symmetry

MDPI

Article

Enhanced LDR Detail Rendering for HDR Fusion by TransU-Fusion Network

Bo Song [1], Rui Gao [1], Yong Wang [2] and Qi Yu [1,*]

[1] State Key Laboratory of Electronic Thin Films and Integrated Devices, School of Integrated Circuit Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China; songbo@imagedesign.com.cn (B.S.); gaorui@imagedesign.com.cn (R.G.)
[2] Chengdu Image Design Technology Co., Ltd., 171# Hele 2nd Street, Chengdu 610213, China
* Correspondence: qiyu@uestc.edu.cn; Tel.: +86-28-8320-5248

Abstract: High Dynamic Range (HDR) images are widely used in automotive, aerospace, AI, and other fields but are limited by the maximum dynamic range of a single data acquisition using CMOS image sensors. High dynamic range images are usually synthesized through multiple exposure techniques and image processing techniques. One of the most challenging task in multiframe Low Dynamic Range (LDR) images fusion for HDR is to eliminate ghosting artifacts caused by motion. In traditional algorithms, optical flow is generally used to align dynamic scenes before image fusion, which can achieve good results in cases of small-scale motion scenes but causes obvious ghosting artifacts when motion magnitude is large. Recently, attention mechanisms have been introduced during the alignment stage to enhance the network's ability to remove ghosts. However, significant ghosting artifacts still occur in some scenarios with large-scale motion or oversaturated areas. We proposea novel Distilled Feature TransformerBlock (DFTB) structure to distill and re-extract information from deep image features obtained after U-Net downsampling, achieving ghost removal at the semantic level for HDR fusion. We introduce a Feature Distillation Transformer Block (FDTB), based on the Swin-Transformer and RFDB structure. FDTB uses multiple distillation connections to learn more discriminative feature representations. For the multiexposure moving scene image fusion HDR ghost removal task, in the previous method, the use of deep learning to remove the ghost effect in the composite image has been perfect, and it is almost difficult to observe the ghost residue of moving objects in the composite HDR image. The method in this paper focuses more on how to save the details of LDR image more completely after removing the ghost to synthesize high-quality HDR image. After using the proposed FDTB, the edge texture details of the synthesized HDR image are saved more perfectly, which shows that FDTB has a better effect in saving the details of image fusion. Futhermore, we propose a new depth framework based on DFTB for fusing and removing ghosts from deep image features, called TransU-Fusion. First of all, we use the encoder in U-Net to extract image features of different exposures and map them to different dimensional feature spaces. By utilizing the symmetry of the U-Net structure, we can ultimately output these feature images as original size HDR images. Then, we further fuse high-dimensional space features using Dilated Residual Dense Block (DRDB) to expand the receptive field, which is beneficial for repairing over-saturated regions. We use the transformer in DFTB to perform low-pass filtering on low-dimensional space features and interact with global information to remove ghosts. Finally, the processed features are merged and output as an HDR image without ghosting artifacts through the decoder. After testing on datasets and comparing with benchmark and state-of-the-art models, the results demonstrate our model's excellent information fusion ability and stronger ghost removal capability.

Keywords: HDR fusion; DFTB; DRDB; U-Net; ghosting artifact

Citation: Song, B.; Gao, R.; Wang, Y.; Yu, Q. Enhanced LDR Detail Rendering for HDR Fusion by TransU-Fusion Network. *Symmetry* **2023**, *15*, 1463. https://doi.org/10.3390/sym15071463

Academic Editor: Sergei D. Odintsov

Received: 28 May 2023
Revised: 10 July 2023
Accepted: 18 July 2023
Published: 23 July 2023

1. Introduction

With the development of computer graphics, there is an increasing demand for high-quality images in daily life, HDR imaging technology is becoming more and more

widely used. Compared to low dynamic range images, HDR images have a higher dynamic range, closer to what the human eye sees, and contain more information. There are high application degree and development potential in industries such as film, gaming, VR, military, and medical imaging. The traditional method of obtaining HDR images is through special HDR cameras, which are too expensive for ordinary people [1,2]. Therefore, research has turned to obtaining HDR images from regular low dynamic range images. The most common method is to use a digital camera to capture multiple differently exposed LDR images and merge them into one HDR image by algorithms. Although general merging algorithms perform well in static scenes, they produce ghosting artifacts when the camera or the scene is in motion. Recently, there have been two main categories of HDR ghost-free synthesis algorithms.

Traditional algorithms: Methods based on motion detection are commonly used to detect motion areas and remove motion pixels in multi-frame LDR synthesis. This algorithm can achieve good results when the motion area is small, but when the motion area exceeds a certain threshold, the removed area is too large, resulting in significant loss of synthesized image information. Alignment-based methods commonly use optical flow to align the motion scene in the LDR image. In some specific scenes, it can align large moving objects, but when the motion scene contains overexposed and underexposed areas, it is still affected by the accuracy of the optical flow estimation of the moving object pixels, resulting in significant ghosting.

Deep learning algorithms: With the rapid development and powerful performance of deep learning, algorithms based on convolutional neural networks have shown better performance in repairing image details and removing ghosting than traditional algorithms. Kalantari et al. [3] directly added CNN for fusion after aligning with optical flow. Wu et al. [4] used U-Net [5] and ResNet separately for HDR fusion. Yan et al. [6] introduced an attention module for image alignment and then used the DRDB [7] module for image fusion. Niu et al. [8] used GAN and deep supervision for fusion. These works have greatly improved the removal of ghosting and image repair in overexposed areas compared to traditional algorithms. However, when overexposed areas overlap with dynamic scenes, these works still produce significant ghosting.

In this work, a transformer-based U-Net is employed to parallel process deep feature maps of images globally and locally. This can composite three LDR images with different exposures into one high-quality HDR image without the need for specific image alignment modules. Use U-Net to introduce symmetry into the network structure and information flow mode of this method. Among them, the encoder and decoder are structurally imaging the Symmetric relation of the image, so that features can be completely acquired and transferred on different scales. Jumping connections enable direct information transmission between low-level and high-level features, which effectively improves the network's perception of features at different scales and helps to reconstruct details and edge information. ViT [9,10] has rapidly developed due to its excellent long-range modeling ability and outstanding induction bias with increasing data specificity. Since the scarcity and difficulty in collecting datasets for multi-frame ghost-free HDR synthesis tasks, the training dataset used is small and lacks sufficient data specificity to some extent. However, in the help of the low-pass filtering properties of MSA in ViT [11], it exhibits excellent and promising performance in ghost removal. In this article, we draw on the strengths and weaknesses of previous methods [6,12], retaining the method of using U-Net for feature extraction and image reconstruction, omitting the attention mechanism, and transforming the ghost removal HDR reconstruction process into an integrated process. For the multiexposure moving scene image fusion HDR ghost removal task, in the previous method, the use of deep learning to remove the ghost effect in the composite image has been very perfect, and it is almost difficult to observe the ghost residue of moving objects in the composite HDR image. The method in this paper focuses more on how to save the details of LDR image more completely after removing the ghost to synthesize high-quality HDR image. After using the proposed FDTB, the edge texture details of the synthesized HDR image are

saved more perfectly, which shows that FDTB has a better effect in saving the details of image fusion. The use of FDTB instead of the nonlocal mechanism has further improved the ghosting effect, which is the improvement strategy of this article. The novel feature distillation mechanism FDTB proposed in this article extracts global features through ViT and utilizes feature channel compression and expansion to map rough image features with ghosts and clean image features with detailed information into different feature spaces, thus separating the two and achieving the effect of removing ghosts.

