Selective TransHDR: Transformer-based selective HDR Imaging using Ghost Region Mask

Jou Won Song*[1], Ye-In Park*[1], Kyeongbo Kong[2], Jaeho Kwak[1], and Suk-Ju Kang[1]

[1] Department of Electronic Engineering, Sogang University, Seoul, Korea
{wn5649,yipark06,resky1111,sjkang}@sogang.ac.kr
[2] Department of Media communication, Pukyong National University, Busan, Korea
kbkong@pknu.ac.kr

Abstract. The primary issue in high dynamic range (HDR) imaging is the removal of ghost artifacts afforded when merging multi-exposure low dynamic range images. In the weakly misaligned region, ghost artifacts can be suppressed using convolutional neural network (CNN)-based methods. However, in highly misaligned regions, it is necessary to extract features from the global region because the necessary information does not exist in the local region. Therefore, the CNN-based methods specialized for local features extraction cannot obtain satisfactory results. To address this issue, we propose a transformer-based selective HDR image reconstruction network that uses a ghost region mask. The proposed method separates a given image into ghost and non-ghost regions, and then, selectively applies either the CNN or the transformer. The proposed selective transformer module divides an entire image into several regions to effectively extract the features of each region for HDR image reconstruction, thereby extracting the whole information required for HDR reconstruction in the ghost regions from the entire image. Extensive experiments conducted on several benchmark datasets demonstrate the superiority of the proposed method over existing state-of-the-art methods in terms of the mitigation of ghost artifacts.

1 Introduction

Typical digital cameras can only capture luminance within a limited dynamic range due to sensor limitations. Therefore, low dynamic range (LDR) images with 8-bit depth obtained by these cameras have significant underexposed and overexposed regions, thereby yielding large data loss compared to the real scene. A lot of studies have been conducted to recover lost data from LDR images and to generate 10-bit or 12-bit high dynamic range (HDR) images that can provide a wide illuminance range. Multi-exposure image fusion is the most common HDR reconstruction method; LDR images with different exposure values are obtained

* equal contribution

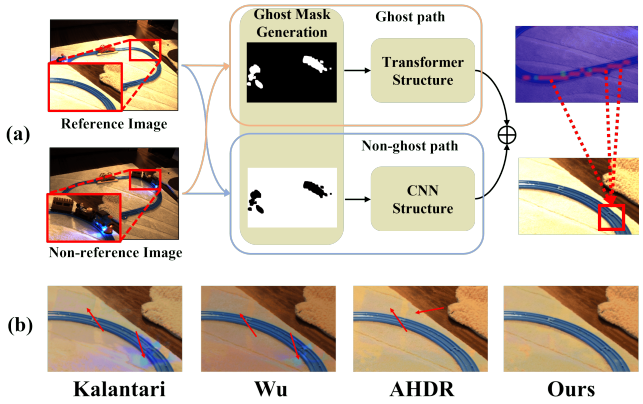

Fig. 1. (a) Overview of the proposed selective transformer module, and (b) sample images generated by the proposed and the state-of-the-art methods in Tursun et al's [34] dataset. The proposed selective transformer module separates the ghost regions from the non-ghost regions, and selectively uses one of the CNN and the transformer.

by cameras with a limited dynamic range and merged into an HDR image after alignment of the LDR images [4, 11, 12, 29, 40].

However, in most cases, the aligned LDR images are difficult to obtain in the acquisition process of several LDR images with different exposure values due to the motion of a camera or an object. If an HDR image is obtained using these unaligned LDR images, ghost and blur artifacts occur in the HDR image [13, 25, 22, 31, 1, 19]. To solve this problem, optical-flow methods [7, 10] or patch-based methods [30] have been presented.

Although these methods can resolve some scenarios with large motions, ghost artifacts still exist. Recently, many deep neural network (DNN)-based methods have been proposed for reconstructing HDR images [37, 39, 8, 9, 21]. Typically, DNN-based methods utilize convolution neural networks (CNNs) to extract spatial features. These methods have higher performance than existing methods. However, these CNN-based methods are difficult to explicitly remove ghost artifacts due to the limitation of CNN specialized for local feature extraction from unaligned LDR images, and hence, it is difficult to consider global information required for HDR reconstruction from an entire region. As shown in Fig. 1 (a), the red box regions of the reference image and the non-reference image are in the same location, but contain different information. The rail is visible in the reference image, but it is occluded by the train in the non-reference image. That is, it is impossible to extract rail information from the non-reference image due to the movement of the object. Therefore, in Fig. 1 (b), Kalantari et al.'s [18], Wu et al.'s [37], and Yan et al.'s [38] yielded final images with ghost artifacts in the motion region.

To solve this limitation of CNN, the use of transformer structure in various tasks of computer vision has been studied [43, 3]. In the transformer structure [35], images are split into patches, and the attention weights between all patches

are extracted. Therefore, it is possible to extract features of patches that are far from the reference patch, allowing global information to be taken into account. However, as confirmed by the vision transformer [6], when the amount of data is limited, the performance improvement cannot be expected due to lack of inductive bias compared to CNN. Furthermore, in the non-ghost region, for aligned LDR images, the transformer structure is relatively inefficient because the local region has important information, so we need to focus more on that region.

