A Unified HDR Imaging Method with Pixel and Patch Level

Qingsen Yan[†1], Weiye Chen[†2], Song Zhang[†2], Yu Zhu[1], Jinqiu Sun[1], Yanning Zhang[1] *
[1]Northwestern Polytechnical University, [2]Xidian University

Abstract

Mapping Low Dynamic Range (LDR) images with different exposures to High Dynamic Range (HDR) remains nontrivial and challenging on dynamic scenes due to ghosting caused by object motion or camera jitting. With the success of Deep Neural Networks (DNNs), several DNNs-based methods have been proposed to alleviate ghosting, they cannot generate approving results when motion and saturation occur. To generate visually pleasing HDR images in various cases, we propose a hybrid HDR deghosting network, called HyHDRNet, to learn the complicated relationship between reference and non-reference images. The proposed HyHDRNet consists of a content alignment subnetwork and a Transformer-based fusion subnetwork. Specifically, to effectively avoid ghosting from the source, the content alignment subnetwork uses patch aggregation and ghost attention to integrate similar content from other non-reference images with patch level and suppress undesired components with pixel level. To achieve mutual guidance between patch-level and pixel-level, we leverage a gating module to sufficiently swap useful information both in ghosted and saturated regions. Furthermore, to obtain a high-quality HDR image, the Transformer-based fusion subnetwork uses a Residual Deformable Transformer Block (RDTB) to adaptively merge information for different exposed regions. We examined the proposed method on four widely used public HDR image deghosting datasets. Experiments demonstrate that HyHDRNet outperforms state-of-the-art methods both quantitatively and qualitatively, achieving appealing HDR visualization with unified textures and colors.

1. Introduction

Natural scenes cover a very broad range of illumination, but standard digital camera sensors can only measure a limited dynamic range. Images captured by cameras often have saturated or under-exposed regions, which lead to terrible visual effects due to severely missing details. High Dynamic Range (HDR) imaging has been developed to address these limitations, and it can display richer details. A common way of HDR imaging is to fuse a series of differently exposed Low Dynamic Range (LDR) images. It can recover a high-quality HDR image when both the scene and the camera are static, however, it suffers from ghosting artifacts on dynamic objects or hand-held camera scenarios.

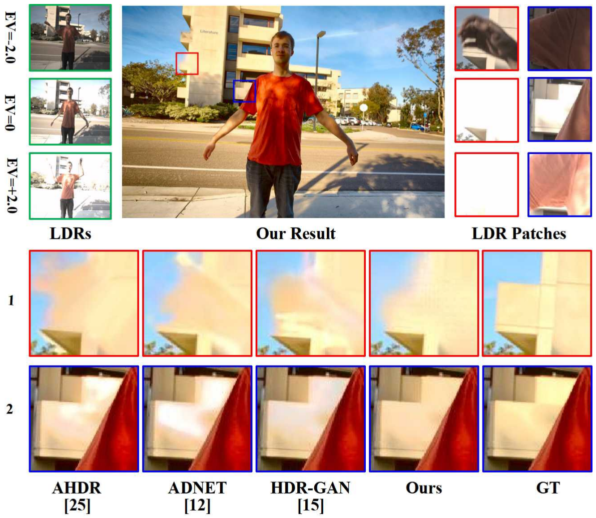


Figure 1. Our approach produces high-quality HDR images, leveraging both patch-wise aggregation and pixel-wise ghost attention. The two modules provide complementary visual information: patch aggregation recovers patch-level content of the complex distorted regions and ghost attention provides pixel-level alignment.

Several methods have been proposed to alleviate these problems, including alignment-based methods [1, 7, 20], rejection-based methods [3, 8, 16, 18, 33] and patch-based methods [5, 13, 19]. The alignment-based methods employ global (*e.g.*, homographies) or non-rigid alignment (*e.g.*, optical flow) to rectify the content of motion regions, but they are error-prone and cause ghosts due to saturation and occlusion. The rejection-based methods attempt to remove the motion components from the exposed images and replace them with the content of the reference image. Al-

*† The first three authors contributed equally to this work. This work was partially supported by NSFC (U19B2037, 61901384), Natural Science Basic Research Program of Shaanxi (2021JCW-03, 2023-JC-QN-0685), the Fundamental Research Funds for the Central Universities (D5000220444), and National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology. Corresponding author: Jinqiu Sun.

though these methods achieve good quality for static scenes, they discard the misalignment regions, which causes insufficient content in moving regions. The patch-based methods utilize patch-level alignment to transfer and fuse similar content. While these patch-based methods achieve better performance, they suffer from high computational costs.

With the rise of Deep Neural Networks (DNNs), many works directly learn the complicated mapping between LDR and HDR using a CNN. In general, these models follow the paradigm of alignment before fusion. The non-end-to-end DNN-based methods [6, 22] first align LDR images with optical flow or homographies, and then fuse aligned images to generate HDR images. The alignment approaches are error-prone and inevitably cause ghosting artifacts when complex foreground motions occur. Based on the attention-based end-to-end method AHDRNet [25, 26] which performs spatial attention to suppress motion and saturation, several methods [2, 12, 27, 30, 31] have been proposed to remove ghosting artifacts. The spatial attention module produces attention maps and element-wise multiplies with non-reference features, thus the model removes motion or saturated regions and highlights more informative regions.

However, the success of these methods relies on noticeable variations between reference and non-reference frames. These methods perform well in the marginal areas, even if there is a misalignment in the input images. Unluckily, spatial attention produces unsatisfactory results when motion and saturation are present simultaneously (see Figure 1). The reason can be attributed to that spatial attention uses element-wise multiplication, which only considers the features in the same positions. For example, in the reference frame of Figure 1 (*i.e.*, LDR with EV=0), the information in the over-exposed regions is unavailable, spatial attention can only rely on the non-saturated information of the same position (*i.e.*, moving regions) in the non-reference frame due to element-wise multiplication. Therefore, recovering the content of the moving and saturated regions is challenging. Finally, this limitation of spatial attention causes obvious ghosting artifacts in these complex cases.

