CRNet: A Detail-Preserving Network for Unified Image Restoration and Enhancement Task

Kangzhen Yang[1†] Tao Hu[1†] Kexin Dai[1] Genggeng Chen[2] Yu Cao[3]
Wei Dong[2] Peng Wu[1] Yanning Zhang[1] Qingsen Yan[1*]
[1]Northwestern Polytechnical University [2]Xi'an University of Architecture and Technology
[3]Xi'an Institute of Optics and Precision Mechanics of CAS

Abstract

In real-world scenarios, images captured often suffer from blurring, noise, and other forms of image degradation, and due to sensor limitations, people usually can only obtain low dynamic range images. To achieve high-quality images, researchers have attempted various image restoration and enhancement operations on photographs, including denoising, deblurring, and high dynamic range imaging. However, merely performing a single type of image enhancement still cannot yield satisfactory images. In this paper, to deal with the challenge above, we propose the Composite Refinement Network (CRNet) to address this issue using multiple exposure images. By fully integrating information-rich multiple exposure inputs, CRNet can perform unified image restoration and enhancement. To improve the quality of image details, CRNet explicitly separates and strengthens high and low-frequency information through pooling layers, using specially designed Multi-Branch Blocks for effective fusion of these frequencies. To increase the receptive field and fully integrate input features, CRNet employs the High-Frequency Enhancement Module, which includes large kernel convolutions and an inverted bottleneck ConvFFN. Our model secured third place in the first track of the Bracketing Image Restoration and Enhancement Challenge, surpassing previous SOTA models in both testing metrics and visual quality.

1. Introduction

In real photography scenarios, images captured are often subject to image degradation such as blur and noise due to lighting conditions and exposure time limitations. Moreover, due to typical sensor limitations, people usually obtain low dynamic range images. To address these issues, various methods have been studied for deblurring, denoising, and HDR imaging. However, these methods often focus only on individual specific tasks, and the resulting images are still unsatisfactory. To address this issue, establishing a model that can handle the Unified Image Restoration and Enhancement Task is crucial.

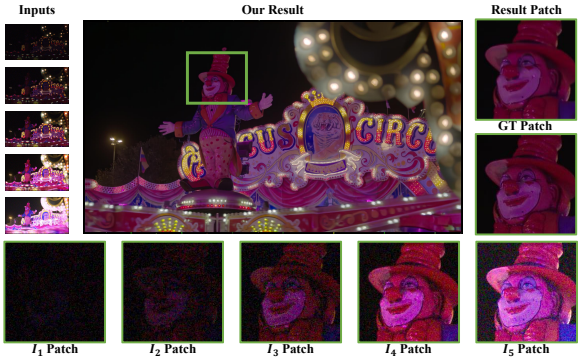

Figure 1. Our CRNet can finely restore the details lost in images from multiple exposure inputs and enhance them into HDR image.

In previous studies, various single-image restoration methods have been proposed, such as denoising [1, 5, 13, 22, 54, 55], deblurring [7, 28, 32, 41, 54], and high dynamic range reconstruction [9, 20]. However, these methods often perform poorly due to limitations such as insufficient information from a single image and exposure time constraints. Consequently, people have started exploring methods for multi-exposure image restoration and enhancement. For instance, some studies [6, 19, 31, 39, 53, 57] utilize dual-exposure images for image restoration, while others [15, 16, 24, 33, 36, 47–52] employ multi-exposure images to generate HDR images for image enhancement. These methods have made breakthroughs again, but they overlook the importance of edge and texture details in image restoration and enhancement tasks, lacking in the enhancement of high-frequency details. Recently, TMRNet[58] provides a feasible solution for the unified image restoration and en-

*Corresponding author. † The first two authors contributed equally to this work.
This work was partially supported by NSFC (62301432,62306240), NSBRPS (2023-JC-QN-0685, QCYRCXM-2023-057).

hancement tasks. It takes multi-exposure images as input and progressively integrates non-reference frames with reference frames. However, with the deepening of the network and the increase in input images, such fusion methods lead to the gradual forgetting of earlier added frames by the network.

To address the aforementioned issues, we propose the Composite Refinement Network (CRNet) (Example visual effects are depicted in Fig. 1). CRNet takes multi-exposure images as input and feeds them together into subsequent network parts after optical flow alignment, rather than frame by frame. To tackle the insufficient restoration of high-frequency details in image restoration tasks, CRNet explicitly employs pooling layers to separate high and low-frequency information for enhancement, and utilizes a specially designed non-stacked deep Multi-Branch Block for thorough fusion. To better integrate different image features, CRNet adopts the Convolutional Enhancement Block, a pure convolutional module primarily composed of large-kernel convolutions, which increases model's receptive field, along with ConvFFN featuring an inverted bottleneck structure to fully merge features. CRNet not only achieves breakthroughs in visual effects but also attains previous state-of-the-art performance in various evaluation metrics. Our CRNet secured third place in track 1 of the Bracketing Image Restoration and Enhancement Challenge. In summary, the main contributions are as follows:

- CRNet surpasses previous state-of-the-art (SOTA) models in both metrics and visual effects, achieving third place in track 1 of the Bracketing Image Restoration and Enhancement Challenge.
- To address the issue of insufficient high-frequency details in image restoration tasks, we use pooling layers to explicitly separate high and low-frequency information and employ a Multi-Branch Block for fusion.
- CRNet employs the Convolutional Enhancement Block, utilizing large kernel convolutions to increase the receptive field, along with ConvFFN featuring an inverted bottleneck structure for comprehensive feature fusion.