This paper proposes a novel approach that can selectively consider regions where it is effective to apply a transformer using a ghost region mask as a guide. As shown in Fig. 1 (a), unlike the existing methods that utilize the same network structure for all regions, the proposed network adaptively applies the transformer and CNN structures by separating the ghost and non-ghost regions. The proposed ghost region mask-guided transformer module uses the transformer structure to extract important features for the ghost region from the global regions of the non-reference images. Fig. 1 (a) visualizes the attention map of the proposed selective transformer module. The proposed attention map focuses on the visible rail away from the red box where the rail is obscured by the train. The proposed method can extract meaningful information from global regions of the non-reference image, excluding unnecessary information from local regions. Therefore, as shown in Fig. 1 (b), the proposed method affords significantly better reconstruction in terms of the color and details of the ghost regions. The main contributions of this paper can be summarized as follows:

- We propose a novel contents-aware ghost region detector to effectively consider both global and local features focused by the proposed model. This detector distinguishes between ghost and non-ghost regions, and the networks, suitable for each region, are selectively applied.
- We propose a transformer-based selective HDR image reconstruction network to extract the necessary features to restore the ghost region. Our method does not simply apply the transformer, but uses it for the global information analysis and selection that the transformer can be effectively used. Therefore, the proposed selective transformer module can extract important global features from the entire region of each non-reference image for HDR reconstruction of ghost regions.
- Experiments on various datasets validate the superiority of the proposed method compared to existing methods. We also demonstrate that using a transformer adaptively rather than using a single model significantly improves performance by reflecting the characteristics of images with various exposure values.

2 Related Work

2.1 Motion Detection-based HDR Reconstruction

Motion detection-based methods are based on the assumption that the LDR images with different exposures can be globally registered in the HDR coordinate.

These methods can detect moving pixels in the images which are rejected for final weighted HDR fusion [20, 17, 14]. Heo et al. [15] detected motion regions with joint probability densities. Yan et al. [41] used a sparse representation to detect the object motion. However, since these methods ignore unaligned pixels and not all input regions are available, these methods heavily depend on effectiveness of motion detection and cannot expect high performance when the large motion appears.

2.2 Alignment-based HDR Reconstruction

Alignment-based methods focused on aligning LDR images to a reference image, and then, merged them to reconstruct the HDR image [5, 24]. For the alignment, optical flow or patch matching methods are generally used. Bogoni [2] used optical flow to estimate motion vectors. Sen et al. [30] used a patch-based energy minimization method. Hu et al. [16] aligned the images using brightness and gradient consistency in the transformed domain. However, alignment-based methods are sensitive to complex backgrounds and large motions. These methods also requires significantly high execution time.

2.3 CNNs-based HDR Reconstruction

Kalantari et al. [18] introduced neural networks into the alignment-before-merging pipeline for the HDR image generation. Wu et al. [37] proposed an autoencoder that can learn to convert multiple LDR images into a ghost-free HDR image. In [42], multiple LDR images were reconstructed into HDR images using a non-local network [36]. These CNN-based HDR reconstruction methods extract features of unaligned image regions, causing geometric or color distortion. In addition, they are difficult to reconstruct regions with large motions due to the CNN specialized in the local feature extraction. Yan et al. [38] used a spatial attention mechanism to generate ghost-free HDR images. Although the attention map of the spatial attention mechanism deletes unnecessary information, it is difficult to extract features for HDR reconstruction in the global region because it still uses a CNN-based model. Prabhakar et al. [27] proposed an HDR imaging method using bilateral guided upsampler and motion compensation. This method can compensate for ghost regions in LDR images.

2.4 Transformer

Transformers have been actively studied in many tasks of computer vision. Carion et al. [3] used the transformer and CNN structures simultaneously for object detection. Vision transformer demonstrated that the model using a pure transformer achieved the best performance [6]. Zheng et al. [43] proposed an encoder with a transformer structure to solve the problem of the reduced sparse resolution. In a recent study related to our method, Yan et al. [42] used a non-local network to perform the HDR reconstruction. However, this method may still

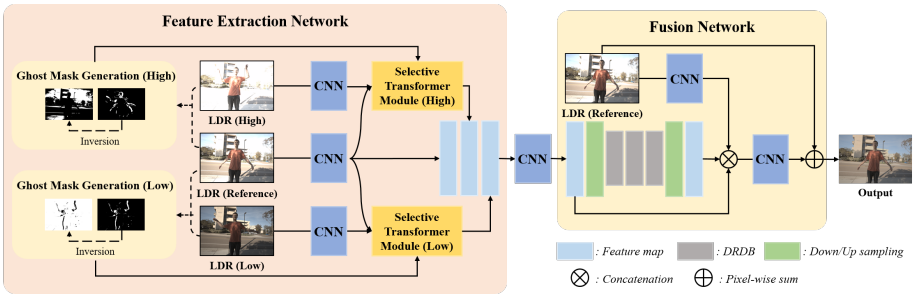

Fig. 2. The proposed method architecture. It consists of a feature extraction network and a fusion network. The selective transformer module selectively applies the transformer and the CNN structures to the ghost and non-ghost regions, respectively, using a proposed ghost region mask. The fusion network is constructed based on a series of dilated residual dense blocks (DRDBs) and multi-scale CNNs. The final HDR result is generated by combining the reference image with the final stage of the fusion network.

extract unnecessary features because the CNN structure is used in the ghost region. The proposed method selectively applies the transformer structure only to the ghost region, which is the region that requires global feature extraction. Therefore, the model can appropriately extract the necessary information for ghost and non-ghost regions.

