Improving Dynamic HDR Imaging with Fusion Transformer

Rufeng Chen[1], Bolun Zheng[1*], Hua Zhang[1*], Quan Chen[1], Chenggang Yan[1], Gregory Slabaugh[2], Shanxin Yuan[2]

[1]Hangzhou Dianzi University, Xiasha No.2 Street, Hangzhou, 310018, Zhejiang, China
[2]Queen Mary University of London, London, UK
{chenrufeng, blzheng, zhangh, chenquan, cgyan}@{hdu.edu.cn}
{g.slabaugh, shanxin.yuan}@{qmul.ac.uk}

Abstract

Reconstructing a High Dynamic Range (HDR) image from several Low Dynamic Range (LDR) images with different exposures is a challenging task, especially in the presence of camera and object motion. Though existing models using convolutional neural networks (CNNs) have made great progress, challenges still exist, *e.g.*, ghosting artifacts. Transformers, originating from the field of natural language processing, have shown success in computer vision tasks, due to their ability to address a large receptive field even within a single layer. In this paper, we propose a transformer model for HDR imaging. Our pipeline includes three steps: alignment, fusion, and reconstruction. The key component is the HDR transformer module. Through experiments and ablation studies, we demonstrate that our model outperforms the state-of-the-art by large margins on several popular public datasets.

Introduction

Dynamic range is used to define the ability of the camera to capture a range of brightness, usually between the lowest and highest values of the same image. Scenes with a large differences in lighting may pose a challenge to capture. If the dynamic range is not large enough and the illumination is too bright, an overexposed image will be produced; and if the scene is too dark, the image will appear underexposed. Both over- and under-exposure will lead to loss of details in the image. While most sensors can record 8-bit, 10-bit, or slightly higher depth images, those that can record 16-bit depth images are too expensive to be widely used in everyday devices, and standard displays only support 8 bit prompting the need for HDR imaging (Meylan, Daly, and Süsstrunk 2006; Dong et al. 2021).

Initial work performing high dynamic range restoration using a single LDR image (An, Ha, and Cho 2012; Akyüz et al. 2007; Banterle et al. 2007; Huo et al. 2014; Rempel et al. 2007; Eilertsen et al. 2017; Endo, Kanamori, and Mitani 2017; Lee, An, and Kang 2018; Zheng et al. 2022b) showed the dynamic range of the image can be extended, but the under- or over-exposed regions are unrecoverable. Therefore researchers began to explore using multiple LDR images at different exposures (e.g. short, medium, long).

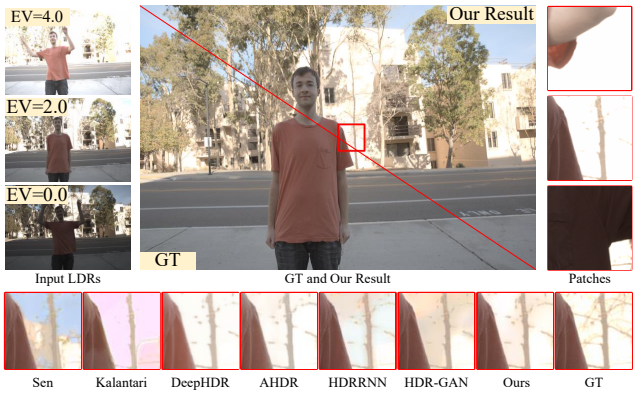

Figure 1: Three LDR images with different exposures are located on the left side. The image in the middle is our result and ground-truth (GT). EV denotes exposure value, which is determined by the exposure time, ISO and f-number.

The task is to synthesize a single HDR image that preserves the details of the scene using multiple LDR images (Debevec and Malik 2008; Jacobs, Loscos, and Ward 2008; Reinhard et al. 2010; Sen et al. 2012; Dai et al. 2021; Li et al. 2022).

However, capturing multiple low dynamic range images at different exposures poses some challenges. This typically requires capturing multiple exposures at different times using a single sensor. However motion due to the camera or objects in the scene will result in misalignment between LDR images. If unaddressed, this misalignment will result in obvious ghosting artifacts (Bogoni 2000; Li et al. 2020; Ma et al. 2017; Zheng et al. 2019) in the merged HDR image.

To solve this problem, many recurrent networks and lightweight networks have been proposed, such as AHDRNet, HDRRNN, Kalantari, NHDRRNet (Yan et al. 2019; Prabhakar, Agrawal, and Babu 2021; Kalantari, Ramamoorthi et al. 2017; Yan et al. 2020). All of these models aim to build higher-performing architecture, and follow a similar design of LDR CNN-based alignment and fusion to reconstruct the HDR image. At present, the proposed methods are aimed at the alignment between images, the reconstruction of HDR images, and the use of various model structures of recurrent neural networks through attention orientation,

*Corresponding author: Bolun Zheng and Hua Zhang.

but they cannot handle the task of LDR-to-HDR well as unresolved motion results in ghosting artifacts, blurring, and color defects. Due to the specific nature of this task, using transformer (Vaswani et al. 2017; Dosovitskiy et al. 2020), which has recently received much attention in computer vision, can be difficult due to hardware and GPU memory limitations. However, traditional convolutional neural network themselves have limitations in terms of receptive field.

In order to solve the above problems, this paper proposes a multi-exposure LDR-to-HDR converter called HDR Fusion Transformer (HFT), which uses a transformer to capture long-range context dependence while ensuring the support of hardware devices. Notably, it uses a "CNN+Transformer" architecture. Specifically, HFT can be divided into three parts: a Shallow Feature Alignment (SFA), a Pyramid Fusion Module (PFM), and an Image Reconstruction Module (IRM). For SFA, more attention is paid to reducing the depth of the features of the middle layer and using Deformable Convolution (DCN) (Dai et al. 2017; Zhao et al. 2021) to correct the problem of alignment between images. In the PFM, a HDR Fusion Module (HFM) is used in three scales and HDR Transformer (HT) is used in the smallest scale, which simultaneously removes the decoder part of traditional transformer to reduce GPU memory consumption by using a multi-head attention mechanism. It is worth noting that HT takes into account the global information of the image, and it has a much larger receptive field than conventional convolution and can extract more useful contextual information. Therefore, HFT can effectively repair the fused defects through long-distance features after image fusion, making it more competitive. For image reconstruction, an novel Channel Attention Dilated Block (CADB) is proposed as the basic feature extraction unit, which can adaptively adjust the weight of each channel to eliminate the ghost caused by misalignment. The main contributions are as follows:

- We propose a new Pyramid Fusion Module (PFM) with Transformer. The HDR Fusion Module (HFM) fuses the higher scale features; while the smallest scale features are fused with Self-Attention Fusion (SAF) which includes a lightweight transformer. With this approach, the features can be fused with less computation and according to the global information.
- We propose a Channel Attention Dilated Block (CADB) to reduce ghosting artifacts.
- We propose HDR Fusion Transformer (HFT), which can better learn non-local features for the HDR fusion.

