Labeled from Unlabeled: Exploiting Unlabeled Data for Few-shot Deep HDR Deghosting

K Ram Prabhakar[1] Gowtham Senthil[1,2*] Susmit Agrawal[1*] R. Venkatesh Babu[1]
Rama Krishna Sai S Gorthi[2]
[1]Indian Institute of Science, Bangalore [2]Indian Institute of Technology Tirupati

Abstract

High Dynamic Range (HDR) deghosting is an indispensable tool in capturing wide dynamic range scenes without ghosting artifacts. Recently, convolutional neural networks (CNNs) have shown tremendous success in HDR deghosting. However, CNN-based HDR deghosting methods require collecting large datasets with ground truth, which is a tedious and time-consuming process. This paper proposes a pioneering work by introducing zero and few-shot learning strategies for data-efficient HDR deghosting. Our approach consists of two stages of training. In stage one, we train the model with few labeled (5 or less) dynamic samples and a pool of unlabeled samples with a self-supervised loss. We use the trained model to predict HDRs for the unlabeled samples. To derive data for the next stage of training, we propose a novel method for generating corresponding dynamic inputs from the predicted HDRs of unlabeled data. The generated artificial dynamic inputs and predicted HDRs are used as paired labeled data. In stage two, we finetune the model with the original few labeled data and artificially generated labeled data. Our few-shot approach outperforms many fully-supervised methods in two publicly available datasets, using as little as five labeled dynamic samples.

1. Introduction

Unlike the human eye, a standard digital camera has limited dynamic range that it can recognize in a scene. All objects beyond the recognizable dynamic range are thresholded to the minimum or maximum pixel intensity value, thus losing their details in the process. High Dynamic Range imaging is an algorithmic solution to this problem. It creates an image with a wider dynamic range than a standard camera image, closer to what human eyes perceive. The generated HDR image contains details in both bright and dark regions.

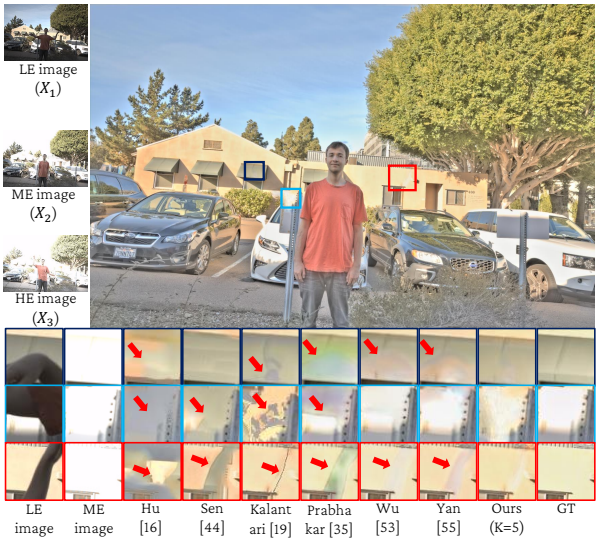

Figure 1. Qualitative results by different methods for an example from Kalantari *et al.* dataset [19]. As shown in the zoomed in boxes, our proposed few-shot approach using only 5 labeled dynamic sequences and pool of unlabeled sequences, generate better results without any artifacts, than existing methods trained with full dataset of 74 labeled dynamic sequences.

To generate an HDR image, multiple images with different exposure values (also known as *exposure stack* or LDR images) are captured and merged. The merging process is simple when the exposure stack's input images are static without any camera or object motion. However, such an assumption is too good to be true in real-world scenarios. Most often, the exposure stack is dynamic, consisting of camera and object motion. Fusing such dynamic exposure stack naively results in undesirable ghosting artifacts. The process of fusing dynamic exposure stack without ghosting artifacts is known as *HDR deghosting*.

A widely followed approach for HDR deghosting is to register the LDRs first and identify moving regions. Once identified, either the motion affected regions are excluded

* equal contribution

in those images, or a chosen reference image is used [10, 13, 17, 22, 33, 39, 59]. Such methods result in only LDR content for moving regions. Another popular approach is to align images using estimated dense correspondence between a reference image and input LDRs, and merging the aligned images. Dense registration techniques like optical flow methods introduce warping artifacts in heavily saturated and occluded regions [3, 11, 21, 48, 51, 60]. Patch-based optimization methods [16, 44] synthesize static sequences from dynamic input sequences and merge them to generate the final HDR image. These methods are computationally expensive and hallucinate false details in heavily saturated regions (see Fig. 1).

Recently proposed deep learning-based methods offer a significant advantage over traditional methods in terms of deghosting quality and computation time [19, 34, 35, 53, 56]. CNNs can learn complex fusion rules using abundant training data with ground truth HDRs. However, collecting a large amount of labeled training data for HDR deghosting is challenging, due to the reasons listed below.

Difficulty in capturing labeled data: Capturing a single labeled sample requires considerable effort. First, a dynamic exposure stack with controlled object motion is captured. Then, a static exposure stack of the same scene, without object motion, is captured with a tripod to generate the ground truth [19]. In this process, it must be ensured that there are no other unexpected or unwanted motions in the scene, such as tree, cloud, or vehicle motions. These constraints are applicable only for very few scene types and human-controllable motion, thus limiting dataset diversity.

Post-capture manual examination: Another major difficulty in collecting large-scale supervised HDR deghosting datasets is the post-capture manual examination. All samples must be carefully examined for any unwanted motion in the static exposure stack. If any sample has such discrepancies present, then that sample cannot be used to generate artifact-free ground truth and hence has to be discarded. This often leads to almost a quarter of the captured samples getting discarded. Kalantari *et al.* [19] discarded 25 samples from the captured 114, and Prabhakar *et al.* [35] had to discard 118 from the captured 700 samples, after manually scrutinizing every single one.

