Ghost-free High Dynamic Range Imaging with Context-aware Transformer

Zhen Liu[1,*], Yinglong Wang[2,*], Bing Zeng[3], and Shuaicheng Liu[3,1,†]

[1] Megvii Technology, Beijing, China
liuzhen03@megvii.com
[2] Noah's Ark Lab, Huawei Technologies, Shenzhen, China
ylwanguestc@gmail.com
[3] University of Electronic Science and Technology of China, Chengdu, China
{zengbing,liushuaicheng}@uestc.edu.cn
*Joint First Author, †Corresponding Author

Abstract. High dynamic range (HDR) deghosting algorithms aim to generate ghost-free HDR images with realistic details. Restricted by the locality of the receptive field, existing CNN-based methods are typically prone to producing ghosting artifacts and intensity distortions in the presence of large motion and severe saturation. In this paper, we propose a novel Context-Aware Vision Transformer (CA-ViT) for ghost-free high dynamic range imaging. The CA-ViT is designed as a dual-branch architecture, which can jointly capture both global and local dependencies. Specifically, the global branch employs a window-based Transformer encoder to model long-range object movements and intensity variations to solve ghosting. For the local branch, we design a local context extractor (LCE) to capture short-range image features and use the channel attention mechanism to select informative local details across the extracted features to complement the global branch. By incorporating the CA-ViT as basic components, we further build the HDR-Transformer, a hierarchical network to reconstruct high-quality ghost-free HDR images. Extensive experiments on three benchmark datasets show that our approach outperforms state-of-the-art methods qualitatively and quantitatively with considerably reduced computational budgets. Codes are available at https://github.com/megvii-research/HDR-Transformer.

Keywords: High Dynamic Range Deghosting, Context-Aware Vision Transformer

1 Introduction

Multi-frame high dynamic range (HDR) imaging aims to generate images with a wider dynamic range and more realistic details by merging several low dynamic range (LDR) images with varying exposures, which can be well fused to an HDR image if they are aligned perfectly [31,32,21,41,23,20]. In practice, however, this

Fig. 1. Visual comparisons with the state-of-the-art methods [33,10,13,37,39,25] on Kalantari *et al* [13]'s dataset. As shown, the patch-match based methods [33,10] and the CNN-based methods [13,37,39,25] fail to remove the long-range ghosts caused by large motion and hallucinate reasonable local details in saturated regions. On the contrary, the proposed HDR-Transformer can effectively remove the ghosting artifacts and produce visual consistent local details.

ideal situation is often undermined by camera motions and foreground dynamic objects, yielding unfavorable *ghosting artifacts* in the reconstructed HDR results. Various methods, commonly referred to as *HDR deghosting algorithms*, have thus been proposed to acquire high-quality ghost-free HDR images.

Traditionally, several methods propose to remove ghosting artifacts by aligning the input LDR images [2,10,14,42] or rejecting misaligned pixels [7,8,27,11,15] before the image fusion. However, accurate alignment is challenging, and the overall HDR effect is diminished when useful information is dropped by imprecise pixel rejection. Therefore, CNN-based learning algorithms have been introduced to solve ghosting artifact by exploring deep features in data-driven manners.

Existing CNN-based deghosting methods can be mainly classified into two categories. In the first category, LDR images are pre-aligned using homography [9] or optical flow [1], and then multi-frame fusion and HDR reconstruction are performed using a CNN [13,29,28,37]. However, homography cannot align dynamic objects in the foreground, and optical flow is unreliable in the presence of occlusions and saturations. Hence, the second category proposes end-to-end networks with implicit alignment modules [39,19,4] or novel learning strategies [25,30] to handle ghosting artifacts, achieving state-of-the-art performance. Nonetheless, the restraints appear when confronted with long-range object movements and heavy intensity variations. Fig. 1 shows a representative scene where

large motions and severe saturations occur, producing unexpected ghosting and distortion artifacts in the results of previous CNN-based methods. The reason lies in the intrinsic locality restriction of convolution. CNN needs to stack deep layers to obtain a large receptive field and is thus ineffective to model long-range dependency (e.g., ghosting artifacts caused by large motion) [24]. Moreover, convolutions are content-independent as the same kernels are shared within the whole image, ignoring the long-range intensity variations of different image regions [16]. Therefore, exploring content-dependent algorithms with long-range modeling capability is demanding for further performance improvement.

Vision Transformer (ViT) [6] has recently received increasing research interest due to its superior long-range modeling capability. However, our experimental results indicate two major issues that hinder its applications on HDR deghosting. On the one hand, Transformers lack the inductive biases inherent to CNN and therefore do not generalize well when trained on insufficient amounts of data [6,16], despite the fact that available datasets for HDR deghosting are limited as gathering huge numbers of realistic labeled samples is prohibitively expensive. On the other hand, the neighbor pixel relationships of both intra-frame and inter-frame are critical for recovering local details across multiple frames, while the pure Transformer is ineffective for extracting such local context.