3 Proposed Method

This paper proposes the novel transformer-based HDR image reconstruction and ghost mask generation to extract ghost regions in multi-exposure images. In the proposed method, the LDR image is divided into ghost and non-ghost regions, and features corresponding to these regions are extracted. Our goal is to reconstruct a ghosting-free HDR image using the given LDR images (L_1, L_2, L_3). We also use a middle exposure image (L_2) as a reference image. As used in several researches [18, 42, 37, 28], we convert the LDR images to corresponding HDR representations through gamma correction. As confirmed in [18], LDR images are effective in detecting noise or saturated regions, and HDR images are used to measure content deviations from the reference image. We use LDR images and mapped HDR images together as inputs and pixel values are all normalized to [0, 1].

As shown in Fig. 2, the entire network comprises of a feature extraction network, which extracts the necessary features and a fusion network, which combines the extracted features to construct an HDR image. The feature extraction network consists of a ghost mask generation, which detects ghost regions to generate ghost region masks, and ghost region mask-guided transformer modules, which extract features from ghost and non-ghost regions. First, the feature extraction network extracts the feature from every LDR image through convolution layers. The extracted features are used as inputs for the ghost region

mask-guided transformer module. In addition, the proposed method generates a mask that separates the regions by comparing the reference and non-reference images. In the fusion network, similar to the structure used in AHDR [38], the dilated residual dense block (DRDB) is applied and configured. In the following sub-sections, we describe the ghost region mask generation and the ghost region mask-guided transformer.

3.1 Feature Extraction Network

Proposed Ghost Region Mask Generation This module generates ghost and non-ghost region masks in both underexposed and overexposed images. Unlike the non-ghost region, the same location of the two images has different information in the ghost region. Therefore, the information required for the ghost region in the reference image must be determined from the other regions of the non-reference image. Fig. 3 shows the ghost region mask generation process between the low-exposure image (I_{non}) and reference images (I_{ref}). In the first step, the average filtering is used to blur three multi-exposure images. In the weak ghost region, the features required for HDR image reconstruction can be sufficiently extracted using the CNN structure. Therefore, the average filter is applied to select only the large motion region. The pixel value difference between blurred images decreases in the weakly misaligned regions. In the second step, the reference image is transformed into the same luminance space as the non-reference image through histogram matching. Through this process, the luminance of the reference and non-reference images becomes similar, thereby decreasing the pixel value difference for all regions other than the ghost regions. The ghost mask is determined by applying a pre-determined threshold to the pixel difference value (The experiments for changing this threshold are added in the supplemental material.). However, when histogram matching is performed, the saturation region may be falsely detected as the ghost region due to the luminance difference between the reference and non-reference images in the saturation region. Therefore, we add a process of removing the saturation region of the non-reference image from the ghost mask. Even if the saturation region includes the ghost region, the process of removing the saturation region is not a problem to extract the ghost region features because there is no information for HDR imaging in the saturation region of the non-reference image. Finally, the opening operation is performed to remove a noise region caused by weakly misaligned region. Therefore, the small noise region of the ghost mask is removed and only the strong misaligned region remains. The proposed ghost region mask (G_i) can be calculated by

$$G_i = |L_i - K(L_i, L_2)|, \text{if } i = 1, 3, \tag{1}$$

where K is the operation for histogram matching. Finally, to compensate for the undetected ghost regions and train for various ghost masks, the kernel size of the erosion is set to 11 and the kernel size of the dilation is set to a value between 11 and 17 in the opening operation by our experiments. The kernel size of dilation is

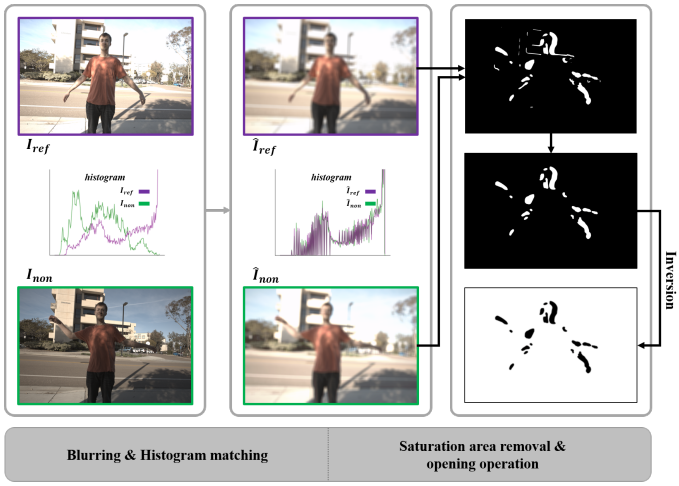

Fig. 3. The proposed ghost region mask generation method. First, image blurring is performed on the reference image and non-reference images to exclude weak ghost regions. After that, the luminance of the non-reference image is converted to the same as the reference image through histogram matching. The ghost mask is generated using the difference between the two images. Finally, opening operation and saturation region removal are performed to remove noise from the generated ghost mask.

fixed to 15 during the inference process. The non-ghost region mask is generated by inverting the ghost region mask (The experiments of several kernel sizes are added in the supplemental material.).