TransU-fusion mainly consists of two parts: three parallel encoders for feature extraction, collecting shallow and deep features of images with different exposures; the deep features are fused by concatenating input DRDB and transformer. DRDB processes local information of deep features to fuse image information at semantic and abstract level, repair image details, while transformer processes global information to remove ghosting artifacts in deep features and use it as a motion structural template for upsampling, followed by upsampling operations in the decoder, and finally fused by a CNN block. By aligning and fusing long-range deep feature information with local pixel information, TransU-fusion repairs and enhances image details through skip-connections, resulting in high-quality ghost-free HDR images. As shown in Figure 1, compare with other models, the main contributions of our method can be summarized as follows:

Figure 1. (**a**) Represents three LDR images with different exposure levels. (**b**) is the output of perposed method after tone mapping. (**c**) is the image region containing oversaturated regions and significant motion in different LDR images. (**d**–**g**) Comparison details of generated image for each model [6,8,12]. (**h**) Ground truth. Comparing the proposal method with the reference method, the results of the proposal method are relatively better. For example, on awnings and doors (blue and red arrows mark the positions), reference methods have bright light invading image details, while the proposed method retains more details.

- This work proposes a novel method, called TransU-fusion, for HDR reconstruction from multiexposure LDR images using a combination of U-Net and Transformer.

The method is capable of reconstructing high-quality and ghost-free HDR images even when LDR images contain large foreground motions and oversaturated areas.
- The proposed Transformer structure, FDTB, distills image features extracted from the Encoder and splits ghost from effective image information, showing better performance than previous methods.
- Experimental results on three benchmarks demonstrate that the proposed model prevails over state-of-the-art HDR models.

2. Related Work

The main related works we have summarized as two kind of methods, reconstructing HDR image with or without using deep learning neural network.

2.1. HDR Reconstruction Algorithms without Deep Learning Methods

Pixel rejection methods: This method marks each pixel as a static region or moving object based on global image registration, thereby rejecting erroneous single pixels. Grosch et al. [13] used color differences in input images to define an error map that aided in generating ghost-free HDR images. Jacobs et al. [14] detected misaligned regions through a weighted variance measure. Pece et al. [15] identified ghost regions by computing a median threshold bitmap from input LDR images. Zhang et al. [16] proposed to detect misaligned regions by analyzing image gradients. Heo et al. [17] computed joint probability density to detect motion areas and then used Gaussian-weighted distance to weight each exposure during merging. However, the pixel rejection method reduces useful image information that is important for reconstructing HDR images. These methods usually result in unsatisfactory outcomes.

Alignment before merging: These methods align non-reference images to the reference one before merging them into an HDR image. Bogoni et al. [18] used optical flow to align input LDR images. Kang et al. [19] estimated flow through various optical flow variants and used a specialized merging method to reject artifacts. Zimmer et al. [20] reconstructed HDR images by registering LDR images with optical flow found by minimizing an energy function consisting of gradient and smoothness terms. Gallo et al. [21] proposed a fast motion estimation method for small motion images. These methods are more robust than pixel rejection methods but still generate alignment artifacts when challenging cases occur.

Patch-based methods: These methods deal with alignment and HDR merging using a unified optimization system. Sen et al. [22] presented a patch-based energy minimization method to complete the missing details in the reference image from other LDR images in the stack. Hu et al. [23] proposed a smaller patch-based system and optimized image alignment by brightness and gradient consistencies on the transformed domain. Patch-based methods perform better than the above methods. However, when the reference image has large saturated regions or large motions exist in nonreference LDR images, patch-based systems produce unsatisfactory results.

2.2. Deep Learning Based Algorithms

CNN-based methods: With the development of CNN networks, many CNN-based methods have been used for multiframe HDR image synthesis tasks. Kalantari et al. [3] were the first to use CNN in HDR synthesis tasks. They aligned LDR images to the reference image using optical flow and then merged them through a CNN. They proposed a benchmark dataset for multiframe synthesis tasks. Hu et al. [4] first used homography transformation to align the background of LDR images and then learned the mapping from LDR images to HDR images using ResNet or U-Net. Yan et al. [6] proposed a special attention mechanism to align LDR images and generated ghost-free HDR images by fusing LDR image features with DRDB. Niu et al. [8] proposed the first model (as shown in Figure 2) that uses GAN for multiframe HDR image fusion and used deep HDR supervision to generate higher quality HDR images.

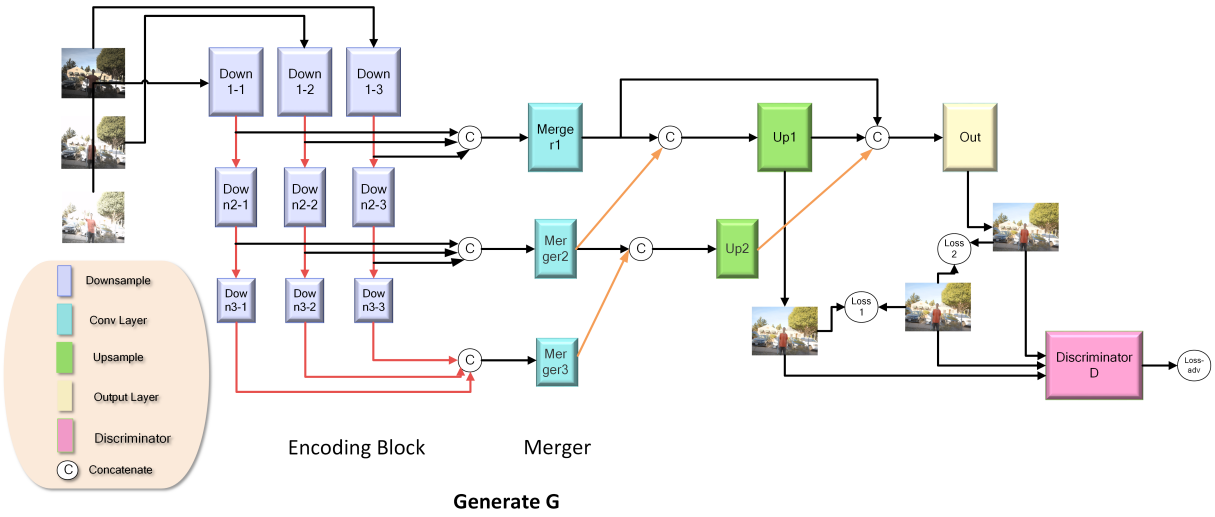

Figure 2. This is the diagram of the HDR-GAN [8] network. This network consists of two modules: Generator G and Discriminator D. As shown in the figure, three LDR images with different exposures are inputted into G, and feature images are extracted through the downsampling module (blue cube). Image alignment is performed through the feature merging module (purple cube) and then inputted into the upsampling module (green cube) to restore the original image size, output final HDR image (yellow cube). The two HDR images generated are calculated with L1&L2 loss and input to Discriminator D (pink cube) to distinguish between the generated image and Ground Truth through GAN.