Related Work

CNN-Based HDR Models

CNNs have been widely use for image restoration (Zhao et al. 2021; Liu et al. 2020b,a; Isobe et al. 2020). Recently, many CNN-based models have been proposed for HDR. For example, Kalantari *et al*. (Kalantari, Ramamoorthi et al. 2017) use an optical flow algorithm to compensate for motion and merge the resulting images using a simple CNN. ADNet (Liu et al. 2021) was proposed to align the dynamic frames with a deformable alignment module. Wu *et al.* (Wu et al. 2018) use homographies to align the background motion prior to fusion. Yan *et al*. use spatial attention to rule out misaligned components and build a deep convolutional neural network to merge features. Prabhakar *et al*. (Prabhakar, Agrawal, and Babu 2021) propose a scalable CNN architecture to efficiently handle the varying LDR inputs. In addition, earlier work by Prabhakar *et al*. (Prabhakar et al. 2020) uses the optical flow of aligned images prior to fusion. Niu *et al*. (Niu et al. 2021) use a GAN to create images and video from the adjustable part of a data stream based on an event camera. Zhang *et al*. (Zhang and Lalonde 2017) use a depth self coding architecture to regress linear and high dynamic range panoramic images from nonlinear, saturated and low dynamic range panoramic images.

Vision Transformer

Transformers, which started out in natural language processing, have made a major breakthrough in NLP, leading the computer vision community to consider their application. Transformer's core idea is the multi-head self-attention mechanism, which can capture long-range information without the limitation of narrow receptive field in traditional CNNs. The pioneering work of the Vision Transformer (ViT) demonstrated the potential of transformers to replace traditional CNNs, by representing 2D image features into a one-dimensional sequence, which can be fed into a Transformer. Transformers have been fully used in image classification (Li et al. 2021; Touvron et al. 2021), target detection (Carion et al. 2020; Wang et al. 2022), super-resolution (Lu et al. 2022; Liang et al. 2021; Yang et al. 2020) and other tasks (Qu et al. 2022; Bai et al. 2022; Liu et al. 2022b,a, 2023). However, in the LDR-to-HDR task, transformers have not yet been applied due to the limitation of hardware devices and insufficient GPU memory. Our goal was to develop an effective HDR Fusion Transformer for LDR-to-HDR reconstruction.

Proposed Method

As shown in Figure 2, the HDR Fusion Transformer (HFT) is mainly composed of three parts: Shallow Feature Alignment (SFA), the Pyramid Fusion Module (PFM), and the Image Reconstruction Module (IRM). We define [L_{-1},L_0,L_1] and H as the input and output, where L_0 is the reference image, L_{-1},L_0 and L_1 are the supporting images. In SFA, the features of the two supporting images are aligned with the features of the reference image.

$$F_i = \text{SFA}(L_r, L_i) \tag{1}$$

where L_r denotes the reference image, L_i denotes the supporting image, SFA denotes the shallow feature alignment layer. F_i is the extracted shallow aligned feature, which is then used as the input to the PFM module.

$$P = \text{PFM}(F_{-1}, F_0, F_1) \tag{2}$$

where P denotes the fused features, which are are sent to the IRM for HDR image reconstruction.

$$H = \text{IRM}(P) \tag{3}$$

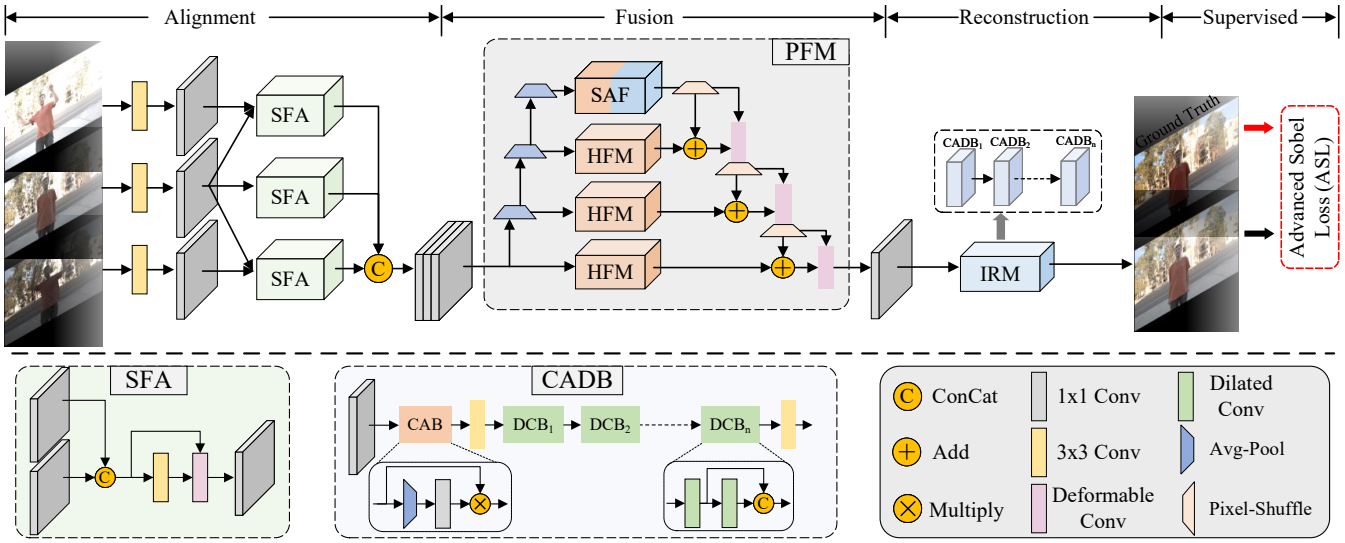

Figure 2: The architecture of the proposed HDR Fusion Transformer (HFT). Among them, SFA, PFM, IRM and CADB stand for the Shallow Feature Alignment, Pyramid Fusion Module, Image Reconstruction Module and Channel Attention Dilated Block, respectively.

