SMAE: Few-shot Learning for HDR Deghosting with Saturation-Aware Masked Autoencoders

Qingsen Yan[†1]   Song Zhang[†2]   Weiye Chen[†2]   Hao Tang[3]   Yu Zhu[1]
Jinqiu Sun[1]   Luc Van Gool[3]   Yanning Zhang[1*]
[1]Northwestern Polytechnical University   [2]Xidian University   [3]CVL, ETH Zurich

Abstract

Generating a high-quality High Dynamic Range (HDR) image from dynamic scenes has recently been extensively studied by exploiting Deep Neural Networks (DNNs). Most DNNs-based methods require a large amount of training data with ground truth, requiring tedious and time-consuming work. Few-shot HDR imaging aims to generate satisfactory images with limited data. However, it is difficult for modern DNNs to avoid overfitting when trained on only a few images. In this work, we propose a novel semi-supervised approach to realize few-shot HDR imaging via two stages of training, called SSHDR. Unlikely previous methods, directly recovering content and removing ghosts simultaneously, which is hard to achieve optimum, we first generate content of saturated regions with a self-supervised mechanism and then address ghosts via an iterative semi-supervised learning framework. Concretely, considering that saturated regions can be regarded as masking Low Dynamic Range (LDR) input regions, we design a Saturated Mask AutoEncoder (SMAE) to learn a robust feature representation and reconstruct a non-saturated HDR image. We also propose an adaptive pseudo-label selection strategy to pick high-quality HDR pseudo-labels in the second stage to avoid the effect of mislabeled samples. Experiments demonstrate that SSHDR outperforms state-of-the-art methods quantitatively and qualitatively within and across different datasets, achieving appealing HDR visualization with few labeled samples.

1. Introduction

Standard digital photography sensors are unable to capture the wide range of illumination present in natural scenes, resulting in Low Dynamic Range (LDR) images that often
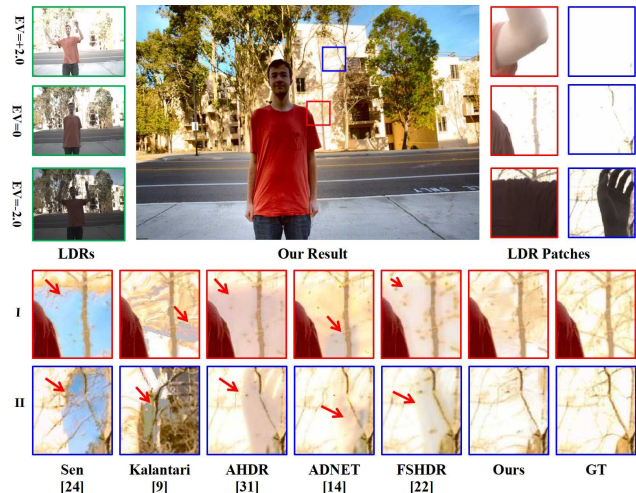

Figure 1. The proposed method generates high-quality images with few labeled samples when compared with several methods.

suffer from over or underexposed regions, which can damage the details of the scene. High Dynamic Range (HDR) imaging has been developed to address these limitations. This technique combines several LDR images with different exposures to generate an HDR image. While HDR imaging can effectively recover details in static scenes, it may produce ghosting artifacts when used with dynamic scenes or hand-held camera scenarios.

Historically, various techniques have been suggested to address such issues, such as alignment-based methods [3,10,27,37], patch-based methods [8,15,24], and rejection-based methods [5, 11, 19, 20, 35, 40]. Two categories of alignment-based approaches exist: rigid alignment (*e.g.*, homographies) that fail to address foreground motions, and non-rigid alignment (*e.g.*, optical flow) that is error-prone. Patch-based techniques merge similar regions using patch-level alignment and produce superior results, but suffer from high complexity. Rejection-based methods aim to eliminate misaligned areas before fusing images, but may result in a loss of information in motion regions.

As Deep Neural Networks (DNNs) become increasingly prevalent, the DNN-based HDR deghosting methods

[9, 33, 36] achieve better visual results compared to traditional methods. However, these alignment approaches are error-prone and inevitably cause ghosting artifacts (see Figure 1 Kalantari's results). AHDR [31, 32] proposes spatial attention to suppress motion and saturation, which effectively alleviate misalignment problems. Based on AHDR, ADNET [14] proposes a dual branch architecture using spatial attention and PCD-alignment [29] to remove ghosting artifacts. All these above methods directly learn the complicated HDR mapping function with abundant HDR ground truth data. However, it's challenging to collect a large amount of HDR-labeled data. The reasons can be attributed to that 1) generating a ghost-free HDR ground truth sample requires an absolute static background, and 2) it is time-consuming and requires considerable manpower to do manual post-examination. This generates a new setting that only uses a few labeled data for HDR imaging.

Recently, FSHDR [22] attempts to generate a ghost-free HDR image with only few labeled data. They use a preliminary model trained with a large amount of unlabeled dynamic samples, and a few dynamic and static labeled samples to generate HDR pseudo-labels and synthesize artificial dynamic LDR inputs to improve the model performance of dynamic scenes further. This approach expects the model to handle both the saturation and the ghosting problems simultaneously, but it is hard to achieve under the condition of few labeled data, especially misaligned regions caused by saturation and motion (see Figure 1 FSHDR). In addition, FSHDR uses optical flow to forcibly synthesize dynamic LDR inputs from poorly generated HDR pseudo-labels, the errors in optical flow further intensify the degraded quality of artificial dynamic LDR images, resulting in an apparent distribution shift between LDR training and testing data, which hampers the performance of the network.