2. Related Work

Burst Restoration and Enhancement Burst images refer to a series of images captured in rapid succession. During the capturing process, issues such as misalignment or blurring may arise due to camera movement or subject motion. The goal of burst image restoration and enhancement is to effectively use algorithms and techniques to process these consecutive images, in order to obtain high-quality, clear image outputs. Current methods[3, 11, 26, 29, 30, 45, 46] typically involve techniques such as denoising, deblurring, image alignment, and super-resolution, with a primary focus on feature alignment and fusion.

In the aspect of denoising, Godard et al.[11] proposed a recursive fully convolutional deep neural network as a "feature accumulator" capable of processing images from a single frame to any number of burst frames. In terms of deblurring, Aittala et al.[2] introduced a novel convolutional architecture that can equally process information from all frames in a sequence, unaffected by their order. In the aspect of image enhancement, HDR+[35] selects a frame with short exposure as the reference frame to avoid clipped highlights and motion blur, then aligns the other frames to this one before merging. It decides on a per-pixel basis whether to merge the image content or not. Building upon this, there has been work focused on addressing two or more tasks simultaneously. A method involving a deep reparametrization of the maximum a posteriori (MAP)[3] formulation was proposed, which involves reparametrizing the classical MAP into a deep feature space. This approach shows good performance on both super-resolution and burst denoising tasks.

Multi-Exposure HDR Image Reconstruction. HDR imaging refers to the process of reconstructing a High Dynamic Range (HDR) image from a series of Low Dynamic Range (LDR) images with different exposures. A key challenge in this task is how to align multiple input frames to address the ghosting problem. Traditional alignment methods primarily utilize alignment-based[4, 17, 44], rejection-based[12, 34, 56], and patch-based methods[14, 27, 38] for alignment. Although they have achieved certain breakthroughs, they often err under conditions of extreme lighting and motion. The emergence of deep learning has provided new ideas for this issue. Some researchers have used CNN-based methods for alignment[16, 21, 33, 47, 48]. Kalantari et al.[16] first aligned images using optical flow and then fused them through a CNN network. Yan et al.[47] introduced a CNN-based spatial attention mechanism to suppress movement and oversaturated areas. Yan et al. [48] designed a non-local module to expand the receptive field for global merging. CNN-based alignment methods have made further progress, but still result in ghosting in overexposed areas due to motion. The advent of Transformers has given researchers new ideas for alignment methods. Song et al.[40] utilized the large receptive field of the Transformer to globally recover motion areas.

3. Proposed Method

We are given a sequence of Raw images $\{R_1, R_2, \ldots, R_N\}$ captured in a dynamic scene, each with a distinct exposure level, and our objective includes simultaneous denoising, deblurring, and HDR reconstruction. The resultant image should closely mirror a reference

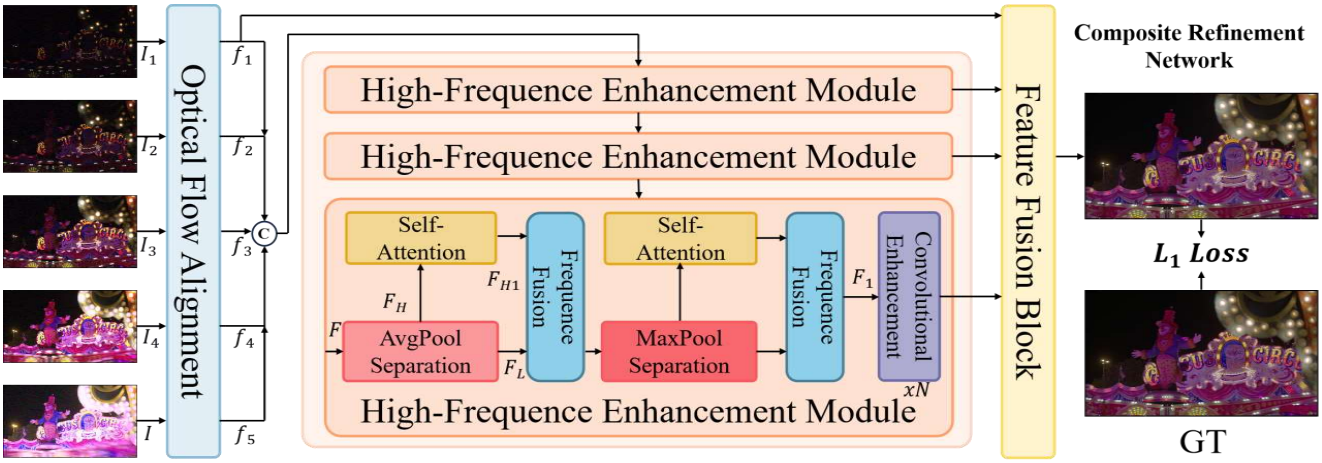

Figure 2. An overview of the CRNet model reveals that it mainly comprises three components: alignment, high-frequency enhancement, and output fusion. In the High-Frequency Enhancement Module, we swiftly separate high and low-frequency features and then utilize cleverly designed Multi-Branch Block to fuse them. Subsequently, we employ a purely Convolutional Enhancement Block to efficiently extract and fuse features for image enhancement.

image, denoted as R_r. Specifically, five Raw images $\{R_1, R_2, R_3, R_4, R_5\}$, ranked by their exposure from lowest to highest, are selected for input, with R_1 serving as the reference. Then, we normalize R_i to $\frac{R_i}{\Delta t_i/\Delta t_1}$, and the term Δt_i refers to the exposure time of the i-th image. Following the recommendations from multi-exposure HDR reconstruction methods [24, 33, 42, 47, 50], we transform the raw images after normalizing with gamma mapping to obtain $\{L_1, L_2, L_3, L_4, L_5\}$:

$$L_i = \left(\frac{R_i}{\Delta t_i/\Delta t_1}\right)^{\gamma}, \tag{1}$$

where γ represents the gamma correction parameter and is generally set to $\frac{1}{2.2}$, the same as TMRNet[58] does in its code.