Shallow Feature Alignment

The Shallow Feature Alignment (SFA) module aligns the features of L_r and L_i through deformable convolution (Liu et al. 2021). Due to the motion between the reference image and supporting image, it is necessary to align the features to minimize ghosting effects.

$$y\left(\mathbf{p}_0\right)=\sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}\left(\mathbf{p}_n\right) \cdot x\left(\mathbf{p}_0+\mathbf{p}_n\right) \tag{4}$$

where y is the result of the convolution, $\mathbf{p_n}$ is the nth pixel and $\mathbf{w}$ is the convolutional kernel, x is the input feature and $\mathcal{R}=\{(-1,-1),(-1,0),...,(0,1),(1,1)\}$ denotes the sampling area at a shifted pixel. However, the traditional convolutions as shown in Eq. 4 are limited by the size of receptive field, which struggles with longer range dependencies. Therefore, we add a learnable offset $\Delta\mathbf{p}_n$ to learn more complete information. The convolution model with the offset can be expressed as:

$$y\left(\mathbf{p}_0\right)=\sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}\left(\mathbf{p}_n\right) \cdot x\left(\mathbf{p}_0+\mathbf{p}_n+\Delta\mathbf{p}_n\right) \tag{5}$$

where $\Delta\mathbf{p}_n$ denotes the offset to be learned.

After the initial alignment, the aligned features are concatenated with L_r. The concatenated features are sent to PFM.

Pyramid Fusion Module

The second difficulty in synthesizing an HDR image is to combine the features of three different exposures. Therefore, we propose a new pyramid model (PFM), which is more capable of high quality fusion than other models.

In HFT, patches surrounding the image can be used as a reference image, so that the real details of the current image block can be fused using the reference features around the image. Due to the large size of the image, we adopt a multi-scale fusion mechanism, which can effectively reference information around the image. In addition, although traditional transformers such as ViT have been applied in the field of computer vision, there is no transformer suitable for HDR reconstruction because of the high GPU memory required. In this paper, an novel lightweight Self-Attention Fusion (SAF) module based on transformer is proposed, and its effectiveness is proved.

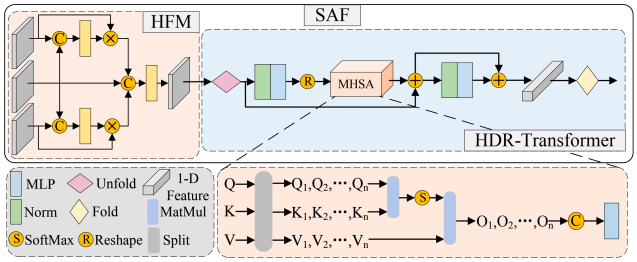

Figure 3: The architecture of the proposed modeule Self-Attention Fusion (SAF) in PFM, which is composed of the HDR Fusion Module and HDR Transformer, respectively. The bottom in the figure is Multi-Head Self Attention.

Figure 2 shows the architecture of the PFM. The input feature F_i is downsampled to a smaller scale, and F_i is fused to the feature through the HFM module. Between two different scales, the input features are down-sampled through average pooling, the processed features (through HFM or SAF) up-sampled through bicubic interpolation. After up-sampling, the feature of the smaller scale S_i is added to the feature B_{i-1} of the larger scale, then the combined features are sent to deformable convolution. HFM and SAF are used for fusion in the first three scales and the last scale, respectively.

The main purpose of SAF is to refine the fused feature by capturing the long-range features after image fusion, to achieve the best fusion effect. Due to the computational cost, the transformer is only used in the coarsest scale. We improve the original transformer as follows: we 1) remove the decoder part of the transformer, 2) simplify the encoder part, and 3) use only features at the bottleneck in the transformer and retain these features.

HDR Fusion Module (HFM) The HFM module's structure is shown in Figure 3. The main purpose of HFM is to initially fuse the features of different exposed images.

The outputs of SFAs are taken as the input of HFM in the first scale. In HFM, F_0 is concatenated with F_i ($i = -1, 1$) respectively, and then the concatenated features are sent into the convolution layer. We speculate that exposure is affected by ambient brightness and belongs to global information. Therefore, in order to capture the missing information of the reference image, we multiply F_i and the convolution features to F_i obtaining the refined feature F_i^t. After the F_{-1}^t and F_1^t are concatenate with F_0, they enter the convolution layer, and H_a is the output of HFM.

$$F_i^t = F_i \cdot \text{Conv}(\text{Cat}(F_i, F_0))|_{i=-1,1} \tag{6}$$

$$H_a = \text{Conv}(\text{Cat}(F_{-1}^t, F_0, F_1^t)) \tag{7}$$

HDR Transformer (HT) ViT divides the two-dimensional image into several small patches and combines them into a one-dimensional representation. This allows the transformer to be applied to visual tasks, but has some disadvantages such as a large demand for training data and a large amount of calculation.

Inspired by ESRT (Lu et al. 2022), we apply an unfold operation to the feature map in HDR-Transformer, as shown in the upper right corner of Figure 3. This turns the original two-dimensional features into a one-dimensional sequence H_o. Considering the shortcomings of VIT (Yu et al. 2022), we focus on reducing the computational burden and complexity of the transformer in the visual domain. We directly map the features into 1D features, which greatly reduces the computation and ensures that the hardware can support subsequent processing. And when used at the smallest scale, it can ensure that the GPU memory can meet its needs. Then, the method of up-sampling using a pixel-shuffle, residual structure and deformable convolution is used to ensure that sufficient information can be retained, so that satisfactory feature information can be obtained.

The traditional transformer requires heavy computation. For the sake of simplicity and efficiency, we remove the decoder part in our HT, instead only retaining the encoder structure as shown in the bottom right of Figure 3. This plot shows the multi-head self-attention module, layer normalization, and MLP. Although batch normalization has beneficial effects when dealing with two-dimensional features, layer normalization is preferred after the two-dimensional features are compressed into one dimension.

In the HDR-Transformer, the 2D features are transformed into 1D sequential features through the unfolding operation and become H_n. As shown in Figure 3, H_n go through layer normalization and MLP layer to obtain three categories of Q, K, V, and each category was segmented into N pieces again. which enters the multi-head self-attention module.