The above analysis makes it very challenging to directly generate a high-quality and ghost-free HDR image with few labeled samples. A reasonable way is to address the saturation problems first and then cope with the ghosting problems with a few labeled samples. In this paper, we propose a semi-supervised approach for HDR deghosting, named SSHDR, which consists of two stages: self-supervised learning network for content completion and sample-quality-based iterative semi-supervised learning for deghosting. In the first stage, we pretrain a Saturated Mask AutoEncoder (SMAE), which learns the representation of HDR features to generate content of saturated regions by self-supervised learning. Specifically, considering that the saturated regions can be regarded as masking the short LDR input patches, inspired by [6], we randomly mask a high proportion of the short LDR input and expect the model to reconstruct a no-saturated HDR image from the remaining LDR patches in the first stage. This self-supervised approach allows the model to recover the saturated regions

with the capability to effectively learn a robust representation for the HDR domain and map an LDR image to an HDR image. In the second stage, to prevent the overfitting problem with a few labeled training samples and make full use of the unlabeled samples, we iteratively train the model with a few labeled samples and a large amount of HDR pseudo-labels from unlabeled data. Based on the pretrained SMAE, a sample-quality-based iterative semi-supervised learning framework is proposed to address the ghosting artifacts. Considering the quality of pseudo-labels is uneven, we develop an adaptive pseudo-labels selection strategy to pick high-quality HDR pseudo-labels (*i.e.*, well-exposed, ghost-free) to avoid awful pseudo-labels hampering the optimization process. This selection strategy is guided by a few labeled samples and enhances the diversity of training samples in each epoch. The experiments demonstrate that our proposed approach can generate high-quality HDR images with few labeled samples and achieves state-of-the-art performance on individual and cross-public datasets. Our contributions can be summarized as follows:

- We propose a novel and generalized HDR self-supervised pretraining model, which uses mask strategy to reconstruct an HDR image and addresses saturated problems from one LDR image.
- We propose a sample-quality-based semi-supervised training approach to select well-exposed and ghost-free HDR pseudo-labels, which improves ghost removal.
- We perform both qualitative and quantitative experiments, which show that our method achieves state-of-the-art results on individual and cross-public datasets.

2. Related Work

2.1. HDR Deghosting Methods

The existing HDR deghosting methods include four categories: alignment-based method, patch-based method, rejection-based method, and CNN-based method.

Alignment-based Method. Rigid or non-rigid registration is mainly used in alignment-based approaches. Bogoni [3] estimated flow vectors to align with the reference images. Kang *et al*. [10] utilized optical flow to align images in the luminance domain to remove ghosting artifacts. Tomaszewska *et al*. [27] used SIFT feature to perform global alignments. Since the dense correspondence computed by alignment methods are error-prone, they cannot handle large motion and occlusion.

Rejection-based Method. Rejection-based methods detect and eliminate motion from static regions. Then they merge static inputs to get HDR images. Grosch *et al*. [5] estimated a motion map and used it to generate ghost-free HDR. Zhang *et al*. [40] obtained a motion weighting map using quality measurement on image gradients. Lee *et al*. [11] and Oh *et al*. [19] detected motion regions using rank min-

imization. However, rejection-based methods remove the misalignment of regions. It will result in a lack of content in moving regions.

Patch-based Method. Patch-based methods use patch-level alignment to merge similar contents. Sen *et al*. [24] proposed a patch-based energy minimization approach that optimizes alignment and reconstruction simultaneously. Hu *et al*. [8] utilized a patch-match mechanism to produce aligned images. Although these methods have good performance, they suffer from high computational costs.

CNN-based Method. Kalantari *et al*. [9] used a CNN network to fuse LDR images that are aligned with optical flow. Wu *et al*. [30] used homography to align the camera motion and reconstructed HDR images by CNN. Yan *et al*. [31] proposed an attention mechanism to suppress motion and saturation. Yan *et al*. [34] designed a nonlocal block to release the constraint of locality receptive field for global merging HDR. Niu *et al*. [18] proposed HDR-GAN to recover missing content using generative adversarial networks. Ye *et al*. [38] proposed multi-step feature fusion to generate ghost-free images. Liu *et al*. [14] utilized the PCD alignment subnetwork to remove ghosts. However, these methods require a large number of labeled samples, which is difficult to collect.

2.2. Few-shot Learning (FSL)

Humans can successfully learn new ideas with relatively little supervision. Inspired by such ability, FSL aims to learn robust representations with few labeled samples. There are three main categories for FSL methods: data-based category [1, 23, 25], which augment the experience with prior knowledge; model-based category [2, 17, 26], which shrinks the size of the hypothesis space using prior knowledge; algorithm-based category [4, 12, 39], which modifies the search for the optimal hypothesis using prior knowledge. For HDR deghosting, Prabhakar *et al*. [22] proposed a data-based category deghosting method, which uses artificial dynamic sequences synthesis for motion transfer. Still, it is hard to handle both the saturation and the ghosting problems simultaneously with few labeled data.

3. The Proposed Method

3.1. Data Distribution

Following the setting of few-shot HDR imaging [22], we utilize 1) N dynamic unlabeled LDR samples $U=\{L_1^U, \ldots, L_N^U\}$, where each L^U consists of three LDRs (X_1^U, X_2^U, X_3^U) with different exposures. 2) M static labeled LDR samples $S=\{L_1^S, \ldots, L_M^S\}$, where each L^S consists of three LDRs (X_1^S, X_2^S, X_3^S) with different exposures and ground truth Y^S. 3) K dynamic labeled LDR samples $D=\{L_1^D, \ldots, L_K^D\}$, where each L^D consists of three LDRs (X_1^D, X_2^D, X_3^D) and ground truth Y^D. Since it is difficult to

collect labeled samples, we set K to be less than or equal to 5, and M is fixed at 5. While it is easy to capture unlabeled samples, N can be arbitrary.

3.2. Model Overview

Generating a non-saturated and ghost-free HDR image with few labeled samples is challenging. It is a proper way to address saturated problems first and then handle ghosting problems. As shown in Figure 2, we propose a semi-supervised approach for HDR deghosting. Our approach consists of two stages: a self-supervised learning network for content completion and a sample-quality-based iterative semi-supervised learning for deghosting. In the first stage, we propose a multi-scale Transformer model based on self-supervised learning with a saturated-masked autoencoder to make it capable of recovering saturated regions. In a word, we randomly mask LDR patches and reconstruct non-saturated HDR images from the remaining LDR patches.

