Alignment-free HDR Deghosting with Semantics Consistent Transformer

Steven Tel[1,*]   Zongwei Wu[1,2,*]   Yulun Zhang[3][†]   Barthélémy Heyrman[4]
Cédric Demonceaux[1,5]   Radu Timofte[2]   Dominique Ginhac[1]

[1] University of Burgundy, CNRS, ICB   [2] Computer Vision Lab, CAIDAS & IFI, University of Würzburg
[3] CVL, ETH Zürich   [4] University of Burgundy, ImViA   [5] University of Lorraine, CNRS, Inria, Loria

{steven.tel; barthelemy.heyrman; cedric.demonceaux; dginhac}@u-bourgogne.fr, {zongwei.wu; radu.timofte}@uni-wuerzburg.de, yulun100@gmail.com

Abstract

High dynamic range (HDR) imaging aims to retrieve information from multiple low-dynamic range inputs to generate realistic output. The essence is to leverage the contextual information, including both dynamic and static semantics, for better image generation. Existing methods often focus on the spatial misalignment across input frames caused by the foreground and/or camera motion. However, there is no research on jointly leveraging the dynamic and static context in a simultaneous manner. To delve into this problem, we propose a novel alignment-free network with a Semantics Consistent Transformer (SCTNet) with both spatial and channel attention modules in the network. The spatial attention aims to deal with the intra-image correlation to model the dynamic motion, while the channel attention enables the inter-image intertwining to enhance the semantic consistency across frames. Aside from this, we introduce a novel realistic HDR dataset with more variations in foreground objects, environmental factors, and larger motions. Extensive comparisons on both conventional datasets and ours validate the effectiveness of our method, achieving the best trade-off on the performance and the computational cost. The source code and dataset are available at https://github.com/Zongwei97/SCTNet.

1. Introduction

While dealing with the full range of illumination in natural scenes, traditional cameras fail to provide images with reasonable quality and result in the loss of information in under-exposed and over-exposed areas. Some specialized hardware devices [18, 29] have been proposed to produce HDR images directly. Unfortunately, they are usually too expensive to be widely adopted. To this end, computational High Dynamic Range (HDR) imaging has recently attracted
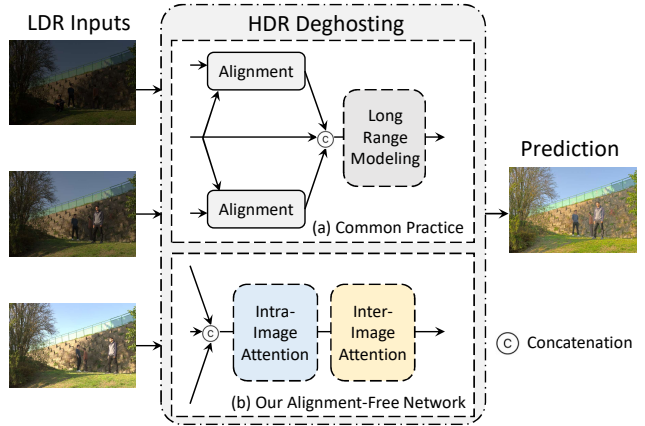

Figure 1. Motivation of our alignment-free network. Most existing methods [15, 27, 8, 24, 34] calibrate the LDR inputs at shallow latent space vis-à-vis the dynamic motion. Differently, we propose an alignment-free network that jointly models the dynamic and static context across frames through attention blocks.

great research attention, aiming to merge multiple Low Dynamic Range (LDR) images with different exposure times to produce an HDR image with increased dynamic range and heightened realism.

Given well-aligned LDR images, traditional imaging methods [4, 25, 7, 10, 31] can produce faithful results. Nevertheless, a perfect alignment is always hard to achieve. In practice, it is very common to have dynamic objects in the foreground and/or with ego motion, resulting in undesirable misalignment between the LDR inputs. Moreover, the movement across frames with different exposure values (EV) may lead to missing content in the over-/under-exposed regions. Therefore, a method that can efficiently leverage information from multiple exposures, while being robust to the misalignment is highly desired.

To tackle the misalignment issue due to the moving object, multiple methods [2, 11] propose to compute the optical flow to model the foreground motion. Several recent methods [10, 17, 3] further integrate optical flow into a learning-based method to align motion between the input

---

[*]Both authors contributed equally to this research.
[†]Corresponding Author: Yulun Zhang

images. However, due to the poor generalization of optical flow methods on multiple exposures, these methods fail to produce reliable results. Other methods [32, 15] compute the affinity matrices between different images with the help of attention modules as shown in Figure 1. Once the multi-EV features are calibrated, they are further merged and fed into deep convolutional neural networks (CNNs) such as [36]. More recently, Niu *et al.* [19] leverages generative adversarial network (GAN) for HDR imaging to deal with ghost artifacts. [8] further merges GAN with an attention module to improve the model efficiency. However, these methods are based on local convolutions, limiting their ability to model and benefit from long-range dependencies. A recent method [15] replaces the deep convolutional network with a transformer and proposes a hybrid CNN-transformer architecture. In general, existing methods [15, 32, 8, 24, 34]often follow the common practice of first computing the cross-image alignment with convolutional attention. Then, the features are further refined with contextualized awareness. Despite the plausible performance, these methods are heavily based on the alignment module for multi-EV feature calibration. Moreover, the SOTA performance is achieved with heavy computational cost, which is undesirable for real-time applications.