Subsequently, we concatenate each L_i with its corresponding R_i to form $\{I_1, I_2, I_3, I_4, I_5\}$, according to the following equation:

$$I_i = \left\{\frac{R_i}{\Delta t_i/\Delta t_1}, L_i\right\}, \tag{2}$$

Then, we concatenate each L_i with its corresponding R_i to form $\{I_1, I_2, I_3, I_4, I_5\}$. Feeding $\{I_1, I_2, I_3, I_4, I_5\}$ into our model, we derive an HDR image free of noise and blur, denoted as $\hat{H}$, according to the following equation:

$$\hat{H} = f(I_1, I_2, I_3, I_4, I_5; \theta), \tag{3}$$

where the function $f(\cdot)$ signifies the imaging network, and θ encapsulates the parameters of the network.

3.1. Overview of the CRNet

As shown in Fig. 2, within CRNet, for the inputs $\{I_1, I_2, I_3, I_4, I_5\}$, we first align five input images to get $\{f_1, f_2, f_3, f_4, f_5\}$, employing an Optical Flow Alignment Block to align the five input images, which includes convolutional blocks for shallow feature extraction and a pre-trained Spynet[37]. Experiments(See in See the experimental section, "The Structure to Fuse Features" part) indicate that the frame-by-frame input method adopted by TMRNet[58] could lead to the network gradually forgetting earlier added images as the network deepens and the number of input images increases, shifting focus to more recently added images and thus degrading image quality. Therefore, we concatenate the five aligned images as the input for subsequent processing, allowing for more comprehensive utilization of the information from each image.

Subsequently, the merged features are sequentially passed through 3 High-Frequency Enhancement Modules. Each module initially utilizes two distinct pooling layers to separate and individually amplify high and low-frequency information[25], followed by N Convolutional Enhancement Blocks. After frequency separation, the valuable high-frequency information is enhanced using Transformer, and the information across frequencies is fused using Multi-Branch Blocks[43], thereby fully restoring the details of the image. Convolutional Enhancement Blocks can be regarded as high-frequency filters, employing large-kernel depth-wise separable convolutions [23] and convolutionalized FFN [8] to increase the receptive field and fully merge features, while further enhancing high-frequency information.

Finally, the outcomes of the 3 High-Frequency Enhancement Sub-networks are fused with the reference frame, which has been aligned using simple convolutional blocks, to output the final result.

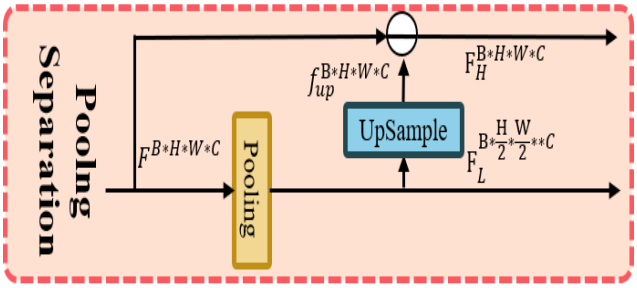

Figure 3. We efficiently separate high and low-frequency information through simple pooling layers.

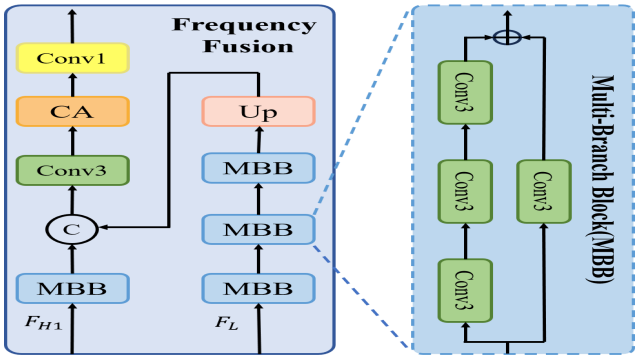

Figure 4. Through asymmetric parallel convolutional groups, the model effectively integrates high and low-frequency information.