In the second stage, we propose a sample-quality-based iterative semi-supervised learning approach that learns to address ghosting problems. We finetune the pretrained model based on the first stage with a few labeled samples. Then, we iteratively train the model with labeled samples and unlabeled samples with pseudo-labels. Considering that the HDR pseudo-labels inevitably contain saturated and ghosting regions, which deteriorate the model performance, we propose an adaptive pseudo-labels selection strategy to pick high-quality HDR pseudo-labels to avoid awful pseudo-labels hampering the optimization process.

3.3. Self-supervised Learning Stage

Input. Considering that there are more saturated regions in the medium (X_2^U) and long exposure frames (X_3^U) of unlabeled data U, we first transform the short exposure frame (X_1^U) into a new medium $(X_{2'}^U)$ and long exposure frames $(X_{3'}^U)$ by exposure adjustment,

$$X_{i'}^U = clip\left(\left(\frac{(X_1^U)^\gamma \times t_i}{t_1}\right)^{\frac{1}{\gamma}}\right), \quad i = 2, 3. \quad (1)$$

Then following previous work [9, 30], we map the LDR input images $X_1^U, X_{2'}^U, X_{3'}^U$ to HDR domain by gamma correction to get H_i,

$$H_{i'} = (X_{i'}^U)^\gamma / t_i. \quad (2)$$

Note that $X_{1'}^U = X_1^U$, t_i denotes the exposure time of LDR image X_i, γ represents the gamma correction parameter, we set γ to 2.2. Then, we concatenate $X_{i'}$ and $H_{i'}$ along the channel dimension to get a 6-channel input $I_i = [X_{i'}, H_{i'}]$, we subsequently mask the input I_i patches to get I_i'. Concretely, we divide the input into non-overlapping patches and randomly mask a subset of these patches with a high mask ratio (75%) (see Figure 3). Note that the patch size
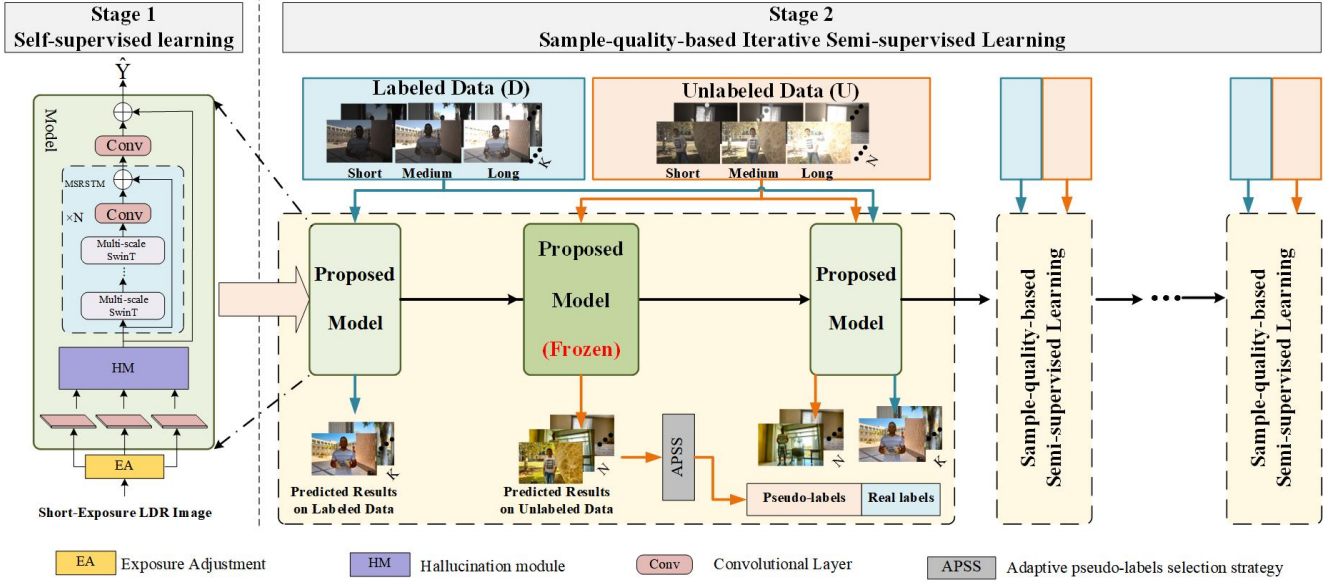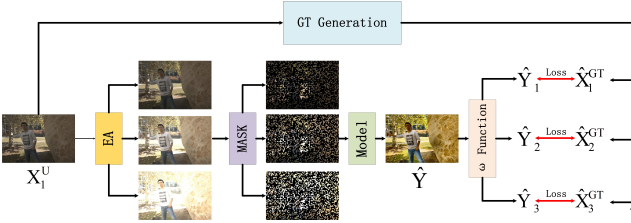
Figure 2. The overview of our framework.

Figure 3. The detailed procedure of Stage 1. To recover the saturated regions, we utilize the short exposure frame as input and ground truth.

is 8×8. Considering that the masking strategy is another way to destruct the saturated regions, we intend the model to learn a robust representation to recover these saturated regions. Finally, $I' = \{I'_1, I'_2, I'_3\}$ is the input of the model.

Model. Our SMAE self-supervised training-based multi-scale Transformer consists of a feature extraction module, hallucination module, and Multi-Scale Residual Swin Transformer fusion Module (MSRSTM). The details in our model are included in the Appendix.

Hallucination Module. We first adopt three convolutional layers to extract shallow feature F_i. Then, we divide the shallow feature F_i into non-overlapping patches $\overline{F_i}$, and map each patch $\overline{F_i}$ into query, keys and values. Subsequently, we calculate the similarity map between $\overline{q}$ and $\overline{k}$, and perform the Softmax function to get the attention weight. Finally, we apply the attention weight to $\overline{v}$ to get F_s^i,

$$\overline{q} = \overline{F_2}W_q, \quad \overline{k_i} = \overline{F_i}W_k, \quad \overline{v_i} = \overline{F_i}W_v, \quad i = 1, 3$$
$$F_s^i = \text{Softmax}(\overline{q}\overline{k_i}^T / \sqrt{d} + b)\overline{v_i}, \tag{3}$$

where b represents a learnable position encoding, d denotes the dimension of $\overline{q}$.