These serve as inputs for the scaled dot-product attention module. The output H_m is multiplied by Q and K and then multiplied by V. After passing through a concatenation, H_m are output of the multi-head self-attention module through the full connection layer. After residual operations are performed on H_n and H_m in the HT,then go through layer normalization and MLP layer, and then perform residual operation with the previous residual result to obtain H_s. The one-dimensional feature H_s output from HT is folded to the two-dimensional feature of the corresponding size image. The specific formula is as follows.

$$Q, K, V = \text{MLP}(\text{Norm}(\text{Unfold}(H_a))) \tag{8}$$

$$S_v = \text{MLP}(\text{SoftMax}(Q \cdot K^T) \cdot V) + H_n \tag{9}$$

$$H_s = \text{Fold}(\text{MLP}(\text{Norm}(S_v)) + S_v) \tag{10}$$

Image Reconstruction Module

After the initial alignment of features, although the ghosting is greatly reduced, it is inevitable that there will be local misalignments that may result in residual ghosting. At the same time, the high dynamic range image must be reconstructed. Image Reconstruction Module (IRM) is designed to address these challenges.

As shown in Figure 2, the IRM is composed of several Channel Attention Dilated Block (CADB) modules whose purpose is to eliminate the small amount of ghosting caused by residual misalignment and reconstruct the HDR image.

In the process of training, HFT cannot effectively distinguish the ghosting caused by feature misalignment from the aligned part of the real scene. CADB can effectively reduce the influence of the features learned from the ghosting in the model at the level of feature channel, so that it can eliminate the ghosting that cannot be solved through alignment to the greatest extent.

The CADB structure is shown in the bottom of Figure 2, with H_s as the input. First, global pooling is carried out to obtain the weight of channels at each layer, and then feature extraction is performed. The extracted feature H is used as the weight of channels at each layer and attached to H_s to achieve the effect of reducing the weight of virtual shadows and obtain ghost-free feature H_w. The formula is as follows:

$$H_w = \text{Conv}(\text{AvgPool}(H_s)) * H_s \tag{11}$$

Then, the features are sent into several dilated residual modules. Due to the small receptive field of the common convolution layer, some local patches of the HDR image require a large range of information for reconstruction. Therefore, dilated convolution is used for concatenation after traditional convolution is applied, and helps make full use of local and non-local feature information. This effectively expands the receptive field and better reconstructs the details of under- and over-exposed regions producing high quality results.

Experiments

Training Loss

Since HDR images are usually displayed after tone mapping, it is more effective to train the network on a tone mapped image than directly in the HDR domain. Given the HDR image H in the HDR domain, we use the μ-law to compress H within the range of [0,1], with μ=5000.

$$\mathcal{T}(H) = \frac{\log(1+\mu H)}{\log(1+H)} \tag{12}$$

where μ is the parameter that defines the amount of compression, and $\mathcal{T}(H)$ represents the tone mapped image.

To train our HFT, we adopt the L1 loss as the base loss function. However, as the L1 loss is a point-wise loss, it does not capture edge information important particularly to minimize ghosting in the HDR reconstruction. Following the novel training strategy of CNN (Zheng et al. 2020, 2021, 2022a), we adopt the Advanced Sobel Loss (ASL) and combine it with the L1 loss to formulate the loss function for to enhance the edge information, which can be expressed as:

$$\text{Loss}(\widehat{Z}, Z) = \mathcal{L}1(\widehat{Z}, Z) + \frac{1}{N}\sum_{N}^{i} \mathcal{ASL}_i\left(\widehat{Z}, Z\right) \tag{13}$$

where $\widehat{Z}$ and Z represent the generated HDR image and the ground truth (GT) image respectively, $\mathcal{L}1$ represents the L1 Loss, N represents the number of Sobel Loss kernels, and $\mathcal{ASL}$ represents Advanced Sobel Loss function. Here four convolution kernels are used to optimize edge information on $\widehat{Z}$ and Z in four directions. The specific optimization process is as follows:

$$\mathcal{ASL}_i\left(\widehat{Z}, Z\right) = \mathcal{L}1(\mathcal{K}_i(\widehat{Z}), \mathcal{K}_i(Z)) \tag{14}$$

where $\mathcal{K}_i$ denotes Sobel loss kernel (Vincent, Folorunso et al. 2009; Gao et al. 2010).

Implementation and Details

Due to the multi-scale architecture, the model must be a multiple of 16, so we zero-pad the image as necessary. A 64-channel, 3×3 convolution kernel is used in the Conv layer. We use an Adam optimizer, with the initial learning rate set to 10^{-4}. LDR images and corresponding images were split into patches of 128×128 size for training, however validation and test images were full resolution. In order to avoid over-fitting in the training stage, patches were randomly rotated for data augmentation. During training, we measured the validation set at using PSNR-μ. If the model performance was not improved after five epochs, the learning rate is halved. When the learning rate is less than 10^{-6}, the training ends. We implemented our HFT using Pytorch on single NVIDIA RTX3090 GPU.

Comparisons with Advanced HDR Models

Datasets and Metrics Kalantari's dataset is used as the basic training data set. All models are trained on this dataset.[1] In addition, we performed supplementary experiments using the Prabhakar's dataset (Prabhakar et al. 2019)

[1]https://cseweb.ucsd.edu/ viscomp/projects/SIG17HDR/

Figure 4: Visual comparisons on the testing data from Kalantari's dataset. We compare the zoomed-in local areas of the HDR images generated by our method with seven other methods, namely Sen, SPD-MEF, Kalantari, DeepHDR, AHDR, HDRRNN and HDR-GAN.

to prove the generalization of the proposed model. Tursun's (Tursun et al. 2016) dataset are widely used in the related studies, which is a dataset without ground true. The PSNR and SSIM of predicted HDR images were measured in the linear domain (-L) and HDR domain (-μ) for quantitative evaluation. We also used HDR-VDP-2 (Mantiuk et al. 2011; Marnerides et al. 2018) as another indicator in the comparison.

