High Quality Reference Feature for Two Stage Bracketing Image Restoration and Enhancement

Xiaoxia Xing[1] HyunHee Park[2] Fan Wang[1]
Ying Zhang[1] Sejun Song[2] Changho Kim[2] Xiangyu Kong[1]
[1]Samsung R&D Institute China-Beijing
[2]Department of Camera Innovation Group, Samsung Electronics
{xx.xing,inextg.park,fan.wang,ying09.zhang,sejun.song,chang-ho.kim,xiangyu.kong}@samsung.com

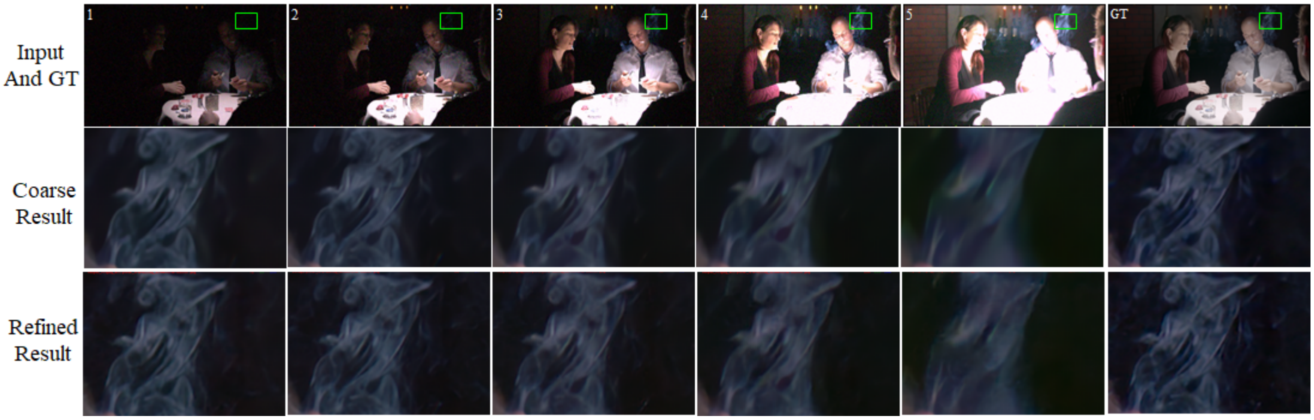


Figure 1. Results of using the different reference image. The first row shows the input images with different exposures that are also the reference image for the first stage training for corresponding columns, with the last column showing the ground truth image. The second and third rows show the results of the first and second stages. The last column shows the results of our proposed reference feature generation method. The first column of the second row is the effect of [48].

Abstract

In a low-light environment, it is difficult to obtain high-quality or high-resolution images with sharp details and high dynamic range (HDR) without noise or blur. To solve this problem, the Bracketing Image Restoration and Enhancement integrates Dnoise, Deblur, HDR Reconstruction, and Super Resolution techniques into a unified framework. However, we find that most methods select the image that aligns with GT as the reference image. Since the details of the reference image are not good enough, seriously affects the feature fusion, which finally leads to details being blurred. To generate a high dynamic range and a high-quality image, we propose a two-stage Bracketing method named RT-IRE. In the first stage, we generate the high-quality reference feature to guide feature fusion, remove the degradation, and reconstruct HDR to get coarse results. The second stage learns the residuals between the coarse result and the GT, which further enhances and generates details. Extensive experiments show the effectiveness of the proposed module. In particular, RT-IRE won two champions in the NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge.

1. Introduction

It is very difficult to obtain high-quality or high-resolution images with sharp details and high dynamic range (HDR) without noise or blur in low-light environments. If the exposure time is too short, underexposed photos will be obtained with noisy and dark areas invisible. Conversely, if the exposure time is too long, objects or camera shake will cause motion blur and the bright areas will be overexposed.

The single image enhancement approach is not sufficient to remove noise and blur while reconstructing the HDR and super-resolution, so more researchers use multi-images to solve this problem. The sub-pixel displacement between multi-images caused by camera motion can be beneficial for denoise and SR. So most Burst method [2, 12] utilizes multi consecutive frames with the same exposure time, which achieve the good effect for super-resolution (SR) and denoise, but cannot reconstruct HDR. Long exposure images are less noisy, so it has a significant positive effect on denoise. Short-exposure images have less motion blur that can assist in deblurring. Under-exposed areas in the short exposure image may be well exposed in the long exposure image, and over-exposed areas in the long exposure image may be clear in the short exposure image. So It is possible to reconstruct HDR using multi-exposure images, but current methods have limitations, some [9, 22] can denoise but cannot deblur, and some [34] can SR but cannot denoise.

Recent study [48] has proposed the BracketIRE method, which integrates the four tasks(i.e., Denoise, Deblur, HDR reconstruction, and SR) into a unified framework to generate sharp, high dynamic range, and high-resolution images with multi-exposure images. The input is multiple images with different exposures. The exposure time is increased sequentially, so the blur becomes more serious, the under-exposure areas become less and the over-exposure areas become more, as shown in the first row of Fig. 1. The solution for this class of methods is basically the same: firstly, multiple images are aligned at the feature level according to the reference image, then the aligned features are fused and finally reconstructed to get a high-quality image.