MSRSTM. To merge more information from different exposure regions, inspired by [13], we propose a Multi-Scale Residual Swin Transformer Module (MSRSTM). First, F_s^1, F_s^2, F_s^3 is concatenated along the channel dimension to get the input of MSRSTM. Note that F_s^2 denotes F_2. Then, MSRSTM merges a long range of information from different exposure regions. MSRSTM consists of multiple multi-scale Swin Transformer layers (STL), a few convolutional layers, and a residual connection. Given the input feature $F_{out,i}^{N-1}$ of i-th MSRSTM, the output $F_{out,i}^N$ of MSRSTM can be formulated as follows :

$$F_{STL,i}^N = Conv((Concat(STL_i^{N,l_1}(F_{out,i}^{N-1}), \\ STL_i^{N,l_2}(F_{out,i}^{N-1}), STL_i^{N,l_3}(F_{out,i}^{N-1})), \tag{4}$$

$$F_{out,i}^N = Conv(F_{STL,i}^N) + F_{out,i}^{N-1}, \tag{5}$$

where $STL_i^{N,l_j}(\cdot)$ represents the N-th Swin Transformer layer of the l_j scale in the i-th MSRSTM, $F_{out,i}^{N-1}$ denotes the input feature of the N-th Swin Transformer layer in the i-th MSRSTM.

Loss Function. Since unlabeled samples do not have HDR ground truth labels, we calculate the self-supervised loss in the LDR domain. We first use function ω to transform the predicted HDR image $\hat{Y}$ to short, medium, and long exposure LDR images $\hat{Y_i}$,

$$\hat{Y_i} = \omega(\hat{Y}) = (\hat{Y} \times t_i)^{\frac{1}{\gamma}}. \tag{6}$$

To recover the saturated regions, we transform the short exposure frame (since the predicted HDR in this stage is aligned to the short exposure frame) to new short, medium, and long exposure frames by ground truth generation. Then, we regard the new exposure frames as the ground truth X_i^{GT} of the model,

$$X_i^{GT} = \left(\frac{(X_1^U)^\gamma \times t_i}{t_1}\right)^{\frac{1}{\gamma}}, \quad i = 1, 2, 3. \qquad (7)$$

Finally, we calculate L_1 self-supervised loss between $\hat{Y}_i$ and X_i^{GT},

$$L_{SSL} = \sum_{i=1}^{3} ||\hat{Y}_i - X_i^{GT}||_1. \qquad (8)$$

3.4. Semi-supervised Learning Stage

Finetune. At the beginning of this stage, to improve the saturated regions and further learn to handle ghosting regions, we first finetune the pretrained model with a few dynamic samples D and static labeled samples S. Here we apply μ-law to map the linear domain image to the tonemapped domain image,

$$T(x) = \frac{log(1 + \mu x)}{log(1 + \mu)}, \qquad (9)$$

where $T(x)$ is the tonemap function, μ=5000. Then we calculate the reconstruction loss L_{recon} and perceptual loss L_{percep} between the predicted HDR $\hat{Y}_0^D, \hat{Y}_0^S$ and ground truth HDR Y_0^D, Y_0^S,

$$L_{recon} = ||T(\hat{Y}_0^D) - T(Y_0^D)||_1 + ||T(\hat{Y}_0^S) - T(Y_0^S)||_1, \qquad (10)$$

$$L_{percep} = ||\phi_{i,j}(T(\hat{Y}_0^D)) - \phi_{i,j}(T(Y_0^D))||_1 \\ + ||\phi_{i,j}(T(\hat{Y}_0^S)) - \phi_{i,j}(T(Y_0^S))||_1, \qquad (11)$$

$$L_{finetune} = L_{recon} + \lambda L_{percep}, \qquad (12)$$

where $\phi_{i,j}$ represents the j-th convolutional layer and the i-th max-pooling layer in VGG19, λ=$1e^{-2}$.

Iteration. To prevent the overfitting problem with a few labeled training samples and exploit unlabeled samples, we further generate the pseudo-labels $\hat{Y}_t^U$ of unlabeled data. Concretely, we iteratively and adaptively train the model with a few dynamic and static samples D and S and a large number of unlabeled samples U. Specifically, at timestep t, we use model N_t to predict the pseudo-labels $\hat{Y}_t^U$ of unlabeled data. Then, we train the model N_t with a few labeled and pseudo-labeled samples to get the model N_{t+1} at timestep $t + 1$. Note that we use finetune model to generate unlabel HDR pseudo-labels $\hat{Y}_0^U$ at timestep t=0. Finally, at each timestep in the refinement stage, we calculate the reconstruction loss and perceptual loss as follows,

$$L_{Iteration} = L_{recon,t+1}^D + L_{recon,t+1}^S + \sum_{i=1}^{N} W_{t+1}^{U_i} L_{recon,t+1}^{U_i}$$

$$+ \lambda(L_{percep,t+1}^D + L_{percep,t+1}^S + \sum_{i=1}^{N} W_{t+1}^{U_i} L_{percep,t+1}^{U_i}), \qquad (13)$$

where λ=$1e^{-2}$. $W_{t+1}^{U_i}$ is the weight factor of unlabeled data U_i. To get loss weight $W_{t+1}^{U_i}$, please refer to the next section in detail.