We observe that previous works [24, 34, 3, 32, 27] often focus on dealing with the motion changes across different images. However, motion changes are highly correlated, not decoupled, with the static contexts, *i.e.*, the semantic information of the holistic environment during imaging. Inspired by this observation, in this paper, we propose SCTNet, a novel Semantics Consistent Transformer network, which consists of two parts: intra-image spatial feature modeling and inter-EV semantic intertwining. Specifically, instead of computing the attention at a shallow stage, as shown in Figure 1, we directly investigate the transformer network to explore the spatial correlation between the merged multi-EV inputs. This spatial feature modeling operation aims to not only compute the contextualized awareness within a single image but also explore the spatial correlation across LDR inputs. However, directly aggregating the multi-input features is challenging since the different exposures may have different spatial distributions. Thus, we complement the spatial attention block with our new semantic-consistent cross-frame attention block. These two attention blocks work in a consecutive manner to make sure that both the foreground motion and the static context are well-learned and merged simultaneously in an end-to-end manner.

Moreover, to fully benefit from the popular non-local attention mechanism, we introduce a novel dataset for HDR deghosting that adheres to the standard acquisition approach put forth by [10]. Our dataset is collected to fulfill the demand for larger training and evaluation data required to implement vision transformers in HDR deghost-

ing tasks. The dataset comprises diverse global and local motion and illumination ranges to generate more versatile and all-encompassing HDR deghosting solutions. To make algorithms more robust, we have expanded our dataset to include a wider range of scenes, including twilight and night scenes, in contrast to the conventional dataset [10] with only day scenes. To summarize, our contributions are threefold:

- We propose a unified framework, SCTNet, that jointly models the foreground motion and static semantic context simultaneously, both of which are essential for HDR deghosting.
- We introduce a realistic supervised HDR deghosting dataset presenting a richer variety of scenes with more variation in the types of motion and more environmental factors than existing benchmark datasets.
- We demonstrate the effectiveness of our approach through extensive experiments on both the Kalantari dataset [10] as well as our proposed dataset. Our SCTNet performs favorably against other state-of-the-art techniques by reaching the best trade-off between performance and computational cost.

2. Related works

HDR Deghosting Methods: Existing HDR deghosting algorithms can be categorized into two groups: Patch Match-Based and Deep Learning methods.

Patch Match-Based Methods: One category involves pixel rejection techniques. These methods focus on globally aligning LDR images and then discarding misaligned pixels. Approaches include error map creation based on color differences by Grosch *et al.* [6], motion area identification using threshold bitmaps by Pece *et al.* [20], and weighted intensity variance analysis by Jacobs *et al.* [9]. Other methods propose gradient-domain weight maps by Zhang *et al.* [35] and probability maps by Khan *et al.* [12]. However, such methods often lead to suboptimal HDR results due to pixel rejection causing information loss.

Another set of methods focuses on motion registration. These involve aligning non-reference LDR images with the reference LDR image prior to merging. Techniques include optical flow-based motion vector prediction by Bogoni [2], luminance domain transformation using exposure time and optical flow-based motion compensation by Kang *et al.* [11], and optical flow-based registration followed by HDR reconstruction by Zimmer *et al.* [37]. Additionally, Sen *et al.* [25] introduced a patch-based energy minimization for simultaneous alignment and HDR reconstruction. Hu *et al.* [7] enhanced image alignment through brightness and gradient consistency optimization.

Deep Learning Solutions: To address occlusions in patch-based methods, convolutional neural networks (CNNs) gained prominence due to their ability to handle such

challenges more effectively. Kalantari *et al.* [10] pioneered the integration of a sequential neural network into the alignment-before-merging pipeline using optical flow. Wu *et al.* [31] adopted a direct domain translation network for ghost-free HDR generation. Further solutions utilized attention mechanisms to better handle long-range ghosting effects. Yan *et al.* [32] employed spatial attention, while Yan *et al.* [33] used a non-local network for merging. Niu *et al.* [19] introduced a GAN-based training scheme for high-quality ghost-free HDR generation.

Transformers have emerged as a promising solution due to their self-attention mechanism. Song *et al.* [26] proposed a transformer-based selective HDR reconstruction network, and Liu *et al.* [15] introduced a hierarchical network combining Transformers and convolutional channel attention to capture global and local dependencies, respectively.

HDR Deghosting Datasets: The first dataset with HDR labels for training supervised HDR deghosting networks has been proposed by Kalantari *et al.* [10]. Each scene is composed of a set of three LDR input images of the same scene separated by two or three exposure values (EV). Motion in each input image is achieved using a human subject moving in the foreground of a static scene. For each scene, the corresponding HDR ground truth is generated from a static set of three input images (in which the subject is requested to stay still during the acquisitions), using a triangle weighting function similar to that proposed by Debevec *et al.* [4]. The middle exposure input from the static bracket is used as the middle input of the sample. This method enables the creation of realistic HDR ground truth. However, acquiring a large number of samples can be a tedious and time-consuming process, as it involves multiple iterations to ensure that scenes without any motion are selected for ground truth generation. For example, Prahakbar *et al.* [23] utilized this method to create a dataset. Nonetheless, the resulting ground truth images often contain ghosting artifacts.

Another approach for generating a three-input HDR deghosting dataset was introduced by the NTIRE HDR challenge [21]. To create a more diverse dataset, the HDR ground-truth images were obtained from the work of Froelich *et al.* [5], who captured a comprehensive collection of HDR videos using a professional two-camera rig with a semitransparent mirror for the purpose of evaluating HDR displays. Since these images do not have the required LDR input images, the corresponding LDR counterparts were generated synthetically using image formation models that include noise sources.