Vision Transformer: With the introduction of ViT [10], the Transformer achieved tremendous success in the field of image tasks, such as image classification, recognition, and segmentation. Liu et al. [24] introduced Swin-Transformer, which greatly reduced the model's parameter size while maintaining or even surpassing ViT's performance in some areas.

Our strategy is inspired by [6,12,24], combining the advantages of CNN and transformer.

3. Proposed Algorithm

For the task of creating ghost-free HDR images by merging multiple dynamic LDR images, we followed the approach used in previous literature [3,6] and selected three LDR images with varying exposure levels (i.e., $I_i, i = 1,2,3$) as input. We aligned the images with respect to the central-exposure frame I_2. Prior to being fed into the network, the LDR images were mapped onto the HDR domain, producing $(H_i, i = 1,2,3)$. Mapping LDR images onto the HDR domain was beneficial in detecting misalignments during processing, whereas LDR images were more effective in detecting noise or saturated areas [3]. We applied gamma correction to the I_i images to obtain the corresponding H_i.

$$H_i = \frac{I_i^{\gamma}}{t_i}, \ \forall i = 1,2,3, \tag{1}$$

is the exposure time of I_i, and γ is the gamma correction parameter, whose value is greater than 1 and in this article is set to 2.2. Based on the strategy proposed in [3], we concatenate I_i and H_i to obtain a 6-channel input $X_i = (I_i, H_i)$ for the network. The HDR image is then generated by applying the function $F(\cdot)$ to X_i.

$$\hat{H} = F(X_i; \theta), \ \ i = 1,2,3 \tag{2}$$

$\hat{H}$ is the output of the three-channel HDR image, and θ represents the network parameters. In this paper, we present an end-to-end model that does not require any preliminary alignment of the original image before entering the network.

3.1. FDTB

In this paper, we introduce a Feature Distillation Transformer Block (FDTB), based on the Swin-Transformer and RFDB structure. As shown in Figure 3, FDTB uses multiple distillation connections to learn more discriminative feature representations [25]. FDTB is a novel feature distillation module based on ViT. Extracting global features through FDTB, compressing and restoring feature image channels, mapping rough feature images containing ghosts and clean feature images containing image edge details to feature spaces of different dimensions to separate the two. At the same time, the low-pass filter features of ViT [11] can also help FDTB filter high-frequency information, thereby suppressing the residual ghost effect.

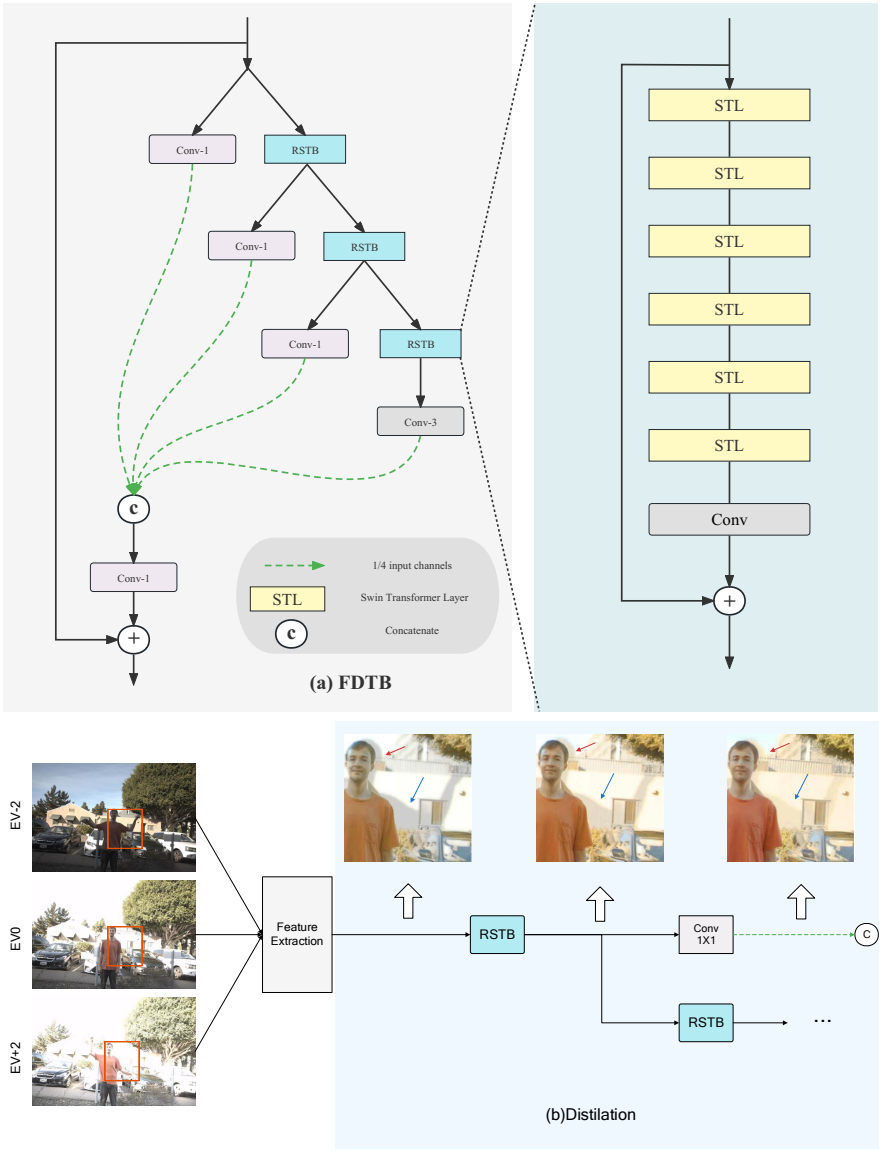

Figure 3. (**a**) For the schematic diagram of FDTB structure, the input image features are subjected to RSTB distillation and dimensionality reduction through a conv structure, compressing the number of channels to 1/4 of the input channel number. Then, each layer of distillation features are merged and combined with the input features for residual analysis before output. RSTB is composed of multilayer Swin-Transformer Layers and residual structures. (**b**) representing the distillation process schematic, the input image is extracted through feature extraction and input into the distillation structure. After passing through multiple layers of RSTB, the ghost shadows (as arrows show) in the synthesized image are gradually eliminated and image details are supplemented.