3.1.1 Frequence Separation and Fusion

In the unified image restoration and enhancement task, enhancing high-frequency feature information is crucial, and one effective method is to separately amplify high-frequency and low-frequency features. Traditional separation techniques, such as Fourier transforms, entail significant computational overhead and are not straightforwardly integrated into networks. To minimize the computational expenses associated with segregating high and low-frequency features, we employ pooling layers[25] for this task (as shown in the Fig. 4). In the CRNet, we specifically utilized average pooling and max pooling. Specifically, a pooling layer down-samples the input feature maps with dimensions $B \times H \times W \times C$ to obtain the low-frequency features F_L at a reduced resolution $F_L^{B\times\frac{H}{2}\times\frac{W}{2}\times C}$. Bilinear interpolation is then used to upsample these features back to the original dimensions, $f_{\text{up}}^{B\times H\times W\times C}$. The high-frequency features F_H of F are subsequently computed by subtracting F_L from the original features F. This approach enables us to quickly capture the image's high and low-frequency characteristics, as described by the equation below:

$$F_L = \text{Pooling}(F), \tag{4}$$

$$F_H = F - \text{Upsample}(F_L), \tag{5}$$

where Pooling refers to the down-sampling operation using a pooling layer, and Upsample refers to the up-sampling operation using bilinear interpolation. After obtaining the explicitly extracted high-frequency features F_H, we opt to globally enhance them using Self-Attention mechanism to fully exploit the high-frequency information and get F_{H_1}. However, simply fusing features across different frequency domains through basic convolutional blocks could lead to information loss and insufficient fusion. Therefore, to adequately merge high-frequency and low-frequency information, we primarily utilize the Multi-Branch Block (See in Fig. 4).

The Multi-Branch Block consists of a dual-path convolutional assembly, with each path employing a different number of convolutions. The first branch, equipped with three convolutional kernels, focuses more on the high-frequency details of the image, while the second branch, with a single convolutional kernel, pays more attention to the low-frequency content and contours of the image[43]. This allows our network to effectively meet the requirements for merging high and low-frequency features.

Overall, we first pass the high-frequency and low-frequency features through several Multi-Branch Blocks individually, then upsample the low-frequency information to match the dimensions of the high-frequency features, concatenate them together, and finally fuse them through channel attention and simple convolution. The specific formulas are as follows:

$$H = \text{MBB}(F_{H1}), \tag{6}$$

$$L = \text{MBB}(\text{MBB}(\text{MBB}(F_L))), \tag{7}$$

$$\text{Out} = \text{Conv}_{1\times1}(\text{CA}(\text{Conv}_{3\times3}(\text{Concatenate}(\text{Up}(L), H)))), \tag{8}$$

where MBB represents the Multi-Branch Block, $\text{Conv}_{1\times1}$ represents a 1×1 convolution, $\text{Conv}_{3\times3}$ represents a 3×3 convolution, CA represents channel attention, and Up represents bilinear interpolation upsampling.

3.1.2 Convolutional Enhancement

To improve the receptive field and thoroughly integrate the features from inputs, we employ the Convolutional Enhancement Block in our network (see Fig. 5). This purely convolutional module utilizes 7×7 depth-wise separable convolutions to achieve a broad receptive field and features an inverted bottleneck structure ConvFFN for thorough information extraction. Additionally, it acts as a high-pass

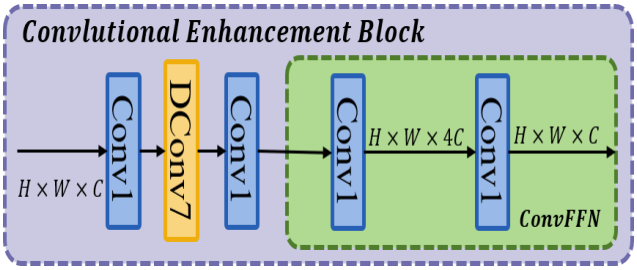

Figure 5. Our Convolutional Enhancement Block utilizes depth-wise separable convolutions with large kernels to achieve a large receptive field. Additionally, it leverages ConvFFN with an inverted bottleneck structure to enhance feature fusion capability.

filter, facilitating efficient content fusion from the five input frames and implicitly enhancing high-frequency information. The formula for the entire process is as follows:

$$F_2 = \text{ConvFFN}(\text{Conv}_{1\times1}(\text{DConv}_{7\times7}(\text{Conv}_{1\times1}(F_1)))) \tag{9}$$

Here, $\text{Conv1}_{1\times1}$ represents a 1×1 convolution, $\text{DConv}_{7\times7}$ denotes depth-wise separable 7×7 convolution, and each convolution layer is followed by GELU activation for non-linear activation.

4. Experiments

4.1. Experiments Settings

Datasets. The dataset we utilized is the training set from track 1 of the Bracketing Image Restoration challenge. This dataset is constructed based on HDR videos to simulate the motion and blur effects of multi-exposure images. Initially, HDR videos from Froehlich et al[10] were extracted and processed by frame interpolation and conversion into Bayer-pattern HDR raw sequences. Subsequently, degradation operations were introduced to generate multi-exposure images, including downsampling, grouping, blur synthesis, conversion to LDR images, and noise addition. Finally, the dataset comprises 1355 data pairs, with 1045 pairs used for training and 290 pairs for testing.

Training Loss. Following the approach of TMRNet, we employ the μ-law to map the resulting image from the linear domain to the tone-mapping domain:

$$T(x) = \frac{\log(1+\mu x)}{\log(1+\mu)}, \tag{10}$$

where μ=5000.

Given the result $\hat{H}$ from CRNet and the ground truth H, we conduct pixel-level L1 loss in the tone-mapping domain as follows:

$$L = \|T(H) - T(\hat{H})\|_1, \tag{11}$$

Evaluation Metrics. We compute five commonly used metrics for evaluation, namely PSNR-μ and SSIM-μ, where 'L' denotes the linear domain, and 'μ' denotes the tone-mapping domain. In the μ-law tone-mapping, $\mu = 5000$. In calculating the PSNR metric, the maximum value is set to 1. In the calculation of the SSIM metric, the maximum dynamic range is 1.