To this end, we propose a novel Context-Aware Vision Transformer (CA-ViT), which is formulated to concurrently capture both global and local dependencies with a dual-branch architecture. For the global branch, we employ a window-based multi-head Transformer encoder to capture long-range contexts. For the local branch, we design a local context extractor (LCE), which extracts the local feature maps through a convolutional block and selects the most useful features across multiple frames by channel attention mechanism. The proposed CA-ViT, therefore, makes local and global contexts work in a complementary manner. By incorporating with the CA-ViT, we propose a novel Transformer-based framework (termed as HDR-Transformer) for ghost-free HDR imaging.

Specifically, the proposed HDR-Transformer mainly consists of a feature extraction network and an HDR reconstruction network. The feature extraction network extracts shallow features and fuses them coarsely through a spatial attention module. The early convolutional layers can stabilize the training process of the vision Transformer and the spatial attention module helps to suppress undesired misalignment. The HDR reconstruction network takes the proposed CA-ViT as basic components and is constituted hierarchically. The CA-ViTs model both long-range ghosting artifacts and local pixel relationship, thus helping to reconstruct ghost-free high-quality HDR images (an example is shown in Fig. 1) without the need of stacking very deep convolution blocks. In summary, the main contributions of this paper can be concluded as follows:

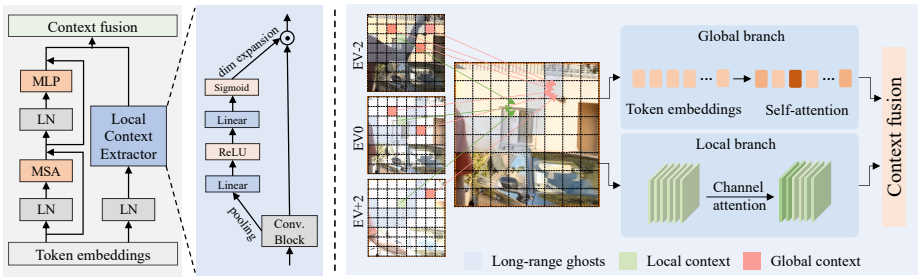
- We propose a new vision Transformer, called CA-ViT, which can fully exploit both global and local image context dependencies, showing significant performance improvements over prior counterparts.
- We present a novel HDR-Transformer that is capable of removing ghosting artifacts and reconstructing high-quality HDR images with lower compu-

tational costs. To our best knowledge, this is the first Transformer-based framework for HDR deghosting.
– We conduct extensive experiments on three representative benchmark HDR datasets, which demonstrates the effectiveness of HDR-Transformer against existing state-of-the-art methods.

2 Related Work

2.1 HDR Deghosting Algorithms

We summarize existing HDR deghosting algorithms into three categories, i.e., motion rejection methods, image registration methods, and CNN-based methods.
Motion rejection methods Methods based on motion rejection proposed first to register the LDR images globally and then reject the pixels which are detected as misaligned. Grosch *et al.* generated an error map based on the alignment color differences to reject mismatched pixels [8]. Pece *et al.* detected motion areas using a median threshold bitmap for input LDR images [27]. Jacobs *et al.* identified misaligned locations using weighted intensity variance analysis [11]. Zhang *et al.* [41] and Khan *et al.* [15] proposed to calculate gradient-domain weight maps and probability maps for the LDR input images, respectively. Additionally, Oh *et al.* presented a rank minimization method for the purpose of detecting ghosting regions [26]. These methods frequently produce unpleasing HDR results due to the loss of useful information while rejecting pixels.
Motion registration methods Motion registration methods rely on aligning the non-reference LDR images to the reference one before merging them. Begoni *et al* proposed using optical flow to predict motion vectors [2]. Kang *et al.* transferred the LDR picture intensities to the luminance domain based on the exposure time and then estimated optical flow to account for motion [14]. Zimmer *et al.* reconstructed the HDR image by first registering the LDR images with optical flow [42]. Sen *et al.* presented a patch-based energy minimization method that simultaneously optimizes alignment and HDR reconstruction [33]. Hu *et al.* proposed to optimize the image alignment using brightness and gradient consistencies on the transformed domain [10]. Motion registration methods are more robust than motion rejection methods. However, when large motions occur, this approach generates visible ghosting artifacts.
CNN-based methods Several CNN-based methods have been recently proposed. Kalantari *et al.* proposed the first CNN-based method for multi-frame HDR imaging of dynamic scenes. They employed a CNN to blend the LDR images after aligning them with optical flow [13]. Wu *et al.* developed the first non-flow-based framework by formulating HDR imaging as an image translation problem [37]. Instead of using explicit alignment, Yan *et al.* adopted a spatial attention module to address ghosting artifacts [39]. Prabhakar *et al.* proposed an efficient method to generate HDR images with bilateral guided upsampler [28] and further explored zero and few-shot learning for HDR Deghosting [30]. Lately, Niu *et al.* proposed the first GAN-based framework for multi-frame HDR imaging [25]. The approaches based on CNNs demonstrate superior capabilities and



(a) Context-aware Vision Transformer (b) CA-ViT for HDR deghosting

Fig. 2. Illustration of the proposed CA-ViT. As shown in Fig. 2 (a), the CA-ViT is designed as a dual-branch architecture where the global branch models long-range dependency among image contexts through a multi-head Transformer encoder, and the local branch explores both intra-frame local details and inner-frame feature relationship through a local context extractor. Fig. 2 (b) depicts the key insight of our HDR deghosting approach with CA-ViT. To remove the residual ghosting artifacts caused by large motions of the hand (marked with blue), long-range contexts (marked with red), which are required to hallucinate reasonable content in the ghosting area, are modeled by the self-attention in the global branch. Meanwhile, the well-exposed non-occluded local regions (marked with green) can be effectively extracted with convolutional layers and fused by the channel attention in the local branch.

achieve state-of-the-art performance. However, ghosting artifacts can still be observed when confronted with large motion and extreme saturation.