RSTB stands for residual swin transformer blocks [26]. It is used to extract deep features F_{DF} in this paper, consisting of K layers of STL and a 3×3 conv layer. $F_i(i = 1, 2, 3 \ldots, K)$ represents the intermediate

$$F_i = H_{STL_i}(F_{i-1}),\ i = 1, 2, 3, \ldots, K$$

$$F_{DF} = H_{CONV}(F_K) + F_\theta \tag{3}$$

$H_{STL_i}(.)$ refers to the i-th layer of STL. We added a conv structure at the end of the STL network to introduce the inductive bias of the conv operation into the transformer-based network, resulting in better stability when extracting and processing deep features of HDR images. The core of FDTB is a progressive refinement module (PRM) similar to IMDB [27], which is similar to RFDB. The split operation in IMDB is decoupled and replaced with an RSTB and 3×3 conv layer to process deep image features input from the encoder downsampling, extracting the depth texture information and semantic information in the features. Meanwhile, low-pas filtering is carried out on the motion features that are inconsistent with the reference frame's dynamic scene to achieve a ghost-free effect. The features are then output and enter the next RSTB or DL. The features that enter the DL layer undergo information distillation and dimensionality reduction to obtain complete processed features. The features entering the next RSTB are sent to the next distillation step to repeat the above steps. $I_i(i = 1, 2, 3, \ldots, T,)$ represents the intermediate feature maps output after RSTB processing. The flowchart of the block can be described as follows:

$$I_{distilled_1}, I_{coarse_1} = DL_1(I_0), L_{RSTB_1}(I_0),$$

$$I_{distilled_i}, I_{coarse_i} = DL_i\left(I_{coarse_i-1}\right), L_{RSTB_i}\left(I_{coarse_i-1}\right), i = 2, 3, \ldots, T-1,$$

$$I_{distilled_T} = DL_T(I_{coarse_T-1}) \tag{4}$$

L_{RSTB_i} represents the ith RSTB, DL_1 represents the ith 3×3 conv, $I_{distilled_i}$ denotes the ith distilled feature, and I_{coarse_i} is the coarse feature that needs to be further processed by the subsequent layers. Finally, all distilled features are concatenated along the channel direction and served as the output of the PRM module:

$$I_{distilled} = Concat(I_{distilled_1}, I_{distilled_2}, \ldots, I_{distilled_T}) \tag{5}$$

3.2. Overall Architecture

We propose the TransU-fusion network for ghost removal in HDR images, with the structure shown in Figure 4. Similar to [8,12], we choose U-Net as the backbone network structure since U-Net has shown excellent generalization and powerful performance in many computer vision tasks. U-Net is a structurally symmetrical and concise encoder–decoder network with skip-connection. The encoder downsamples the image to obtain image features at different scales. The shallow features contain HDR image details and flat region information, while deep features contain texture and semantic information. We perform parallel global and local-level information processing on deep features, and the resulting image features are restored to the original size through the decoder with skip-connections to supplement image details, thereby generating a ghost-free and complete HDR image. Our network structure consists of three parts: encoder, merger, and decoder. Since we are synthesizing HDR images from three frames, the encoder provides three downsampling channels corresponding to three different exposures. Different encoders learn different parameters from their corresponding exposures, and then the three image features output by the encoders are concatenated in the channel direction and input into the merger. The merger consists of two parts: Dilated Residual Dense Blocks (DRDB) [6] and Distilled Feature Transformer (DFTB). The concatenated image features are processed in parallel through two modules. The first module enters DRDB after passing through a 3×3 CNN layer and processes the features at a local-level, which can enhance image texture information and perform hallucination on overexposed areas and occlusion to

supplement the missing image information in that region. The second module enters DFTB after one layer of 3×3 CNN to reduce the transformer parameters required for processing. In DFTB, the deep features are processed at a global level. Inputting into RSTB, the image semantic information is fused and the features are low-pass filtered [11] to distill out the ghost-free and information-fused features. After the 3×3 CNN downsampling, the remaining image features continue to enter the next RSTB for processing until all features have been processed. The features output by DRDB and DFTB are concatenated in the channel direction and input into the decoder for upsampling to restore the original size of the image. Meanwhile, the shallow image texture features and detailed features extracted by the encoder enter the decoder of the same layer through skip-connections to supplement image detail information and overexposed/underexposed regions. Finally, after passing through one layer of 3×3 CNN as the restoration layer, the ghost-free HDR image is output.

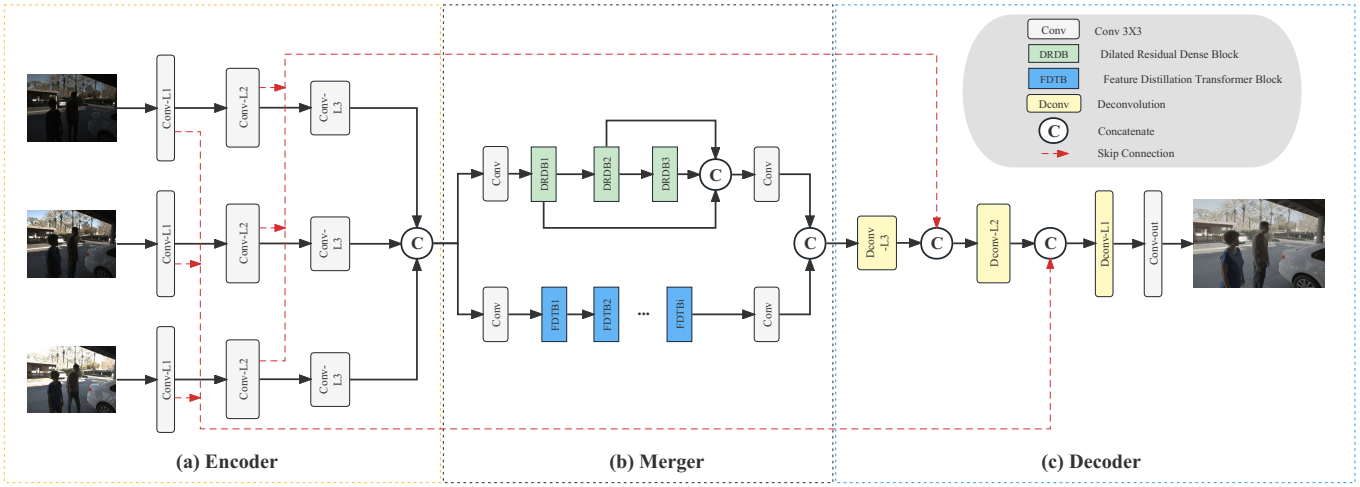

Figure 4. This is a schematic diagram of the Trans-U Fusion network structure. The network consists of an encoder for feature extraction, a merger for deep feature distillation and fusion, and a decoder for reconstructing HDR images. The encoder extracts image features at different scales through downsampling and sends deep features into the merger for distillation. The merger consists of a dual branch system, where a module composed of three DRDBs processes local features through residual connections, and a module composed of six FDTBs processes global features in parallel, achieving the goal of fully extracting image details and allowing the FDTB module to distill ghost features. The output features are reconstructed into HDR images by combining the decoder layer with image details supplemented by skip connections at the channel level. The final output HDR image is displayed after tone mapping.