Reference image selection is very important, and most methods select the image that aligns with GT as the reference image. For example, BracketIRE [48] uses the first image as the reference, which has the aligned GT and the shortest exposure time, the weakest blur, but with the most under-exposure areas and worst noise. Since the details of the reference image are not good enough, especially in the under-exposure areas, it seriously affects the feature alignment and fusion, which finally leads to details blurred in the bright areas. As shown in the first result of the second row in Fig. 1. The first and second images in this training data are almost aligned, and the quality of the second image is significantly better than the first image. So we changed the reference image directly. The result of using the second image as the reference is shown in the second row and second column in Fig. 1. The result is significantly better than the method of using the first image as the reference image. The third image is the best quality of all the input images, but the results of using it as a reference image are not perfect because it has a large misalignment with GT.

An end-to-end one-stage approach is effective in reconstructing HDR and reducing noise in the image. However, the results tend to be too smooth and lack detailed textures. To address this issue, we propose a two-stage approach. In the first stage, we use multi-exposure images to reconstruct HDR and remove noise. In the second stage, we enhance the images and generate detailed textures. Our primary contributions include:

- We propose a dual-branch framework for generating reference features, where one branch removes degradations such as noise and blur, while the other preserves details, which drastically improves the quality of the reference features.
- We introduce a two-stage solution pipeline to first reconstruct HDR and remove noise, and then further improve and generate detailed textures.
- Extensive experiments are conducted to demonstrate the effectiveness of our proposed method. We won two champions in the NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge.

2. Related Work

Burst Image Restoration and Enhancement. Compared with single image processing, burst-based image restoration [1, 13, 15, 16, 21, 39, 40] employs multiple consecutive frames to achieve superior performance. The inherent randomness and independence of inter-frame noise [38] play a crucial role in denoising tasks, prompting the development of numerous methods [14, 26, 32] designed to exploit this property for effective burst image denoising. [32] proposes an end-to-end burst denoising pipeline to jointly utilize high-resolution and high-frequency features derived from wavelet transforms. BPN [42] predicts a global low-dimensional basis set for large denoising kernels to achieve effective burst denoising. Similar to the denoising task, burst image deblurring [1, 40] relies on the fusion of available information in multiple frames to restore the underlying clear image. Even though individual frames may be very blurry, they still retain some information about the sharp image [1]. [30] performs a local relative ranking of frames through a novel blur estimation bi-variable function, showing superior results compared with other burst image deblurring methods. Similarly, burst image super-resolution [2, 39] offers the possibility to reconstruct rich details by combining high-frequency information from multiple low-resolution views of the same scene. In contrast to handling specific degradation, [16] is proposed to process joint denoising and HDR recon

Dual-Exposure Image Restoration. Several methods [7, 20, 27, 33, 50] utilize short-long exposure image pairs for image restoration. High noise levels and color distortion issues in the short-exposure image may be well eliminated in the long-exposure one, while motion-blurred areas in the long-exposure images may be sharp in the short-exposure one. LSD_2 [27] performs joint image denoising and deblurring by leveraging information from short-long exposure images and adjusting their contributions based on the prevailing conditions. Considering RGB images as input, this method fails to account for noise or blur in the imaging pipeline, resulting in unsatisfactory outcomes when applied

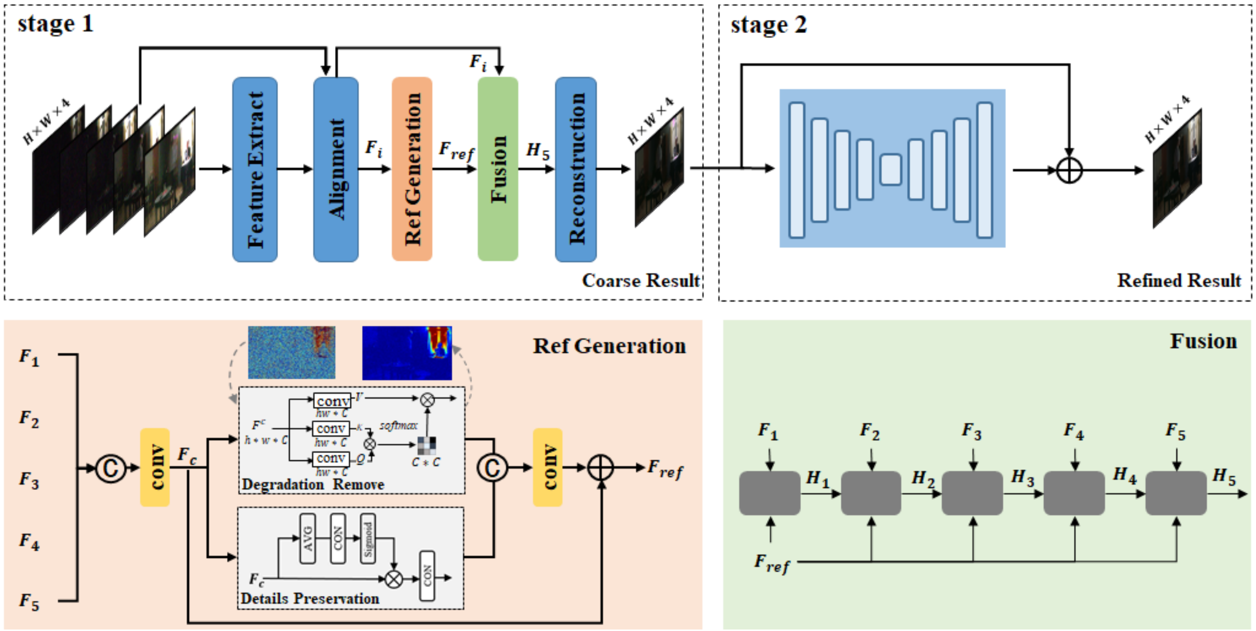


Figure 2. Overall architecture of RT-IRE. RT-IRE consists of two modifications to improve feature fusion and HDR reconstruction coarse result. For feature fusion, we introduce the Ref Generation module to obtain high-quality reference feature F_{ref} to guide feature fusion. In addition, on top of the first stage, we add the second stage to refine the coarse result into the refined result.