Implementation Details. During the training process, images are randomly cropped to patches with a size of 128×128, and are subject to random flipping and rotation. In CRNet, N is set to 10. We used the PyTorch framework and adopted the AdamW optimizer with β_1 = 0.9 and β_2 = 0.999. The initial learning rate was set to 10^{-4}, using StepLR for learning rate decay by multiplying it by 0.5 every 80 epochs. The model was trained on the synthetic dataset provided by the BracketIRE Task on 4 A100 GPUs for a total of 500 epochs, amounting to 3 days of training. We utilized Hugging Face's Accelerator for parallel training.

4.2. Comparison with the State-of-the-art Methods

Table 1. The evaluation results on the Bracketing Image Restoration and Enhancement Challenge[58]'s dataset for Track1 and Track2. The best and second-best results are highlighted in **Bold** and Underline, respectively.

Track1 Results							
Models	**AHDRNet**	**CA-VIT**	**HyHDR**	**SCTNet**	**TMRNet**	**Kim's**	**Ours**
PSNR-μ	36.32	36.54	37.4	36.90	<u>38.19</u>	37.93	**39.03**
SSIM-μ	0.927	0.9341	0.940	0.944	<u>0.949</u>	0.945	**0.950**
Track2 Results							
Models	**AHDRNet**	**CA-VIT**	**HyHDR**	**SCTNet**	**TMRNet**	**Kim's**	**Ours**
PSNR-μ	28.17	28.18	28.59	28.28	<u>28.91</u>	28.33	**29.45**
SSIM-μ	0.842	0.848	0.849	0.846	<u>0.857</u>	0.849	**0.871**

To further validate the superiority of our model, we compared CRNet with several state-of-the-art deep learning models AHDRNet [47], TMRNet[58] Transformer-based CA-ViT [24], SCTNet[42], Kim's[18] and HyHDR [50] quantitatively and qualitatively on track 1 of the Bracketing Image Restoration and Enhancement Challenge dataset. As shown in Tab. 1, our approach outperforms the compared models in the testing metrics. From Fig. 6(a), it can be observed that in the bonfire scene, where the background color is too dark and the flame area is too bright, and the fluttering bonfire presents foreground object motion, which is extremely challenging. Thanks to the proposed high-frequency enhancement module, our model successfully delineates clear and distinct edges in the flame's boundary region and suppresses ghosting, while other models exhibit edge discontinuities or redundancies. Additionally, in the facial area (see in Fig. 6(b)), the facial contour, especially the nose area, is particularly challenging. Our CRNet successfully restores the contour of the nose, whereas other models lack contour on the left side of the nose or exhibit