3.3. Training

Before being displayed, HDR images usually need to undergo tone mapping. Although some TM strategies that work well have been proposed, they are often too complex or not differentiable. Here, we use the μ-law introduced in [3] to perform TM on HDR images and calculate the loss, since the μ-law is differentiable:

$$T(x) = \frac{\log(1 + \mu x)}{\log(1 + \mu)} \tag{6}$$

$\mu = 5000$ decides the extent of compression, and $T(x)$ is the TM-ed HDR image. Inspired by [28], we use perceptual loss [29] to optimize our network. Perceptual loss is widely used in image restoration tasks, improving image quality at a feature level rather than a pixel level. The used loss function consists of two parts:

$$\mathcal{L}(\theta, H) = \mathcal{L}_2(\theta, H) + \lambda_p \mathcal{L}_p(\theta, H), \tag{7}$$

where θ denotes the parameters in TransU-fusion, H is the estimated HDR image, and λ_p is a hyperparameter set to 0.01. $\mathcal{L}_2(\cdot)$ refers to the MSE loss, defined as follows:

$$\mathcal{L}_2(\theta, H) = ||T(H) - T(\hat{H})||_2, \tag{8}$$

$\hat{H}$ is the GT image of H. $\mathcal{L}_p(\cdot)$ is:

$$\mathcal{L}_p(\theta, H) = \sum_i ||V_i(T(H) - V_i(T(\hat{H}))||_1 \tag{9}$$

$V_i(\cdot)$ refers to the feature map extracted from the pretrained VGG-16 [30], where i indicates the ith layer of VGG-16. In this loss function, MSE focuses on improving the pixel-level details of the image, while perceptual loss improves the image's contrast and structural similarity at an abstract level.

4. Experiments

4.1. Dataset and Metrics

Proposed approach builds upon the work of previous HDR image restoration methods such as HDR-GAN [8], HDRi With LFM [4], Ghost-free [6], and Nonlocal [12]. To train our model, we utilized the Kalantari [3] dataset, which includes 74 sets of training images and 15 sets of testing images. For each group of image data, there are 3 LDR images with varying exposures (-2, 0, +2 or -3, 0, +3) and one HDR image as the ground truth (GT). To augment our training data, we applied random rotation and 90-degree flipping operations to the 512×512 image patches.

The testing phase is conducted on the Kalantari dataset, as well as the datasets used in Sen [22] and Tursun [31] to verify the generalization of our model.The evaluation metrics used in our experiments include PSNR-μ, (PSNR after μ-law) PSNR-l (PSNR without μ-law), SSIM-μ, SSIM-l, and HDR-VDP-2. PSNR measures the pixelwise signal-to-noise ratio between the generated HDR image and the GT image. SSIM measures the structural similarity between the two images. HDR-VDP-2 is a metric designed specifically for evaluating the quality of HDR imagery. Overall, our choice of dataset and augmentation techniques were inspired by the effectiveness demonstrated in prior works. The Kalantari dataset is widely used for evaluating HDR image restoration algorithms and our augmentation techniques served to further improve the generalization of our model. Our TransU-fusion network is implemented using PyTorch, and we use the ADAM optimizer with $beta_1 = 0.9$, $beta_2 = 0.999$, and $\epsilon = 1 \times e^{-8}$. The initial learning rate is set to $1 \times e^{-3}$, and we train the network from scratch with a batch size of 8. During the initial epochs, we use L1-loss as the loss function, while we switch to the perceptual loss function later on. We trained our network on a NVIDIA 3090 GPU and it costs about 5 days.

4.2. Ablation Studies

4.2.1. Model Architecture

As shown in Table 1, we conducted a detailed analysis on the TransU model and tested the importance and performance of its different network components. Through an ablation study, we analyzed the impact of varying components of the deghosting module (FDTB position) and U-Net related hyperparameters (U-Net layers). The experimental variables tested were:

- w/o Transformer: Compare results by removing the FDTB part present in the original TransU model and using it as a control variable.
- w/Transformer: Replaced the FDTB part with the same number of vanilla transformers to analyze the contribution of the transformer in image fusion.
- Swin-Transformer: Replaced the transformer with the same number of Swin-Transformers to compare their deghosting performance.
- FDTB: Replaced the Swin-Transformer with FDTB to test its effectiveness in deghosting and information fusion capabilities.

Figure 5 shows that the transformer module was highly effective in removing ghosting effects from the images. Comparing the w/o transformer with w/transformer, it was evident that the transformer effectively suppressed ghosting effects. When comparing w/transformer with Swin-Transformer, it was observed that Swin-Transformer had fewer parameters but with an equally effective deghosting capability. In comparison, by inheriting Swin-T's fewer parameters and excellent deghosting abilities, FDTB showed stronger image fusion ability for oversaturated image areas and displayed more image details.