to real images. To alleviate this problem, LSFNet [7] designs a novel method to synthesize realistic short- and long-exposure raw images by simulating the imaging pipeline in a low-light environment and proposes a new fusion network to deal with the problems of low-light image fusion. D2HNet [50] presents a two-phase network to address the domain gap between training data and real photos. However, due to the limited dynamic range of dual-exposure images, their application in HDR imaging is extremely limited.

Multi-Exposure HDR Image Reconstruction. Multi-exposure images, which contain rich information about the scene, are widely utilized for HDR image reconstruction [18, 24, 28, 31, 35, 47]. To achieve this, [18] first aligns the low and high-exposure images to the medium-exposure one using optical flow and then uses the aligned images for high-quality HDR reconstruction. Recently, HDR-Transformer [24] employs a spatial attention module to suppress misalignment and designs the context-aware vision transformer for high-quality HDR imaging. Most existing methods are trained on images without any degradation and may not fit the real-world data with noise or blur. Mobile-HDR [23] establishes an HDR image dataset captured by mobile phone cameras and presents a cross-attention based alignment module to perform joint HDR denoising. Meanwhile, [10] further introduces motion blur for joint HDR denoising and deblurring by learning spatiotemporal distortion models. Considering more realistic degradations in low-light environments, BracketIRE [48] utilizes the complementary potential of multi-exposure images to deal with image denoising, deblurring, HDR reconstruction, and SR, achieving SOTA performance on both synthetic and real-world datasets.

3. Method

We propose a two-stage method for Bracketing Image Restoration and Enhancement. In the first stage, our method aims to produce a high-quality reference feature to guide feature fusion, remove degradation, and reconstruct HDR, thereby obtaining coarse results. Then, in the second stage, our method works to refine the output of the first module and generate a restored result that contains much more detail than before.

3.1. Preliminary

Our method employs bracketing photography to accomplish multi-frame denoising, deblurring, and super-resolution tasks, yielding clear, high dynamic range, and high-resolution images. RT-IRE is designed to perform denoising, deblurring, and HDR reconstruction, while RT-IRE+ expands on these capabilities to include the super-resolution (SR) task.

Specifically, RT-IRE uses multi-exposure noisy and blur RAW images $\{X_i\}_{i=1}^{N}$, $X_i \in H \times W \times 4$ to produce a clear and high-resolution Raw image as H. Different from RT-IRE, the RT-IRE+ model takes low-resolution images as input, with a shape of $\frac{H}{s} \times \frac{W}{s} \times 4$. Here, N represents the number of input frames, s indicates the super-resolution factor and i denotes the raw image captured with exposure time t_i and $t_i < t_{i+1}$. Following existing multi-exposure HDR reconstruction techniques [10, 16, 24, 28], we normalize X_i to $\frac{X_i}{t_i/t_1}$, aligning brightness across all frames and then concatenate it with its gamma-transformed image.

$$X_i^{in} = Concat(\frac{X_i}{t_i/t_1}, (\frac{X_i}{t_i/t_1})^{\gamma}). \quad (1)$$

In our implementation, γ signifies the gamma correction, conventionally set at $1/2.2$. Subsequently, these concatenated images are inputted into the model.

3.2. Reference Feature Generation

TMRNet [48] adopts Basicvsr++ [5] architecture, a promising network for expanding into more sophisticated systems due to its simplicity and adaptability. It is mainly used for video super-resolution tasks, where the first frame is employed as the reference frame to guide the network learning and direct the acquisition of knowledge from other frames in each iteration. In video super-resolution, the degradation of each frame is similar, with most frames experiencing only blurry changes. However, for bracket image restoration and enhancement, each frame has different levels of exposure, noise, and blur. From the first to the last frame, the noise decreases while the blur gradually increases. The first frame has low blur but significant noise, making it unsuitable as the reference frame due to potential interference in unclear areas during feature fusion.

To overcome this issue, we propose two approaches to obtain F_{ref} with less noise and blur for use in the feature fusion stage: selecting a high-quality frame as the reference frame, called 2ndRF, and generating a high-quality reference feature, denoted as GRF. Our experiments have shown that these minimal redesigns can produce robust and efficient results without additional bells and whistles.

3.2.1 Selecting High-quality Reference Frame

We chose the second frame as the reference frame since it has less noise and acceptable blur, making it easier to avoid interference from the reference feature noise on the fused feature. It is important to note that the second frame has very little motion with ground truth, which means it has less impact on any misalignment with the ground truth image. Using the second frame as the reference frame gives a PSNR improvement of up to 0.93 dB, making it superior to using the first frame as the base frame.

3.2.2 Generating Reference Feature

The method mentioned above may not apply when there is significant motion between the high-quality input and ground truth frames. The reference frame feature F_{ref} in Fig. 2 plays a guiding role in the process of feature fusion, so generating high-quality reference features will yield results similar to the performance of the above method. To achieve this, we have developed a lightweight module that fuses features from five frames as the reference feature. This approach contains more information than selecting a single reference feature as the guiding feature, as the five frame features have good quality in different regions.