2.2 Vision Transformers

Transformers have achieved huge success in the field of natural language processing [36,5], where the multi-head self-attention mechanism is employed to capture long-range correlations between word token embeddings. Recently, ViT [6] has shown that a pure Transformer can be applied directly to sequences of non-overlapping image patches and performs very well on image classification tasks. Liu *et al.* developed Swin Transformer, a hierarchical structure where cross-window contexts are captured through the shift-window scheme [18]. Chen *et al.* built IPT, a pretrained Transformer model for low-level computer vision tasks [3]. Liang *et al.* extended the Swin Transformer for image restoration and proposed SwinIR, achieving state-of-the-art performance on image super-resolution and denoising [16]. Unlike CNN-based methods, our approach is inspired by [18,16] and built on Transformers.

3 Method

3.1 CA-ViT

Unlike prior vision Transformers that adopt the pure Transformer encoder, we propose a dual-branch context-aware vision Transformer (CA-ViT), which ex-

plores both the global and local image information. As depicted in Fig. 2 (a), the proposed CA-ViT is constructed with a global Transformer encoder branch and a local context extractor branch.

Global Transformer Encoder For the global branch, we employ a window-based multi-head Transformer encoder [6] to capture long-range information. The Transformer encoder consists of a multi-head self-attention (MSA) module and a multi-layer perceptron (MLP) with residual connection.

Considering the input token embeddings $E \in \mathbb{R}^{H\times W\times D}$, the global context branch can be formulated as:

$$\begin{aligned} E &= MSA(LN(E)) + E, \\ CTX_{global} &= MLP(LN(E)) + E, \end{aligned} \tag{1}$$

where LN denotes LayerNorm, and CTX_{global} denotes the global contexts captured by the Transformer encoder.

Local Feature Extractor For the local branch, we design a local context extractor (LCE) to extract local information CTX_{local} from adjacent pixels and select cross-channel features for fusion, which is defined as:

$$CTX_{local} = LCE(LN(E)). \tag{2}$$

Specifically, for the token embeddings E normalized with an LN layer, we first reshape them into $H \times W \times D$ features and use a convolution block to extract local feature maps f_{local}. The local features are then average pooled to a shape of $1\times1\times D$, and the channel-wise weights ω are calculated from two linear layers followed by a ReLU and a sigmoid activation layer, respectively. Afterward, the useful feature maps are selected through a channel-wise calibration from the original local features f_{local}, i.e.,

$$\begin{aligned} f_{local} &= Conv(LN(E)), \\ \omega &= \sigma_2(FC(\sigma_1(FC(f_{local})))), \\ CTX_{local} &= \omega \odot f_{local}, \end{aligned} \tag{3}$$

where σ_1 and σ_2 denote the ReLU and sigmoid layer, and FC denotes the linear layer. As a result, the local context branch not only adds the locality into the Transformer encoder, but also identifies the most informative local features across multiple frames for feature fusion.

Finally, a context fusion layer is employed to combine the global and local contexts. Although other transformation functions (e.g., linear or convolution layer) can be used to implement the context fusion layer, in this paper, we simply merge the contexts by element-wise addition to reduce the influence of additional parameters.

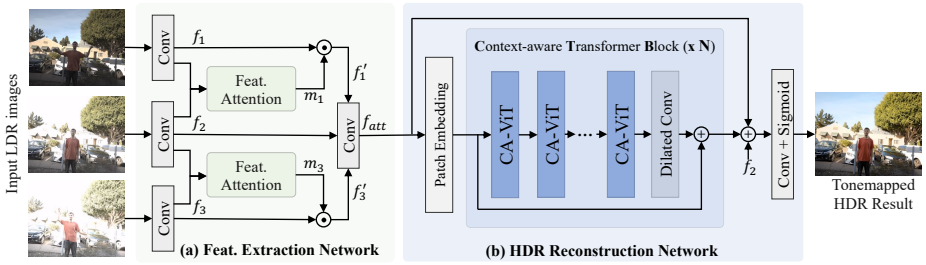


Fig. 3. The network architecture of HDR-Transformer. The pipeline consists of two stages: (a) The feature extraction network first extracts the coarse features through a spatial attention module. (b) The extracted features are then fed into the HDR reconstruction network to recover the HDR results. The HDR reconstruction network consists of several Context-aware Transformer Blocks (CTBs), which take the proposed CA-ViT as basic components.