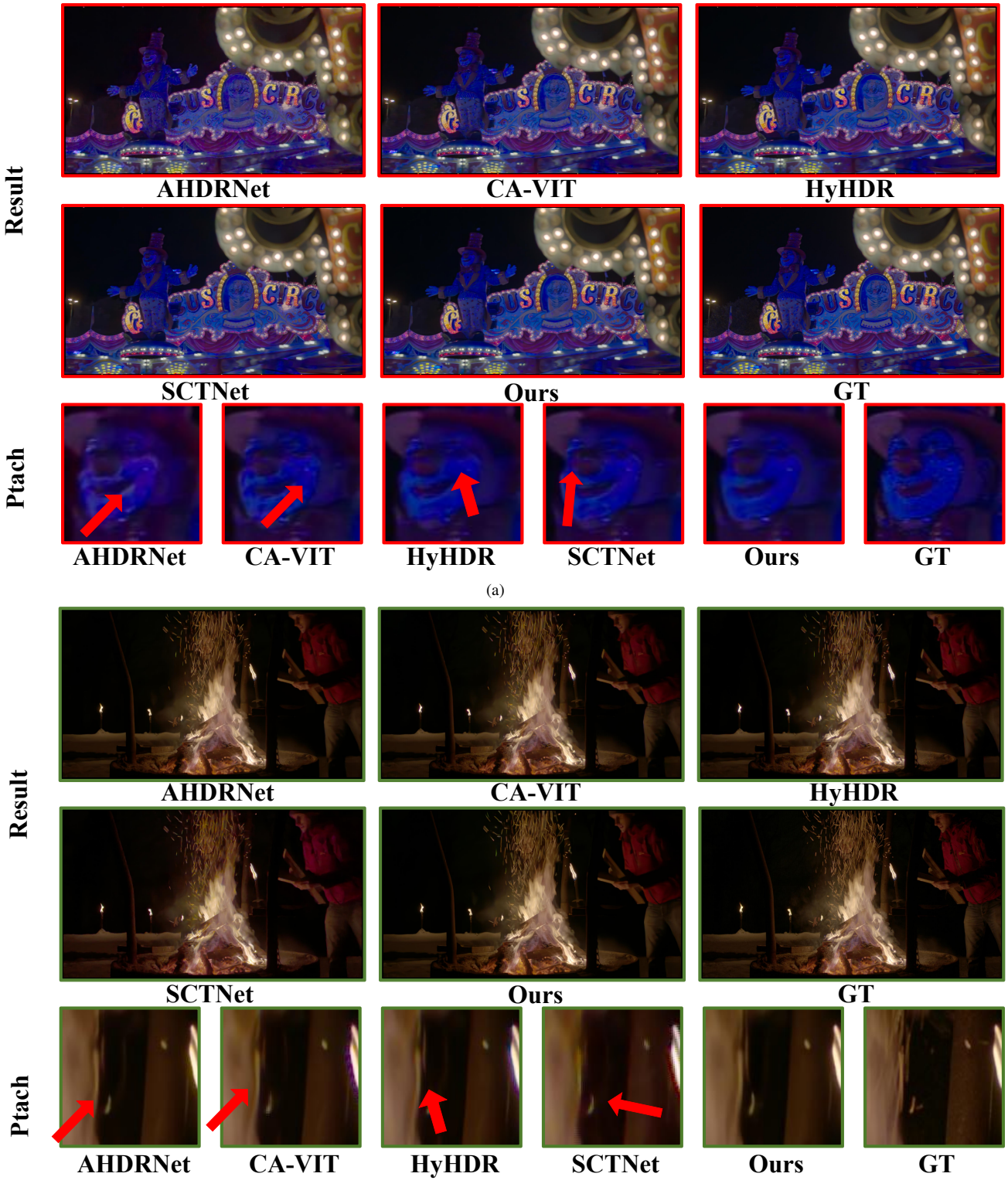

Figure 6. Examples of comparisions on the track 1 of the Bracketing Image Restoration and Enhancement Challenge dataset.

Table 2. Results[59] on track 1 of the Bracketing Image Restoration and Enhancement Challenge. Quantitative metrics are calculated on the full images. #FLOPs, inference time, and GPU memory are measured when generating a 1920×1080 raw image. The test uses NVIDIA RTX A6000 GPU to calculate the inference time and adopts the THOP [60] tool to calculate #FLOPs. The ranking is based on the PSNR metric of the full images.The best and second-best results are highlighted in **Bold** and Underline, respectively.

Rank	Team	Full Images PSNR ↑	#Params (M)	#FLOPs (T)	Time (s)	Memory (GB)
1	SRC-B	40.54	94.34	48.238	3.102	20
2	MegIRE	39.78	19.75	30.751	2.383	16
3	UPN1(CRNet)	39.03	13.32	**10.409**	**1.090**	**6**
4	CVG	38.78	13.29	21.340	7.518	11
5	FZU_DXW	38.46	14.04	22.283	1.829	16
-	TMRNet [58]	38.19	13.29	21.340	1.874	15

unclear delineation on the right side. This indicates a significant improvement of our CRNet over previous methods, especially in scenes with too dark or too bright conditions, in high-frequency detail regions, and also surpasses previous state-of-the-art methods in numerical metrics. Furthermore, we introduced an upsampling module similar to that of TMRNet into the base of CRNet and conducted tests on track 2 of the Bracketing Image Restoration and Enhancement Challenge dataset. As shown in Tab. 1, our approach outperforms the compared models in the testing metrics.

4.3. Results of track 1 of the Bracketing Image Restoration and Enhancement Challenge

As shown in Tab. 2, we achieved the third place in track 1 of the Bracketing Image Restoration and Enhancement Challenge. Compared to other ranking models, our CRNet significantly outperforms in terms of FLOPS, inference time, and GPU memory usage.computational costs.

4.4. Ablation Studies

Table 3. The Ablation study of CRNet.'-' indicates that a certain structure is not applicable, and another module with similar computational complexity is embedded to replace it.

Models	**PSNR-μ**	**SSIM-μ**
CRNet	**39.03**	**0.949**
CRNet-Frequency Separation	38.71	0.922
CRNet-Multi-Branch Block	38.64	0.937
CRNet-Large Kernel Conv	38.76	0.941
CRNet-ConvFFN	38.69	0.933
CRNet+Recurrent Structure	38.79	0.945

To validate the effectiveness of each module in our CRNet, we conducted a series of ablation experiments(See in Tab. 3):

- **Model1:** We use the full CRNet.
- **Model2:** We do not perform frequency separation; instead, we directly pass the original feature maps through self-attention and merge them with the results after downsampling.
- **Model3:** In the frequence fusion part, we do not utilize the Multi-Branch Block structure; instead, we simplify it to four consecutive convolutional layers.
- **Model4:**We replace the 7×7 convolution in the Convolutional Enhancement with three 3×3 convolutions.
- **Model5:** We do not employ the inverted bottleneck structure in ConvFFN. Instead, we use two 1×1 convolutions without changing the number of channels.
- **Model6:** While keeping the model's parameter count almost identical, we arrange all the High-Frequency Enhancement Modules in CRNet according to the recurrent structure of TMRNet, sequentially feeding images into them. The output of the last High-Frequency Enhancement Module is then combined with the reference frame and input into the Feature Fusion Block.

Figure 7. From the figure, it can be observed that our CRNet successfully restores the edges of the branches clearly after using frequency separation.**FS** represents for Frequence Separation Block.

Frequence Separation.

As shown in Table 3, Model 1 outperforms Model 2 by 0.32 in PSNR-μ and 0.027 in SSIM-μ. This indicates that utilizing frequency separation to extract high-frequency features and enhancing them selectively significantly improves the image quality. As illustrated in Fig. 7, after employing

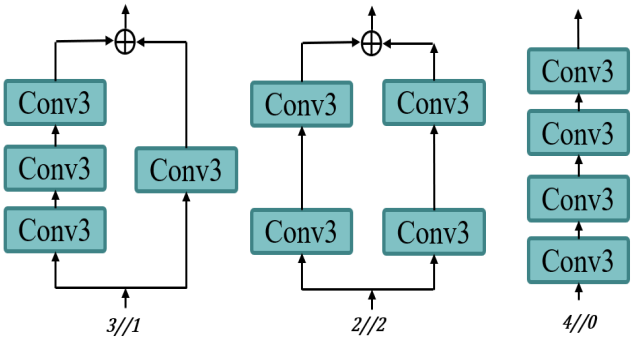

Figure 8. "\\" represents parallel inputs. As shown in the figure, we attempted three different ways of allocating convolutional blocks.