To strike a delicate balance between suppressing noise and blur and synthesizing detailed information, generating the reference feature involves two branches: one focuses on denoising and deblurring while the other preserves detailed information about the original aligned features F_i. The architecture depicted in Fig. 2 is inspired by self-attention module [36] and channel attention mechanisms [17]. The first 1×1 convolution decreases the number of feature channels to reduce computational complexity. Including the residual connection preserves signal integrity in the network, while the last 1×1 convolution balances noise reduction and signal retention, with parameters trained alongside the entire network.

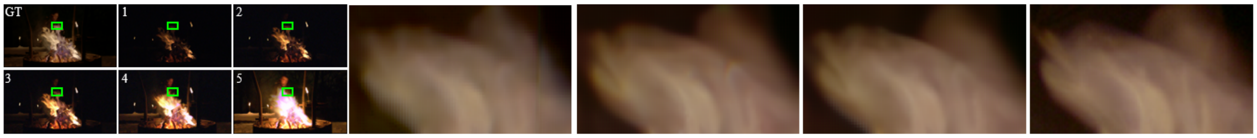
For the Degradation Remove module, denoising and deblur are achieved based on attention module [36]. The attention is regarded as non-local operation [37], which can also be traced back to classic algorithms like non-local means [4] and BM3D [11]. Their efficacy stems from abundant redundant information in high-dimensional feature maps, allowing these algorithms to eliminate significant noise and blur. Concretely, we used transposed attention with linear complexity to reduce computational complexity. Traditional self-attention operates across spatial dimensions, while transposed attention applies attention across feature dimensions. Ref Generation module as shown in Fig. 2, this branch helps to decrease the noise in the features.

The Detail Preservation module dynamically selects essential features based on their significance in the spatial dimension by the channel mechanism. As shown in Fig. 2, we extract information using an average pooling layer, a convolution layer, and a sigmoid activation. This information describes the importance of features and is used to select feature maps. A 1×1 convolution layer is then used to guide the convolution layer in preserving the essential features while dropping less informative ones.

Last, the concatenation of features from Degradation Remove and Details Preservation is then processed through convolution layers to reduce dimension, resulting in reference features that are both high-quality and detailed.

3.3. Two-Stage Structure

Our approach consists of two primary modules: one for fusion and structure reconstruction, and the other for detailed enhancement.



(a) GT and Input Frames (b) Burstormer [12] (c) Kim *et al.* [19] (d) RT-IRE (e) GT

Figure 3. Visual comparison on the Challenge dataset [49] of RT-IRE task. Our method restores more details.

In the first stage of the network, we use the TMR-NET [48], which consists of four essential steps: feature extraction, alignment, feature fusion, and image reconstruction. The primary purpose of this stage is to combine and reconstruct high dynamic range images by integrating aligned features. We do this by merging high-quality areas from each frame to restore the overall image structure. However, the reconstructed images from the first stage may miss out on some image details. To address this issue, we introduce a Refinement Module based on the lightweight network NAFNet [8]. The module helps the network to focus more on learning the image details by learning residuals on the coarse results of the first stage. This ensures detail enhancement while preserving the performance of the first stage. Our experimental results show an improvement of 2.35 dB PSNR for the RT-IRE task and 5.35 dB PSNR for the RT-IRE+ involving super-resolution tasks. The refinement module exhibits a more significant impact on super-resolution tasks than on image reconstruction alone.

Previous research [3, 43–45] has demonstrated that networks tend to learn structural information first and then focus on learning detailed information. However, our network is designed to learn specific information at different stages. This includes low-frequency structural cues in the first stage and high-frequency fine details in the second stage, which makes learning easier for the network and significantly reduces training complexity.

3.4. Training Loss

A cost function based on linear high dynamic range (HDR) values may give too much importance to high luminance levels, which could lead to overlooking significant differences within lower luminance ranges. To avoid this, most deep learning systems that focus on HDR prediction use objective functions based on tone-mapped luminance values. In this context, we calculate the loss using the widely used τ-law function in the tone-mapped domain.

$$\mathcal{T}(x) = \frac{\log(1+\mu x)}{\log(1+\mu)} \tag{2}$$

where $\mathcal{T}(x)$ is the tonemapped HDR image, and we set τ to 5000. In contrast to prior approaches [41, 46, 48] that rely on pixel-wise loss metrics such as L_1 or L_2 error, we adopt the PSNR loss for our pixel-level evaluation. The loss term is defined as follows:

$$\mathcal{L} = -\mathrm{PSNR}(\mathcal{T}(H), \mathcal{T}(\hat{H})) \tag{3}$$

This PSNR function term is defined as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{4}$$

where MAX denotes the maximum possible value of the signal, and MSE stands for Mean Squared Error.

4. Experimental Results

4.1. Experimental Settings

Dataset NTIRE 2024 Challenge on Bracketing Image Restoration and Enhancement [48] contains 1,045 data pairs from 31 scenes as the training dataset and 290 pairs from the other four scenes as the test dataset. The low-quality multiple-exposure images are synthesized by applying various processing techniques and degradations to high-quality videos, including frame interpolation, conversion of RGB videos to raw space with Bayer pattern, 4× bicubic downsampling (optional) serving for involving super-resolution task, blur synthesis, and the introduction of noise. Each data pair includes five low-quality frames in raw space, and their corresponding ground truth is aligned with the first frame.