3.2 HDR Deghosting

The task of deep HDR deghosting aims to reconstruct a ghost-free HDR image through deep neural networks. Following most of the previous works [13,37,39], we consider 3 LDR images (i.e., $I_i, i = 1, 2, 3$) as input and refer to the middle frame I_2 as the reference image. To better utilize the input data, the LDR images $\{I_i\}$ are first mapped to the HDR domain using the gamma correction, generating the gamma-corrected images $\{\check{I}_i\}$:

$$\check{I}_i = \frac{(I_i)^\gamma}{t_i}, \quad i = 1, 2, 3, \tag{4}$$

where t_i denotes the exposure time of I_i, and γ is the gamma correction parameter, which is set to 2.2 in this paper. We then concatenate the original LDR images $\{I_i\}$ and the corresponding gamma-corrected images $\{\check{I}_i\}$ into a 6-channels input $\{X_i\}$. This strategy is suggested in [13] as the LDR images help to detect the noisy or saturated regions, while the gamma-corrected images are helpful for detecting misalignments. Finally, the network $\Phi(\cdot)$ is defined as:

$$\hat{I^{\mathrm{H}}} = \Phi(X_i; \theta), \quad i = 1, 2, 3, \tag{5}$$

where $\hat{I^{\mathrm{H}}}$ denotes the reconstructed HDR image, and θ is the network parameters to be optimized.

Instead of stacking very deep CNN layers to obtain a large receptive field as existing CNN-based approaches, we propose the HDR-Transformer to handle HDR deghosting. Our key insight is that, with the specifically-designed dual-branch CA-ViT, the long-range ghosting can be well modeled in the global branch, and the local branch helps to recover fine-grained details. We describe the architecture of the proposed HDR-Transformer in the next section.

3.3 Overall Architecture of HDR-Transformer

As illustrated in Fig. 3, the overall structure of our proposed HDR-Transformer mainly consists of two components, i.e., feature extraction network (Fig. 3 (a)) and HDR reconstruction network (Fig. 3 (b)). Given three input images, we first extract the spatial features through a spatial attention module. The extracted coarser features are then embedded and fed into the Transformer-based HDR reconstruction network, generating the reconstructed ghost-free HDR image.

Feature Extraction Network The early convolution layers help to stabilize the training process of Vision Transformers [38]. For the input images $X_i \in \mathbb{R}^{H\times W\times 6}, i = 1, 2, 3$, we first extract the shallow features $f_i \in \mathbb{R}^{H\times W\times C}$ by three separate convolution layers, where C is the number of channels. Then, we concatenate each non-reference feature (i.e., f_1 and f_3) with the reference feature f_2 and calculate the attention maps m_i through a spatial attention module $\mathcal{A}$:

$$m_i = \mathcal{A}(f_i, f_2), \quad i = 1, 3, \tag{6}$$

The attention features $f_i^{'}$ are computed by multiplying the attention maps m_i by the non-reference features f_i, i.e.,

$$f_i^{'} = f_i \odot m_i, \quad i = 1, 3, \tag{7}$$

where $\odot$ denotes the element-wise multiplication. The spatial attention module has been proved to effectively reduce undesired contents caused by foreground object movements [39,19]. The convolution layers in the attention module can also increase the inductive biases for the subsequent Transformer layers.

HDR Reconstruction Network As shown in Fig. 3, the HDR reconstruction network is mainly composed of several context-aware Transformer blocks (CTBs). The input of the first CTB $f_{att} \in \mathbb{R}^{H\times W\times D}$ is obtained from $f_1^{'}$, f_2, and $f_3^{'}$ and embedded into token embeddings, where D denotes the embed dimension. The HDR result is reconstructed by N subsequent CTBs and a following convolution block. We also adopt the global skip connection to stabilize the optimization process.

Context-aware Transformer Block As illustrated in Fig. 2 (b), when suffering occlusion caused by large object movements and heavy saturation, long-range context is required for removing the corresponding ghosting regions and hallucinating reasonable content, while the non-occluded areas can be fused well by the convolutional layers. To this end, we develop the context-aware Transformer block (CTB) by taking the proposed CA-ViT as the basic component.

For clarity, each CTB contains M CA-ViTs. For the n-th CTB with the input of $F_{n,0}$, the output of the m-th CA-ViT can be formulated as:

$$F_{n,m} = \mathcal{C}_{n,m}(F_{n,m-1}), \quad m = 1, 2, ..., M, \tag{8}$$

where $C_{n,m}(\cdot)$ denotes the corresponding CA-ViT. Then, we feed the output of the M-th CA-ViT into a dilated convolution layer. The dilated convolutional layer is employed to increase the receptive field of the context range. We also adopt the residual connection in each CTB for better convergence. Consequently, the output of the n-th CTB is formulated as:

$$F_n = DConv(F_{n,M}) + F_{n,0}, \tag{9}$$

where $DConv(\cdot)$ denotes the dilated convolutional layer, and M and N are empirically set to 6 and 3, respectively.