3. Our Dataset

3.1. Comparison with state-of-the-art HDR datasets

In Table 2, we highlight the strengths of our dataset over existing HDR datasets. Specifically, compared to the widely

Table 1. Comparison against existing benchmarks. Our proposed benchmark is more complete, providing more realistic and challenging samples for HDR deghosting.

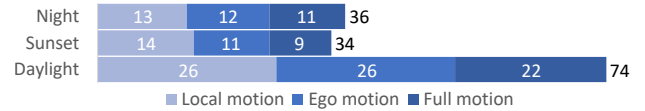| Dataset	P.A	R.W	U.G.T	V.L.C.	I.C.
Kalantari *et al.* [10]	✓	✓	✓	✗	✓
Prabhakar *et al.* [23]	✗	✓	✗	✗	✓
NTIRE21 [21]	✓	✗	✓	✓	✗
Ours	✓	✓	✓	✓	✓
Attribute Description:					
P.A.	Publicly Accessible				
R.W.	Real World dataset				
U.G.T.	Unghosted Ground Truth				
V.L.C.	Variety of Light Conditions				
I.C.	Information Consistency				

Figure 2. Scene and motion distributions of our proposed dataset.

used Kalantari [10] dataset, our dataset offers a wider variety of scenes including day, twilight, and night scenarios, while Kalantari only includes scenes captured during the day. Moreover, the variety of movements in our samples is richer because our dataset includes not only the human subjects as foreground motion but also ego motion and combined human subjects/camera motion. Besides, our dataset contains more scene samples for both training and testing, *i.e.*, 108/36 training/testing samples compared to 74/15 in the Kalantari dataset. It is also noteworthy that the scene variation of Kalantary dataset is limited to the daytime and only 3 among the 15 testing samples contain extremely light camera motion, restricting its evaluative capacity. Conversely, our dataset addresses these limitations with a more **balanced scene and motion distribution**, as shown in Figure 2. We collected samples from diverse light conditions (day, sunset, and night) and encompassed **richer motion types** (local, ego, and full motions) in equal proportions, for both train/test samples. This diverse motion range facilitates a comprehensive assessment of HDR methods across challenging scenes.

During recent challenges [21], other HDR samples have been released. However, the associated LDR images are generated from HDR images with different sources of synthetic noise, resulting in non-faithful inputs. As shown in Figure 3, there may exist information in the ground truth (GT) images, while being not retrievable from the input LDR images. Finally, the recent Prabhakar dataset [23] claims to have more than 500 samples, but, until now, only 32 of them are publicly available. Moreover, as shown in Figure 3, the GT images from the Prabhakar dataset may contain ghosts, which severely limits its application.

In contrast to these state-of-the-art datasets, our dataset is more realistically complete because it was built from high-

Figure 3. Comparison with existing datasets. (A) NTIRE dataset [21] may contain content in the ground truth (GT) images while not being available from the LDR images. (B) Prabhakar [23] dataset generates the GT image with ghosts, which is mainly due to the background motion in the LDR images. (C) Our dataset contains challenging samples with different kinds of motion. From left to right: local motion where only people are moving, ego motion where only the camera is moving, and full motion where both foreground and camera are moving. Please zoom in for more details.

quality LDR inputs and artifact-free HDR ground truth. Thanks to the large variation in both motion types and lighting conditions, our dataset can be naturally used for analyzing the generalization capability across different scenarios.

3.2. Acquisition process

Our objective is to produce a set of LDR images with controlled motion and the GT HDR image. Before any acquisition, we first select the exposure time of the reference image to be certain to get the most information on the scene. For the acquisition process, we use bracketings of 9 images spaced by ± i-EV difference where $i \in 1, 2, 3, 4$ around the reference exposure. Using 9 exposure values guarantees to cover the full dynamic range of a scene. A tripod-mounted camera is used to perform sequentially the 9 acquisitions of a static scene, for which we check that no parasitic motion exists. All the images are captured in raw format with a resolution of 4256 x 2832 pixels using a Nikon D700 camera. Motion in each scene is controlled by a stop motion technique including 3 stops of 9 images. In order to provide an enhanced dataset including greater opportunities to accurately assess the deghosting capabilities of current and future HDR solutions, we propose three different kinds of motions for each scene, as illustrated in Figure 3:
Local motion: The scene is static. Movement between frames is introduced by people or objects in the foreground;
Ego motion: Background and foreground of the scene are static, the deghosting challenge is introduced in all of the pixels of the scene due to the camera ego motion;
Full motion: Composed of both ego and local motions, these scenarios are the most challenging.

3.3. Sample generation

Since there is no motion between the 9 images of each stop, the method proposed by Debevec [4] is applied to produce a ground truth HDR image (named HDR_9) for the second stop by merging the 9 input images. Following the common HDR practice with only 3 LDR inputs, we also have to select the 3 best images from the 9 available candidates in each stop, while maintaining the most information possible. For this purpose, the reference image and the images with ± i-EV difference, $i \in 1, 2, 3, 4$, are selected to produce the different HDR_{3_i} images. In order to preserve the informative features from each input image, the triangle functions proposed by [10] are used as blending weights w_n to generate the ground truth HDR images:

$$\hat{H}(p) = \frac{\sum_{n=1}^{3} w_n(p) H_n(p)}{\sum_{n=1}^{3} w_n(p)}, \tag{1}$$

where the resulting HDR image $\hat{H}$ at the pixel p is the sum of the projected HDR inputs H_n obtained following the gamma projection function described in Eq. 2.