To generate high-quality HDR images in various cases, we propose a Hybrid HDR deghosting Network, named HyHDRNet, to establish the complicated alignment and fusion relationship between reference and non-reference images. The proposed HyHDRNet comprises a content alignment subnetwork and a Transformer-based fusion subnetwork. For the content alignment subnetwork, inspired by patch-based HDR imaging methods [5, 19], we propose a novel Patch Aggregation (PA) module, which calculates the similarity map between different patches and selectively aggregates useful information from non-reference LDR images, to remove ghosts and generate content of saturation and misalignment. While the traditional patch-based HDR imaging methods have excellent performance but have the following drawbacks: 1) low patch utilization ratio caused by reusing the same patches, which leads to insufficient content during fusion, 2) structural destruction of images when transfering patches, 3) high computational complexity in full resolution. To this end, our Patch Aggregation mechanism 1) aggregates multiple patches which improves the patch utilization ratio 2) aggregates patches instead of exchanging them to maintain structural information, 3) calculates a similarity map within a window to reduce computational complexity. These advantages promote the network to remedy the content of saturated and motion regions(See Figure 9), other patch-based HDR imaging methods cannot achieve this goal. In a word, our PA module (patch level) discovers and aggregates similar patches within a large receptive field according to the similarity map, thus it can recover the content inside the distorted regions. To further avoid ghosting, we also employ a ghost attention module (pixel level) as a complementary branch for the PA module, and propose a gating module to achieve mutual guidance of these two modules in the content alignment subnetwork. In addition, unlike previous methods using DNN structure in the feature fusion stage which has static weights and only merges the local information, we propose a Transformer-based fusion subnetwork that uses Residual Deformable Transformer Block (RDTB) to model long-range dependencies of different regions. The RDTB can dynamically adjust weights and adaptively merge information in different exposure regions. The experiments demonstrate that our proposed method achieves state-of-the-art performance on public datasets. The main contributions of our work can be summarized as follows:

- We propose a hybrid HDR deghosting network to effectively integrate the advantage of patch aggregation and ghost attention using a gating strategy.
- We first introduce the patch aggregation module which selectively aggregates useful information from non-reference LDR images to remove ghosts and generate content for saturation.
- A novel residual deformable Transformer block is proposed, which can adaptively fuse a large range of information to generate high-quality HDR images.
- We carry out both qualitative and quantitative experiments, which show that our method achieves state-of-the-art results over four public benchmarks.

2. Related work

The related work includes four categories: alignment-based method, rejection-based method, patch-based method and CNN-based method.

Alignment-based method. Alignment methods utilize rigid or non-rigid registration to match reference images

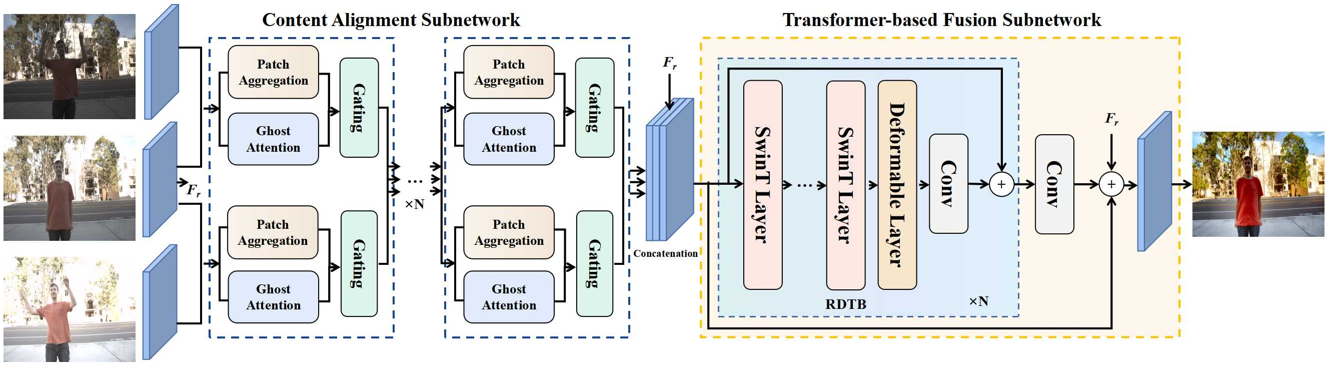


Figure 2. The illustration of HyHDRNet which consists of a content alignment subnetwork and a Transformer-based fusion subnetwork. The content alignment subnetwork uses patch aggregation and ghost attention to integrate similar content from other LDR images with patch level and suppress undesired components with pixel level. A gating module is leveraged to achieve mutual guidance between patch-level and pixel-level. Then the Transformer-based fusion subnetwork uses Residual Deformable Transformer Block (RDTB) to adaptively merge information for different exposed regions to obtain a high-quality HDR image.

densely. Bogoni *et al*. [1] computed flow vectors for alignment and used pattern selective fusion to obtain HDR images. Kang *et al*. [7] aligned images of video using optical flow in the luminance domain and fused the aligned images to remove ghosting artifacts. Tomaszewska *et al*. [20] employed SIFT feature to eliminate global misalignments, which can reduce blurring and artifacts. Although alignment methods can find dense correspondence, they are easy to fail in large motion and occlusion regions.

