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

Qingsen Yan†, Song Zhang†, Weiye Chen†, Hao Tang, Yu Zhu
Jinqiu Sun, Luc Van Gool, Yanning Zhang†
Northwestern Polytechnical University, Xidian University, 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. Unlike 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 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
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. Bogni [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 saturated reconstructed HDR images by CNN. Kalantari 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: database-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 $(X_{1,i}^U, X_{2,i}^U, X_{3,i}^U)$ 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 $(X_{1,i}^S, X_{2,i}^S, X_{3,i}^S)$ 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 $(X_{1,i}^D, X_{2,i}^D, X_{3,i}^D)$ 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,

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

$$H_i = (X_{i}^U)^{\gamma} / t_i.$$  \hspace{1cm} (2)

Note that $X_{1}^U = X_{2}^U$, $t_i$ denotes the exposure time of LDR image $X_i$, $\gamma$ represents the gamma correction parameter, we set $\gamma$ 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 $I\_$. 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
is $8 \times 8$. Considering that the masking strategy is another way to destroy the saturated regions, we intend the model to learn a robust representation to recover these saturated regions. Finally, $I'_i = \{I'_1, I'_2, I'_3\}$ 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 $F'_i$, and map each patch $F'_i$ into query, keys and values. Subsequently, we calculate the similarity map between $q$ and $k$, and perform the Softmax function to get the attention weight. Finally, we apply the attention weight to $v$ to get $F_{is}$.

$$q_i = F'_2 W_q, \quad k_i = F'_1 W_k, \quad v_i = F'_1 W_v, \quad i = 1, 3$$

$$F_{is} = \text{Softmax}(\frac{q_i k_i^T}{\sqrt{d} + b} v_i),$$

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

**MSRSTM.** To merge more information from different exposure regions, inspired by [13], we propose a Multi-Scale Residual Swin Transformer Module (MSRSTM). First, $F_{1s}, F_{2s}, F_{3s}$ is concatenated along the channel dimension to get the input of MSRSTM. Note that $F_{2s}$ 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:

$$F_{STL,i}^N = \text{Conv}((\text{Concat}(STL_{i,j}^{N,1}(F_{out,i}^{N-1})), STL_{i,j}^{N,2}(F_{out,i}^{N-1})), STL_{i,j}^{N,3}(F_{out,i}^{N-1})), (4)$$

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

where $STL_{i,j}^{N,1}(\cdot)$ represents the $N$-th Swin Transformer layer of the $i$-th 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 $\omega$ to transform the predicted HDR image $\hat{Y}$ to short, medium, and long exposure frames by ground truth generation. Then, we regard the new exposure frames as the ground truth $X_{GT}$ of the model.

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

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 $X_{GT}'$ of the model.
\[ X_i^{GT} = \left( \frac{(X_i^T)\gamma + t_i}{t_1} \right)^\frac{1}{\gamma}, \quad i = 1, 2, 3. \]  

Finally, we calculate \( L_1 \) self-supervised loss between \( \hat{Y}_t \) and \( X_i^{GT} \),
\[
L_{SSL} = \sum_{i=1}^{3} ||\hat{Y}_t - X_i^{GT}||_1.
\]  

### 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 \( \mu \)-law to map the linear domain image to the tonemapped domain image,

\[
T(x) = \frac{\log(1 + \mu x)}{\log(1 + \mu)},
\]

where \( T(x) \) is the tonemap function, \( \mu = 500 \). Then we calculate the reconstruction loss \( L_{recon} \) and perceptual loss \( L_{percep} \) between the predicted HDR \( Y_0^D, \hat{Y}^D_0 \) and ground truth HDR \( Y_0^D, Y_0^S \).

\[
L_{recon} = ||T(\hat{Y}_0^D) - T(Y_0^D)||_1 + ||T(\hat{Y}_0^S) - T(Y_0^S)||_1.
\]

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

\[
L_{finetune} = L_{recon} + \lambda L_{percep},
\]

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

**Iteration.** To prevent the overfitting problem with a few labeled training samples and exploit unlabeled samples, we further generate the pseudo-labels \( X_i^U \) 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 unlabeled HDR pseudo-labels \( \hat{Y}_t^U \) at timestep \( t=0 \). Finally, at each timestep in the refinement stage, we calculate the reconstruction loss and perceptual loss as follows,

\[
L_{Iteration} = L_{recon,t+1}^D + L_{recon,t+1}^S + \sum_{i=1}^{N} \frac{W_{t+1} U_i}{U_i} L_{recon,t+1}^i + \lambda (L_{percep,t+1}^D + L_{percep,t+1}^S + \sum_{i=1}^{N} \frac{W_{t+1} U_i}{U_i} L_{percep,t+1}^i),
\]

where \( \lambda = 1e^{-2} \). \( W_{t+1} U_i \) 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 \( Y_t^D \) and \( Y_t^S \). Then we use function \( \omega \) to map the predicted HDR image to medium exposure image \( Y_t^{D,SSL} \) and calculate the loss between \( Y_t^{D,SSL} \) and original medium exposure LDR image \( X_{2,t} \) in well exposure regions to get \( L_{select,t}^{D,SSL} \).

\[
L_{select,t}^{D,SSL} = ||\text{mask}(\omega(\hat{Y}_t^D)) - \text{mask}(\omega(Y_{2,t}^D))||_1 + ||\text{mask}(\omega(\hat{Y}_t^S)) - \text{mask}(\omega(Y_{2,t}^S))||_1,
\]

where \( \text{mask}(\cdot) \) denotes masking the over and under-exposure regions. Subsequently, we sort all patches’ losses, and adopt \( \sigma(\cdot, \cdot) \) function to get \( \beta \) percentile (85th) loss as a selection threshold \( \tau_t \).

\[
\tau_t = \sigma(L_{select,t}^{D,SSL}, \beta).
\]

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 \( \omega \) function to map \( \hat{Y}_t^U \) to medium exposure to get \( Y_t^U \) and calculate the loss between \( Y_t^U \) and original medium exposure LDR image \( X_{2,t} \) to get \( L_{select,t}^{U} \).