3.4 Loss Function

As HDR images are typically viewed after tonemapping, we compute the loss in the tonemapped domain using the commonly used μ-law function:

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

where $\mathcal{T}(x)$ is the tonemapped HDR image, and we set μ to 5000. Unlike previous methods [13,37,39] that only adopt the pixel-wise loss (e.g., l_1 or l_2 error), we utilize l_1 loss and perceptual loss to optimize the proposed HDR-Transformer. Given the estimated HDR image $I^{\hat{H}}$ and the ground truth HDR image I^H, the l_1 loss term is defined as:

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

The perceptual loss [12] is widely used in image inpainting [17] for better visual quality improvements. We also apply the perceptual loss to enhance the quality of the reconstructed HDR images:

$$\mathcal{L}_p = \sum_j \| \Psi_j(\mathcal{T}(I^H)) - \Psi_j(\mathcal{T}(I^{\hat{H}})) \|_1, \tag{12}$$

where $\Psi(\cdot)$ denotes the activation feature maps extracted from a pre-trained VGG-16 network [34], and j denotes the j-th layer. We analyze the effectiveness of the perceptual loss in our ablation study (Sec. 4.3). Eventually, our training loss function $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_r + \lambda_p \mathcal{L}_p, \tag{13}$$

where λ_p is the hyper-parameter and we set it to 0.01.

4 Experiments

4.1 Dataset and Implementation Details

Datasets Following previous methods [37,39,40,25], we train our network on the widely used Kalantari *et al.*'s dataset [13], which consists of 74 samples for

Table 1. Quantitative comparison between previous methods and ours on Kalantari *et al.* [13]'s test set. We use PSNR, SSIM, and HDR-VDP-2 as evaluation metrics. The '-μ' and '-l' refers to values calculated on the tonemapped domain and the linear domain, respectively. All values are the average over 15 testing images and higher better. The best results are highlighted and the second best are underlined.

Metrics	Methods								
	Sen12 [33]	Hu13 [10]	Kalantari17 [13]	DeepHDR [37]	AHDRNet [39]	NHDRRNet [40]	HDR-GAN [25]	SwinIR [16]	HDR-Transformer Ours
PSNR-μ	40.80	35.79	42.67	41.65	43.63	42.41	<u>43.92</u>	43.42	**44.32**
PNRR-l	38.11	30.76	41.23	40.88	41.14	41.43	41.57	<u>41.68</u>	**42.18**
SSIM-μ	0.9808	0.9717	0.9888	0.9860	0.9900	0.9877	<u>0.9905</u>	0.9882	**0.9916**
SSIM-l	0.9721	0.9503	0.9846	0.9858	0.9702	0.9857	<u>0.9865</u>	0.9861	**0.9884**
HDR-VDP-2	59.38	57.05	65.05	64.90	64.61	61.21	<u>65.45</u>	64.52	**66.03**

training and 15 samples for testing. Each sample from Kalantari *et al.*'s dataset comprises three LDR images with exposure values of $\langle -2, 0, +2 \rangle$ or $\langle -3, 0, +3 \rangle$, as well as a ground truth HDR image. During the training, we first crop patches of size 128×128 with a stride of 64 from the training set. We then apply rotation and flipping augmentation to increase the training size. We quantitatively and qualitatively evaluate our method on Kalantari *et al.*'s testing set. We also conduct evaluations on Sen *et al.* [33]'s and Tursun *et al.* [35]'s datasets to verify the generalization ability of our method.

Evaluation Metrics We use PSNR and SSIM as evaluation metrics. To be more precise, we calculate PSNR-l, PSNR-μ, SSIM-l, and SSIM-μ scores between the reconstructed HDR images and their corresponding ground truth. The '-l' and '-μ' denote the linear and tonemapped domain values, respectively. Given that HDR images are typically displayed on LDR displays, metrics in the tonemapped domain more accurately reflect the quality of the reconstructed HDR images. Additionally, we conduct evaluations using the HDR-VDP-2 [22], which is developed specifically for evaluating the quality of HDR images.

Implementation Details Our HDR-Transformer is implemented by PyTorch. We use the ADAM optimizer with an initial learning rate of 2e-4 and set β_1 to 0.9, β_2 to 0.999, and ϵ to 1e-8, respectively. We train the network from scratch with a batch size of 16 and 100 epochs enables it to converge. The whole training is conducted on four NVIDIA 2080Ti GPUs and costs about two days.