Network Details Our solution pipeline comprises TMR-NET [48] as the first-stage framework and NAFNet [8] as the second-stage network. Compared to the original TMR-NET [48], we only increased the channel number from 64 to 96. For the second stage, we follow the same settings of the model architecture in NAFNet [8].

Training Details In the first stage of our experiment, we use input patch sizes of 256×256 for the RT-IRE task and 64×64 for the RT-IRE+ task. In the second stage, the input patch size is the same for both tasks, at 512×512. The batch size is set to 4. We use the AdamW optimizer [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.01, and a cosine annealing strategy. During the first stage, the learning rate gradually decreases from an initial rate of 1×10^{-4} to 1×10^{-6} over 400 epochs. Similarly, in the second stage, the learning rate decreases from an initial rate of 3×10^{-5} to 1×10^{-6} over 600 epochs. All experiments use PyTorch [29] on a single Nvidia RTX A100 GPU. Additionally, we initialize model weights in RT-IRE+ using the weights from RT-IRE.

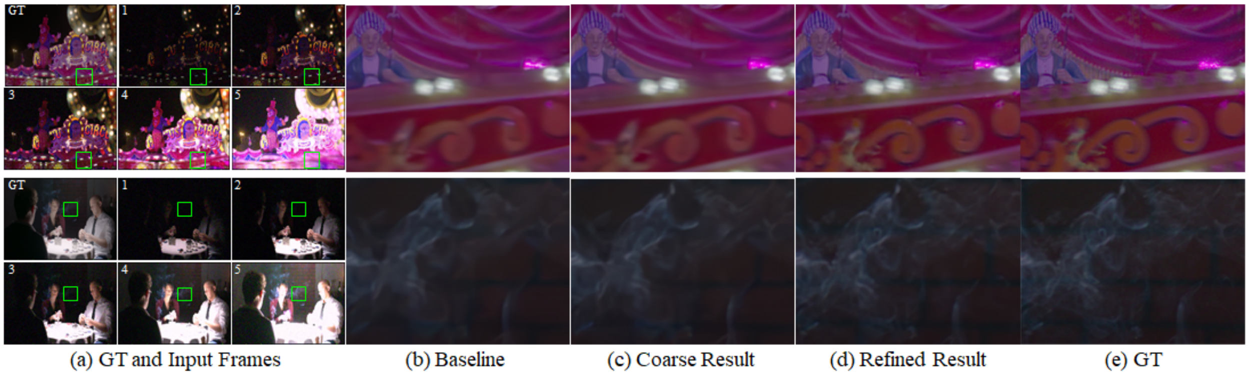


Figure 4. Visual comparison on Challenge dataset [49] of RT-IRE. Our refined results improve edge sharpness and smoke details.

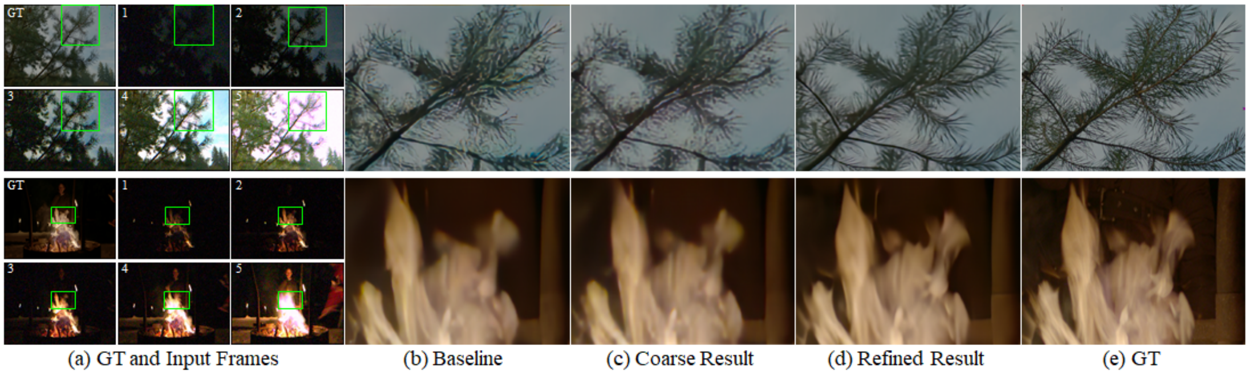


Figure 5. Visual comparison on Challenge dataset [49] of RT-IRE+. Our refined results restore more details in tree and fire.

Table 1. Quantitative comparison with PSNR for the full image.

Methods	Burstormer [12]	Kim *et al*. [19]	RT-IRE
PSNR	38.57	38.60	**38.64**

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

Reconstructing images by integrating information from multiple frames mainly involves feature extraction, alignment, and fusion. Currently, two main frameworks separate alignment and fusion, while the other integrates them. We have chosen one method from each of these two advanced frameworks for comparison: Burstormer [12] belonging to the former and Kim *et al*. [19] belonging to the latter. The Burstormer [12] based on Transformer leverages multi-scale features for alignment and feature fusion, while Kim *et al*. [19] utilizes attention mechanism [36] for implicit alignment and feature fusion from different frames. To ensure a fair evaluation, we adjusted their architectures to accommodate inputs with five frames and retrained them using the 2024 NTIRE Bracketing Image Restoration and Enhancement Challenge data. In addition, to eliminate the impact of the reference frame on the results, we use the same reference frame with the same order of 5-frame inputs. As shown in Tab. 1, by leveraging the powerful fusion feature capability of BasicVSR++ [5], simply by increasing the network size achieves a slight improvement compared to the other two frameworks. Moreover, visual comparison results in Fig 3 show that our method performs better in restoring details.

