feature space alignment laid a solid foundation for subsequent aggregation. When we introduced the learnable global spatiotemporal aggregation module after recurrent aggregation, the PSNR increased by 0.24dB. Overall, these experiments demonstrated that each component proposed in our LGSTANet played a significant role. Our approach proved effective in restoring and enhancing bracket images.

Effectiveness of LGSTA Module. To show the advantages of the proposed LGSTA module. We conducted ablation experiments on different variants of it. As shown in the Tab. 5, instead of utilizing the Global Average Pooling (GAP) and activation function processing method, only concatenation aggregation of base and complementary frames was employed, resulting in a decrease in PSNR due to the failure to effectively extract beneficial information from the complementary frames. To further thoroughly validate the effectiveness of the design. We replaced GAP with Global Max Pooling (GMP) to demonstrate that using GAP can achieve higher performance. Simultaneously, we observe that w/ modulator can bring a performance improvement of 0.16 dB, which reveals the effectiveness of the modulator. We designed a non-weight-sharing modulator to cope with different types of degradation disturbances. When we replaced the non-weight-sharing adaptive modulator with a weight-sharing adaptive modulator, we observed a decrease in PSNR of 1.24 dB. Our analysis indicated that this was because using only one modulator couldn't adapt to multiple degradations, and it could introduce noise, leading to a performance decline. To thoroughly validate the effectiveness of the design, we conducted an ablation study on the impact of different activation functions on performance, as shown in Tab. 6.

Figure 6. Visual comparison of different variants of LGSTA.

Table 5. Ablation studies of different variants of LGSTA.

Method	PSNR↑	SSIM↑	LPIPS↓	Params
Only concat	37.57	0.9430	0.117	13.315M
GAP→GMP	37.77	0.9470	0.115	14.041M
w/o Modulator	37.79	0.9456	0.121	14.041M
Share Modulator	36.71	0.9402	0.128	14.041M
Ours	37.95	0.9469	0.117	14.041M

Table 6. Ablation studies activation function effects.

Method	PSNR↑	Method	PSNR↑	Method	PSNR↑
celu	37.89	prelu	37.66	selu	37.84
elu	37.89	relu	37.89	tanh	37.90
lrelu	37.79	relu6	37.89	gelu	37.95

5. Conclusion

In this paper, we proposed LGSTANet, an aggregated restoration network designed for the restoration and enhancement of bracket images. Utilizing multi-frame bracketed images as input, we achieved efficient restoration and enhancement of various degradation types such as noise, blur, low dynamic range (LDR), and low resolution. We employed an architecture consisting of alignment, recurrent aggregation, and reconstruction, and introduced a learnable global spatio-temporal adaptive aggregation module to aggregate complementary information from different frames. Additionally, we proposed an ultra-lightweight adaptive restoration modulator to adapt to various degradation perturbations, thereby achieving the restoration of details and textures. Our method surpassed other state-of-the-art methods and achieved highly competitive results in both tracks of the NTIRE2024 BIRE Challenge. Moreover, our method ranked first in inference speed and memory usage in Track 2, demonstrating its practical applicability in real-world scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62171133, in part by the Artificial Intelligence and Economy Integration Platform of Fujian Province, and the Fujian Health Commission under Grant 2022ZD01003.

References

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

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

[3] Jiezhang Cao, Yawei Li, Kai Zhang, Jingyun Liang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 3

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 2, 3, 5

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 2, 3, 4, 5

[6] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 3

[7] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4196–4205, 2021. 2

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3, 5

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[10] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022. 1, 3, 5, 7

[11] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burstormer: Burst image restoration and enhancement transformer. *CVPR*, 2023. 1, 2, 5, 7

[12] Deniz Engin, Anil Genç, and Hazim Kemal Ekenel. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 825–833, 2018. 2

[13] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *Proceedings of the European conference on computer vision (ECCV)*, pages 538–554, 2018. 3

[14] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 2

[15] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020. 2

[16] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 3

[17] Jungwoo Kim and Min H Kim. Joint demosaicing and deghosting of time-varying exposures for single-shot hdr imaging. In *ICCV*, 2023. 5, 7

[18] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8878–8887, 2019. 1, 2

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

[20] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 3, 4

[21] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2507–2515, 2017. 3

[22] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6): 3064–3073, 2018. 2

[23] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *ECCV*, 2022. 1, 5, 7

[24] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on Computer Vision*, pages 344–360. Springer, 2022. 1, 3

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

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466, 2022. 3

[28] 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. 3

[29] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from

multi-exposed ldr images with large motions. *IEEE TIP*, 2021. 1, 5, 7

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[31] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 4

[32] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018. 3

[33] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2808–2817, 2020. 2

[34] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 3

[35] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 1, 2

[36] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginhac. Alignment-free hdr deghosting with semantics consistent transformer. In *ICCV*, 2023. 1, 2, 5, 7

[37] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginhac. Alignment-free hdr deghosting with semantics consistent transformer. *arXiv preprint arXiv:2305.18135*, 2023. 3

[38] 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. 3

[39] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision*, pages 4799–4807, 2017. 1, 2

[40] Roger Y Tsai and Thomas S Huang. Multiframe image restoration and registration. *ACVIP*, 1:317–339, 1984. 3

[41] 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. 3, 4

[42] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 2, 3, 5

[43] Renlong Wu, Zhilu Zhang, Shuohao Zhang, Hongzhi Zhang, and Wangmeng Zuo. Rbsr: Efficient and flexible recurrent network for burst super-resolution. In *PRCV*, 2023. 5, 7

[44] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018. 3

[45] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. An implicit alignment for video super-resolution. *arXiv preprint arXiv:2305.00163*, 2023. 3

[46] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *CVPR*, 2019. 1, 5, 7

[47] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 3

[48] Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. Fmanet: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. *arXiv preprint arXiv:2401.03707*, 2024. 3

[49] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2301–2310, 2020. 3

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 3

[51] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3, 6

[52] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018. 2

[53] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1, 2

[54] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 1, 2

[55] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Zifei Yan, and Wangmeng Zuo. Bracketing is all you need: Unifying image restoration and enhancement tasks with multi-exposure images. *arXiv preprint arXiv:2401.00766*, 2024. 5, 6, 7

[56] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Wangmeng Zuo, Radu Timofte, et al. Ntire 2024 challenge on bracketing im-

age restoration and enhancement: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 7

[57] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 191–207. Springer, 2020. 2, 3, 5

[58] Xingyu Zhou, Leheng Zhang, Xiaorui Zhao, Keze Wang, Leida Li, and Shuhang Gu. Video super-resolution transformer with masked inter&intra-frame attention. *arXiv preprint arXiv:2401.06312*, 2024. 3