We compare our results with previous state-of-the-art methods, including three patch-based methods Sen, Hu, SPD-MEF (Sen et al. 2012; Hu et al. 2013; Ma et al. 2017) and eight CNN based methods, Kalantari (Kalantari, Ramamoorthi et al. 2017), DeepHDR (Wu et al. 2018), AHDRNet (Yan et al. 2019), HDR-GAN (Niu et al. 2021), Prabhakar (Prabhakar et al. 2020), HFNet (Xiong and Chen 2021), HDRRNN (Prabhakar, Agrawal, and Babu 2021). We generated Sen, SPD-MEF, DeepHDR, AHDRNet, HDRRNN, HDR-GAN results and compared them, and we replicated them for all other methods (if open source was available). For methods without publicly available code we did not reproduce their results and instead use the results reported in their papers. The HDR-VDP-2 score is only for reference because it may vary depending on the parameters assessed and is not described in the details of the paper. Quantitative evaluation was calculated for five indicators.

Quantitative evaluations: as shown in the Table 1, our HFT achieves excellent performance across all indicators, and reaches a new SOTA on PSNR-μ, PSNR-L and SSIM-μ. Based on Kalantari's benchmark dataset, the performance of HFT is qualitatively compared with other models. As shown in Figure 4, our experimental results show that HFT is better than other methods in reconstructing the dynamic range including darker and brighter regions, better preserving details and colors more realistically matching the ground truth.

Experiments on Additional Datasets In addition to Kalantari's dataset, we also performed experiments on Prab-

Methods	Publication	Quantitative Results					Computational Costs	
		PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	HDR-VDR-2	Params(M)	Time(s)
Sen	TOG 2012	40.95	38.31	0.982	0.972	56.72	-	73.41
Hu	CVPR 2013	32.18	31.88	0.970	0.969	55.24	-	103.57
SPD-MEF	TIP 2017	43.34	40.77	0.986	0.986	61.84	-	13.29
Kalantari	TOG 2017	42.74	41.22	0.988	0.985	60.51	-	-
DeepHDR	ECCV 2018	41.65	40.86	0.986	0.986	61.21	13.57	0.28
AHDRNet	CVPR 2019	43.62	41.03	0.990	0.989	62.30	1.24	0.94
Prabhakar	ECCV 2020	43.08	41.68	-	-	62.21	-	-
HDR-GAN	TIP 2021	43.92	41.57	0.991	0.987	65.45	-	-
HFNet	ACM MM 2021	44.28	41.48	-	-	62.33	2.70	-
HDRRNN	TCI 2021	42.82	41.68	0.990	0.990	-	-	0.47
HFT	-	**44.45**	**42.14**	**0.992**	0.988	66.32	4.19	0.10

Table 1: Quantitative results of our HFT method compared with other advanced methods on Kalantari's dataset.

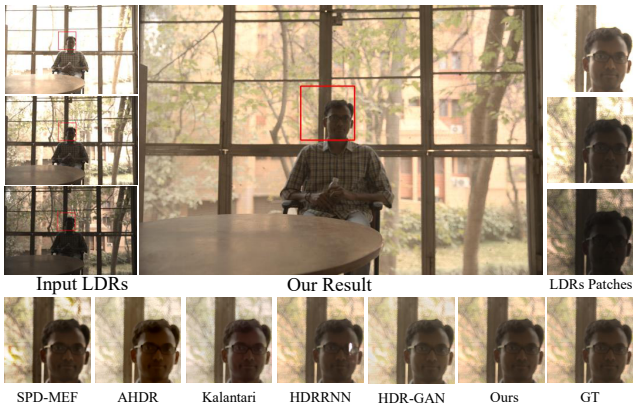

Figure 5: Visual comparison on Prabhakar's dataset.

hakar's dataset[2]. Quantitative results are shown in Table 2 and qualitative results are shown in Figure 5 and Figure 6. 'Our Cross Result' denotes the model is trained on Kalantari's dataset. The results on Prabhakar's dataset are similar to those on Kalantari's dataset, with our method outperforming others.

Additionally, we also provide the results of all compared results on our hand-captured images in Figure 7. From the visual comparison, only our method is able to produce a ghosting-free result, while the ghosting effects more or less exist in results produced by the other compared methods.

Ablation Study

We verify the key parts of the HFT model, including (1) the initial alignment is improved by using the SFA module compared with concatenating the two LDR images directly; (2) the use of the SAF module in the fourth scale of PFM instead of the original HFM module; and (3) CADB has the advantage over ordinary DRDB in the IRM module. We also

[2]https://val.cds.iisc.ac.in/HDR/ICCP19/, MIT License

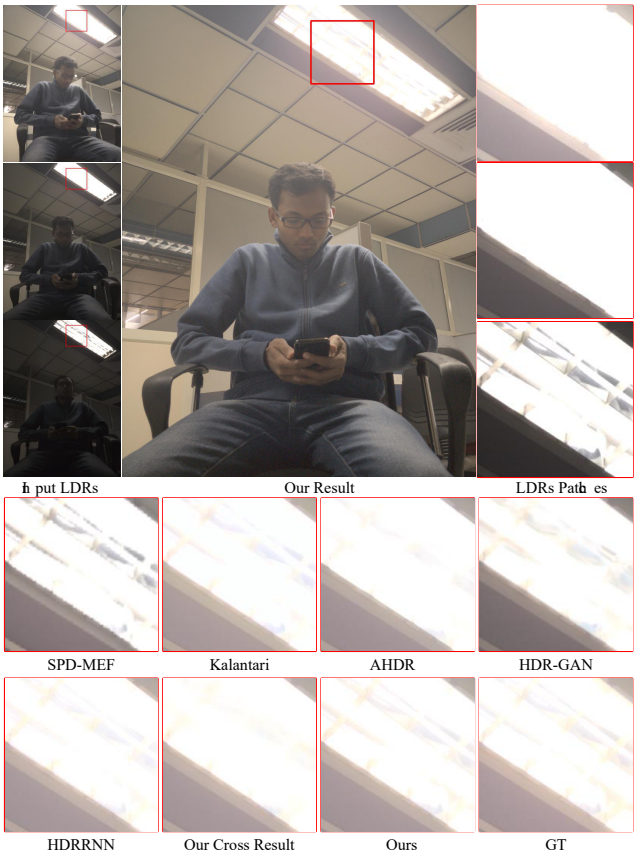

Figure 6: Visual comparison on Prabhakar's dataset.

explore the number of CADB in IRM required, and the gap between the L1 and DSL in the loss function.