By computing the HDR-VDP2 [16] metrics between HDR_9 and each of the HDR_{3_i} images, we choose the HDR_{3_i} producing the best score as the ground truth HDR image of the sample scene, leading to 200 samples. To ensure high-quality of our dataset, we invited five viewers for quality control and reject the unfaithful results. Our final benchmark contains 108/36 training/testing samples.

4. Our Method

4.1. Problem Setup

We define L_k as the input sequence of k LDR images from a dynamic scene with different exposures, HDR imaging aims at reconstructing an HDR image $\hat{H}$ aligned to a selected reference image L_R. Following the common practice, we chose $k = 3$ as the number of input images, and select the middle exposure L_2 input as the reference input. To increase the robustness of our solution to exposure differences between input frames, we compute the respective projection of each LDR input frame into the HDR domain
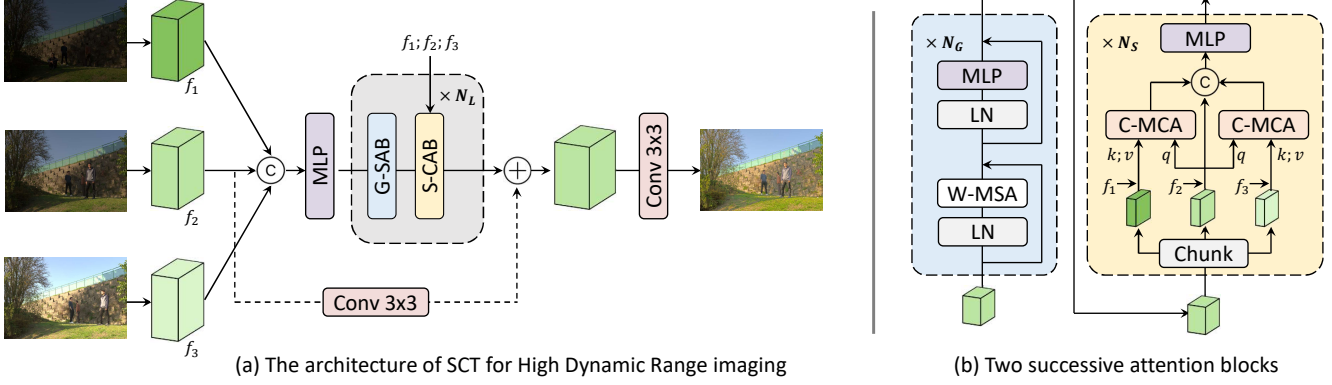
(a) The architecture of SCT for High Dynamic Range imaging   (b) Two successive attention blocks

Figure 4. The architecture of our proposed SCTNet for HDR deghosting. (b) The inner structure of two successive attention blocks G-SAB and S-CAB. The G-SAB leverages the spatial awareness to deal with the intra-image correlation to model the dynamic motion, while our S-CAB block attends to the channel direction and enables the inter-image intertwining to enhance the semantic consistency across frames.

using the gamma encoding function, as described in Eq. 2:

$$H_i = \frac{L_i^\gamma}{t_i}, \quad \gamma = 2.2, \tag{2}$$

where t_i is the exposure time corresponding to the image i, $H_i \in \mathbb{R}^{3 \times H \times W}$ is the gamma-projected input of the corresponding image I_i. By concatenating the LDR images L_i with H_i, we obtain $I_i \in \mathbb{R}^{6 \times H \times W}$ for each EV input.

4.2. Network Architecture

Shallow Features Extraction: Given the input $I_i \in \mathbb{R}^{6 \times H \times W}$, we first extract shallow features f_i with convolution. Unlike existing methods that compute alignment matrices between the shallow features, we directly merge them and feed the fused features into our attention blocks.

Global Spatial Attention Block: To model the spatial correlation between different frames, we adopt the window-based multi-head self-attention following [14, 13, 15]. Technically, the merged features are first fed into the patch partition for tokenization. Then, the sequence of tokens is fed into the spatial attention block, which contains layer norm (LN), window-based multi-head self-attention (W-MSA), and a point-wise multi-layer perceptron (MLP). For the jth layer, $j \in \{1, ..., L\}$, it takes the sequence z_{j-1} as input, and outputs the new sequence z_j:

$$\hat{z}_j = W\text{-}MSA(LN(z_{j-1})) + z_{j-1};$$
$$z_j = MLP(LN(\hat{z}_j)) + \hat{z}_j. \tag{3}$$