4.2 Comparison with State-of-the-art Methods

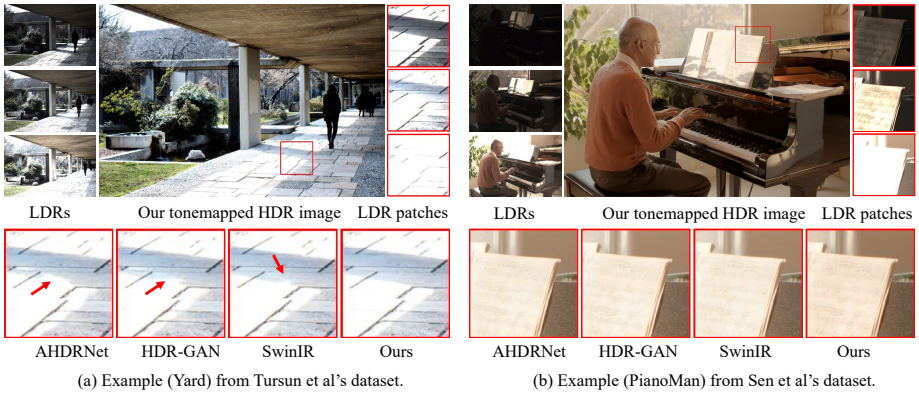
Results on Kalantari *et al.*'s Dataset We first compare the results of the proposed HDR-Transformer with several state-of-the-art methods, which include two patch match based methods (Sen *et al.* [33] and Hu *et al.* [10]) and five CNN-based methods (Kalantari *et al.* [13], DeepHDR [37], AHDRNet [39], NHDRRNet [40], and HDR-GAN [25]). We also compare with a tiny version of SwinIR [16] as the original one fails to converge on the limited dataset. Among the deep learning-based methods, Kalantari *et al.* [13] adopt optical flow to align the input LDR images while DeepHDR [37] aligns the background using homography. In contrast, the left approaches and our HDR-Transformer don't require

Fig. 4. More visual comparisons between the proposed method and state-of-the-art methods [33,10,13,37,39,25] on Kalantari *et al.* [13]'s dataset.

any pre-alignment. We report the quantitative and qualitative comparison results as this testing set contains ground truth HDR images.

Quantitative results Table 1 lists the quantitative results. For the sake of fairness, the results of prior works are borrowed from HDR-GAN [25], and all results are averaged over 15 testing samples from Kalantari *et al.*'s dataset. Several conclusions can be drawn from Table 1. Firstly, all deep learning-based algorithms have demonstrated significant performance advantages over patch match based methods. Secondly, the pure Transformer encoder adopted in SwinIR doesn't perform well for the aforementioned reasons. Thirdly, the proposed HDR-Transformer surpasses the recently published HDR-GAN [25] by up to 0.6dB and 0.4dB in terms of PSNR-l and PSNR-μ, respectively, demonstrating the effectiveness of our method.

Qualitative results For fair comparisons, all qualitative results are obtained using the codes provided by the authors and tonemapped using the same settings in Photomatix Pro. Fig. 4 illustrates an intractable scene that contains saturations and large motion. The first row shows the input LDR images, our tonemapped HDR result, and the corresponding zoomed LDR patches from left to right. The second row lists the compared HDR results, where the two comparison locations are highlighted in red and blue, respectively. As can be seen, the red boxed area suffers heavy intensity variation within the three input LDR images and causes long-range saturation. Previous approaches remove the ghosting artifacts induced by slight head movements but fail to hallucinate the details of the saturation regions on the face, resulting in color distortions and inconsistent details. The blue boxed patches show a large motion region caused by the hand,



(a) Example (Yard) from Tursun et al's dataset. (b) Example (PianoMan) from Sen et al's dataset.

Fig. 5. Comparison results on the datasets without ground truth. Scenes are obtained from the Tursun *et al.* [35]'s and the Sen *et al.* [33]'s datasets. Our approach generates better results in the saturated boundary and hallucinates more high-frequency details when suffering heavy intensity variation.

patch match based methods fail to discover the correct regions, and CNN-based methods fail to handle the long-range motion, leading to ghosting artifacts in the reconstructed HDR image. On the contrary, The proposed HDR-Transformer reconstructs ghost-free results while hallucinating more visually pleasing details in these areas.

Results on the Datasets w/o Ground Truth To validate the generalization ability of our method, we conduct evaluations on Sen *et al.* [33]'s and Tursun *et al.* [35]'s datasets. As illustrated in Fig. 5, we report the qualitative results as both datasets have no ground truth HDR images. As seen in Fig. 5 (a), When suffering long-range saturation, the CNN-based algorithms AHDRNet [39] and HDR-GAN [25] produce undesired distortions in saturated boundaries. The Transformer-based method SwinIR [16] performs better but still contains noticeable distortion as the inefficiency of local context modeling. On the contrary, the proposed HDR-Transformer generates more precise boundaries (best to compare with the corresponding LDR patches), demonstrating the context-aware modeling ability of our method. Fig. 5 (b) shows a scene where the piano spectrum gets saturated. Previous methods lose the high-frequency details and produce blurry results, while our approach hallucinates more details than them.

Analysis of Computational Budgets We also compare the inference times and model parameters with previous works. As shown in Table 2, the patch match based methods [33,10] take more than 60 seconds to fuse a 1.5MP LDR sequence. Among the CNN-based methods, Kalantari *et al.* [13] costs more time than the left non-flow based methods because of the time-consuming optical flow preprocess. DeepHDR [37] and NHDRRNet [40] consume fewer inference times

Table 2. The inference times and parameters of different methods. Part of the values are from [40]. The '-' denotes the patch match based methods have no parameters.