As shown in Table 3, we quantitatively compared the effects of the three modules relative to the whole model, listing eight cases respectively in Kalantrai's and Prabhakar's dataset. The SFA, HT, and CADB are all required to achieve

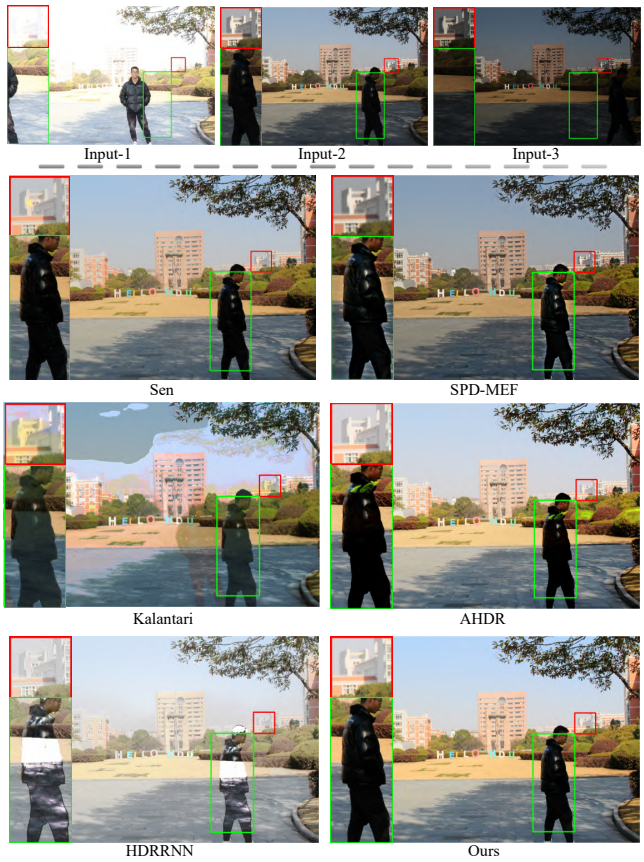

Figure 7: Visual comparison on our hand-captured image.

Methods	Quantitative Results			
	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L
Kalantari	35.63	32.50	0.961	0.969
DeepHDR	38.03	34.40	0.971	0.977
AHDR	38.65	35.28	0.973	0.980
SCHDR	36.08	32.74	0.959	0.967
Prabhakar	38.30	34.98	0.970	0.978
HDRRNN	39.03	36.38	**0.975**	0.983
HFT	**40.12**	**37.64**	0.971	**0.987**

Table 2: Quantitative comparison on Prabhakar's dataset.

the best results. Figure 8 provides visual results of our ablation studies. We compared all combinations of the three modules to objectively demonstrate the role of each module and prove its value.

Loss Function Table 4 shows the results of using the Kalantari's dataset on several IRM modules using different loss functions (L1 or ASL). As shown in Table 4, we conducted quantitative ablation experiments on the number of CADB in IRM and the loss function used. Under the condition that other parameters remain unchanged, the ASL loss

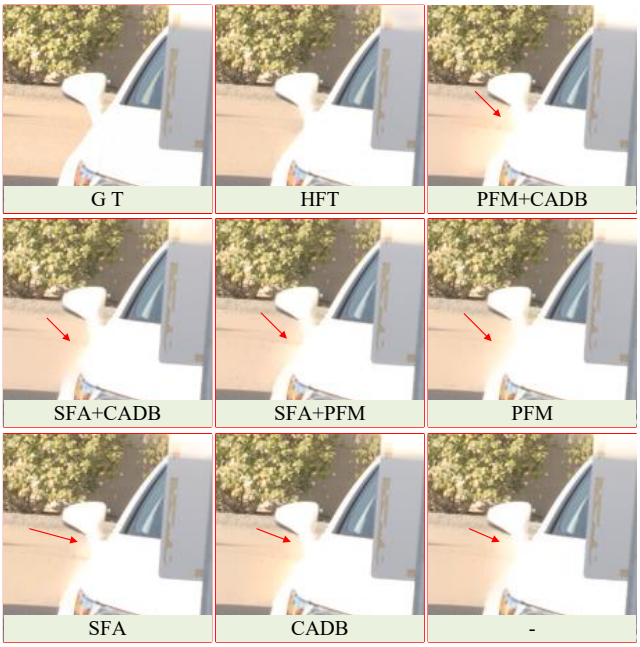

Figure 8: The comparison of seven ablation experiments with GT, HFT's result and all comparison methods in ablation experiments. The lower part of the image indicates the included module.

outperforms the L1 loss, and the model achieves the best results when the number of CADBs is 3. Qualitative results shown in Figure 9 demonstrate that compared to the L1 loss, the ASL better preserves details and more faithfully reconstructs colors.

Shallow Feature Alignment The main purpose of Shallow Feature Alignment module is to better align the moving image to the reference image and reduce ghosting as much as possible in the initial stage. We perform an ablation study by removing the SFA and replacing it with an ordinary concatenation of the two input feature maps.

HDR Transformer The purpose of using HDR Transformer at the smallest scale is to repair the defects of the fused features with long-distance information. We conducted an ablation study which removes the HT at the minimum scale.

Channel Attention Dilated Block Our IRM module is composed of several CADBs, which reconstruct the HDR image and eliminate ghosting. We performed a study to explore the number of CADBs on the performance. As shown in Figure 10, we compare the changes of features when they pass through CADB, and intuitively show that the CADB is successful in eliminating ghosting artifacts which caused by misalignment of features.

Conclusion

In this paper, we propose a combined CNN and Transformer model for end-to-end HDR generation. Input multi-exposure LDR images are combined to produce a high quality HDR

Modules			Kalantari's				Prabhakar's			
SFA	HT	CADB	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L
-	-	-	44.06	41.83	0.991	0.987	39.03	36.60	0.970	0.981
-	✓	-	44.19	41.90	0.991	0.988	39.45	37.31	0.971	0.983
✓	-	-	44.23	41.71	0.991	0.988	39.14	36.86	0.970	0.980
-	-	✓	44.15	41.75	0.991	0.988	39.67	37.34	0.971	0.986
-	✓	✓	44.15	41.69	0.991	0.987	40.07	37.63	0.971	0.986
✓	-	✓	44.04	41.76	0.991	0.988	39.81	37.33	0.971	0.987
✓	✓	-	44.26	41.70	0.991	0.987	39.53	37.48	0.971	0.984
✓	✓	✓	**44.45**	**42.14**	**0.992**	**0.988**	**40.15**	**37.75**	**0.971**	**0.987**

Table 3: Ablation experiments on the Kalantari's and Prabhakar's dataset were performed to compare the effects of using three modules (SFA, HT, and CADB).