Rejection-based method. Rejection methods identify and remove motion regions of inputs after global registration operation, then fuse static regions to reconstruct HDR images. Grosch *et al*. [3] marked motion regions and used the predicted color's error map to obtain ghost-free HDR. Pece *et al*. [18] employed the median threshold bitmap of the input image to reject motion regions. Zhang *et al*. [33] used quality measures based on image gradients to generate a motion weighting map. Several methods [8, 16, 29] utilized rank minimization to detect the moving region and reconstruct ghost-free HDR images. However, pixel rejection abandons the misaligned regions, which causes insufficient content in moving regions.

Patch-based method. The patch-based methods use patch-wise alignment between the exposure images for deghosting. Sen *et al*. [19] proposed a patch-based energy minimization method, and it optimizes alignment and reconstruction jointly. Hu *et al*. [5] propagated intensity and gradient information iteratively using a coarse-to-fine schedule. Ma *et al*. [13] proposed a structural patch decomposition approach that decomposes an image patch into signal strength, signal structure and mean intensity components to reconstruct ghost-free images. But these methods do not compensate for saturation and suffer from high computational costs.

CNN-based method. Kalantari *et al*. [6] used a convolutional neural network to fuse LDR images after optical flow alignment. Wu *et al*. [22] defined HDR imaging as an image translation problem, and used an image homography transformer to align the camera motion. Yan *et al*. [25] leveraged an attention mechanism to suppress undesired information to merge ghost-free HDR. Yan *et al*. [28] designed a nonlocal block to learn the constraint of locality receptive field for global merging HDR. Niu *et al*. [15] proposed HDR-GAN to synthesize missing content using Generative Adversarial Networks. Xiong *et al*. [24] formulated HDR imaging as two problems, ghost-free fusion and ghost-based restoration. Ye *et al*. [32] proposed multi-step feature fusion and comparing/selecting operations to generate ghost-free images. Liu *et al*. [12] utilized the PCD pyramid alignment subnetwork to register in multiple layers. These methods still use pixel-level information to complete HDR imaging, which neglects the content of the patch level.

3. Proposed method

3.1. Problem Definition

Given a sequence of LDR images $\{L_1, L_2, ..., L_N\}$ with different exposures in dynamic scenes, our goal is to merge them into an HDR image without ghosting artifacts. The estimated HDR image will be structurally similar to a reference image L_r. Following previous works [6, 25], we utilize three LDR images (L_1, L_2, L_3) as input, and set L_2 as reference image.

Following [6, 25], we first map the LDR images L_i to the HDR domain using gamma correction:

$$H_i = L_i^{\gamma}/t_i, \tag{1}$$

where t_i denotes the exposure time of LDR image L_i, γ represents the gamma correction parameter, we set γ to 2.2.

Then we concatenate L_i and H_i along the channel dimension to get a 6-channel input $X_i = [L_i, H_i]$. Given inputs X_1, X_2, X_3, our model produces an HDR image $\hat{H}$ by:

$$\hat{H} = f(X_1, X_2, X_3; \theta), \tag{2}$$

where $f(\cdot)$ represents the HDR imaging network, θ is the parameters of the network.

3.2. Overview

As shown in Figure 2, the proposed HyHDRNet consists of two subnetworks, a content alignment subnetwork and a Transformer-based fusion subnetwork. The content alignment subnetwork first extracts shallow features from LDR sequence using three individual convolutional layers. Then the features utilize Patch Aggregation (PA) and Ghost Attention (GA) modules to separately aggregate useful patches from non-reference LDR images for complex cases and identify the misaligned components for easy cases. To take full advantage of these two modules, we use the gating module to recover both contents of saturated regions and sharp edges in motion and saturation. Considering that the missed content can be filled up using the information of neighborhood regions from the non-reference image, the PA module can search and assemble similar patches to complete misaligned or saturated regions, and the GA module is a complementary branch that removes ghosting with pixel level.

The Transformer-based fusion subnetwork aims to generate high-quality ghost-free HDR images from the extracted features of LDR inputs. We first utilize the features extracted from the content alignment subnetwork as input, then use several Swin Transformer layers and deformable layers to model long-range dependencies and dynamically merge useful information with a larger receptive field. Finally, we adopt convolutional layers to obtain a 3-channel HDR image.

3.3. Content Alignment Subnetwork

Shallow Feature Extraction Module. Given three 6-channel LDR images $X_i \in R^{H\times W\times 6}$, i =1, 2, 3, we adopt three convolutional encoders to extract the shallow feature $F_i \in R^{H\times W\times C}$:

$$F_i = e(X_i), i = 1, 2, 3 \tag{3}$$

Patch Aggregation Module. Traditional patch-based HDR imaging methods have excellent performance, but are limited by a low patch utilization ratio which causes insufficient content during fusion and has high computational complexity in full resolution.

In order to overcome the shortcoming of element-wise multiplication in spatial attention, the PA module selectively aggregates information by calculating the similarity map between patches. As shown in Figure 3, we first divide the shallow feature $F_i \in R^{R\times W\times C}$ into HW/M^2 non-overlapping $M \times M$ patches $\hat{F}_i$. We map each patch $\hat{F}_i$ to query $\hat{q}$, keys $\hat{k}_i$ and values $\hat{v}_i$:

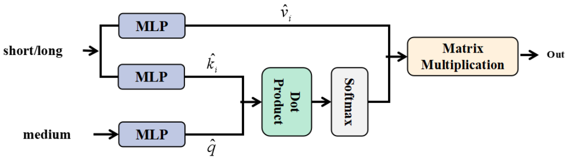


Figure 3. The structure of patch aggregation module. The PA module calculates the similarity map between different patches and selectively aggregates useful information from non-reference LDR images to remove ghosts and generate content of saturation and misalignment.