APSS. Since the HDR pseudo-labels inevitably contain saturated and ghosted samples, we propose an Adaptive Pseudo-labels Selection Strategy (APSS) to pick well-exposed and ghost-free HDR pseudo-labels to avoid awful pseudo-labels hampering the optimization process. Specifically, at timestep t, we use model N_t to predict HDR images with dynamic and static labeled samples $\hat{Y}_t^D$ and $\hat{Y}_t^S$. Then we use function ω to map the predicted HDR image to medium exposure image $\tilde{Y}_t^{D \cup S}$ and calculate the loss between $\tilde{Y}_t^{D \cup S}$ and original medium exposure LDR image $X_{2,t}^{D \cup S}$ in well exposure regions to get $L_{select,t}^{D \cup S}$,

$$L_{select,t}^{D \cup S} = ||mask(\omega(\hat{Y}_t^D)) - mask(\omega(X_{2,t}^D))||_1 \\ + ||mask(\omega(\hat{Y}_t^S)) - mask(\omega(X_{2,t}^S))||_1, \qquad (14)$$

where $mask(\cdot)$ denotes masking the over and under-exposure regions. Subsequently, we sort all patches' losses, and adopt $\sigma(\cdot, \cdot)$ function to get β percentile (85th) loss as a selection threshold τ_t,

$$\tau_t = \sigma(L_{select,t}^{D \cup S}, \beta). \qquad (15)$$

Furthermore, we use model N_t to predict pseudo-labels $\hat{Y}_t^U$ of unlabeled samples, similar to the operation of labeled data mentioned above. We then use ω function to map $\hat{Y}_t^U$ to medium exposure to get $\tilde{Y}_t^U$ and calculate the loss between $\tilde{Y}_t^U$ and original medium exposure LDR image $X_{2,t}^U$ to get $L_{select,t}^U$=$\{L_{select,t}^{U_1}, L_{select,t}^{U_2}, \ldots, L_{select,t}^{U_N}\}$. If the current loss $L_{select,t}^{U_i}$ is greater than τ_t, we consider the pseudo-label to be of poor quality, which has more saturated and ghosted regions. Then we will give a lower weight which tends to decay linearly in the next training iteration.

$$L_{select,t}^U = ||mask(\omega(\hat{Y}_t^U)) - mask(\omega(X_{2,t}^U))||_1, \quad (16)$$

$$m_t^U = max(L_{select,t}^U), \qquad (17)$$

$$W_{t+1}^{U_i} = \begin{cases} 1 & L_{select,t}^{U_i} \leq \tau_t \\ \frac{m_t^U - L_{select,t}^{U_i}}{m_t^U - \tau_t} & L_{select,t}^{U_i} > \tau_t \end{cases} \qquad (18)$$

where $X_{2,t}^U$ is the unlabeled medium exposure image in timestep t, m_t^U is the largest selection loss of unlabeled samples in timestep t, $W_{t+1}^{U_i}$ is the weight factor of U_i smaple in the $t + 1$ training iteration.

4. Experiments

Datasets. We train all the methods on two public datasets, Kalantari's [9] and Hu's dataset [7]. Kalantari's dataset includes 74 training samples and 15 testing samples. Three different LDR images in a sample are captured with exposure biases of {-2, 0, +2} or {-3, 0, +3}. Hu's dataset is captured at three exposure levels (*i.e.*, {-2, 0, +2}). There are
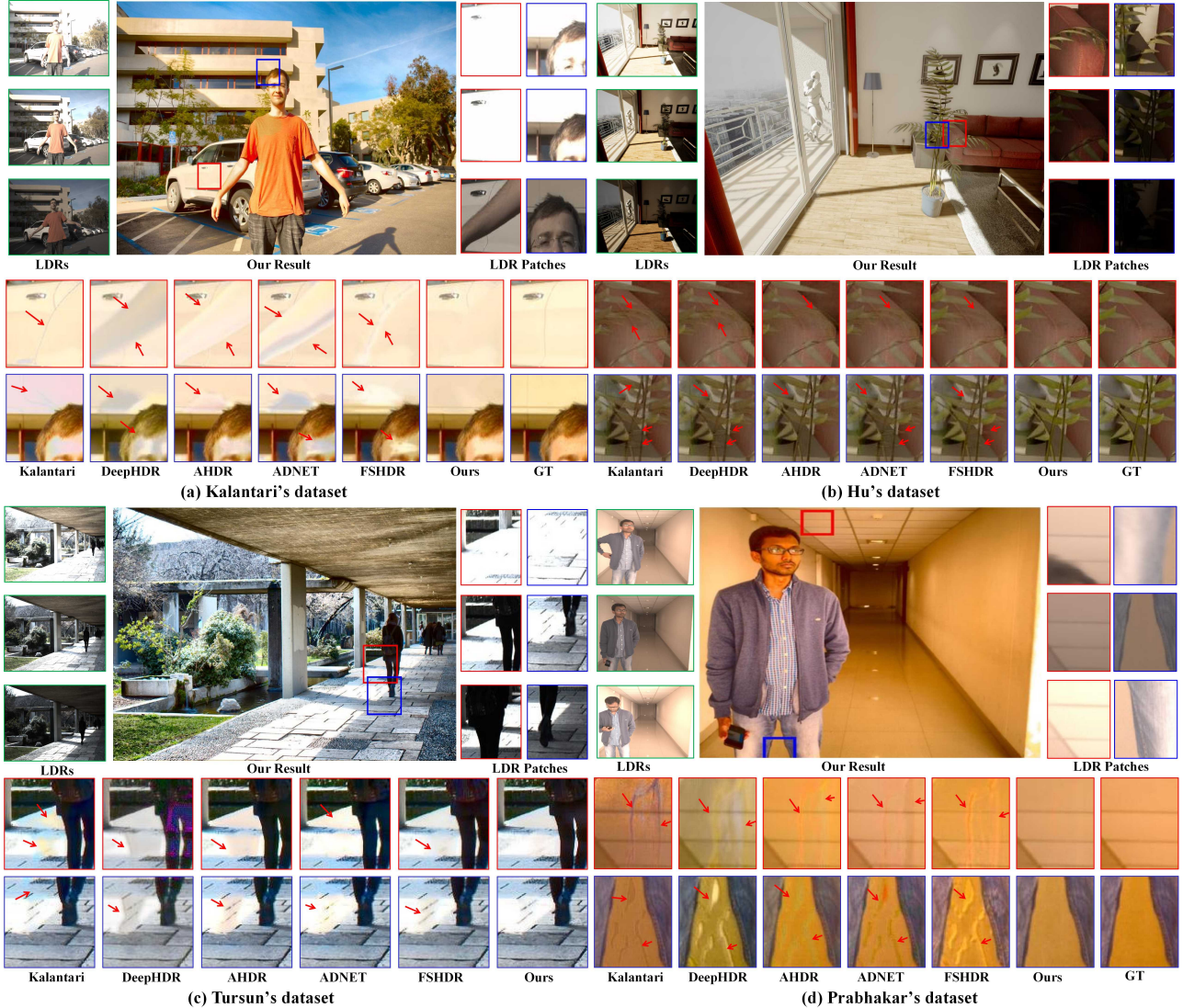
Figure 4. Examples of Kalantari's [9] and Hu's [7] datasets (top row) and Tursun's [28] and Prabhakar's [21] datasets (bottom row). Note that we directly evaluate the methods on Tursun's and Prabhakar's datasets with the checkpoint trained on Kalantari's dataset.