Method	Sen12 [33]	Hu13 [10]	Kalantari17 [13]	DeepHDR [37]	AHDRNet [39]	NHDRRNet [40]	HDR-GAN [25]	HDR-Transformer Ours
Environment	CPU	CPU	CPU+GPU	GPU	GPU	GPU	GPU	GPU
Time(s)	61.81s	79.77s	29.14s	0.24s	0.30s	0.31s	0.29s	0.15s
Parameters(M)	-	-	0.3M	20.4M	1.24M	38.1M	2.56M	1.22M

Table 3. Quantitative results of the ablation studies. BL: the baseline model, CA-ViT: the proposed Context-aware Vision Transformer, SA: the spatial attention module, $\mathcal{L}_p$: the perceptual loss term.

BL	CA-ViT	SA	$\mathcal{L}_p$	PSNR-μ	PSNR-l	HDR-VDP-2
✓				43.42	41.68	64.52
✓	✓			44.03	41.99	65.94
✓		✓		43.77	41.78	65.30
✓	✓	✓		44.26	42.09	65.97
✓	✓	✓	✓	44.32	42.18	66.03

but need huge amounts of parameters. AHDRNet [39] and HDR-GAN [25] have a better balance of performance and efficiency by taking advantage of their well-designed architectures. In contrast, HDR-Transformer outperforms the state-of-the-art method HDR-GAN [25] with only half computational budgets.

4.3 Ablation Study

To analyze the effectiveness of each component, we conduct comprehensive ablation studies on Kalantari *et al.* [13]'s dataset. We report the PSNR and HDR-VDP-2 scores for quantitative comparison.

Ablation on the network architecture For the network design, we compare the proposed CA-ViT, the adopted spatial attention (SA) module, and the overall HDR-Transformer with the baseline model. Specifically, we design the following variants:

- **Baseline**. We take a tiny version of SwinIR [16], which is constituted with vanilla Transformer encoders, as our baseline model. The baseline model keeps comparable network parameters and the same training settings as our proposed HDR-Transformer.
- + **CA-ViT**. This variant replaces the vanilla Transformer encoder used in the baseline model with the proposed Context-aware Vision Transformer.
- + **SA**. In this variant, we add a spatial attention (SA) module to fuse the shallow features extracted from the three input LDR images.
- + **CA-ViT** + **SA**. The overall network of the proposed HDR-Transformer.

Table 3 summarizes the quantitative results of our ablation study. The first row in Table 3 shows that directly applying the Transformer to HDR deghosting

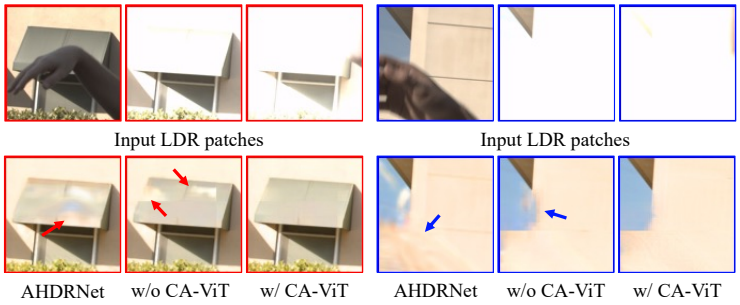


Fig. 6. Qualitative results of our ablation study on the proposed CA-ViT.

does not perform well. By comparing the first four rows, several conclusions can be drawn. On the one hand, the CA-ViT and SA both improve the performance, but the benefit from CA-ViT is more significant than SA. We conclude the reasons in two folds. Firstly, the inductive biases introduced by the convolution layers in the CA-ViT or SA help the Transformer be better optimized in limited data. Moreover, by incorporating the CA-ViT into each Transformer encoder, both the global and local contexts are explored, resulting in better capabilities of long-range ghosting removal and local details reconstruction. The qualitative results in Fig. 6 also demonstrate our conclusions. On the other hand, the performance is further improved by combining all the components, which proves the effectiveness of the HDR-Transformer's pipeline design.

Ablation on losses We also conduct experiments to verify the effectiveness of the perceptual loss by training the HDR-Transformer from scratch both with and without the perceptual loss term. Comparing the last two rows in Table 3, we can see that the adopted perceptual loss improves the performance of the proposed HDR-Transformer.

5 Conclusions

In this paper, we have proposed a dual-branch Context-aware Vision Transformer (CA-ViT), which overcomes the lack of locality in vanilla ViTs. We have extended the standard ViTs by incorporating a local feature extractor, and therefore both global and local image contexts are modeled concurrently. Furthermore, we have introduced the HDR-Transformer, a task-specific framework for ghost-free high dynamic range imaging. The HDR-Transformer incorporates the benefits of Transformers and CNNs, where the Transformer encoder and the local context extractor are used to model the long-range ghosting artifacts and short-range pixel relationship, respectively. Extensive experiments have demonstrated that the proposed method achieves state-of-the-art performance.

Acknowledgement This work was supported by National Natural Science Foundation of China under grants No. (61872067, 62031009 and 61720106004).