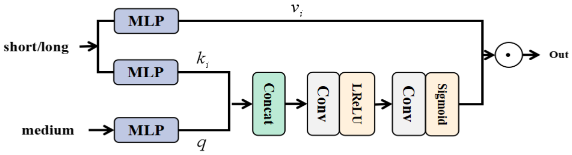


Figure 4. The structure of ghost attention module. The GA module utilizes spatial attention to identify the misaligned components.

$$\hat{q} = \hat{F}_r W_q, \hat{k}_i = \hat{F}_i W_k, \hat{v}_i = \hat{F}_i W_v, i = 1, 3 \tag{4}$$

where $\hat{F}_r$ is the features of the reference image.

Then we calculate the similarity map between non-reference feature $\hat{k}_i$ and reference feature $\hat{q}$, and apply the softmax function to the similarity map. Finally, we aggregate patches from non-reference feature $\hat{v}_i$ according to the similarity map, and obtain the output feature F^i_{pa}.

$$F^i_{pa} = Softmax(\hat{q}\hat{k}_i^T/\sqrt{d} + B)\hat{v}_i, \tag{5}$$

where B is a learnable position encoding, d is the dimension of $\hat{q}$. Note that, we also leverage the shifted windows [10] to achieve interaction of each patch.

Ghost Attention Module. Spatial Attention [25] can suppress motion and saturation well at the pixel level. As shown in Figure 4, we design a ghost attention which first maps the shallow feature F_i to query, keys and values. F_r is mapped from X_2, query q is mapped from F_r, keys k_i and values v_i are mapped from non-reference frame F_i, i =1,3. Note that W_q, W_k, W_v are shared in PA and GA.

$$q = F_r W_q, k_i = F_i W_k, v_i = F_i W_v, i = 1, 3 \tag{6}$$

Then, spatial attention $a(\cdot)$ calculate the attention between the non-reference feature k_i and reference feature q.

$$A_i = a(q, k_i), i = 1, 3 \tag{7}$$

Finally, we obtain the output feature F_{ga} by element-wise multiplication of non-reference feature v_i and attention map

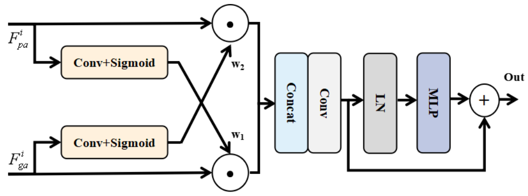


Figure 5. The structure of gating module. The gating module achieves mutual guidance of PA and GA modules in the content alignment subnetwork.

A_i, $\odot$ represents element-wise multiplication.

$$F^i_{ga} = v_i \odot A_i, i = 1, 3 \tag{8}$$

Gating Module. In order to integrate the advantage of PA and GA modules, we propose a gating module to achieve mutual guidance between these two modules. As shown in Figure 5, given the output features F_{pa} and F_{ga} from the PA module and GA module, we use φ function to predict the weight w_1 and w_2 for F_{ga} and F_{pa}, respectively. Then the element-wise multiplicated features are concatenated as inputs of a convolutional layer. Then we adopt residual connection, LN layer and MLP layer to obtain the output F_{out}, denoted by:

$$F_{out} = F_{gating} + LN(MLP(F_{gating})), \tag{9}$$

$$F_{gating} = Conv(Concat(F_{ga} \odot \varphi(F_{pa}), F_{pa} \odot \varphi(F_{ga}))), \tag{10}$$

where $LN(\cdot)$ is a layernorm layer, φ consists of a convolution and a sigmoid function.

3.4. Transformer-based Fusion Subnetwork

Inspired by SWIN-IR [9] and Deformable Attention Transformer [23], we use several Residual Deformable Transformer Blocks (RDTB) to dynamically merge features for generating high-quality HDR image. Compared with [23], we design a Window-based Deformable Transformer Layer (WDTL) to maintain the performance and decrease the computational cost, simultaneously. In addition, we also utilize channel attention in FFN to make the WDTL block converge more quickly and smooth the loss landscape.

Residual Deformable Transformer Block. As shown in Figure 2, our RDTB consists of Swin Transformer Layers (STL), Window-based Deformable Transformer Layers (WDTL), convolutional layers and residual connection. Given the input feature F^0_i of i-th RDTB, the output of RDTB can be formulated as:

$$R_{out} = Conv(WDTL(STL^N_i)) + F^0_i, \tag{11}$$

where $WDTL(\cdot)$ denotes Deformable Layer, $STL^N_i(\cdot)$ denotes the output of N-th Swin Transformer layer in i-th RDTB. STL includes multi-head intra-/inter-window self-attention and Feed-Forward Network (FFN), more details can be found in [9]. We will describe the details of WDTL in the following.

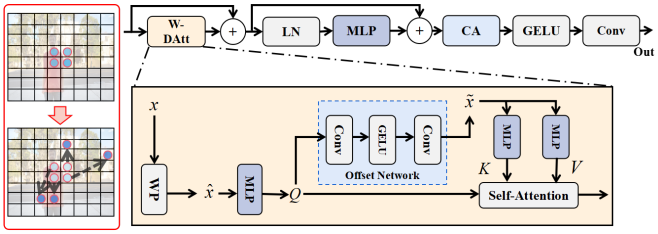


Figure 6. The window-based deformable Transformer layer. We uniformly place a group of reference points on the feature map, and the offsets of these reference points are learned from the query by the offset network. Then the deformed keys and values are projected from the sampled features with the deformed points, the transformed features are obtained by the self-attention block.