4.3. NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge Result

RT-IRE obtained two champions in the NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge [49]. Our model set a new record of 40.54 dB in the Tack 1 and 34.26 dB in the Tack 2. The competition results can be seen in Tab. 2 and Tab. 3. In Track 1, our method outperformed the second place by 0.76 dB and the baseline method by 2.35 dB. In Tack 2, our method outperformed the second place by 3.67 dB and the baseline method by 5.35 dB. The remarkable performance in the competition highlights the generalizability and effectiveness of RT-IRE.

4.4. Ablation Study

We started by measuring the impact of the proposed components by gradually integrating them into the baseline. As shown in Table 4, each component significantly improved the PSNR. Specifically, in RT-IRE, the PSNR increased from 0.2 dB to 2.35 dB, while in RT-IRE+, the increase ranged from 0.7 dB to 5.35 dB.

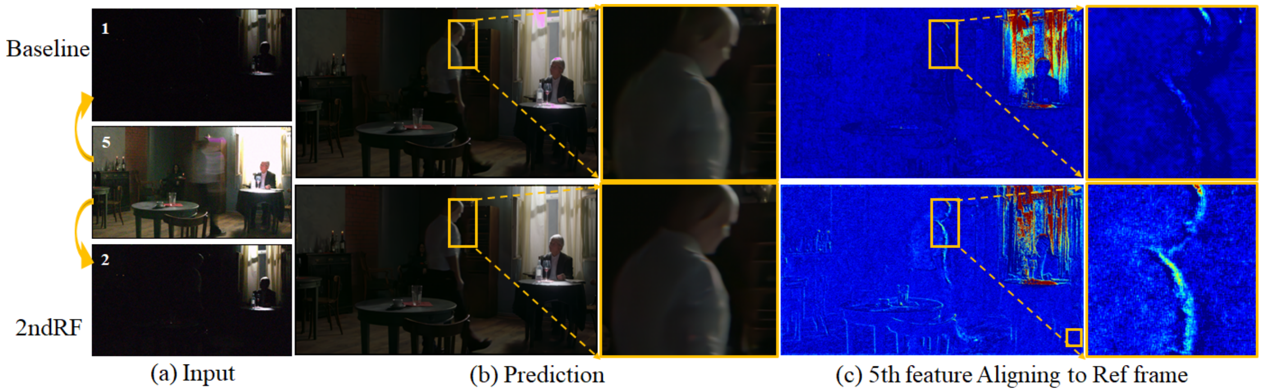


Figure 6. Visualization of feature maps in the alignment stage using different reference frames. The first row represents the baseline method using the first frame as the reference frame. The second row represents the 2ndRF method using the second frame as the reference image. The first column from top to bottom represents the first, second, and fifth frames. The second column displays the prediction results. The fourth column displays aligned features that align the fifth frame feature with the reference image.

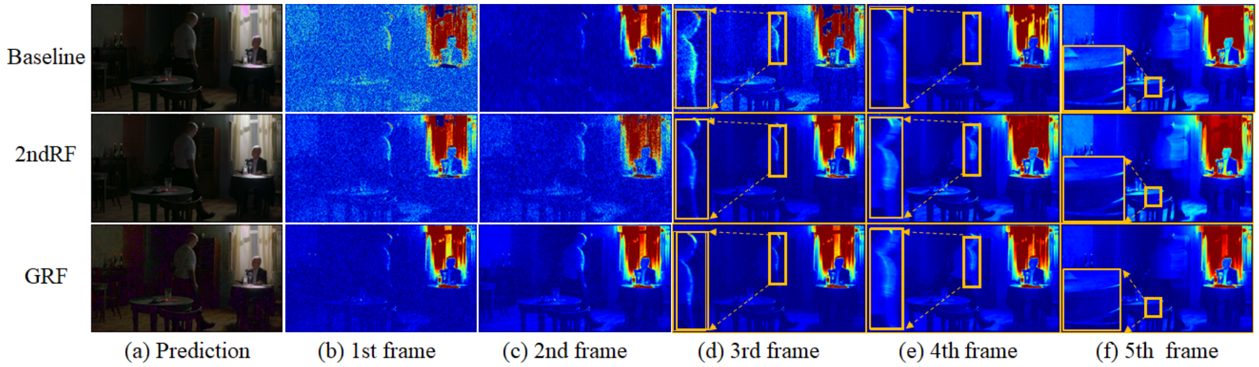


Figure 7. Visualization of feature maps in the fusion stage of different methods. The first row represents the baseline method using the first frame as the reference frame, the second row represents the second frame as the reference image, and the third row represents our proposed reference feature generation method. The first column represents the prediction. From the second column to the last column, the fused feature maps are shown from the first frame to the fifth frame.

Table 2. Quantitative results on Track 1 of Bracketing Image Restoration and Enhancement Challenge. #FLOPs and inference time are measured when generating a 1920 × 1080 RAW image. Using NVIDIA RTX A6000 GPU calculates the inference time and adopting THOP [51] toolkit calculates #FLOPs. The ranking is based on the PSNR metric of the full images.