Ghost Region Mask-Guided Selective Transformer When an object or a camera moves, a reference image (a middle exposure image) and non-reference image (high and low exposure images) have different pixel information in the area where motion occurs. To address this problem, the proposed method employs a transformer-based module to extract important global information from entire non-reference images. As shown as Fig. 4, the selective transformer module consists of a transformer-based ghost path, which extracts the ghost region features of the reference image, and a CNN-based non-ghost path, which extracts the non-ghost region features of the reference image.

We construct the selective transformer module by applying 1 layer of cross attention. The selective transformer module first uses a CNN layer to extract the query (Q) from reference image features and key (K) and value (V) from non-reference image features. The transformer structure in this module only works on ghosted regions of the reference image. Therefore, as shown as Fig. 4, the reference image feature is multiplied by the ghost region mask to generate Q remaining only the ghost region features. In the following process, the similarity between the ghost region in a reference image and an entire non-reference image is calculated to select the best regional information. For this, all image features are unfolded into p-sized patches. Therefore, Q and K can be

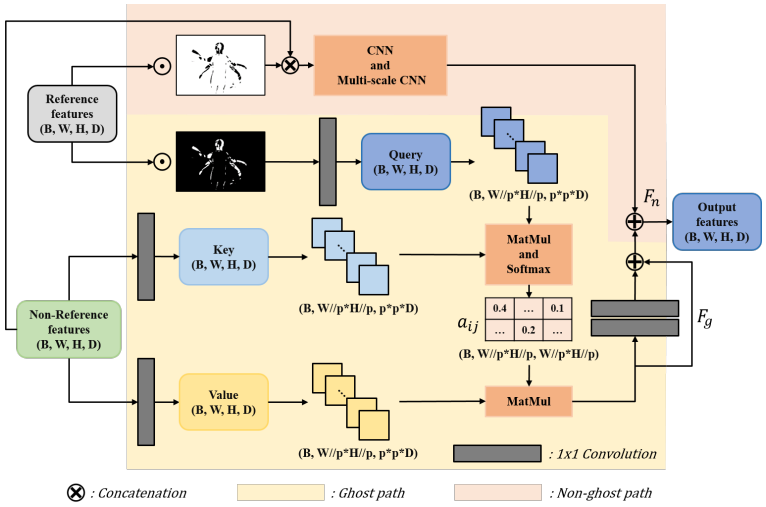

Fig. 4. The proposed ghost region mask-guided transformer module. a_{ij} denotes the attention map calculated from K and Q. F_g and F_n represent the ghost and non-ghost region features, respectively. B and p represents batch and patch sizes, respectively, and (H, W, D) is the size of the feature.

represented as $(W/p \times H/p)$ number of vectors with size $(p \times p \times D)$, denoted as $q_i (i \in [1, (W/p \times H/p)]$ and $k_j (j \in [1, (W/p \times H/p)]$, respectively. The weight a_{ij} between these patches is calculated through the dot product of q_i and k_j. a_{ij} is defined as follows:

$$a_{ij} = softmax(\frac{q_i k_j^T}{\sqrt{p \times p \times D}}). \tag{2}$$

Also, features are extracted from the non-reference regions that are the most relevant to the reference image patch in the ghost region generated using the weight a_{ij}. The extracted features use two convolution layers to yield an output, and the ghost region mask is multiplied to the output so that it does not affect the other regions. The output of ghost region is as follows:

$$F_g = Conv(Relu(Conv(a_{ij} v_i))) + a_{ij} v_i, \tag{3}$$

where $Conv$ and v_i denote a convolutional layer and image patches of V, respectively, and F_g represent the ghost region features. Furthermore, in the non-ghost path, features from the reference and non-reference images are concatenated for feature extraction in the non-ghost region. Since non-ghost regions are aligned or weakly misaligned, the features required for HDR reconstruction can be extracted even if a conventional CNN structure is used. We use the multi-scale CNN [42, 32] to extract detailed local features. The multi-scale CNN concatenates outputs of each layer by configuring CNNs with different kernel sizes in parallel. The concatenated output is transformed into same-sized features as the input channel and added to the input features. Each convolution layer uses kernel sizes of 1, 3, and 5 with the ReLU activation function. The proposed method

employs a 3×3 convolution layer and three multi-scale convolution layers for the non-ghost region feature extraction. The features extracted from two regions of ghost and non-ghost are combined into one feature using the pixel-wise addition. The properties of each region are preserved as no overlap exists between the regions. As a result, the transformer module is selectively applied for the feature extraction in the ghost and non-ghost regions.