Figure 9: ASL loss compared to L1 loss with the ground truth as reference. Obviously, the model with ASL preserves the edge details and color better.

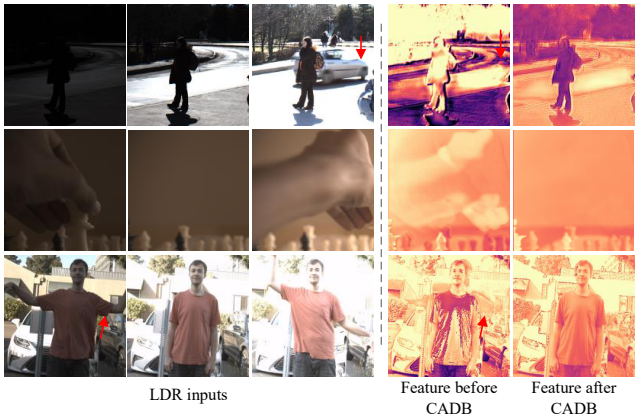

Figure 10: The three columns pf images on the left are the same position with different exposures in three different datasets(Tursun et al. 2016; Prabhakar et al. 2019; Kalantari, Ramamoorthi et al. 2017) respectively. The two columns of images on the right are the features before and after CADB.

Loss & N	Quantitative Results			
	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L
ASL & 3	**44.45**	**42.14**	**0.992**	**0.988**
ASL & 2	43.18	41.82	0.991	0.988
ASL & 1	44.35	41.94	0.992	0.988
L1 & 3	43.58	41.38	0.990	0.987
L1 & 2	**44.04**	**41.60**	**0.991**	**0.987**
L1 & 1	42.10	40.64	0.990	0.986

Table 4: Quantitative comparison of using two loss functions (ASL and L1) on model performance, where N denotes the number of CADB in IRM.

image with preserved details, colors and minimal ghosting. To achieve this, we introduced a module for preliminary alignment (SFA). Then we propose PFM, in which a transformer is used for the first time in the HDR problem to learn the discontinuous information of features at different scales. Finally, we proposed the IRM to reconstruct the HDR features with minimal ghosting. Experiments show that our approach has strong performance and produces high quality HDR images. Source code for HFT is available at https://github.com/Chenrf1121/HFT

Acknowledgements

This work is supported by the National Nature Science Foundation of China (U21B2024, 62001146). This work is also supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grants GK229909299001-009, and the National Nature Science Foundation of China 62271180.

References

Akyüz, A. O.; Fleming, R.; Riecke, B. E.; Reinhard, E.; and Bülthoff, H. H. 2007. Do HDR displays support LDR content? A psychophysical evaluation. *ACM Transactions on Graphics (TOG)*, 26(3): 38–es.

An, J.; Ha, S. J.; and Cho, N. I. 2012. Reduction of ghost effect in exposure fusion by detecting the ghost pixels in saturated and non-saturated regions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1101–1104. IEEE.

Bai, Y.; Yang, X.; Liu, X.; Jiang, J.; Wang, Y.; Ji, X.; and Gao, W. 2022. Towards End-to-End Image Compression and Analysis with Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 104–112.

Banterle, F.; Ledda, P.; Debattista, K.; Chalmers, A.; and Bloj, M. 2007. A framework for inverse tone mapping. *The Visual Computer*, 23(7): 467–478.

Bogoni, L. 2000. Extending dynamic range of monochrome and color images through fusion. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, 7–12. IEEE.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dai, T.; Li, W.; Cao, X.; Liu, J.; Jia, X.; Leonardis, A.; Yan, Y.; and Yuan, S. 2021. Wavelet-Based Network For High Dynamic Range Imaging. *arXiv preprint arXiv:2108.01434*.

Debevec, P. E.; and Malik, J. 2008. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, 1–10.

Dong, X.; Hu, X.; Li, W.; Wang, X.; and Wang, Y. 2021. MIEHDR CNN: Main Image Enhancement based Ghost-Free High Dynamic Range Imaging using Dual-Lens Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1264–1272.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eilertsen, G.; Kronander, J.; Denes, G.; Mantiuk, R. K.; and Unger, J. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM transactions on graphics (TOG)*, 36(6): 1–15.

Endo, Y.; Kanamori, Y.; and Mitani, J. 2017. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6): 177–1.

Gao, W.; Zhang, X.; Yang, L.; and Liu, H. 2010. An improved Sobel edge detection. In *2010 3rd International conference on computer science and information technology*, volume 5, 67–71. IEEE.

Hu, J.; Gallo, O.; Pulli, K.; and Sun, X. 2013. HDR deghosting: How to deal with saturation? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1163–1170.

Huo, Y.; Yang, F.; Dong, L.; and Brost, V. 2014. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30(5): 507–517.

Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.; Xu, C.; Li, Y.-L.; Wang, S.; and Tian, Q. 2020. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8008–8017.

Jacobs, K.; Loscos, C.; and Ward, G. 2008. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, 28(2): 84–93.

Kalantari, N. K.; Ramamoorthi, R.; et al. 2017. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4): 144–1.

Lee, S.; An, G. H.; and Kang, S.-J. 2018. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *proceedings of the European Conference on Computer Vision (ECCV)*, 596–611.

Li, H.; Ma, K.; Yong, H.; and Zhang, L. 2020. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29: 5805–5816.

Li, W.; Xiao, S.; Dai, T.; Yuan, S.; Wang, T.; Li, C.; and Song, F. 2022. SJ-HDˆ 2R: Selective Joint High Dynamic Range and Denoising Imaging for Dynamic Scenes. *arXiv preprint arXiv:2206.09611*.

Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; and Van Gool, L. 2021. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Liu, L.; An, J.; Liu, J.; Yuan, S.; Chen, X.; Zhou, W.; Li, H.; Wang, Y.; and Tian, Q. 2023. Low-Light Video Enhancement with Synthetic Event Guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, L.; Liu, J.; Yuan, S.; Slabaugh, G.; Leonardis, A.; Zhou, W.; and Tian, Q. 2020a. Wavelet-based dual-branch network for image demoiréing. In *European Conference on Computer Vision*, 86–102. Springer.