Window-based Deformable Transformer Layer. As shown in Figure 6, given input feature x, we first put x into Window-based Deformable self-Attention (WDAtt). We use Window Partition (WP) to divide the input into local windows $\hat{x}$, then we map them to query embedding Q with W'_q. Afterwards, we use a lightweight network $O_{offset}(\cdot)$ which consists of two convolutional layers and an activation function to predict the offsets Δo of each window, as shown in Figure 6. We then get the deformable windows $\tilde{x}$ and map them to embedding K and V.

$$Q = \hat{x}W'_q, \Delta o = O_{offset}(Q), \tilde{x} = \Phi(\hat{x}; p + \Delta o), \tag{12}$$

$$K = \tilde{x}W'_k, V = \tilde{x}W'_v. \tag{13}$$

Finally, we obtain transformed features X_{tr} using Q, K, V.

$$X_{tr} = Softmax(QK^T/\sqrt{D})V, \tag{14}$$

where Φ denotes the bilinear interpolation function, D represents the dimension of Q.

In order to make the WDTL block converge more easily and make the loss landscape more smoothing, inspired by [17], we utilize channel attention in the FFN layer to get X_{DT}, denoted by:

$$X'_{tr} = MLP(LN(X_{tr})) + X_{tr}, \tag{15}$$

$$X_{DT} = Conv(GELU(CA(X'_{tr}))), \tag{16}$$

where $CA(\cdot)$ is the channel attention block, $GELU(\cdot)$ denotes the GELU non-linearity function.

3.5. Training Loss

Since the HDR images are displayed in the tonemapped domain, we use μ-law [6] to map the image from the linear domain to the tonemapped domain:

$$T(x) = \frac{log(1 + \mu x)}{log(1 + \mu)}, \tag{17}$$

where $T(x)$ is the tonemapped function, $\mu = 5000$.

Given the estimated result $\hat{H}$ of HyHDRNet and the ground truth H, we calculate the tonemapped per-pixel loss (first term) and perceptual loss (second term) as follows:

$$L_{total} = ||T(H) - T(\hat{H})||_1 + \lambda||\phi_{i,j}(T(H)) - \phi_{i,j}(T(\hat{H}))||_1 \tag{18}$$

where $\phi_{i,j}$ represents the j-th convolutional feature extracted from VGG19 after the i-th max-pooling operation, $\lambda = 1e^{-2}$ is a weighting hyperparameter.

4. Experiments

4.1. Experimental Settings

Datasets. All methods are trained on two public datasets, Kalantari's dataset [6] and Hu's dataset [4]. Kalantari's dataset is captured in the real world, including 74 training and 15 testing samples. Three different LDR images in a sample are captured with exposure biases of {-2, 0, +2} or {-3, 0, +3}. Hu's dataset is a sensor-realistic synthetic dataset from a game engine, which is captured at three exposure levels (*i.e.*, {-2, 0, +2}). We use the dynamic scene images in Hu's dataset. Following [4], we choose the first 85 samples for training, and the remainder 15 samples for testing. To verify generalization performance, we evaluate all methods on Sen's dataset [19] and Tursun's dataset [21] (without ground truth).

Evaluation Metrics. We calculate five common metrics used for testing, *i.e.*, PSNR-L, PSRN-μ, SSIM-L, SSIM-μ and HDR-VDP-2 [14], where '-L' denotes linear domain, '-μ' represents tonemapping domain.

Implementation Details. We apply 8 × 8 patch size in the PA module. The layer numbers of STL and RDTB are 6 and 3. The window size of WDTL is 8 × 8. In the training stage, we crop the 128 × 128 patches with stride 64 for the training dataset. We use Adam optimizer, and set the batch size and learning rate as 4 and 0.0002, receptively. We drop the learning rate by 0.1 every 50 epochs. And we set β_1=0.9, β_2=0.999 and ϵ=1e^{-8} in Adam optimizer. We implement our model using PyTorch with 2 NVIDIA GeForce 3090 GPUs and train for 150 epochs.

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

To evaluate our model, we perform quantitative and qualitative experiments comparing with several state-of-the-art deep learning-based methods: Kalantari's method [6], DeepHDR [22], AHDRNet [25], NHDRR [28], HDR-GAN [15], ADNet [12], APNT [2], CA-ViT [11].