The proposed method extracts two features of 64 channel from two non-reference images using the above process. The final output is generated using a concatenation of these features and the feature of the reference image.

3.2 Fusion Network

The fusion network uses the features extracted from the feature extraction network to reconstruct the HDR image. As shown in Fig. 2, concatenated three features in the feature extraction network are combined into one feature through a 1×1 convolution and three multi-scale CNNs, and then, they are downsampled using maxpooling. Then, three DRDBs are used for the sufficient receptive field. Since the DRDB consists of dilated convolutions, the information for HDR reconstruction can be extracted using a large receptive field. Three multi-scale CNNs and a transposed CNN are used to generate features of the same size as the HDR image. Finally, as shown in Fig. 2, the HDR image is reconstructed with three features concatenated by the skip connection and reference image feature. The loss function used to reconstruct the HDR image are as follows:

$$\begin{aligned} L(H, \hat{H}) =& \| \, T(H) - \, T(\hat{H}) \, \|_2 \\ &+ \| \, M(T(H)) - \, M(T(\hat{H})) \, \|_2, \end{aligned} \tag{4}$$

where H and $\hat{H}$ stands for the ground truth HDR image and the reconstructed HDR image, respectively. $M(\cdot)$ stands for the operation to extract the edge map computing the difference between adjacent pixels, and $T(\cdot)$ and $\| \cdot \|_2$ denote the tone mapping using the μ-law and l_2 norm, respectively. Detailed information on the training and network architectures is provided in the supplementary material.

4 Experiments

4.1 Datasets

We used the Kalantari et al.'s [18] dataset to validate the performance of our method. It consists of 74 training and 15 test samples. Each sample includes three unaligned images with different exposure biases of $\{-2, 0, +2\}$ or $\{-3, 0, +3\}$. For the training, we employed randomly cropped 256×256 sized patches from the full images and applied random rotation and flip to diversify the training samples. To verify the generalization ability of the proposed method, we also used Tursun et al.'s [34] datasets used in several other papers [38, 28, 26]. Since these datasets do not contain ground truths of HDR images, we only displayed the tone-mapped HDR images of the proposed and conventional methods.

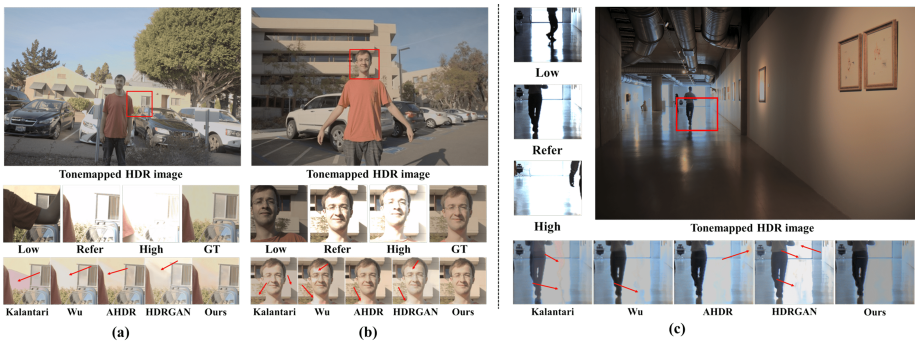

Fig. 5. Comparison of qualitative results of the proposed method with the state-of-the-art methods in (a), (b) the Kalantari et al.'s [20] dataset and (c) the Tursun et al.'s [33] dataset

4.2 Evaluation metrics

The proposed method was evaluated based on five metrics. Since HDR images may be displayed on LDR screens, the quality of the tone-mapped images needs to be checked. Therefore, we measured peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) on the μ-law mapped images (PSNR-μ and SSIM-μ). We also measured PSNR and SSIM on the linear domain (PSNR-L and SSIM-L). Finally, we performed quantitative evaluations by calculating HDR-VDP-2 [23] designed to evaluate the HDR image quality.

4.3 Qualitative results

Using the Kalantari et al.'s dataset, we compared our proposed method with several state-of-the-art methods. In Fig. 5, the first row displays the tone-mapped HDR images. The second row displays the LDR images and the ground truth; they are enlarged images of the red boxes in the images of the first row. As shown in Fig. 5 (a), since the background region of the reference image (Refer) is mostly saturated, features from the low-exposure image (Low) should be extracted. However, due to the large motion of objects in the non-reference images, many details in the background are obscured, thereby making the extraction of the necessary features difficult. Therefore, the method of Kalantari et al. [18] could not completely exclude the region where the arm movement occurred in the LDR image with low-exposure value. It was confirmed that the resulting image reflected the pixel information of the corresponding region, resulting in ghost artifacts. The methods of Wu et al. [37], AHDR [38], and HDRGAN [26] succeeded in removing the regions where the arm movements occurred using LDR images with low-exposure values. However, weak ghost artifacts were observed due to the CNN limitation focusing on extracting local information. In contrast, our method used a transformer structure to extract features that are the most relevant to the reference image patch from non-reference image regions without