If the current loss \( L_{select,t}^{U} \) is greater than \( \tau_t \), 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.

\[
L_{select,t}^{U} = ||\text{mask}(\omega(\hat{Y}_t^U)) - \text{mask}(\omega(X_{2,t}^U))||_1,
\]

\[
m_t^U = \max(L_{select,t}^{U}),
\]

\[
W_{t+1} U_i = \begin{cases} 
1 & \frac{L_{select,t}^{U} - L_{select,t}^{U_i}}{m_t^U - \tau_t} \\
\frac{L_{select,t}^{U} - L_{select,t}^{U_i}}{m_t^U - \tau_t} & \frac{L_{select,t}^{U} > \tau_t} \end{cases}
\]

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 \( U_i \) 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 (\{ -2, 0, +2 \}). There are
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-\( \mu \), SSIM-L, SSIM-\( \mu \), and HDR-VDP-2 [16], where ‘-L’ denotes linear domain, ‘-\( \mu \)’ denotes tonemapping domain.

**Implementation Details.** The window size in MSRSTM is \( 2 \times 2, 4 \times 4 \) and \( 8 \times 8 \). In the training stage, we crop the 128 \( \times 128 \) 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 \( \beta_1=0.9 \), \( \beta_2=0.999 \), and \( \epsilon=1e^{-8} \) 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.**
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 proposed 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 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-\(\mu\) for 5-way-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 second-best methods 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-\(\mu\) for 5-way-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>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Setting</th>
<th>Kalantari</th>
<th>DeepHDR</th>
<th>AHDRNet</th>
<th>ADNet</th>
<th>FSHDR</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalantari</td>
<td>PSNR-(l)</td>
<td>5way-5shot</td>
<td>39.37 (\pm) 0.12</td>
<td>38.25 (\pm) 0.29</td>
<td>40.61 (\pm) 0.10</td>
<td>40.78 (\pm) 0.15</td>
<td>41.39 (\pm) 0.12</td>
<td>41.54 (\pm) 0.10</td>
</tr>
<tr>
<td></td>
<td>PSNR-(\mu)</td>
<td>39.86 (\pm) 0.19</td>
<td>38.62 (\pm) 0.27</td>
<td>41.05 (\pm) 0.32</td>
<td>40.93 (\pm) 0.38</td>
<td>41.40 (\pm) 0.13</td>
<td>41.61 (\pm) 0.08</td>
<td></td>
</tr>
<tr>
<td>Hu</td>
<td>PSNR-(l)</td>
<td>5way-5shot</td>
<td>41.36 (\pm) 0.25</td>
<td>40.73 (\pm) 0.66</td>
<td>46.37 (\pm) 0.76</td>
<td>46.81 (\pm) 0.81</td>
<td>47.13 (\pm) 0.13</td>
<td>47.41 (\pm) 0.12</td>
</tr>
<tr>
<td></td>
<td>PSNR-(\mu)</td>
<td>38.95 (\pm) 0.14</td>
<td>39.92 (\pm) 0.22</td>
<td>43.42 (\pm) 0.44</td>
<td>43.79 (\pm) 0.48</td>
<td>43.99 (\pm) 0.27</td>
<td>44.24 (\pm) 0.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PSNR-(l)</td>
<td>5way-1shot</td>
<td>36.83 (\pm) 0.62</td>
<td>38.49 (\pm) 1.07</td>
<td>42.37 (\pm) 1.42</td>
<td>42.41 (\pm) 1.20</td>
<td>42.50 (\pm) 0.87</td>
<td>42.55 (\pm) 0.44</td>
</tr>
</tbody>
</table>

| Kalantari | PSNR-\(l\) | 5way-5shot | 38.17 | 40.94 | 0.9711 | 0.9780 | 64.71 | 33.35 | 31.48 | 0.9634 | 0.9531 | 66.39 |
| | PSNR-\(\mu\) | 40.51 | 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 | 0.9893 | 67.16 | 42.99 | 41.30 | 0.9912 | 0.9903 | 72.18 |
| | 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 |
| Hu | PSNR-\(l\) | 5way-5shot | 38.67 \(\pm\) 0.43 | 37.82 \(\pm\) 0.86 | 44.04 \(\pm\) 0.80 | 44.75 \(\pm\) 0.84 | 44.99 \(\pm\) 0.23 | 45.04 \(\pm\) 0.16 |
| | PSNR-\(\mu\) | 36.83 \(\pm\) 0.62 | 38.49 \(\pm\) 1.07 | 42.37 \(\pm\) 1.42 | 42.41 \(\pm\) 1.20 | 42.50 \(\pm\) 0.87 | 42.55 \(\pm\) 0.44 |
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 ceiling. 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-$l$. 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 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 Auto-Encoder 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