85 training samples and 15 testing samples in Hu's dataset. We train all comparison methods with the same set of images. Concretely, we randomly choose $K \in \{1, 5\}$ dynamic labeled samples and $Q=5$ static labeled samples for training in all methods. Furthermore, for each K, we evaluate all methods for 5 runs denoted as 5-way in Table 1. In addition, since FSHDR [22] and our method exploit unlabeled samples, we also use the rest of the dataset samples as unlabeled data U. Finally, to verify generalization performance, we evaluate all methods on Tursun's dataset [28] that does not have ground truth and Prabhakar's dataset [21].

Evaluation Metrics. We calculate five common metrics used for testing, *i.e.*, PSNR-L, PSRN-μ, SSIM-L, SSIM-μ, and HDR-VDP-2 [16], where 'L' denotes linear domain, '-μ' denotes tonemapping domain.

Implementation Details. The window size in MSRSTM is

2×2, 4×4 and 8×8. In the training stage, we crop the 128×128 patches with stride 64 for the training dataset. We use the Adam optimizer, and set the batch size and learning rate as 4 and 0.0005, receptively. And we set β_1=0.9, β_2=0.999, and ϵ=$1e^{-8}$ in the Adam optimizer. We implement our model using PyTorch with 2 NVIDIA GeForce 3090 GPUs and train for 200 epochs.

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

To evaluate our model, we carry out quantitative and qualitative experiments comparing with several state-of-the-art methods, including patch-based classical methods: Sen [24], Hu [8], and deep learning-based methods: Kalantari [9], DeepHDR [30], AHDRNet [31], ADNet [14], FSHDR [22]. We use the codes provided by the authors.

Evaluation on Kalantari's and Hu's Datasets. In Fig-

Table 1. The evaluation results on Kalantari's [9] and Hu's [7] datasets. The best and the second best results are highlighted in **Bold** and Underline, respectively.

Dataset	Metric	Setting	Kalantari	DeepHDR	AHDRNet	ADNet	FSHDR	Ours
Kalantari	PSNR-l	5way-5shot	39.37±0.12	38.25±0.29	40.61±0.10	40.78±0.15	41.39±0.12	**41.54**±0.10
	PSNR-μ	5way-5shot	39.86±0.19	38.62±0.27	41.05±0.32	40.93±0.38	41.40±0.13	**41.61**±0.08
	PSNR-l	5way-1shot	36.94±0.44	36.67±0.67	38.83±0.39	38.96±0.35	41.04±0.11	**41.14**±0.11
	PSNR-μ	5way-1shot	37.33±1.21	37.01±1.68	39.15±1.04	39.08±1.06	41.13±0.07	**41.25**±0.05
Hu	PSNR-l	5way-5shot	41.36±0.25	40.73±0.66	46.37±0.76	46.88±0.81	47.13±0.13	**47.41**±0.12
	PSNR-μ	5way-5shot	38.95±0.14	39.92±0.22	43.42±0.44	43.79±0.48	43.98±0.27	**44.24**±0.17
	PSNR-l	5way-1shot	38.67±0.43	37.82±0.86	44.64±0.80	44.75±0.84	44.94±0.23	**45.04**±0.16
	PSNR-μ	5way-1shot	36.83±0.62	38.49±1.07	42.37±1.42	42.41±1.20	42.50±0.87	**42.55**±0.44

Table 2. Further evaluation results on Kalantari's [9], Hu's [7] and Prabhakar's datasets [21]. The best and the second best results are highlighted in **Bold** and Underline in each setting, respectively.