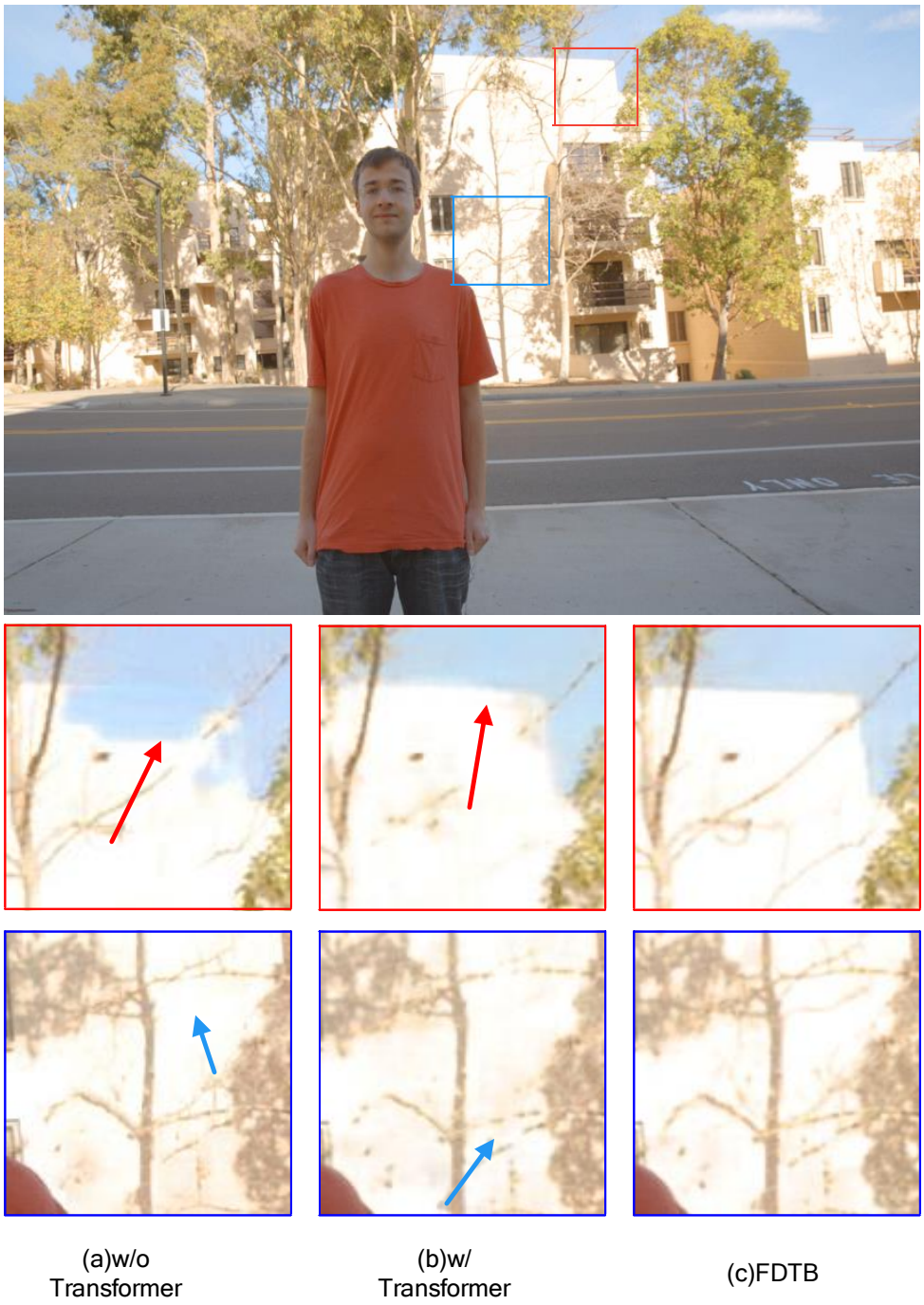

Figure 5. Shows the results for the ablation experiment. (**a**) The HDR image generated without the use of ViT structure, resulting in severe loss of image content. (**b**) The HDR image generated by adding ViT structure, the generated image is relatively complete, but the edge details of the image are severely lost. (**c**) It is an HDR image generated using FDTB by out method, with complete image details.

Table 1. This table shows the PSNR-μ, PSNR-L, and HDR-VDP-2 scores of different model structures.

Structure	PSNR-μ	PSNR-L	HDR-VDP-2
w/o ViT	42.29	40.54	62.32
Vanilla ViT	43.34	41.15	64.52
RSTB	43.68	41.56	64.97
FDTB	**43.87**	**41.83**	**65.83**

Layers of U-Net: In the architecture of U-Net, the number of encoder and decoder layers is a crucial parameter that directly affects the model's number of parameters, inference speed, and performance [5,32]. Therefore, we evaluated the impact of the number of U-Net layers on HDR fusion effect and inference time in the article, as shown in Figure 6 (y-axis for PSNR, x-axis for inference time, and 4 different type markers representing 1–4 layers) and Table 2.

Table 2. This table shows the corresponding PSNR-μ, inference time, and HDR-VDP-2 for U-Net layers from 1 to 4.

Layers	PSNR-μ	Time(s)	HDR-VDP-2
1	43.12	**0.11**	64.59
2	43.29	0.16	65.21
3 (Proposed Method)	**43.87**	0.23	**65.83**
4	43.56	0.45	64.92

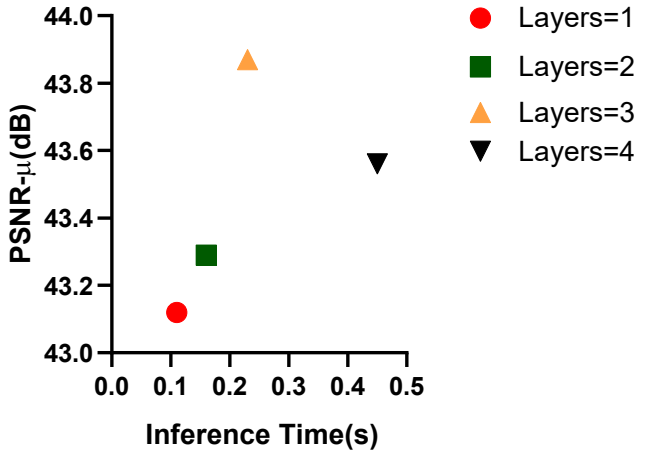

Figure 6. Shows the relationship between the number of U-Net layers, inference time, and the generated image PSNR. It can be seen that as the number of layers increases, the inference time increases, and the PSNR increases. It reaches the highest position when the layer number is 3. Considering trade-off of the inference time and performance, the best layer number decided as 3 layers in this article.

4.2.2. Study on Loss Function

In our experiment, we compared the performance of different loss functions during the pretraining stage of the TransU model. Table 3 displays the quantitative comparison results. Consistent with "Loss functions for image restoration with neural networks", we found that L2 loss is better at preserving image details. Additionally, L2 loss led to faster convergence and higher PSNR values in our model, contributing to greater stability.

Table 3. This table shows the corresponding image objective indicator data under different loss functions.

Loss Function	PSNR-μ	PSNR-L	HDR-VDP-2
L1	43.54	41.03	63.40
L2	43.68	41.11	64.69
L2 + Perception Loss (Proposed)	**43.87**	**41.83**	**65.83**

4.3. Comparison with SOTA Methods

4.3.1. Test on Kalantari et al.'s Dataset [3]

We compared our TransU-fusion method with several state-of-the-art (SOTA) methods, including two traditional patch-based algorithms based on patch-match (Sen [22] and Hu [23]) and three deep learning algorithms (AHDRNet [6], NHDRRNet [12], Kalantari [3], and HDR-GAN [8]). Among these deep learning methods, Kalantari's approach uses optical flow for image alignment before utilizing CNNs. AHDRNet uses attention structures for alignment and DRDB structures for image fusion. NHDRRNet utilizes U-Net for feature extraction and nonlocal structures for processing global features. Transformer-hdr applies attention structures for alignment and transformer structures for image fusion. In contrast, our TransU model does not require a prealignment or specific alignment module to address ghosting effects in the input images. We analyzed the above models both quantitatively and qualitatively on the testing set with ground truth available. Table 4 shows the quantitative results. We retrained the models of AHDRNet, NHDRRNet, and HDR-GAN based on the original author's code and methods to obtain the reproduced image results of each model.