Team	PSNR	#Params (M)	#FLOPs (T)	Time (s)
Samsung	40.54	94.34	48.238	3.102
MegIRE	39.78	19.75	30.751	2.383
UPN1	39.03	13.32	10.409	1.090
TMRNet [48]	38.19	13.29	21.340	1.874

Table 3. Quantitative results on Track 2 of Bracketing Image Restoration and Enhancement Challenge. #FLOPs and inference time are measured when generating a 1920 × 1080 RAW image. Using NVIDIA RTX A6000 GPU calculates the inference time and adopting THOP [51] toolkit calculates #FLOPs. The ranking is based on the PSNR metric of the full images.

Team	PSNR	#Params (M)	#FLOPs (T)	Time (s)
Samsung	34.26	95.00	5.285	0.813
NWPU	30.59	13.37	1.426	0.887
FZU_DXW	29.82	14.34	1.500	0.493
TMRNet [48]	28.91	13.58	1.441	0.489

Crop Black Border The images we use have a black border of 5 pixels. However, we remove this border before feeding the images into the network during training. The issue is that when we use the entire image during testing, we can obtain inaccurate boundary estimates. To tackle this problem, we crop the border of the input image during testing.

For the RT-IRE task, we remove five pixels from the edges of the input image. Next, we add five pixels to the input image by reflecting the existing pixels. Finally, to match the edge pixel values of the input, we set the 5-pixel widths of the output edge to zero. This process has been shown to significantly improve the metrics, as demonstrated in Table 4. The results show that RT-IRE achieves a 0.20 dB PSNR improvement.

Table 4. Accuracy on test images in the Challenge Dataset [49]. The results presented in the bracket are compared with the Baseline.

Methods						PSNR	
Baseline	Crop Black Border	GRF	2ndRF	96 Channels	Second Stage	RT-IRE	RT-IRE+
✓						38.19	28.91
✓	✓					38.39 (+0.20)	29.61 (+0.7)
✓	✓	✓				39.01 (+0.82)	30.04 (+1.13)
✓	✓		✓			39.12 (+0.93)	30.16 (+1.25)
✓	✓			✓		38.76 (+0.57)	30.64 (+1.73)
✓	✓		✓	✓		39.14 (+0.95)	31.36 (+2.45)
✓	✓		✓	✓	✓	**40.54 (+2.35)**	**34.26 (+5.35)**

For the RT-IRE+ task, we remove a 2-pixel border from the input and then add 2 pixels to the input by reflecting existing pixels. Lastly, we pad the output with 16 pixels instead of 5, which is necessary for the super-resolution to work effectively. This also significantly improves the metrics, as shown in Table 4, where RT-IRE+ obtains a 0.70 dB PSNR improvement.

Reference Feature We further provide some qualitative comparisons to understand the contributions of the proposed Reference Feature Generation: 2ndRF and GRF methods, which are more noticeable in regions containing fine details and complex textures in Figs. 6 and 7.

The 2ndRF mainly affects two parts of the network: alignment and fusion modules. The second frame contains less noise than the first, reducing optical flow errors between the current frame and the reference in the alignment module. As shown in Fig. 6, when aligning the fifth frame with a large amount of blur with the reference frame, the alignment feature obtained by 2ndRF is more precise at edges compared to the baseline. The aligned features of the baseline contain blurry edges because the warping of the fifth feature into the first frame is performed using the wrong optical flow. For the fusion module, the feature of the second frame with less noise can also provide more helpful guidance for feature fusion than the feature of the first frame. As shown in Fig 7, all fused features from the first to five frames of 2ndRF have less noise than the baseline. The noise from the first frame feature may interfere with the current feature fusion. As shown in Tab. 4, 2ndRF obtains a 0.93 dB PSNR improvement compared to the baseline.

The GRF algorithm generates a reference future to enhance the reference feature of the baseline. This is useful when there is significant motion between the high-quality input and ground truth frames. The GRF reference feature combines all frame features into a reference, which contains more information than selecting a single reference feature as the guiding feature. As illustrated in Fig. 7, the fused feature of GRF is much clearer than that of 2ndRF. However, GRF does not perform better than 2ndRF in the alignment module. Maybe that's why GRF has a lower score in RT-IRE and RT-IRE+ than 2ndRF as shown in Tab. 4. Therefore, we opted for the 2ndRF approach during the challenge to achieve a high score.

Two-Stage Structure As shown in Figs. 4 and 5, the refined output is better quality than the initial coarse result. The refined output has more details and clearer edges. As shown in Table 4, RT-IRE and RT-IRE+ have PSNR improvements of 2.35 dB and 5.35 dB, respectively, compared to the baseline. The two-stage method outperforms the one-stage method in both visual results and metric scores.

5. Conclusion

Existing multi-image processing methods usually focus only on restoration or enhancement, or only on SR or HDR reconstruction, which is insufficient for obtaining the high-quality image with sharp details in low-light conditions. The current bracket image restoration and enhancement method can be solved, but the fusion effect is seriously affected due to the reference image selection. This paper proposes a two-stage approach, firstly using multi-exposure images to generate high-quality reference features to guide feature fusion in order to get a clean image with a high dynamic range and then further enhancing and generating detailed textures in the second stage. Extensive experiments show that our model achieves the best performance. Our model won two champions in the NTIRE 2024 Bracketing Image Restoration and Enhancement Challenge.