		Kalantari					Hu				
		PSNR-l	PSNR-μ	SSIM-l	SSIM-μ	HV2	PSNR-l	PSNR-μ	SSIM-l	SSIM-μ	HV2
S_1	Sen	38.57	40.94	0.9711	0.9780	64.71	33.58	31.48	0.9634	0.9531	66.39
	Hu	30.84	32.19	0.9408	0.9632	62.05	36.94	36.56	0.9877	0.9824	67.58
	FSHDR	40.97	41.11	0.9864	0.9827	67.08	42.15	41.14	0.9904	0.9891	71.35
	Ours (K=0)	**41.12**	**41.20**	**0.9866**	**0.9868**	**67.16**	**42.99**	**41.30**	**0.9912**	**0.9903**	**72.18**
S_2	Ours (K=1)	41.14	41.25	0.9866	0.9869	67.20	45.04	42.55	0.9938	0.9928	73.23
	Ours (K=5)	41.54	41.61	0.9879	0.9880	67.33	47.41	44.24	0.9974	0.9936	74.49
S_3	Kalantari	41.22	41.85	0.9848	0.9872	66.23	43.76	41.60	0.9938	0.9914	72.94
	DeepHDR	40.91	41.64	0.9863	0.9857	67.42	41.20	41.13	0.9941	0.9870	70.82
	AHDRNet	41.23	41.87	0.9868	0.9889	67.50	49.22	45.76	0.9980	0.9956	75.04
	ADNET	41.31	41.80	0.9871	0.9883	67.57	**50.38**	46.79	0.9987	0.9948	**76.32**
	FSHDR	**41.79**	41.92	0.9876	0.9851	67.70	49.56	45.90	0.9984	0.9945	75.25
	Ours	41.68	**41.97**	**0.9889**	**0.9895**	**67.77**	50.31	**46.88**	**0.9988**	**0.9957**	76.21
S_4	Kalantari	25.87	21.44	0.8610	0.9176	60.00	10.23	16.95	0.6903	0.8346	49.10
	DeepHDR	25.92	21.43	0.8597	0.9170	60.02	25.48	20.86	0.9215	0.8354	66.83
	AHDRNet	26.62	22.08	0.8737	0.9238	58.89	11.44	17.84	0.6732	0.8389	52.79
	ADNET	25.76	21.39	0.8686	0.8217	60.36	10.86	18.09	0.6915	0.8399	49.28
	FSHDR	**28.03**	22.01	0.8751	0.9203	60.53	12.82	19.37	0.7442	0.8347	55.34
	Ours	27.91	**22.45**	**0.8764**	**0.9252**	**61.02**	**30.29**	**21.56**	**0.9440**	**0.8456**	**67.07**
S_5	Kalantari	31.24	33.10	0.9527	0.9593	63.99	19.82	18.63	0.7679	0.8742	59.50
	DeepHDR	30.75	29.01	0.9244	0.9223	63.26	19.84	18.70	0.7698	0.8752	59.48
	AHDRNet	31.84	33.49	**0.9588**	0.9606	64.40	**20.80**	20.51	0.8259	0.9136	**59.79**
	ADNET	31.08	33.50	0.9536	0.9636	63.88	20.78	20.80	**0.8268**	0.9173	59.71
	FSHDR	32.70	32.24	0.9553	0.9465	64.37	20.23	19.71	0.7929	0.9026	59.63
	Ours	**32.72**	**34.49**	0.9586	**0.9713**	**64.45**	20.69	**21.96**	0.8257	**0.9207**	59.76

ure 4 (a) and (b), we compare our method with other state-of-the-art methods in the 5-shot scenario. Due to insufficient labeled samples, large motion, and saturation, most comparing methods suffer from color distortion and ghosting artifacts in these two datasets. Kalantari's method and DeepHDR produce undesirable artifacts and color distortion (see Figure 4 (a)(b)). There are two reasons behind that: misalignment of optical flow and homographies and the lack of labeled data. Although AHDRNet and ADNET are proposed to suppress motion and saturation with attention mechanisms, they cannot reconstruct ghost-free HDR images with few labeled samples. They also produce severe ghosting artifacts (see the red block in Figure 4 (a)(b)). FSHDR exploits unlabeled data to alleviate ghosts under the constraint of a few labeled samples, but it is difficult to handle both ghosting and saturation problems simultaneously. We can see that FSHDR still suffers from ghosting artifacts which leaves an obvious hand artifact in the car (see the red block in Figure 4 (a)). Thanks to the pro-

posed SMAE and sample-quality-based iterative learning strategy, which first address the saturation problems using SMAE and then adaptively sample well-exposed and ghost-free pseudo-labels to handle ghosting problems, we can reconstruct ghost-free HDR images with only a few labeled samples.

The quantitative results under the constraint of few shot scenarios on two dataset are shown in Table 1. We report means and 95% margin of variations for 5 and 1 shot cases across 5 runs. Our method achieves state-of-the-art performance on all metrics of two datasets, while most other methods perform poorly with only a few labeled samples. We show that our proposed method surpasses second-best method by 0.15db and 0.21db in terms of PSNR-l and PSNR-μ for 5way-5shot setting on Kalantari's dataset, and it also improves by 0.28db and 0.26db for 5way-5shot setting on Hu's dataset. For 5way-1shot setting, our method consistently outperforms other approaches on two datasets.

In addition, as shown in Table 2, we further compare our

Table 3. Ablation study of 5 shot scenario on Kalantari's dataset.

#	Model	PSNR-l	PSNR-μ	HDR-VDP-2
B1	SSHDR	**41.54**	**41.61**	**67.33**
B2	Stage2Net	41.31	41.43	67.21
B3	w/o APSS	41.49	41.45	67.29
B4	AHDR*	41.48	41.51	67.30
B5	FSHDR*	41.41	41.43	67.26
B6	Vanilla-AHDR	40.61	41.05	66.95
B7	Vanilla-FSHDR	41.39	41.40	67.25

method with major HDR deghosting approaches in zero-shot setting S_1, few-shot setting S_2, and fully supervised setting S_3. Note that we use all the dynamic labeled samples without static and unlabeled samples for plain training in setting S_3. Our zero-shot approach outperforms other methods in zero-shot setting on two datasets. It also outperforms some 5-shot and fully supervised methods in most metrics. Finally, our few-shot and fully supervised approaches achieve state-of-the-art performance among two datasets.

Evaluation Generalization Across Different Datasets.
We compare our method against other approaches on Kalantari's, Hu's, Tursun's, and Prabhakar's datasets to verify generalization performance. We directly evaluate the methods with the checkpoint trained on Kalantari's dataset and show the qualitative results on Tursun's and Prabhakar's datasets in Figure 4 (c)(d). More results are included in the Appendix. In Figure 4 (c), since the lady's motion is large, all the comparison methods cannot remove the ghosting artifacts. In Figure 4 (d), the comparison methods have obvious color distortion and ghosting artifacts on the floor and in the ceil. It shows that other methods have poor generalization performance across different datasets. All these methods address both the saturation and ghosting problems simultaneously. They cannot learn a robust representation to reconstruct a high-quality HDR image. Thanks to our SMAE and sample-quality-based iterative learning strategy, we can learn a robust representation to recover saturated regions and remove ghosting artifacts.

In Table 2, setting S_4 denotes that we utilize the checkpoint trained on Kalantari's or Hu's dataset under 5 shot scenario to evaluate on Hu's or Kalantari's dataset reversely. Setting S_5 represents that we train on Kalantari's or Hu's dataset under 5 shot scenario and evaluate on Prabhakar's dataset. Our method achieves better numerical performance in terms of PSNR-l and PSNR-μ. It demonstrates that our method generalizes well across different datasets.