Table 4. This table shows the objective indicator parameters obtained by different methods in the Kalantari [3] test set.

Method	PSNR-μ	PSNR-L	SSIM-μ	SSIM-L	HDR-VDP-2
Sen's Algorithm [22]	40.80	38.11	0.9808	0.9721	59.38
Hu's Algorithm [23]	35.79	30.76	0.9717	0.9503	57.05
Kalantari [3]	42.67	41.23	0.9888	0.9846	65.05
AHDRNet [6]	43.63	41.14	0.9900	0.9702	64.61
NHDRRNet [12]	42.41	41.43	0.9877	0.9857	61.21
HDR-GAN [8]	**43.92**	41.57	**0.9905**	0.9865	65.45
Proposed <TransU Fusion>	43.87	**41.83**	0.9904	**0.9876**	**65.83**

To ensure fairness, the generated images were tonemapped using the same method. If there were any discrepancies between our displayed results and the original results from the authors, we defaulted to the original results. As previous studies [1] have repeatedly demonstrated the significant difference between patch-matched traditional methods and deep learning methods in terms of fusion effect, we will not elaborate on this in our qualitative analysis in this study. All test data used Kalantari's testing set, which contains some highly challenging samples. For example, in the high-exposure images, there are many overexposed areas and significant foreground motion changes. Table 4 shows that the deep learning methods outperformed the patch-based methods, Sen's and Hu's, in all aspects. Our method exceeded previous SOTA performance, such as Kalantari and AHDRNet. However, when compared to HDR-GAN (the latest SOTA), there was a slight difference of approximately 1% in the evaluation parameters. Although our method scored slightly lower in the evaluation parameters, the performance of the deghosted images generated by our model was comparable or even better than that of the other models in some areas.

In Figure 7, the first row shows three LDR images with low, medium, and high exposures on the left and the HDR images tonemapped by our model in the middle, and the comparison area on the right. The LDR images have significant foreground motion in the red and blue areas, while the high-exposure images have large overexposed areas that

make ghost removal and reducing the overexposure challenging for the model. The second row shows comparisons of the selected image regions in the HDR images generated by different models. Deep learning methods, including ours, performed well in removing the ghosts caused by significant foreground motion and exhibited minor differences in the detailed areas. AHDRNet and NHDRRNet left some slight ghosting residues in the synthesized images in the red and blue areas. HDR-GAN removed the ghosting well in the red area but there was slight overexposure in the arrow direction, and the blue area had a similar problem with the windows. The red area of the LDR image had a slight head movement, but there were many overexposed areas in the middle-exposure image alignment. Therefore, in AHDRNet and NHDRRNet, although the synthesized image of the character's head had no residual ghosting effect, it could not hallucinate the details of the character's head and inherited the overexposed areas of the LDR image. Our method performed well in this aspect. In the blue area, when facing large overexposed areas in the high-exposure LDR image, the other methods showed missing scene details or color deviation, while our method synthesized the image details completely. Comparing our method with others, our method performed better in terms of deghosting and hallucinating image details in Kalantari's testing set.

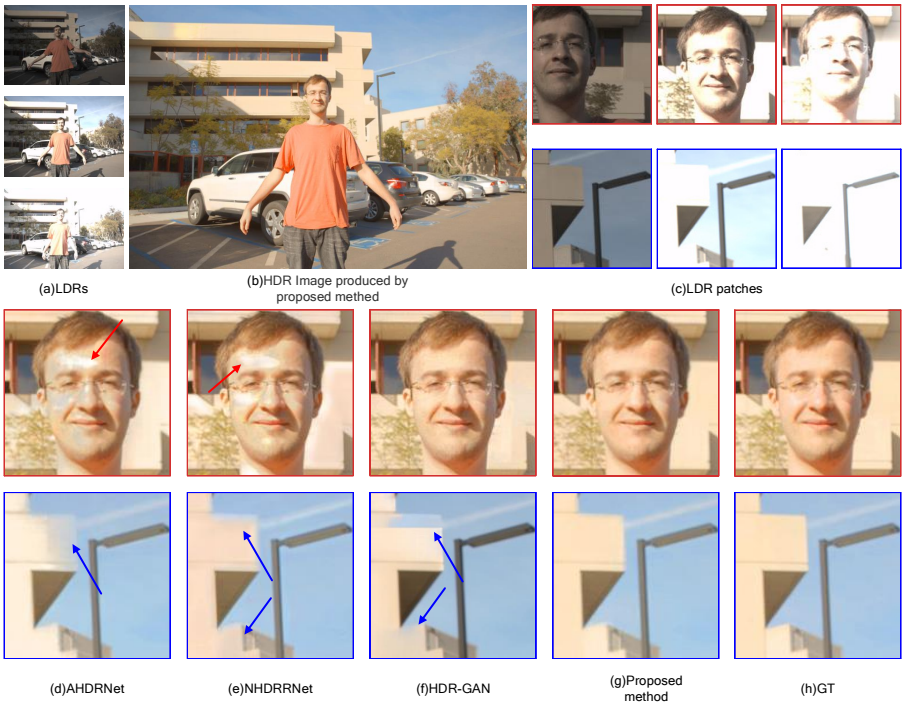

Figure 7. The test results for using the Kalantari dataset. The upper half (**a**–**c**) represents the HDR image and LDR patches after TM output from our LDR image. The lower half (**d**–**g**) shows the comparison of our method with the local regions of the generated images from other networks [6,8,12]. (**h**) represents the ground truth. It can be seen that our network can generate high-quality HDR images, whether in supersaturated or dynamic regions.

4.3.2. Test on Other Dataset

We also validated our method on Sen's [22] and Tursun's [23] datasets to test its generalization. As shown in Figure 8, to distinguish the exposure levels from Kalantari's dataset, we tested images with exposure levels of $[-4, -2, 4]$. We can see that in the red area, after fusing the overexposed high-exposure LDR and the underexposed reference frame, AHDRNet and NHDRRNet generated images that had missing high-frequency information in the spectrum and blurred details, while our method preserved more high-frequency information and had complete details. In the red box, after fusing with AHDRNet, part of the edge texture of the shadow was eroded by the overexposed area, and NHDRNet had a

similar problem. Our method had relatively clear and complete shadow edge details. These results demonstrate that our method has good generalization ability for different scenes and exposure levels and excellent ability to preserve image details in overexposed areas.