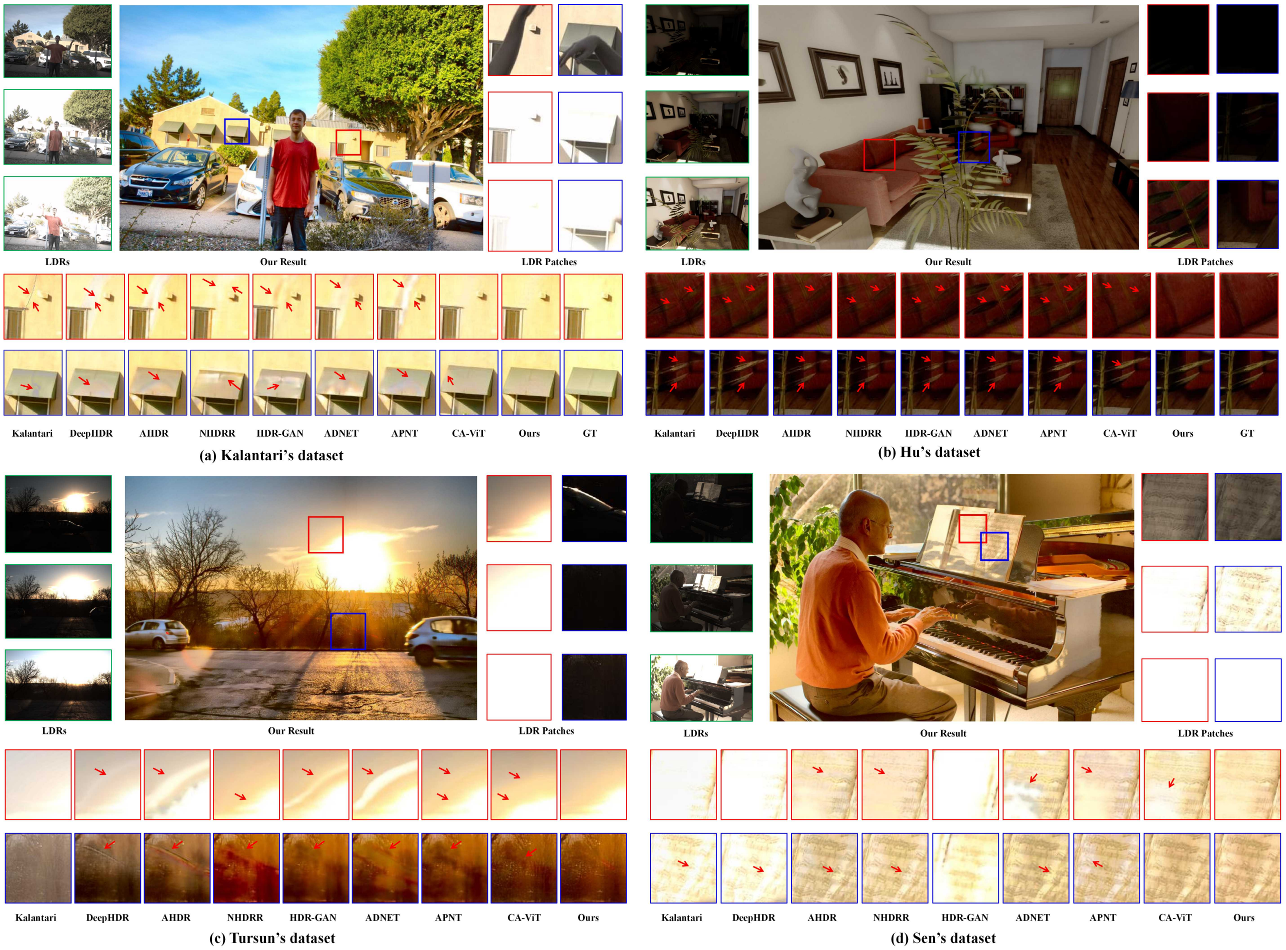
4.2.1 Datasets w/ Ground Truth.

As shown in Figure 7 (a) and (b), these two datasets have some challenging samples that cover a large area of foreground motions and over/under-exposed regions. Due to intractable large motion and occlusion, most comparing approaches produce ghosting artifacts in these areas. Kalantari's method and DeepHDR cannot handle the motion of the background due to error-prone alignments (*i.e.*, optical flow and tomographies), which cause undesirable ghosting (See the red block in Figure 7 (a)(b)). Since NHDRR and HDR-GAN do not have explicit alignment, they cannot recover the details in the saturated areas, therefore generating color distortions. Although AHDRNet and ADNet alleviate ghosting artifacts using ghost attention, they also suppress some useful neighborhood information, and cannot reconstruct large motion overlapped with over/under-exposure cases (See the blue block in Figure 7 (a) and (b)). APNT generates ghost regions and blurred edges because of the limitation of the patch-matching strategy. Without patch-level alignment, CA-ViT also produces ghosts. (See the red block in Figure 7 (a) and blue block in Figure 7 (b)). Thanks to the proposed patch aggregation and gating module, the proposed HyHDRNet not only integrates several similar patch features to complete misaligned or saturated regions, but also recovers sharp edges in motion or saturation.

The quantitative results for the proposed HyHDRNet on two datasets are shown in Table 1. The proposed HyHDRNet achieves state-of-the-art performance consistently on all five metrics of two datasets. We show that the improvement of HyHDRNet is very obvious against other previous SOTA methods. HyHDRNet surpasses the second-best methods by 0.32 db, 0.29 db in terms of PSNR-μ and PSNR-L and on Kalantari's dataset [6], and it also improves by 0.36 db and 0.74 db in terms of PSNR-μ and PSNR-L on Hu's dataset [4].

4.2.2 Evaluation on Datasets w/o Ground Truth.

We compare the proposed HyHDRNet with other approaches on Tursun's [21] and Sen's [19] datasets, which do not have ground truth. Visual comparisons are shown in Figure 7 (c)(d). As can be seen, most methods cannot recover the large area of saturation region and large motion. In Figure 7 (c) (red zoomed-in block), the halo effect of the sun is evident in the results of other methods, causing saturated ghosting artifacts. The over-exposure problem also exists in Figure 7 (d). Thanks to the proposed patch aggregation module and deformable transformer, HyHDRNet can assemble similar patches to fill up the missed content and adaptively fuse useful texture in a larger receptive field (See the red block in Figure 7 (c)). For large motion reconstruction problems, other methods have ghosting artifacts of moving car in the blue block of Figure 7 (c). Because the gating module can leverage both patch-wise aggregation and pixel-wise ghost attention, HyHDRNet can generate pleasing ghost-free results.



(a) Kalantari's dataset

(b) Hu's dataset

(c) Tursun's dataset

(d) Sen's dataset

Figure 7. Examples of Kalantari's dataset [6] and Hu's dataset [4] (top row) and Tursun's dataset [21] and Sen's dataset [19] (bottom row) datasets.

Table 1. The evaluation results on Kalantari's [6] and Hu's [4] datasets. The best and the second best results are highlighted in **Bold** and Underline, respectively.