4.2. Ablation Studies

We conduct ablation studies on Kalantari's dataset under the condition of 5 shot scenario across 5 runs and analyze the importance of each component. We use the following variants of our whole SSHDR model: 1) SSHDR: The full model of SSHDR network trained with two entire stages. 2) Stage2Net: The model only trained in the second
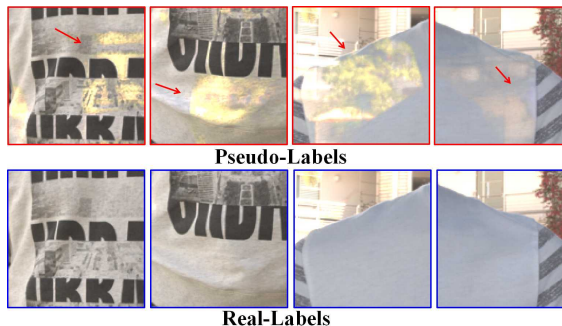

Pseudo-Labels
Real-Labels
Figure 5. Visual results of poor pseudo-labels.

stage without SMAE pre-training. 3) w/o APSS: The model trained with two stages without using sample-quality-based pseudo-labels selection strategy. 4) AHDR*: The AHDR model is trained with our proposed two stages strategy. 5) FSHDR*: Our model is trained with the FSHDR strategy. 6) Vanilla-AHDR: The vanilla AHDR model trained in 5 shot scenario. 7) Vanilla-FSHDR: The vanilla FSHDR model trained with 5 labeled samples.

SMAE Pre-training. As shown in Table 3, the performance of Stage2Net is significantly decreased compared with SSHDR. Since the SMAE learns a robust representation to generate content of saturated regions, it helps to improve the saturated regions. In a word, it demonstrates that the SMAE pre-training stage is an effective mechanism.

Pseudo-labels Selection Strategy. Since the sample-quality-based pseudo-labels selection strategy can exclude saturated and ghosted samples (see Figure 5), the model can be guided in a correct optimization direction which is effective for ghost removal. When we remove the pseudo-labels selection strategy, the performance of the model without APSS is dropped.

Two Stages Strategy. In Table 3, we report the performance of AHDR*. It achieves a significant increment compared with the vanilla AHDR model, which demonstrates the effectiveness of the overall two stages strategy.

Proposed Model Architecture. When we replace our two stages strategy with the FSHDR strategy, the numerical results increase compared with FSHDR. It shows that our proposed model architecture is also sound.

5. Conclusion

We propose a novel semi-supervised deghosting method for few-shot HDR problem via two stages of completing saturation and deghosting. In the first stage, a Saturated Mask AutoEncoder is proposed to learn a robust representation and reconstruct a non-saturated HDR image with a self-supervised mechanism. In the second stage, we propose an adaptive pseudo-label selection strategy to avoid the effects of mislabeled samples. Finally, our approach shows superiority over the existing state-of-the-art methods.

References

[1] Sagie Benaim and Lior Wolf. One-shot unsupervised cross domain translation. *Advances in neural information processing systems (NIPS)*, 31, 2018. 3

[2] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *Advances in neural information processing systems (NIPS)*, 29, 2016. 3

[3] L. Bogoni. Extending dynamic range of mono-chrome and color images through fusion. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 7–12, 2000. 1, 2

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning (ICML)*, pages 1126–1135. PMLR, 2017. 3

[5] Thorsten Grosch. Fast and robust high dynamic range image generation with camera and object movement. In *IEEE Conference of Vision , Modeling and Visualization (VMV)*, 2006. 1, 2

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2

[7] Jinhan Hu, Gyeongmin Choe, Zeeshan Nadir, Osama Nabil, Seok-Jun Lee, Hamid Sheikh, Youngjun Yoo, and Michael Polley. Sensor-realistic synthetic data engine for multi-frame high dynamic range photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 516–517, 2020. 5, 6, 7

[8] Jun Hu, O. Gallo, K. Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1163–1170, 2013. 1, 3, 6

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

[10] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High dynamic range video. *ACM Transactions on Graphics*, 22(3):319–325, 2003. 1, 2

[11] Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE signal processing letters*, 21(9):1045–1049, 2014. 1, 2

[12] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *International Conference on Machine Learning (ICML)*, pages 2927–2936. PMLR, 2018. 3

[13] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1833–1844, 2021. 4

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

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

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

[17] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*, 2016. 3

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

[19] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1219–1232, 2015. 1, 2

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

[21] K Ram Prabhakar, Rajat Arora, Adhitya Swaminathan, Kunal Pratap Singh, and R Venkatesh Babu. A fast, scalable, and reliable deghosting method for extreme exposure fusion. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2019. 6, 7

[22] K Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R Venkatesh Babu, and Rama Krishna Sai S Gorthi. Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4875–4885, 2021. 2, 3, 6

[23] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5822–5830, 2018. 3

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

[25] Pranav Shyam, Shubham Gupta, and Ambedkar Dukkipati. Attentive recurrent comparators. In *International conference on machine learning (ICML)*, pages 3173–3181. PMLR, 2017. 3

[26] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems (NIPS)*, 28, 2015. 3

[27] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. In *International Conference in Central Europe on Computer Graphics and Visualization (WSCG)*, 2007. 1, 2

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

[29] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 0–0, 2019. 2

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

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

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

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

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

[35] Qingsen Yan, Jinqiu Sun, Haisen Li, Yu Zhu, and Yanning Zhang. High dynamic range imaging by sparse representation. *Neurocomputing*, 269:160–169, 2017. 1

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

[37] Qingsen Yan, Yu Zhu, and Yanning Zhang. Robust artifact-free high dynamic range imaging of dynamic scenes. *Multimedia Tools and Applications*, 78:11487–11505, 2019. 1

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

[39] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *Advances in neural information processing systems (NIPS)*, 31, 2018. 3

[40] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2011. 1, 2