Semantic-Consistent Attention Block: As shown in Figure 4, we propose a semantic-consistent attention block (S-CAB) to interact between different frames. Specifically, taking the spatially refined feature z_j, we first split it into three sub-group z'_{ii}, where each group is linked with shallow features of each EV input f_i by a residual addition. We obtain the linked feature z'_i by:

$$z'_i = z'_{ii} + f_i. \tag{4}$$

Then, we apply the channel-wise multi-head cross-attention (C-MCA) between them. Technically, we compute the query from the reference-linked feature, while the keys and values are generated from others. For simplicity, we take z'_1 and z'_2 as examples. Let q_2 be the query from z'_2 and $(k_1; v_1)$ be the key and value generated from z'_1, the output z'_{12} of our C-MCA becomes:

$$z'_{12} = C - MCA(q_2, k_1, v_1) = softmax(\frac{q_2 \times k_1^T}{\sqrt{d_k}})v_1, \tag{5}$$

where d_k is the scaling factor. Similarly, we can obtain the cross-calibrated feature z'_{32} by applying C-MCA on z'_2 and z'_3. Finally, z'_{12}, z'_{22}, and z'_{32} are merged and fed into an FFN to generate the final output of our S-CAB. The computational cost of our C-MCA is:

$$\Omega(C - MCA) = C^2 \times N/9, \tag{6}$$

where C is the channel dimension and N is the spatial resolution. By extending the inter key-query correlation to cross-frame key-query correlation, our S-CAB block suggests a natural way to aggregate multi-EV features. Moreover, our cross-attention is applied on the channel axis, which can deeply preserve and explore the semantic consistency between different frames.

Skip Connection: After the attention blocks, we build a skip connection from the reference feature f_2. The skip connection is realized through a convolution. Finally, we use another convolution to generate the final HDR output.

4.3. Loss function

Our network is trained end-to-end. We first supervise our network with $\mathcal{L}_1$ loss function. Note that Kalantari *et al*. [10] has shown that computing the loss in the HDR domain leads to the appearance of discoloration effects. To avoid this, we compute the $\mathcal{L}_1$ loss in the tone-mapped do-

Table 2. Quantitative comparison with state-of-the-art methods on the Kalantari [10] testing samples. l-PSNR and l-SSIM are computed in the linear domain while μ-PSNR and μ-SSIM are computed after μ-law tone mapping (Eq. 7). PU-PSNR and PU-SSIM are computed by applying the encoding function proposed in [1].

Method	μ-PSNR	PU-PSNR	l-PSNR	μ-SSIM	PU-SSIM	l-SSIM	HDR-VDP2
Hu *et al.* [7]	35.79	34.96	30.76	0.9717	0.9615	0.9503	57.05
Sen *et al.* [25]	40.80	40.47	38.11	0.9808	0.9775	0.9721	59.38
DHDRNet[10]	42.67	41.83	41.23	0.9888	0.9832	0.9846	65.05
DeepHDR[31]	41.65	41.35	40.88	0.9860	0.9815	0.9858	64.90
AHDRNet[32]	43.63	42.93	41.14	0.9900	0.9849	0.9702	64.61
NHDRRNet[33]	42.41	42.97	41.43	0.9877	0.9855	0.9857	61.21
CEN-HDR[28]	43.05	43.24	40.53	0.9908	0.9821	0.9856	64.34
HDRGAN[19]	43.92	44.03	41.57	0.9905	0.9851	0.9865	65.45
HDR-Transformer[15]	44.32	44.23	42.18	0.9916	0.9924	0.9884	66.03
Ours	**44.49**	**44.27**	**42.29**	**0.9924**	**0.9927**	**0.9887**	**66.65**

main using the μ-law function:

$$T(H) = \frac{log(1 + \mu H)}{log(1 + \mu)}, \quad \mu = 5000, \qquad (7)$$

where H is the linear HDR image and μ is the amount of compression. Our l_1 loss for HDR imaging becomes:

$$\mathcal{L}_1 = \|T(H) - T(\hat{H})\|_1, \qquad (8)$$

where T is the μ-law tone mapping function, H and $\hat{H}$ are the prediction and the ground truth in the HDR domain.

Additionally, following [15], we use an auxiliary perceptual loss $\mathcal{L}_p$ for supervision. The perceptual loss measures the difference between the feature representations of the output image and the ground truth image at multiple layers of the pre-trained CNN, realized by computing the mean squared error between the feature maps at each layer. The motivation for such a loss is to assess the dissimilarity between the predicted HDR image and the actual ground truth image. Formally, we have our perceptual loss:

$$\mathcal{L}_p = \sum_j \|\Phi(T(H)) - \Phi(T(\hat{H}))\|_1, \qquad (9)$$

where Φ are the feature maps extracted from a pre-trained VGG-16 network with j indicating the j-th layer. Finally, the training loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_1(T, T_{GT}) + \alpha \cdot \mathcal{L}_p(T, T_{GT}), \qquad (10)$$

where α is a hyper-parameter set to 0.01.

5. Experiments

5.1. Implementation Details

Datasets: We first evaluate our model on the conventional Kalantari [10] benchmark composed of 74 training scenes and 15 testing scenes. We follow the common practice of cropping patches of size 128×128 with a stride of 64 from

the training set. We use rotation and flipping as data augmentation. We also evaluate our proposed dataset which is composed of 108 training samples and 36 testing samples. Each sample represents the acquisition of a dynamic scene caused by large foreground or camera motions. The same dataset preparation and augmentation methods are applied.

Implementation overview: We implement our SCTNet on PyTorch with ADAM optimizer with an initial learning rate of 2e-4, β_1 to 0.9, β_2 to 0.999, and ϵ to 1e-8. We train the network on 2 V100 GPUs. All the counterparts' results are trained or downloaded from the official resources.

Metrics: We use common metrics such as PSNR and SSIM, as well as HDR-VDP2 metric which was specifically designed for the HDR task. We compute PSNR and SSIM evaluation in linear and tone-mapped domains, denoted as $l-$ and $\mu-$, respectively. We also compute PU-PSNR/SSIM in the tone-mapped domain as proposed in [1]. We set the display peak luminance to 100cd/m², a 1000:1 contrast, and an ambient light to 10 lux.