Limitations In low light conditions, input images often have inherent noise that significantly affects the accuracy of the alignment module. The inaccurate alignment makes it difficult to make the most of the cues from multi frames, resulting in overly smoothed details as shown in Fig. 6. This, in turn, causes the feature fusion module to lose detailed information. Our GRF aims to merge complementary information from multiple features as the reference feature to improve feature fusion. However, it does not address the issue of inaccurate alignment caused by noise and blur. In the future, we explore the problem by either selecting more suitable reference frames or enhancing images before alignment, similar to the Clean Module in RealBasicVSR [6].

References

[1] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *ECCV*, pages 731–747, 2018. 2

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, pages 9209–9218, 2021. 2

[3] Qingwen Bu, Dong Huang, and Heming Cui. Towards building more robust models with frequency bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4402–4411, 2023. 5

[4] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 60–65. Ieee, 2005. 4

[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. 4, 6

[6] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 8

[7] Meng Chang, Huajun Feng, Zhihai Xu, and Qi Li. Low-light image restoration with short-and long-exposure raw pairs. *IEEE TMM*, 24:702–714, 2021. 2, 3

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33, 2022. 5

[9] Yiheng Chi, Xingguang Zhang, and Stanley H Chan. Hdr imaging with spatially varying signal-to-noise ratios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5724–5734, 2023. 2

[10] Uğur Çoğalan, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Hdr denoising and deblurring by learning spatio-temporal distortion models. *arXiv preprint arXiv:2012.12009*, 2020. 3, 4

[11] 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. 4

[12] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burstormer: Burst image restoration and enhancement transformer. In *CVPR*, pages 5703–5712. IEEE, 2023. 2, 6

[13] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *ECCV*, pages 538–554, 2018. 2

[14] Shi Guo, Xi Yang, Jianqi Ma, Gaofeng Ren, and Lei Zhang. A differentiable two-stage alignment scheme for burst image reconstruction with large shift. In *CVPR*, pages 17472–17481, 2022. 2

[15] Shi Guo, Xi Yang, Jianqi Ma, Gaofeng Ren, and Lei Zhang. A differentiable two-stage alignment scheme for burst image reconstruction with large shift. In *CVPR*, pages 17472–17481, 2022. 2

[16] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM TOG*, 35(6):1–12, 2016. 2, 4

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

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

[19] 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. 6

[20] 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 TOG*, 41(4):1–16, 2022. 2

[21] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *ICCV*, pages 2370–2379, 2021. 2

[22] 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. 2

[23] Shuaizheng Liu, Xindong Zhang, Lingchen Sun, Zhetong Liang, Hui Zeng, and Lei Zhang. Joint hdr denoising and fusion: A real-world mobile hdr image dataset. In *CVPR*, pages 13966–13975, 2023. 3

[24] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *ECCV*, pages 344–360. Springer, 2022. 3, 4

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[26] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, pages 2502–2510, 2018. 2

[27] Janne Mustaniemi, Juho Kannala, Jiri Matas, Simo Särkkä, and Janne Heikkilä. lsd_2 - joint denoising and deblurring of short and long exposure images with cnns. In *BMVC*, 2020. 2

[28] 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*, 30: 3885–3896, 2021. 3, 4

[29] 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. 5

[30] Fidel Alejandro Guerrero Peña, Pedro Diamel Marrero Fernández, Tsang Ing Ren, Jorge de Jesus Gomes Leandro, and Ricardo Massahiro Nishihara. Burst ranking for blind multi-image deblurring. *IEEE TIP*, 29:947–958, 2019. 2

[31] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. Ntire 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *CVPRW*, pages 691–700, 2021. 3

[32] Xuejian Rong, Denis Demandolx, Kevin Matzen, Priyam Chatterjee, and Yingli Tian. Burst denoising via temporally shifted wavelet transforms. In *ECCV*, pages 240–256. Springer, 2020. 2

[33] Shayan Shekarforoush, Amanpreet Walia, Marcus A Brubaker, Konstantinos G Derpanis, and Alex Levinshtein. Dual-camera joint deblurring-denoising. *arXiv preprint arXiv:2309.08826*, 2023. 2

[34] Xiao Tan, Huaian Chen, Kai Xu, Yi Jin, and Changan Zhu. Deep sr-hdr: Joint learning of super-resolution and high dynamic range imaging for dynamic scenes. *IEEE Transactions on Multimedia*, 25:750–763, 2021. 2

[35] 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*, pages 12836–12845, 2023. 3

[36] 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. 4, 6

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4

[38] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *CVPR*, pages 2758–2767, 2020. 2

[39] 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 *ICCV*, pages 13233–13242, 2023. 2

[40] Patrick Wieschollek, Bernhard Schölkopf, Hendrik PA Lensch, and Michael Hirsch. End-to-end learning for image burst deblurring. In *ACCV*, pages 35–51. Springer, 2017. 2

[41] 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. 5

[42] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *CVPR*, 2020. 2

[43] Zhiqin John Xu and Hanxu Zhou. Deep frequency principle towards understanding why deeper learning is faster. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10541–10550, 2021. 5

[44] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

[45] Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *arXiv preprint arXiv:2201.07395*, 2022. 5

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 5

[47] Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. A unified hdr imaging method with pixel and patch level. In *CVPR*, pages 22211–22220, 2023. 3

[48] 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, 2, 3, 4, 5, 7

[49] 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. 5, 6, 8

[50] Yuzhi Zhao, Yongzhe Xu, Qiong Yan, Dingdong Yang, Xuehui Wang, and Lai-Man Po. D2hnet: Joint denoising and deblurring with hierarchical network for robust night image restoration. In *ECCV*, pages 91–110. Springer, 2022. 2, 3

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