Liu, L.; Xie, L.; Zhang, X.; Yuan, S.; Chen, X.; Zhou, W.; Li, H.; and Tian, Q. 2022a. TAPE: Task-Agnostic Prior Embedding for Image Restoration. In *European Conference on Computer Vision*.

Liu, L.; Yuan, S.; Liu, J.; Bao, L.; Slabaugh, G.; and Tian, Q. 2020b. Self-adaptively learning to demoiré from focused and defocused image pairs. *Advances in Neural Information Processing Systems*, 33: 22282–22292.

Liu, L.; Yuan, S.; Liu, J.; Guo, X.; Yan, Y.; and Tian, Q. 2022b. Siamtrans: zero-shot multi-frame image restoration with pre-trained siamese transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1747–1755.

Liu, Z.; Lin, W.; Li, X.; Rao, Q.; Jiang, T.; Han, M.; Fan, H.; Sun, J.; and Liu, S. 2021. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 463–470.

Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; and Zeng, T. 2022. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 457–466.

Ma, K.; Li, H.; Yong, H.; Wang, Z.; Meng, D.; and Zhang, L. 2017. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5): 2519–2532.

Mantiuk, R.; Kim, K. J.; Rempel, A. G.; and Heidrich, W. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4): 1–14.

Marnerides, D.; Bashford-Rogers, T.; Hatchett, J.; and Debattista, K. 2018. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, 37–49. Wiley Online Library.

Meylan, L.; Daly, S.; and Süsstrunk, S. 2006. The reproduction of specular highlights on high dynamic range displays. In *Color and Imaging Conference*, volume 2006, 333–338. Society for Imaging Science and Technology.

Niu, Y.; Wu, J.; Liu, W.; Guo, W.; and Lau, R. W. 2021. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30: 3885–3896.

Prabhakar, K. R.; Agrawal, S.; and Babu, R. V. 2021. Self-gated memory recurrent network for efficient scalable HDR deghosting. *IEEE Transactions on Computational Imaging*, 7: 1228–1239.

Prabhakar, K. R.; Agrawal, S.; Singh, D. K.; Ashwath, B.; and Babu, R. V. 2020. Towards practical and efficient high-resolution HDR deghosting with CNN. In *European Conference on Computer Vision*, 497–513. Springer.

Prabhakar, K. R.; Arora, R.; Swaminathan, A.; Singh, K. P.; and Babu, R. V. 2019. A fast, scalable, and reliable deghosting method for extreme exposure fusion. In *2019 IEEE International Conference on Computational Photography (ICCP)*, 1–8. IEEE.

Qu, L.; Liu, S.; Wang, M.; and Song, Z. 2022. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2126–2134.

Reinhard, E.; Heidrich, W.; Debevec, P.; Pattanaik, S.; Ward, G.; and Myszkowski, K. 2010. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann.

Rempel, A. G.; Trentacoste, M.; Seetzen, H.; Young, H. D.; Heidrich, W.; Whitehead, L.; and Ward, G. 2007. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. *ACM transactions on graphics (TOG)*, 26(3): 39–es.

Sen, P.; Kalantari, N. K.; Yaesoubi, M.; Darabi, S.; Goldman, D. B.; and Shechtman, E. 2012. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6): 203–1.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Tursun, O. T.; Akyüz, A. O.; Erdem, A.; and Erdem, E. 2016. An objective deghosting quality metric for HDR images. In *Computer Graphics Forum*, volume 35, 139–152. Wiley Online Library.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vincent, O. R.; Folorunso, O.; et al. 2009. A descriptive algorithm for sobel image edge detection. In *Proceedings of informing science & IT education conference (InSITE)*, volume 40, 97–107.

Wang, Y.; Zhang, X.; Yang, T.; and Sun, J. 2022. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2567–2575.

Wu, S.; Xu, J.; Tai, Y.-W.; and Tang, C.-K. 2018. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 117–132.

Xiong, P.; and Chen, Y. 2021. Hierarchical Fusion for Practical Ghost-free High Dynamic Range Imaging. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4025–4033.

Yan, Q.; Gong, D.; Shi, Q.; Hengel, A. v. d.; Shen, C.; Reid, I.; and Zhang, Y. 2019. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1751–1760.

Yan, Q.; Zhang, L.; Liu, Y.; Zhu, Y.; Sun, J.; Shi, Q.; and Zhang, Y. 2020. Deep HDR imaging via a non-local network. *IEEE Transactions on Image Processing*, 29: 4308–4322.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5791–5800.

Yu, F.; Huang, K.; Wang, M.; Cheng, Y.; Chu, W.; and Cui, L. 2022. Width & Depth Pruning for Vision Transformers. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 2022.

Zhang, J.; and Lalonde, J.-F. 2017. Learning high dynamic range from outdoor panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, 4519–4528.

Zhao, H.; Zheng, B.; Yuan, S.; Zhang, H.; Yan, C.; Li, L.; and Slabaugh, G. 2021. CBREN: Convolutional Neural Networks for Constant Bit Rate Video Quality Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zheng, B.; Chen, Q.; Yuan, S.; Zhou, X.; Zhang, H.; Zhang, J.; Yan, C.; and Slabaugh, G. 2022a. Constrained Predictive Filters for Single Image Bokeh Rendering. *IEEE Transactions on Computational Imaging*, 8: 346–357.

Zheng, B.; Chen, Y.; Tian, X.; Zhou, F.; and Liu, X. 2019. Implicit dual-domain convolutional network for robust color image compression artifact reduction. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11): 3982–3994.

Zheng, B.; Pan, X.; Zhang, H.; Zhou, X.; Slabaugh, G.; Yan, C.; and Yuan, S. 2022b. DomainPlus: Cross-Transform Domain Learning towards High Dynamic Range Imaging. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1–10.

Zheng, B.; Yuan, S.; Slabaugh, G.; and Leonardis, A. 2020. Image demoireing with learnable bandpass filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3636–3645.

Zheng, B.; Yuan, S.; Yan, C.; Tian, X.; Zhang, J.; Sun, Y.; Liu, L.; Leonardis, A.; and Slabaugh, G. 2021. Learning frequency domain priors for image demoireing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.