5.2. Evaluation and Comparison

Evaluation on Kalantari *et al*. Dataset: We present the quantitative comparison in Table 2. We compare our method with 9 widely-used HDR methods, varying from early hand-crafted, lightweight, CNN, and Transformer methods. Specifically, Hu *et al.* [7] and Sen *et al.* [25] are based on input patch registration methods. DHDR-Net [10] is one of the pioneering learning-based methods which leverages the optical flow to align the inputs. Similarly, DeepHDR [31] aligns the background feature by the homography and then feeds the aligned inputs into the classical encoder-decoder architecture. NHDRRNet [33] further leverages the non-local attention module during the learning stage. AHDRNet [32] is the first method that introduces an attention block for feature alignment at the shallow stage. Then, the calibrated features are merged and fed into multiple dilated residual dense blocks for feature encoding. HDRGAN [19] is the first GAN-based approach
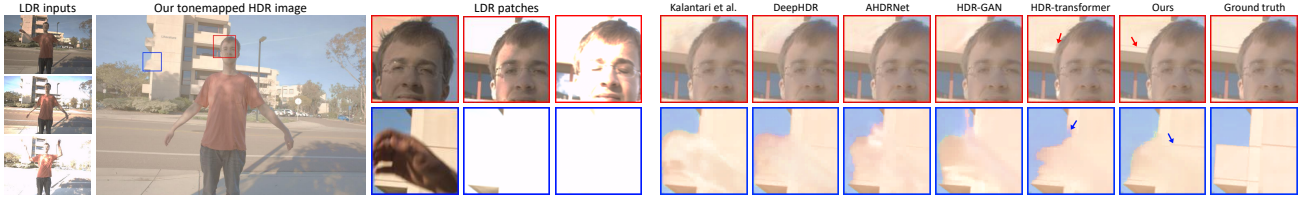
Figure 5. Comparison of our proposed network with state-of-the-art methods on a Kalantari dataset's sample [10]. Thanks to the red patches, we show that our method is the most efficient to recover the ghost from the foreground motion in a bright area of the scene. With blue patches, we exhibit the ability of our solution to recover faithful information in occluded regions. Additionally, as demonstrated in both red and blue patches, our method excels in retrieving the building's texture information. Please zoom in for more details.

Table 3. Quantitative comparison on our dataset. All the models are trained through official implementations. Our method significantly outperforms all counterparts by a large margin. The full comparison can be found in the supplementary material.

Method	μ-PSNR	l-PSNR	μ-SSIM	l-SSIM	HDR-VDP2
Sen *et al.* [25]	39.97	44.21	0.9792	0.9932	67.20
DHDRNet [10]	41.67	47.33	0.9838	0.9961	72.25
AHDRNet [32]	44.16	50.29	0.9896	0.9971	78.12
HDR-Transf. [15]	44.88	51.09	0.9904	0.9981	78.87
Ours	**44.93**	**51.73**	**0.9906**	**0.9981**	**79.53**

Figure 6. Qualitative comparison on our proposed dataset. All the methods are trained using our training set. Our network outperforms all the SOTA counterparts in dealing with ghost artifacts due to the foreground motion. Please zoom in for details.

for HDR merging with a deep supervised HDR method, while HDR-Transformer [15] is the first work by replacing the conventional convolution operations by the transformer architecture. It can be seen that, with the standardized evaluation protocol, our method performs favorably against all the counterparts, setting new records on the conventional Kalantari *et al.* [10] dataset. A qualitative comparison can be found in Figure 5. It can be seen that our method can lead to prediction closer to the ground truth.

Evaluation on Our Dataset : We also conduct comparisons with the most representative HDR methods on our proposed dataset. In Table 3, it can be seen that our method

Table 4. Computational cost comparison of the proposed SCTNet solution against SOTA methods. The number of operations and parameters are evaluated using the script provided by the NTIRE [22] challenge. The input size is set to 1000×1500 pixels. The speed is measured on a Quadro P6000 GPU.

Method	Params. (M)	GMACs	Size (Mb)	Speed (s)
DeepHDR [31]	16.61	1453.70	66.5	0.32
AHDRNet [32]	1.52	2166.69	**6.1**	0.30
HDR-GAN [19]	2.63	778.81	10.6	**0.29**
HDR-Transf. [15]	1.22	981.81	53.4	8.65
Ours	**0.99**	**293.77**	29.1	7.14

sets new records on almost all the metrics, approving our versatility and effectiveness across different datasets. The full performance comparison can be found in the supplementary material. We provide qualitative results of our method in Figure 6. It can be seen that our method can better deal with object movement and generate the HDR output closer to the ground truth. This can be attributed to our attention blocks focusing on inter-image consistency.

Generalization Capability: We provide in Figure 7 the visual comparison on two unseen scenarios. We can observe that our method achieves consistently superior performance compared to all SOTA counterparts. Firstly we lead to better performance in restoring the illumination gradient resulting from the fire motion as shown in Figure7-Left. While dealing with foreground motion as shown in Figure7-Right, Our method outperforms others in recovering the maximum amount of information from the bright areas of the scene. These observations further validate our effectiveness.