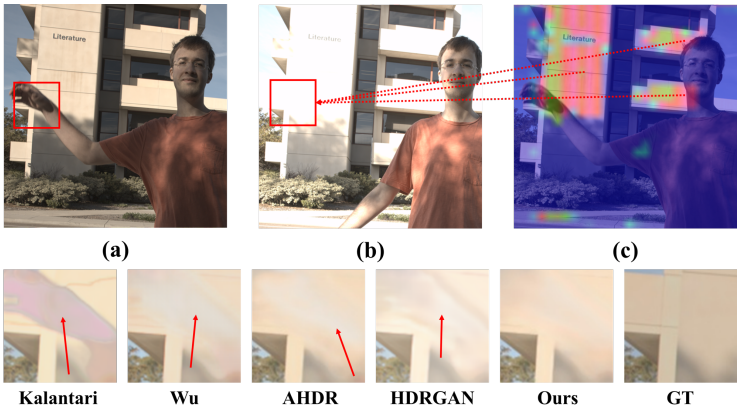

Fig. 6. Comparison of qualitative results of the proposed method with the state-of-the-art methods in Kalantari et al.'s [18] dataset. (a) Non-reference image (low), (b) reference image, and (c) attention region in the non-reference image. The bottom row consists of images enlarged by the red box area in the image above, in order of the state-of-the-art methods, proposed methods, and the ground truth. The red boxes shown in (a) and (b) indicate the same area, but the red box in (a) differs from that in (b). When these areas are merged, ghost artifacts appear. Using an attention map, (c) shows the area of the red box in (b) that is deeply related to (a). The closer the color is to red, the higher is the importance.

focusing on the local regions where motion exists. Therefore, the corresponding ghost region is naturally reconstructed. As shown in Fig. 5 (b), the object is highly saturated in the reference image. Existing methods failed to restore the color information in some regions. However, since our proposed method separately learns the ghost and non-ghost regions, it was optimized to extract the features of non-ghost regions. Therefore, the local features requiring reconstruction were well extracted from the non-reference images, thereby affording high saturation and color reconstruction performance in the corresponding regions.

Additionally, to verify the generalization ability of the proposed HDR imaging method, we evaluated its performance on the Tursun et al's dataset, wherein no ground truth is provided. As shown in Fig. 5 (c), the methods of Kalantari et al., Wu et al., AHDR, and HDRGAN could not completely exclude the motion region information from the non-reference image. Moreover, the saturated regions were not well restored, resulting in a lot of ghost artifacts and poor color restoration. In contrast, the proposed method excluded the interference of motion region information unrelated to HDR reconstruction in the non-reference image. Therefore, our model generated a high-quality HDR image.

Analysis of Attention Maps Generated by the Proposed Network In this section, we visualized the attention map of the typical motion case on the Kalantari et al.'s [18] dataset to verify the effectiveness of the attention map

	Ghost regions		Full image				
Method	PSNR-μ	SSIM-μ	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
Kalantari [20]	38.7431	0.9761	42.7423	0.9877	41.2158	0.9848	60.5088
Wu [37]	40.1244	0.9871	41.6377	0.9869	40.9082	0.9858	60.4955
AHDR [38]	40.9492	0.9883	43.6878	0.9902	41.1613	0.9857	62.0125
Non [42]	-	-	42.4143	0.9877	-	-	61.2107
Robust [28]	-	-	43.8487	0.9906	41.6452	0.9870	62.5495
HDRGAN [26]	41.2814	0.9887	43.9220	0.9905	41.5720	0.9865	63.1245
Proposed	**41.6714**	**0.9890**	**44.0981**	**0.9909**	**41.7021**	**0.9872**	**63.3721**

Table 1. Performance comparison of the proposed and the state-of-the-art methods using PSNR, SSIM, and HDR-VDP-2.

of the proposed transformer module. As shown in Fig. 6, the red box region of the reference image (b) is saturated. Therefore, the information in the region must be extracted from the non-reference image (a) with the low exposure value. However, information cannot be extracted from non-reference images, such as Fig. 6 (a), because the necessary information is deleted due to the movement of the objects. Therefore, Kalantari et al.'s [18], Wu et al.'s [37], AHDR [38], and HDRGAN [26] yielded final images with ghost artifacts in the motion region.

The proposed attention map of Fig. 6 (c) visualized the importance of the pixel region to be referenced in Fig. 6 (a) to restore the red box region in Fig. 6 (b). The proposed attention map focused on the wall region far away from the red box, which is the saturated wall background. Therefore, compared to other state-of-the-art methods, the proposed method afforded significantly better reconstruction in terms of the color and details of the ghost regions.

4.4 Quantitative results

We compared the performance of the proposed model with the state-of-the-art models using the quantitative metrics. Furthermore, to evaluate the performance of the proposed method in the ghost region, we performed evaluations on the full image and ghost regions. The ghost regions are calculated from the proposed ghost region mask. Table 1 denotes the performance on the full image; our method outperformed other methods in terms of PSNR-μ, PSNR-L, SSIM-μ, SSIM-L, and HDR-VDP-2 are 44.0981, 41.7021, 0.9909, 0.9872, and 63.3721, respectively. (Models with performance in bold performed best.)