References

1. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. IJCV **92**(1), 1–31 (2011) 2
2. Bogoni, L.: Extending dynamic range of monochrome and color images through fusion. In: Proc. ICPR. pp. 7–12 (2000) 2, 4
3. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proc. CVPR. pp. 12299–12310 (2021) 5
4. Chung, H., Cho, N.I.: High dynamic range imaging of dynamic scenes with saturation compensation but without explicit motion compensation. In: Proc. CVPR. pp. 2951–2961 (2022) 2
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 5
6. 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 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 3, 5, 6
7. Gallo, O., Gelfandz, N., Chen, W.C., Tico, M., Pulli, K.: Artifact-free high dynamic range imaging. In: Proc. ICCP. pp. 1–7 (2009) 2
8. Grosch, T.: Fast and robust high dynamic range image generation with camera and object movement. Proc. VMV pp. 277–284 (2006) 2, 4
9. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003) 2
10. Hu, J., Gallo, O., Pulli, K., Sun, X.: Hdr deghosting: How to deal with saturation? In: Proc. CVPR. pp. 1163–1170 (2013) 2, 4, 10, 11, 12, 13
11. Jacobs, K., Loscos, C., Ward, G.: Automatic high-dynamic range image generation for dynamic scenes. IEEE Computer Graphics and Applications **28**(2), 84–93 (2008) 2, 4
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proc. ECCV. pp. 694–711 (2016) 9
13. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graphics **36**(4), 144 (2017) 2, 4, 7, 9, 10, 11, 12, 13
14. Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. ACM Trans. Graphics **22**(3), 319–325 (2003) 2, 4
15. Khan, E.A., Akyuz, A.O., Reinhard, E.: Ghost removal in high dynamic range images. In: Proc. ICIP. pp. 2005–2008 (2006) 2, 4
16. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proc. ICCVW. pp. 1833–1844 (2021) 3, 5, 10, 12, 13
17. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proc. ECCV. pp. 85–100 (2018) 9
18. 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: Proc. ICCV. pp. 10012–10022 (2021) 5
19. Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J., Liu, S.: Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In: Proc. CVPRW. pp. 463–470 (2021) 2, 8

20. Ma, K., Duanmu, Z., Zhu, H., Fang, Y., Wang, Z.: Deep guided learning for fast multi-exposure image fusion. IEEE Trans. on Image Processing **29**, 2808–2819 (2019) 1
21. Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., Zhang, L.: Robust multi-exposure image fusion: a structural patch decomposition approach. IEEE Trans. on Image Processing **26**(5), 2519–2532 (2017) 1
22. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graphics **30**(4), 1–14 (2011) 10
23. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: Proc. PG. pp. 382–390 (2007) 1
24. Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Intriguing properties of vision transformers. arXiv preprint arXiv:2105.10497 (2021) 3
25. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. IEEE Trans. on Image Processing **30**, 3885–3896 (2021) 2, 4, 9, 10, 11, 12, 13
26. Oh, T.H., Lee, J.Y., Tai, Y.W., Kweon, I.S.: Robust high dynamic range imaging by rank minimization. IEEE Trans. on Pattern Analysis and Machine Intelligence **37**(6), 1219–1232 (2014) 4
27. Pece, F., Kautz, J.: Bitmap movement detection: Hdr for dynamic scenes. In: Proc. CVMP. pp. 1–8 (2010) 2, 4
28. Prabhakar, K.R., Agrawal, S., Singh, D.K., Ashwath, B., Babu, R.V.: Towards practical and efficient high-resolution hdr deghosting with cnn. In: Proc. ECCV. pp. 497–513. Springer (2020) 2, 4
29. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: Proc. ICCP. pp. 1–8. IEEE (2019) 2
30. Prabhakar, K.R., Senthil, G., Agrawal, S., Babu, R.V., Gorthi, R.K.S.S.: Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In: Proc. CVPR. pp. 4875–4885 (2021) 2, 4
31. Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R.: Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proc. ICCV. pp. 4714–4722 (2017) 1
32. Raman, S., Chaudhuri, S.: Reconstruction of high contrast images for dynamic scenes. The Visual Computer **27**(12), 1099–1114 (2011) 1
33. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. ACM Trans. Graphics **31**(6), 203 (2012) 2, 4, 10, 11, 12, 13
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9
35. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: An objective deghosting quality metric for hdr images. In: Proc. CGF. pp. 139–152 (2016) 10, 12
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. NeurIPS. pp. 5998–6008 (2017) 5
37. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Proc. ECCV. pp. 117–132 (2018) 2, 4, 7, 9, 10, 11, 12, 13
38. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.: Early convolutions help transformers see better. arXiv preprint arXiv:2106.14881 (2021) 8

39. 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: Proc. CVPR. pp. 1751–1760 (2019) 2, 4, 7, 8, 9, 10, 11, 12, 13
40. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Trans. on Image Processing **29**, 4308–4322 (2020) 9, 10, 12, 13
41. Zhang, W., Cham, W.K.: Gradient-directed multiexposure composition. IEEE Trans. on Image Processing **21**(4), 2318–2323 (2011) 1, 4
42. Zimmer, H., Bruhn, A., Weickert, J.: Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In: Proc. CGF. pp. 405–414 (2011) 2, 4