Comparison on the Computational Cost: The computational cost of our method, as well as those of our counterparts, can be found in Table 4. Our method achieves the best trade-off between the performance and the computational cost. It is worth noting that compared to the current SOTA HDR-Transformer [15], our method only requires around 1/3 of the GMACs and 1/2 of the model size. Compared to AHDRNet, as shown in we achieve a notable improvement of +1dB on PSNR while using only 13% of the GMACs and 65% of the parameters, showcasing our efficiency. Our method also compiles quicker than the counterpart, validating the efficiency of our proposed attention blocks.
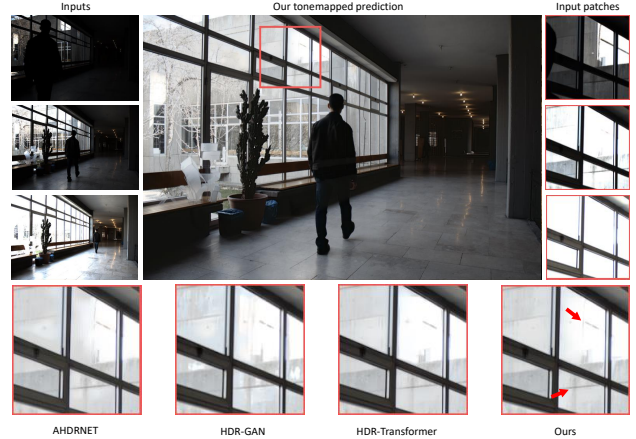
Figure 7. We analyze the generalizability of our solution by evaluating on the unsupervised dataset proposed by Sen *et al*. [30]. For a fair comparison, all networks are trained using our proposed dataset. In the first comparison (left), our network shows superior ability in restoring the illumination gradient resulting from the fire motion. In the second comparison (right), our method outperforms all the counterparts in recovering the maximum amount of information from the bright areas of the scene. Please zoom in for more details.

Table 5. Ablation study on the key components. The comparison is conducted on the Kalantari *et al*. dataset.

Base.	G-SAB	S-CAB	Skip	μ-PSNR	PSNR	μ-SSIM	SSIM
✓				43.42	41.68	0.9882	0.9861
✓	✓			44.04	41.89	0.9921	0.9887
✓	✓	✓		44.12	41.71	0.9923	0.9882
✓	✓	✓	✓	**44.49**	**42.29**	**0.9924**	**0.9887**

Figure 8. I_1-I_{ref} attention visualization.

5.3. Ablation Studies

Key Component Analysis: In this section, we conduct studies to verify the effectiveness and importance of our proposed components: G-SAB, S-CAB, and Skip connection to link with the shallow multi-EV features. The experimental results can be found in Table 5. We gradually add our blocks on top of our baseline (denoted as Base.). It can be seen that each component brings improvement, validating the effectiveness of our proposed method. More ablation studies can be found in the supplementary material.

Attention Visualization: To better understand our proposed attention blocks, we provide feature visualizations between I_1 and the reference frame I_{ref} in Figure 8. It can be seen that: **(1)** Our G-SAB leverages global clues, such as shape and contour, while the S-CAB focuses on feature alignment across input frames. **(2)** The S-CAB leverages semantic clues to compute cross-image affinities, leading to activated attention on misaligned areas, as shown by the **green** bounding boxes. **(3)** The complementary and simultaneous operation of our blocks leads to improved feature modeling for effective deghosting.

6. Conclusion and Future Work

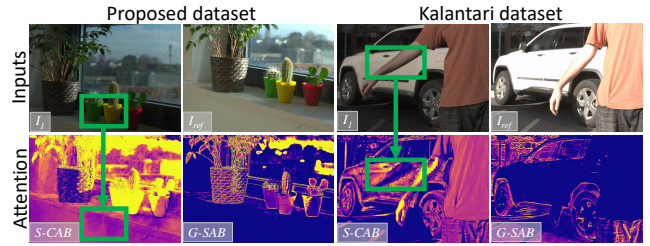In this paper, we present a novel alignment-free network with a Semantics Consistent Transformer for HDR deghost-

ing. Different from existing methods that estimate the image alignment by attention matrix at the shallow stage, our method is unified in that we jointly model the foreground motion and static semantic context simultaneously. We propose two types of attention blocks with intra-/inter-image interaction to directly deal with the alignment issues in the transformer architecture. Furthermore, we propose a novel dataset with more realistic samples and more variations on both lighting conditions and scene motions.

Future work. There are several directions. First, although our method sets new records on both the existing Kalantari benchmark and our dataset, there is room for further improvement, especially in further reducing the computational cost while maintaining the performance. A second direction can be on the domain adaptation. How to improve the performance on one dataset while being trained with the other is yet an open question. We hope our dataset can contribute to the development of methods of better adaptation and generalization capability.

References

[1] Maryam Azimi et al. Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021. 6

[2] L. Bogoni. Extending dynamic range of monochrome and color images through fusion. In *ICPR*, pages 7–12 vol.3, 2000. 1, 2

[3] Sibi Catley-Chandar, Thomas Tanay, Lucas Vandroux, Aleš Leonardis, Gregory Slabaugh, and Eduardo Pérez-Pellitero. Flexhdr: Modeling alignment and exposure uncertainties for flexible hdr imaging. *IEEE TIP*, 31:5923–5935, 2022. 1, 2

[4] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, page 369–378, USA, 1997. ACM Press/Addison-Wesley Publishing Co. 1, 3, 4

[5] Jan Froehlich, Stefan Grandinetti, Bernhard Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. *Proc. SPIE*, 9023, 02 2014. 3