Figure 8. The comparison results from the sen/tursun dataset (which does not provide GT images) were subjectively compared with images generated by other SOTA methods [6,8,12] after TM, and partial comparison area results were presented. The red arrows indicate the compositional differences in the comparison methods.

4.4. Timing Performance

In Table 5, we compared the computing time of our proposed method to previous SOTA models. We calculated the average time taken for 10 images with a resolution of 1500×1000 in the Kalantari dataset on the same GPU, and we only calculated the time taken for the CNN model on the GPU, without considering the time taken for computing optical flow on the CPU. As shown in the table, the computation time of our method is acceptable.

Table 5. This table shows the running times for different methods.

Method	GPU(s)
Kalantari	**0.19**
AHDRNet	0.23
NHDRRNet	0.25
HDR-GAN	0.23
Proposed Method	0.23

5. Conclusions

In this paper, we proposed the feature distillation transformer block (FDTB), which combines the Transformer with distillation structure, to process the features of LDR images with Swin Transformer block and separate the effective information and ghosting information of the image with the distillation structure, thereby achieving simultaneous image synthesis and deghosting. Furthermore, we combined U-Net with FDTB to propose the TransU-Fusion model, which aims to synthesize high-quality HDR images without ghosting for multiframe dynamic scenes. We used U-Net to downsample the image features for deep extraction and processed the local and global information of the image features with DRDB and FDTB. We then restored and supplemented the image details through upsampling to generate high-quality HDR images without ghosting. This model combines the advantages of CNN and Transformer for specific tasks and has achieved SOTA performance in generating image quality after a large number of experiments. The HDR images synthesized through our method can be effectively applied in fields such as automotive, aerospace, AI, etc., helping to achieve the correctness and accuracy of image-based big data processing and machine learning.

Author Contributions: Conceptualization, B.S. and R.G.; methodology, B.S.; software, R.G.; writing—original draft preparation, R.G.; writing—review and editing, B.S.; visualization, R.G.; supervision, Q.Y.; project administration, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not Available

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HDR	High Dynamic Range
LDR	Low Dynamic Range
AI	Artificial Intelligence
LDR	Directory of open access journals
DFTB	Three letter acronym
DRDB	Linear dichroism

VR	Virtual Reality
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
MSA	Multiheaded Self-Attention
FDTB	Feature Distillation Transformer Block
RSTB	Residual Swin Transformer Blocks
ViT	Vision Transformer

References

1. Nayar, S.K.; Mitsunaga, T. High dynamic range imaging: Spatially varying pixel exposures. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hilton Head, SC, USA, 15 June 2000; pp. 472–479.
2. Tumblin, J.; Agrawal, A.; Raskar, R. Why I want a gradient camera. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 103–110.
3. Kalantari, N.K.; Ramamoorthi, R. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* **2017**, *36*, 144. [CrossRef]
4. Wu, S.; Xu, J.; Tai, Y.W.; Tang, C.K. Deep high dynamic range imaging with large foreground motions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 117–132.
5. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
6. Yan, Q.; Gong, D.; Shi, Q.; Hengel, A.V.D.; Shen, C.; Reid, I.; Zhang, Y. Attention-guided network for ghost-free high dynamic range imaging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1751–1760.
7. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
8. Niu, Y.; Wu, J.; Liu, W.; Guo, W.; Lau, R.W.H. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Trans. Image Process.* **2021**, *30*, 3885–3896. [CrossRef] [PubMed]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
11. Park, N.; Kim, S. How do vision transformers work? *arXiv* **2022**, arXiv:2202.06709.
12. Yan, Q.; Zhang, L.; Liu, Y.; Zhu, Y.; Sun, J.; Shi, Q.; Zhang, Y. Deep HDR imaging via a non-local network. *IEEE Trans. Image Process.* **2020**, *29*, 4308–4322. [CrossRef] [PubMed]
13. Grosch, T. Fast and robust high dynamic range image generation with camera and object movement. In *Vision, Modeling and Visualization*; RWTH Aachen: Aachen, Germany, 2006; Volume 3.
14. Jacobs, K.; Loscos, C.; Ward, G. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Comput. Graph. Appl.* **2008**, *28*, 84–93. [CrossRef] [PubMed]
15. Pece, F.; Kautz, J. Bitmap movement detection: HDR for dynamic scenes. In Proceedings of the Conference on Visual Media Production, London, UK, 17–18 November 2010.
16. Zhang, W.; Cham, W.K. Gradient-directed multiexposure composition. *IEEE Trans. Image Process.* **2011**, *21*, 2318–2323. [CrossRef] [PubMed]
17. Heo, Y.S.; Lee, K.M.; Lee, S.U.; Moon, Y.; Cha, J. Ghost-free high dynamic range imaging. In Proceedings of the Computer Vision—ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010; Volume 10, pp. 486–500.
18. Bogoni, L. Extending dynamic range of monochrome and color images through fusion. In Proceedings of the 15th International Conference on Pattern Recognition, ICPR-2000, Barcelona, Spain, 3–7 September 2000; pp. 7–12.
19. Kang, S.B.; Uyttendaele, M.; Winder, S.; Szeliski, R. High dynamic range video. *ACM Trans. Graph. (TOG)* **2003**, *22*, 319–325. [CrossRef]
20. Zimmer, H.; Bruhn, A.; Weickert, J. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. *Comput. Graph. Forum* **2011**, *30*, 405–414. [CrossRef]
21. Gallo, O.; Troccoli, A.; Hu, J.; Pulli, K.; Kautz, J. Locally non-rigid registration for mobile HDR photography. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 49–56.
22. Sen, P.; Kalantari, N.K.; Yaesoubi, M.; Darabi, S.; Goldman, D.B.; Shechtman, E. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.* **2012**, *31*, 1–11. [CrossRef]
23. Hu, J.; Gallo, O.; Pulli, K.; Sun, X. HDR deghosting: How to deal with saturation? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1163–1170.

24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 10012–10022.
25. Liu, J.; Tang, J.; Wu, G. Residual feature distillation network for lightweight image super-resolution. In Proceedings of the Computer Vision—ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; pp. 41–55.
26. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. Electrical Engineering and Systems Science - Image and Video Processing. *arXiv* **2021**, arXiv:2108.10257.
27. Hui, Z.; Gao, X.; Yang, Y.; Wang, X. Lightweight image super-resolution with information multi-distillation network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2024–2032.
28. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [CrossRef]
29. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Tursun, O.T.; Akyüz, A.O.; Erdem, A.; Erdem, E. An objective deghosting quality metric for HDR images. *Comput. Graph. Forum* **2016**, *35*, 139–152. [CrossRef]
32. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018, Proceedings 4*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.