Then, we evaluated the performance results for the ghost regions. The proposed method showed the best performance in all metrics among all methods compared in the region where motion occurred. The CNN-based methods, kalantari et al. [18], Wu et al. [37], and AHDR [38], were difficult to use global features. Moreover, they extracted the local features for ghost regions from non-reference images. In contrast, the proposed method generated a ghost region mask for the region where the motion occurs and applied it to the transformer-based network, so that it was possible to extract information from the relevant regions by searching all regions of the non-referenced image. Therefore, the proposed method had

Method	PSNR-μ	SSIM-μ	PSNR-L	SSIM-L	HDR-VDP-2
Base model (only transformer)	43.3244	0.9882	40.8823	0.9861	60.6124
Base model (only CNN)	43.7274	0.9901	41.2311	0.9866	62.7231
Base model (no mask)	43.8874	0.9903	41.3422	0.9868	62.9245
Base model	**44.0981**	**0.9909**	**41.7021**	**0.9872**	**63.3721**

Table 2. Performance comparison for variants of the proposed model using PSNR, SSIM, and HDR-VDP-2.

excellent performance even with large motions. Our method effectively solved the problem of ghost artifacts in HDR images.

5 Ablation Studies

Ablation study demonstrates the effectiveness of selective transformer module and ghost region mask. We achieved the ablation study by comparing the performance of the following variants of the proposed model, as shown in Table 2.

- **Base model.** All modules of the proposed model are used.
- **Base model (only transformer).** In the selective transformer module of the base model, we replaced multi-scale CNNs of the non-ghost region path with a transformer structure.
- **Base model (only CNN).** In the selective transformer module of the base model, we replaced the transformer structure of the ghost region path with the multi-scale CNN.
- **Base model (no mask).** Instead of using the proposed ghost region mask, features are extracted in parallel using the transformer structure and the CNN structure from the entire input image.

5.1 Effectiveness of Selective Transformer Module

To verify the effectiveness of selectively applying the transformer, we designed two feature extraction networks composed of only CNNs or only transformers, respectively, and compared the performance of these models. As shown in Table 2, the base model consisting of only transformers showed the lowest performance. Due to the nature of the HDR reconstruction task, more important information exists in the local region in the case of the non-ghost region, so the performance of the base model consisting of only CNNs was higher than that of the transformer structure. However, this model failed to reconstruct high-quality HDR images in ghost regions with large motion. In contrast, the base model separated the ghost and non-ghost regions, and selectively applied the CNN and the transformer structures. Therefore, the proposed method could utilize the advantages of both structures. As a result, the base model achieved higher performance than the model using only CNN structure or only transformer structure.

5.2 Effectiveness of Ghost Region Mask

To confirm the effectiveness of the proposed ghost region mask, we performed an ablation experiment without using a ghost region mask on the base model. Therefore, the base model (no mask) used the transformer and the CNN structures to extract features in parallel from the entire image, and fed the combined two features to the fusion network. However, these features may contain information that is not required for each region. This problem can cause ghost artifacts or color distortion in the final HDR image. Therefore, as shown in Table 2, the model without the ghost region mask showed lower performance than the base model.

6 Limitation and Future Work

The proposed method significantly enhanced HDR images in strong ghost regions compared to conventional methods. However, the heuristic module, ghost mask generation, can falsely detect the ghost region if histogram matching is incorrectly performed due to severe saturation regions or camera misalignment. In this case, the proposed model will use the transformer structure in the non-ghost region and may produce a low quality HDR image as confirmed in our ablation study. We will consider these factors in our future work and design a network that outputs refined masks using the generated ghost masks. Through this future work, we will try to configure an end-to-end network including a ghost mask generation module to detect more accurate ghost regions.

7 Conclusion

In this paper, we proposed to selectively apply a network suitable for each region by dividing the image into ghost and non-ghost regions. In the ghost region with large motion, the proposed selective transformer module reconstructed the region well using the transformer structure. This is because the transformer can search the entire region and extract features deeply related to patches of the reference image from the global region of the non-reference image. In the non-ghost region where the LDR images are aligned, the selective transformer module used the CNN structure to effectively extract local features. In addition, through the ablation study, we found that the proposed model outperforms the CNN-only and transformer-only models. Finally, the proposed model provided ghost-free high-quality HDR images with rich details and colors compared to the state-of-the-art models.

Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1004208).