Datasets	Models	Sen	Hu	Kalantari	DeepHDR	AHDRNet	NHDRR	HDR-GAN	ADNet	APNT	CA-ViT	Ours
Kalantari	PSNR-μ	40.95	32.19	42.74	41.64	43.62	42.41	43.92	43.76	43.94	44.32	**44.64**
	PSNR-L	38.31	30.84	41.22	40.91	41.03	41.08	41.57	41.27	41.61	42.18	**42.47**
	SSIM-μ	0.9805	0.9716	0.9877	0.9869	0.9900	0.9887	0.9905	0.9904	0.9898	**0.9916**	0.9915
	SSIM-L	0.9726	0.9506	0.9848	0.9858	0.9862	0.9861	0.9865	0.9860	0.9879	0.9884	**0.9894**
	HDR-VDP-2	55.72	55.25	60.51	60.50	62.30	61.21	65.45	62.61	64.05	66.03	**66.05**
Hu	PSNR-μ	31.48	36.56	41.60	41.13	45.76	45.15	45.86	46.79	46.41	48.10	**48.46**
	PSNR-L	33.58	36.94	43.76	41.20	49.22	48.75	49.14	50.38	47.97	51.17	**51.91**
	SSIM-μ	0.9531	0.9824	0.9914	0.9870	0.9956	0.9945	0.9956	0.9908	0.9953	0.9947	**0.9959**
	SSIM-L	0.9634	0.9877	0.9938	0.9941	0.9980	0.9989	0.9981	0.9987	0.9986	0.9989	**0.9991**
	HDR-VDP-2	66.39	67.58	72.94	70.82	75.04	74.86	75.19	76.21	73.06	77.12	**77.24**

4.3. Ablation Studies

We conduct ablation studies on Kalantari's dataset, and validate the effectiveness of each proposed module in HyHDRNet. We use the following variants of HyHDRNet: 1) **Model1**: As a baseline, we concatenate three images as input to the fusion network that employs the same RSTB

Table 2. The Ablation study on Kalantari dataset

Models	PSNR-μ	PSNR-L	HDR-VDP-2
1.Baseline	43.56	41.72	63.80
2.+GA	43.99	41.95	65.43
3.+PA	44.22	42.03	65.57
4.+GA+PA+Gating	44.49	42.32	65.98
5.+GA+PA+Addition	44.28	42.10	65.79
6.+GA+PA+Concat	44.32	42.16	65.84
7.+GA+PA+Gating+DL	44.60	42.41	66.03
8.A-DRDB	44.06	41.66	65.49
9.Ours	**44.64**	**42.47**	**66.05**

Table 3. The Ablation study on Patch Utilization method

Patch methods	PSNR-μ	PSNR-L	HDR-VDP-2
PM	43.24	40.60	63.10
PA	**44.64**	**42.47**	**66.05**

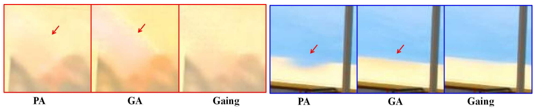


Figure 8. Visual results of ablation study.

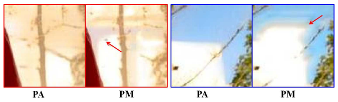


Figure 9. Visual results on PM and PA methods.

setting in SwinIR [9]. 2) **Model2**: We add GA module into model1. 3) **Model3**: PA module is added into model1. 4) **Model4**: We integrate GA, PA and gating modules into model1. 5) **Model5**: We replace gating with addition. 6) **Model6**: We replace gating with concatenation. 7) **Model7**: We add Deformable Layer (DL) into model4. 8) **Model8**: We replace Transformer-based fusion subnetwork with DRDB convolution network [25], named Alignment-DRDB (A-DRDB). 9) **Model9**: The proposed HyHDRNet full model, which adds perceptual loss into model7.

PA and GA modules. As shown in Table 2, compared with Model1, the performance of Model2 and Model3 are obviously improved. It demonstrates that the proposed PA and GA modules are both effective mechanisms for ghost removal. Note that Model3 with PA module achieves better numerical results than Model2, which means that the PA module is a better method for obtaining contents. As shown in Figure 8, the red block results of PA (Model3) do not have ghosting artifacts, which verifies that the PA module can selectively aggregate useful information from non-reference LDR images to remove ghosts and generate content for saturation. However, the PA module will cause blurry edges (see blue block results of PA). Compared with patch aggregation, the results of Model2 (GA) have sharp edges (blue block of Figure 8) but cannot remove ghosts completely. We consider it can be attributed to ghost attention suppress undesired components with pixel level. **Gating module.** Since the gating module can realize mutual guidance of PA and GA modules, the numerical (Table 2) and visual (Figure 8) results are both improved. As shown in Figure 8, since Model4 with GA can achieve mutual guidance of PA and GA modules in the content alignment subnetwork, the results can remove ghosts and hold on sharp edges (see last column), simultaneously. In addition, we also employ Addition (Model5) and Concatenation (Model6) to replace gating, but performance is degraded (See Table 2), which validates the effectiveness of the gating module. **Transformer-based Fusion Subnetwork.** In Table 2, compared Model7 with Model4, deformable layer (DL) obtains a better result. It can be attributed to the advantage of DL which captures more information for the fusion subnetwork. When we replace the Transformer-based Fusion Subnetwork with DRDBs, the numerical results obviously decrease, which demonstrates the effectiveness of the Transformer-based Fusion Subnetwork. **Patch Aggregation vs Patch Matching.** To verify the effectiveness of the patch aggregate module further, we show the qualitative and quantitative results of patch aggregation and patch matching in Figure 9 and Table 3. While the traditional patch matching-based HDR imaging methods have excellent performance, these methods only choose one best patch to reuse the original patches, this operation has a low patch utilization ratio which causes insufficient content during fusion. As shown in Figure 9, since saturated regions are hard to recover content with only one patch, thus the results still have obvious saturation regions. Different from patch matching, patch aggregation can selectively aggregate useful information from non-reference LDR images, which can generate content for saturation.