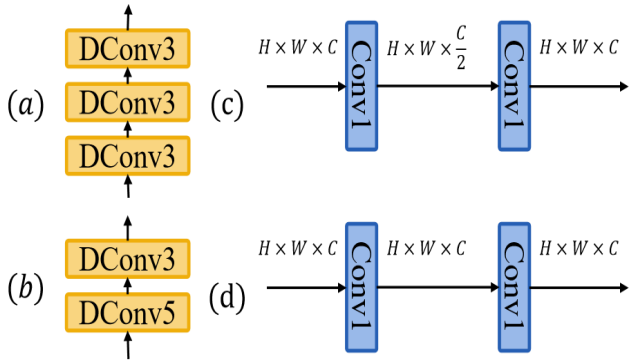

Figure 9. The figure shows the model diagrams of the comparison models included in Tab. 5

frequency separation, our CRNet successfully restores the edges of the branches clearly after using frequency separation.

Multi-Branch Block. As shown in Table 3, Model 1 outperforms Model 3 by 0.39 in PSNR-μ and 0.012 in SSIM-μ. This indicates that our utilization of the Multi-Branch Block for the fusion of high and low-frequency content is highly effective. To further validate the effectiveness of our structure, we conducted more detailed ablation experiments. As illustrated in Fig. 8, we reconfigured the positions of the four convolutional kernels for investigation. As shown in Tab. 4, Model 1 performs the best, implying that our Multi-Branch Block structure achieves better performance under the same computational cost.

Table 4. Ablation for Multi-Branch Block. The allocation scheme of having three convolution blocks on one side and one on the other side appears to be the most effective.

Models	**PSNR-μ**	**SSIM-μ**
3\\1	**39.03**	**0.949**
2\\2	38.82	0.941
4\\0	38.64	0.937

Convolutional Enhancement Block.

As shown in Table 3, Model 1 outperforms Model 4 by 0.27 in PSNR-μ and 0.008 in SSIM-μ. Model 1 outperforms Model 5 by 0.34 in PSNR-μ and 0.016 in SSIM-μ. This indicates that our Convolutional Enhancement Block's large kernel convolution design and ConvFFN fusion of input features are highly effective.

To validate the effectiveness of our Convolutional Enhancement Block structure, we conducted further ablation experiments. (See in Fig. 9)We attempted to replace the 7×7 depthwise separable convolution with either three 3×3 depthwise separable convolutions or a stack of one 5×5 convolution followed by one 3×3 convolution, and we tried to replace the ConvFFN with a normal bottleneck structure or a normal 1×1 convolution chaining structure. As shown in Tab. 5, the experimental results indicate that the 7×7 depthwise separable convolution and the inverse bottleneck ConvFFN adopted by us are more optimal structures for the tasks of image restoration and enhancement.

Table 5. The stuctures presented in the table are depicted in Fig. 9. Models a and b are used to replace the 7×7 depthwise separable convolution, while models c and d are used to replace the inverse bottleneck ConvFFN.

Models	**PSNR-μ**	**SSIM-μ**
CRNet	**39.03**	**0.949**
$3 \times 3 \times 3$(a)	38.76	0.941
5×5+3×3(b)	38.88	0.939
normal bottleneck(c)	38.69	0.933
normal 1×1 convolutions(d)	38.82	0.945

The Structure to fuse features As shown in Table 3, Model 1 outperforms Model 6 by 0.24 in PSNR-μ and 0.005 in SSIM-μ. From this, it is evident that inputting the feature maps together into the High-Frequency Enhancement Module is a better approach than using the Recurrent Structure.

5. Conclusion

We introduced the Composite Refinement Network (CRNet), which uniformly completes image restoration and enhancement. CRNet explicitly separates the frequency domain and employs the Multi-Branch Block for targeted fusion, thereby enhancing high-frequency details in image restoration. Additionally, it utilizes the Convolutional Enhancement Block, leveraging large kernel convolutions and ConvFFN to increase the receptive field and enhance feature fusion capabilities. CRNet surpassed previous state-of-the-art models in both metrics and visual effects, and achieved third place in track 1 of the Bracketing Image Restoration and Enhancement Challenge, with computational costs far below those of other models.

References

[1] Abdelrahman Abdelhamed, Mahmoud Afifi, Radu Timofte, and Michael S Brown. Ntire 2020 challenge on real image denoising: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 496–497, 2020. 1

[2] Miika Aittala and Fredo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *The European Conference on Computer Vision (ECCV)*, 2018. 2

[3] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising, 2021. 2

[4] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, pages 7–12. IEEE, 2000. 2

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019. 1

[6] Meng Chang, Huajun Feng, Zhihai Xu, and Qi Li. Low-light image restoration with short-and long-exposure raw pairs. *IEEE Transactions on Multimedia*, 24:702–714, 2021. 1

[7] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 1

[8] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022. 3

[9] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 1

[10] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, pages 279–288. SPIE, 2014. 5

[11] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising, 2017. 2

[12] Thorsten Grosch et al. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling and Visualization, RWTH Aachen*, 277284 (3):2, 2006. 2