References

1. An, J., Lee, S.H., Kuk, J.G., Cho, N.I.: A multi-exposure image fusion algorithm without ghost effect. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1565–1568. IEEE (2011)
2. Bogoni, L.: Extending dynamic range of monochrome and color images through fusion. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. vol. 3, pp. 7–12. IEEE (2000)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
4. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: ACM SIGGRAPH 2008 classes, pp. 1–10 (2008)
5. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: ACM SIGGRAPH 2008 classes, pp. 1–10 (2008)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
7. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015)
8. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: Hdr image reconstruction from a single exposure using deep cnns. ACM transactions on graphics (TOG) **36**(6), 1–15 (2017)
9. Endo, Y., Kanamori, Y., Mitani, J.: Deep reverse tone mapping. ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA 2017) **36**(6) (Nov 2017)
10. Fleet, D., Weiss, Y.: Optical flow estimation. In: Handbook of mathematical models in computer vision, pp. 237–257. Springer (2006)
11. Granados, M., Ajdin, B., Wand, M., Theobalt, C., Seidel, H.P., Lensch, H.P.: On being 'undigital'with digital cameras: Extending dynamic range by combining differently exposed pictures. In: In Proceedings ofIST. pp. 442–448 (1995)
12. Granados, M., Ajdin, B., Wand, M., Theobalt, C., Seidel, H.P., Lensch, H.P.: Optimal hdr reconstruction with linear digital cameras. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 215–222. IEEE (2010)
13. Granados, M., Kim, K.I., Tompkin, J., Theobalt, C.: Automatic noise modeling for ghost-free hdr reconstruction. ACM Transactions on Graphics (TOG) **32**(6), 1–10 (2013)
14. Grosch, T., et al.: Fast and robust high dynamic range image generation with camera and object movement. Vision, Modeling and Visualization, RWTH Aachen **277284** (2006)
15. Heo, Y.S., Lee, K.M., Lee, S.U., Moon, Y., Cha, J.: Ghost-free high dynamic range imaging. In: Asian Conference on Computer Vision. pp. 486–500. Springer (2010)
16. 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. pp. 1163–1170 (2013)
17. Jinno, T., Okuda, M.: Motion blur free hdr image acquisition using multiple exposures. In: 2008 15th IEEE International Conference on Image Processing. pp. 1304–1307. IEEE (2008)

18. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4), 144–1 (2017)
19. Khan, E.A., Akyuz, A.O., Reinhard, E.: Ghost removal in high dynamic range images. In: 2006 International Conference on Image Processing. pp. 2005–2008. IEEE (2006)
20. Khan, E.A., Akyuz, A.O., Reinhard, E.: Ghost removal in high dynamic range images. In: 2006 International Conference on Image Processing. pp. 2005–2008. IEEE (2006)
21. Khan, Z., Khanna, M., Raman, S.: Fhdr: Hdr image reconstruction from a single ldr image using feedback network. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 1–5 (2019). https://doi.org/10.1109/GlobalSIP45357.2019.8969167
22. Lee, C., Li, Y., Monga, V.: Ghost-free high dynamic range imaging via rank minimization. IEEE signal processing letters **21**(9), 1045–1049 (2014)
23. 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 Transactions on graphics (TOG) **30**(4), 1–14 (2011)
24. Mitsunaga, T., Nayar, S.K.: Radiometric self calibration. In: Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149). vol. 1, pp. 374–380. IEEE (1999)
25. Moon, Y.S., Tai, Y.M., Cha, J.H., Lee, S.H.: A simple ghost-free exposure fusion for embedded hdr imaging. In: 2012 IEEE International Conference on Consumer Electronics (ICCE). pp. 9–10. IEEE (2012)
26. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. In: IEEE Transactions on Image Processing. vol. 30, pp. 3885–3896 (2021)
27. Prabhakar, K.R., Agrawal, S., Singh, D.K., Ashwath, B., Babu, R.V.: Towards practical and efficient high-resolution hdr deghosting with cnn. In: European Conference on Computer Vision. pp. 497–513. Springer (2020)
28. Pu, Z., Guo, P., Asif, M.S., Ma, Z.: Robust high dynamic range (hdr) imaging with complex motion and parallax. In: Proceedings of the Asian Conference on Computer Vision (2020)
29. Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., Myszkowski, K.: High dynamic range imaging: acquisition, display, and image-based lighting. Morgan Kaufmann (2010)
30. 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. **31**(6), 203–1 (2012)
31. Srikantha, A., Sidibé, D.: Ghost detection and removal for high dynamic range images: Recent advances. Signal Processing: Image Communication **27**(6), 650–662 (2012)
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
33. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: The state of the art in hdr deghosting: a survey and evaluation. In: Computer Graphics Forum. vol. 34, pp. 683–707. Wiley Online Library (2015)
34. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: An objective deghosting quality metric for hdr images. In: Computer Graphics Forum. vol. 35, pp. 139–152. Wiley Online Library (2016)

35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
36. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
37. 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). pp. 117–132 (2018)
38. 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. pp. 1751–1760 (2019)
39. Yan, Q., Gong, D., Zhang, P., Shi, Q., Sun, J., Reid, I., Zhang, Y.: Multi-scale dense networks for deep high dynamic range imaging. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 41–50. IEEE (2019)
40. Yan, Q., Sun, J., Li, H., Zhu, Y., Zhang, Y.: High dynamic range imaging by sparse representation. Neurocomputing **269**, 160–169 (2017)
41. Yan, Q., Sun, J., Li, H., Zhu, Y., Zhang, Y.: High dynamic range imaging by sparse representation. Neurocomputing **269**, 160–169 (2017)
42. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Transactions on Image Processing **29**, 4308–4322 (2020)
43. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)