5. Conclusion

We propose an effective HDR imaging method based on content alignment and Transformer-based fusion. In content alignment subnetwork, we employ patch aggregation module to selectively aggregate useful patches from non-reference LDR images. We also propose a novel window-based deformable Transformer block to fuse a large range of information from extracted features. The major advantage of our proposed method is that it can fuse similar content from non-reference LDR images to remove ghosts and generate content for saturation in complex cases. Further, we demonstrate the superiority of our method over existing state-of-the-art methods on four publicly available datasets.

References

[1] L. Bogoni. Extending dynamic range of mono-chrome and color images through fusion. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 7–12, 2000. 1, 3

[2] Jie Chen, Yang Zaifeng, Chan Tsz Nam, Li Hui, Hou Junhui, and Chau. Lap-Pui. Attention-guided progressive neural texture fusion for high dynamic range image restoration. *IEEE Transactions on Image Processing*, 31:2661–2670, 2022. 2, 6

[3] Thorsten Grosch. Fast and robust high dynamic range image generation with camera and object movement. In *IEEE Conference of Vision , Modeling and Visualization*, 2006. 1, 3

[4] Jinhan Hu, Gyeongmin Choe, Zeeshan Nadir, Osama Nabil, Seok-Jun Lee, Hamid Sheikh, Youngjun Yoo, and Michael Polley. Sensor-realistic synthetic data engine for multi-frame high dynamic range photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 516–517, 2020. 6, 7

[5] Jun Hu, O. Gallo, K. Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1163–1170, 2013. 1, 2, 3

[6] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):1–12, 2017. 2, 3, 5, 6, 7

[7] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High dynamic range video. *ACM Transactions on Graphics*, 22(3):319–325, 2003. 1, 3

[8] Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE signal processing letters*, 21(9):1045–1049, 2014. 1, 3

[9] Jingyun Liang, Cao Jiezhang, Sun Guolei, Zhang Kai, Van Gool Luc, and Timofte Radu. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR) Workshop*, pages 1833–1844, 2021. 5, 8

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10012–10022, 2021. 4

[11] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 344–360, 2022. 6

[12] Zhen Liu, Lin Wenjie, Li Xinpeng, Rao Qing, Jiang Ting, Han Mingyan, Fan Haoqiang, Sun Jian, and Liu Shuaicheng. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, pages 463–470, 2021. 2, 3, 6

[13] Kede Ma, Li Hui, Yong Hongwei, Wang Zhou, Meng Deyu, and Zhang. Lei. Robust multi-exposure image fusion: A structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017. 1, 3

[14] Rafat Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. HDR-VDP-2:a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Siggraph*, pages 1–14, 2011. 6

[15] Yuzhen Niu, Wu Jianbin, Liu Wenxi, Guo Wenzhong, and WH Lau. Rynson. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021. 3, 6

[16] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1219–1232, 2015. 1, 3

[17] Namuk Park and Kim Songkuk. How do vision transformers work. In *International Conference on Learning Representations (ICLR)*, 2022. 5

[18] Fabrizio Pece and Jan Kautz. Bitmap movement detection: HDR for dynamic scenes. In *Visual Media Production*, pages 1–8, 2010. 1, 3

[19] Pradeep Sen, Khademi Kalantari Nima, Yaesoubi Maziar, Darabi Soheil, Dan B Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Transactions on Graphics*, 31(6):1–11, 2012. 1, 2, 3, 6, 7

[20] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. In *International Conference in Central Europe on Computer Graphics and Visualization*, 2007. 1, 3

[21] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. An objective deghosting quality metric for HDR images. *Comput. Graph. Forum*, 35(2):139–152, 2016. 6, 7

[22] Shangzhe Wu, Xu Jiarui, Tai Yu-Wing, and Tang. Chi-Keung. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018. 2, 3, 6

[23] Zhuofan Xia, Xuran Pan, Shiji Song, Erran Li Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4803, 2022. 5

[24] Pengfei Xiong and Chen Yu. Hierarchical fusion for practical ghost-free high dynamic range imaging. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4025–4033, 2021. 3

[25] Qingsen Yan, Gong Dong, Shi Qinfeng, van den Hengel Anton, Shen Chunhua, Reid Ian, and Zhang Yanning. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760, 2019. 2, 3, 4, 6, 8

[26] Qingsen Yan, Dong Gong, Javen Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Dual-attention-guided network for ghost-free high dynamic range imaging. *International Journal of Computer Vision*, 130(1):76–94, 2022. 2

[27] Qingsen Yan, Dong Gong, Javen Qinfeng Shi, Anton van den Hengel, Jinqiu Sun, Yu Zhu, and Yanning Zhang. High dynamic range imaging via gradient-aware context aggregation network. *Pattern Recognition*, 122:108342, 2022. 2

[28] Qingsen Yan, Zhang Lei, Liu Yu, Zhu Yu, Sun Jinqiu, Shi Qinfeng, and Zhang Yanning. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020. 3, 6

[29] Qingsen Yan, Jinqiu Sun, Haisen Li, Yu Zhu, and Yanning Zhang. High dynamic range imaging by sparse representation. *Neurocomputing*, 269:160–169, 2017. 3

[30] Qingsen Yan, Bo Wang, Peipei Li, Xianjun Li, Ao Zhang, Qinfeng Shi, Zheng You, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Ghost removal via channel attention in exposure fusion. *Computer Vision and Image Understanding*, 201:103079, 2020. 2

[31] Qingsen Yan, Song Zhang, Weiye Chen, Yuhang Liu, Zhen Zhang, Yanning Zhang, Javen Qinfeng Shi, and Dong Gong. A lightweight network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshop*, pages 824–832, 2022. 2

[32] Qian Ye, Jun Xiao, Kin-man Lam, and Takayuki Okatani. Progressive and selective fusion network for high dynamic range imaging. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5290–5297, 2021. 3

[33] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2011. 1, 3