[13] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1712–1722, 2019. 1

[14] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1163–1170, 2013. 2

[15] Tao Hu, Qingsen Yan, Yuankai Qi, and Yanning Zhang. Generating content for hdr deghosting from frequency view. *arXiv preprint arXiv:2404.00849*, 2024. 1

[16] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 1, 2

[17] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325, 2003. 2

[18] Jungwoo Kim and Min H Kim. Joint demosaicing and deghosting of time-varying exposures for single-shot hdr imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12292–12301, 2023. 5

[19] Wei-Sheng Lai, Yichang Shih, Lun-Cheng Chu, Xiaotong Wu, Sung-Fang Tsai, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Face deblurring using dual camera fusion on mobile phones. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 1

[20] Bruno Lecouat, Thomas Eboli, Jean Ponce, and Julien Mairal. High dynamic range and super-resolution from raw image bursts. *arXiv preprint arXiv:2207.14671*, 2022. 1

[21] Ru Li, Chuan Wang, Jue Wang, Guanghui Liu, Heng-Yu Zhang, Bing Zeng, and Shuaicheng Liu. Uphdr-gan: Generative adversarial network for high dynamic range imaging with unpaired data. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7532–7546, 2022. 2

[22] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Zhijun Tu, Kunpeng Du, Hailing Wang, Hanting Chen, Wei Li, Xiaofei Wang, et al. Ntire 2023 challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1904–1920, 2023. 1

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3

[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, 5

[25] 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, 4

[26] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 2

[27] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017. 2

[28] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1905–1913, 2023. 1

[29] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Gated multi-resolution transfer network for burst restoration and enhancement. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22201–22210. IEEE, 2023. 2

[30] Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks, 2018. 2

[31] Janne Mustaniemi, Juho Kannala, Jiri Matas, Simo Särkkä, and Janne Heikkilä. Lsd–joint denoising and deblurring of short and long exposure images with cnns. 2020. 1

[32] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 1

[33] 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 Transactions on Image Processing*, 30:3885–3896, 2021. 1, 2, 3

[34] Fabrizio Pece and Jan Kautz. Bitmap movement detection: Hdr for dynamic scenes. In *2010 Conference on Visual Media Production*, pages 1–8. IEEE, 2010. 2

[35] K Ram Prabhakar, Rajat Arora, Adhitya Swaminathan, Kunal Pratap Singh, and R Venkatesh Babu. A fast, scalable, and reliable deghosting method for extreme exposure fusion. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2019. 2

[36] K Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R Venkatesh Babu. Towards practical and efficient high-resolution hdr deghosting with cnn. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 497–513. Springer, 2020. 1

[37] 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. 3

[38] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203–1, 2012. 2

[39] Shayan Shekarforoush, Amanpreet Walia, Marcus A Brubaker, Konstantinos G Derpanis, and Alex Levinshtein. Dual-camera joint deblurring-denoising. *arXiv preprint arXiv:2309.08826*, 2023. 1

[40] Jou Won Song, Ye-In Park, Kyeongbo Kong, Jaeho Kwak, and Suk-Ju Kang. Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In *European Conference on Computer Vision*, pages 288–304. Springer, 2022. 2

[41] 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

[42] 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, 5

[43] Huiyuan Tian, Li Zhang, Shijian Li, Min Yao, and Gang Pan. Multi-depth branch network for efficient image super-resolution. *Image and Vision Computing*, page 104949, 2024. 3, 4

[44] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. 2007. 2

[45] Pengxu Wei, Yujing Sun, Xingbei Guo, Chang Liu, Guanbin Li, Jie Chen, Xiangyang Ji, and Liang Lin. Towards real-world burst image super-resolution: Benchmark and method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13233–13242, 2023. 2

[46] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *European conference on computer vision*, pages 352–368. Springer, 2020. 2

[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. 1, 2, 3, 5

[48] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29: 4308–4322, 2020. 2

[49] Qingsen Yan, Dong Gong, Javen Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Dual-attention-guided network for ghost-free high dynamic range imaging. *International Journal of Computer Vision*, pages 1–19, 2022.

[50] Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. A unified hdr imaging method with pixel and patch level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22211–22220, 2023. 3, 5

[51] Qingsen Yan, Tao Hu, Yuan Sun, Hao Tang, Yu Zhu, Wei Dong, Luc Van Gool, and Yanning Zhang. Towards high-quality hdr deghosting with conditional diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[52] Qingsen Yan, Song Zhang, Weiye Chen, Hao Tang, Yu Zhu, Jinqiu Sun, Luc Van Gool, and Yanning Zhang. Smae: Few-shot learning for hdr deghosting with saturation-aware masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5775–5784, 2023. 1

[53] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum. Image deblurring with blurred/noisy image pairs. In *ACM SIGGRAPH 2007 papers*, pages 1–es. 2007. 1

[54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2696–2705, 2020. 1

[55] 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

[56] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2011. 2

[57] Zhilu Zhang, RongJian Xu, Ming Liu, Zifei Yan, and Wangmeng Zuo. Self-supervised image restoration with blurry and noisy pairs. *Advances in Neural Information Processing Systems*, 35:29179–29191, 2022. 1

[58] 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. 1, 3, 5, 7

[59] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Wangmeng Zuo, Radu Timofte, et al. Ntire 2024 challenge on bracketing image restoration and enhancement: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 7

[60] Ligeng Zhu. Thop: Pytorch-opcounter, 2022. https://pypi.org/project/thop. 7