[6] Thorsten Grosch et al. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling and Visualization, RWTH Aachen*, 277284:3, 2006. 2

[7] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *CVPR*, 2013. 1, 2, 6

[8] Yu-Ming Huang, Jui-Chiu Chiang, and Sau-Gee Chen. Hdr-agan: Ghost-free high dynamic range imaging with attention guided adversarial network. In *ICIP*, 2022. 1, 2

[9] Katrien Jacobs, Celine Loscos, and Greg Ward. Automatic high-dynamic range image generation for dynamic scenes. *IEEE CGA*, 28(2):84–93, 2008. 2

[10] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM TOG*, 36(4):144–1, 2017. 1, 2, 3, 4, 5, 6, 7

[11] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM TOG*, 22(3):319–325, jul 2003. 1, 2

[12] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *ICIP*, 2006. 2

[13] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021. 5

[14] 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 *ICCV*, 2021. 5

[15] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7

[16] Rafał 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. *TOG*, 30(4):1–14, 2011. 4

[17] Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *IJCV*, 129:900–920, 2021. 1

[18] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *CVPR*, 2000. 1

[19] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE TIP*, 30:3885–3896, 2021. 2, 3, 6, 7

[20] Fabrizio Pece and Jan Kautz. Bitmap movement detection: Hdr for dynamic scenes. In *CVMP*, 2010. 2

[21] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. Ntire 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *CVPRW*, 2021. 3, 4

[22] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Aleš Leonardis, Radu Timofte, Zexin Zhang, Cen Liu, Yunbo Peng, Yue Lin, Gaocheng Yu, Jin Zhang, Zhe Ma, Hongbin Wang, Xiangyu Chen, Xintao Wang, Haiwei Wu, Lin Liu, Chao Dong, Jiantao Zhou, Qingsen Yan, Song Zhang, Weiye Chen, Yuhang Liu, Zhen Zhang, Yanning Zhang, Javen Qinfeng Shi, Dong Gong, Dan Zhu, Mengdi Sun, Guannan Chen, Yang Hu, Haowei Li, Baozhu Zou, Zhen Liu, Wenjie Lin, Ting Jiang, Chengzhi Jiang, Xinpeng Li, Mingyan Han, Haoqiang Fan, Jian Sun, Shuaicheng Liu, Juan Marín-Vega, Michael Sloth, Peter Schneider-Kamp, Richard Röttger, Chunyang Li, Long Bao, Gang He, Ziyao Xu, Li Xu, Gen Zhan, Ming Sun, Xing Wen, Junlin Li, Jinjing Li, Jinjing Li, Chenghua Li, Chenghua Li, Ruipeng Gang, Ruipeng Gang, Fangya Li, Fangya Li, Chenming Liu, Chenming Liu, Shuang Feng, Fei Lei, Rui Liu, Junxiang Ruan, Tianhong Dai, Wei Li, Zhan Lu, Hengyan Liu, Peian Huang, Guangyu Ren, Yonglin Luo, Chang Liu, Qiang Tu, Sai Ma, Yizhen Cao, Steven Tel, Barthelemy Heyrman, Dominique Ginhac, Chul Lee, Gahyeon Kim, Seonghyun Park, An Gia Vien, Truong Thanh Nhat Mai, Howoon Yoon, Tu Vo, Alexander Holston, Sheir Zaheer, and Chan Y. Park. Ntire 2022 challenge on high dynamic range imaging: Methods and results. In *CVPRW*, 2022. 7

[23] K. Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R. Venkatesh Babu. Towards practical and efficient high-resolution hdr deghosting with cnn. In *ECCV*, 2020. 3, 4

[24] Zhiyuan Pu, Peiyao Guo, M Salman Asif, and Zhan Ma. Robust high dynamic range (hdr) imaging with complex motion and parallax. In *ACCV*, 2020. 1, 2

[25] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust Patch-Based HDR Reconstruction of Dynamic Scenes. *ACM TOG*, 31(6):203:1–203:11, 2012. 1, 2, 6, 7

[26] Jou Won Song, Ye-In Park, Kyeongbo Kong, Jaeho Kwak, and Suk-Ju Kang. Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In *ECCV*, 2022. 3

[27] Xiao Tan, Huaian Chen, Kai Xu, Yi Jin, and Changan Zhu. Deep sr-hdr: Joint learning of super-resolution and high dynamic range imaging for dynamic scenes. *IEEE TMM*, 2021. 1, 2

[28] Steven Tel, Barthélémy Heyrman, and Dominique Ginhac. Cen-hdr: Computationally efficient neural network for real-time high dynamic range imaging. In *ECCVW*, 2022. 6

[29] Jack Tumblin, Amit Agrawal, and Ramesh Raskar. Why i want a gradient camera. In *CVPR*, 2005. 1

[30] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. An objective deghosting quality metric for hdr images. *Computer Graphics Forum*, 35(2):139–152, 2016. 8

[31] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. End-to-end deep hdr imaging with large foreground motions. In *ECCV*, 2018. 1, 3, 6, 7

[32] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. *CVPR*, 2019. 2, 3, 6, 7

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

[34] 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 *CVPRW*, 2022. 1, 2

[35] Wei Zhang and Wai-Kuen Cham. Gradient-directed multi-exposure composition. *IEEE TIP*, 21(4):2318–2323, 2011. 2

[36] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 2

[37] Henning Zimmer, Andres Bruhn, and Joachim Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. *Comput. Graph. Forum*, 30:405–414, 04 2011. 2