Furthermore, existing datasets are limited in the diversity of camera parameters used to capture them, such as ISO and exposure levels. However, collecting new large datasets with ground truth for different settings is cumbersome and highly inconvenient. Due to the above reasons, HDR deghosting CNNs are limited by the diversity, scale, and training dataset settings. The above limitations offer all the more reason to explore data-efficient Deep HDR Deghosting methods.

We address these limitations by proposing zero and few-shot learning strategies for HDR deghosting while using a pool of unlabeled dynamic exposure stacks. Many diverse unlabeled samples can be effortlessly captured without worrying about collecting ground truth for the same. It does not require a tripod since we do not have to capture a corresponding static exposure stack. Also, it does not require post-capture scrutiny, as it is not constrained and can thus include any diverse motion or scene.

Our approach consists of two stages of training. In the first stage, we train a model with a supervised loss for few labeled dynamic samples and use a self-supervised loss for the unlabeled samples. Then, we use the trained model to predict HDRs for the unlabeled samples and call them as predicted HDRs. Since the HDR images predicted by the model will inherently contain errors, they cannot be treated as proper ground truth for the unlabeled samples. Therefore, we generate artificial dynamic input images that correspond to the predicted HDR images and use them along with few labeled dynamic images to improve the model in second stage. In summary, the main contributions of our paper are as follows:

- To the best of our knowledge, this is the first work to explore zero-shot and few-shot learning with unlabeled images for Deep HDR Deghosting.
- We propose a novel method to generate labeled dynamic training data from unlabeled dynamic data.
- Our method trained with only 5 labeled dynamic samples and unlabeled samples achieves comparable, if not better, results than existing methods trained on complete datasets in a supervised fashion, on two publicly available datasets.
- We perform comprehensive experiments and ablation studies to demonstrate the significance of various components of our proposed approach.

2. Related works

Over the past two decades, many different methods have been proposed for HDR deghosting. The methods proposed in the literature can be classified into four major categories.

The first category of methods assumes that only a few pixels were affected by the motion, and majority are static [6, 15, 22, 25, 29, 31, 36, 38, 52]. Static pixels are merged using standard static exposure fusion rules. Whereas the moving pixels are merged either by using only static images in those regions or using a chosen reference image. Grosch [13] method uses brightness consistency criteria in the pixel domain to identify moving pixels. Gallo *et al.* [10] improved upon [13] by comparing patches instead of pixels in the log domain. The method by Reinhard *et al.* [41] threshold the weighted irradiance map variance to locate moving pixels.

The second category of methods align the input images to a chosen reference image using alignment techniques [3, 21, 60]. The aligned static images are merged using stan-

dard HDR fusion methods. Methods such as Ward [51] and Tomaszewska *et al.* [48] use rigid alignment techniques to compensate for global camera motion. More advanced methods use non-rigid alignment techniques to estimate dense correspondence. The method by Bogoni *et al.* [3] uses optical flow computed in a multi-scale fashion to align input images. Jinno and Okuda [18] use Markov Random Field to predict dense correspondence map for aligning. Gallo *et al.* [12] proposed a fast method by computing flow only at sparse locations. They then interpolate it to generate the final dense flow. This category of methods cannot handle complex motion and cannot synthesize new details in occluded regions.

The third category of methods synthesizes static exposure stack from the input dynamic exposure stack [16, 37, 44]. The synthesized static stack will be structurally similar to a chosen reference image but have contents borrowed from other images in saturated regions of the reference image. Hu *et al.* [16] use the patch-match algorithm to generate the aligned sequences. Sen *et al.* [44] use a multi-image bidirectional similarity metric for the same. These methods introduce artifacts for images with extreme exposure profiles, and they are computationally exorbitant.

The last category of methods uses CNNs to generate final HDR images. Kalantari *et al.* [19] align the inputs with optical flow and fuses the aligned images with a CNN. Wu *et al.* [53] train a CNN to fuse unaligned input images directly. Prabhakar *et al.* [35] proposed a scalable CNN feature aggregation architecture that can fuse an arbitrary number of input images. Yan *et al.* [55] use an attention mechanism to select useful features from other non-reference input images. In [57], Yan *et al.* propose to use a non-local module to find a correlation between input feature maps. Similarly, Yan *et al.* [56] use multi-scale CNN architecture to generate accurate results. Recently, Prabhakar *et al.* [34] proposed a computationally efficient method that uses bilateral guided upsampler to generate high-resolution output. Single image HDR reconstruction methods such as Endo *et al.* [8] and Eilertsen *et al.* [7] train a CNN to reconstruct saturated details in a single image.

Few-shot learning (FSL): FSL is a rapidly growing research area focused on learning generalizable representations for new classes/tasks with only few supervised samples. FSL approaches use different methods such as meta-learners [9, 40, 42, 46], distance-based classifiers [45, 49], and embedding learning [2, 47]. However, FSL for HDR deghosting has not been explored before.

Self-Supervised Learning (SSL): SSL investigates the use of unlabeled data to learn better representations [1, 24, 30, 32, 58]. One class of SSL methods generate pseudo labels for input images and use them to train on augmented versions of the same input [4, 14, 26]. This strategy is best suited for classification tasks, where the pseudo (or predicted) label remains the same for both original and augmented input. However, for image reconstruction tasks such as HDR deghosting, a crude pseudo-label cannot be used as a ground truth for augmented versions of the input. In fact, this penalizes the learning model in regions where predicted HDR contains even minor artifacts (refer Sec. 5). To overcome this mismatch, we propose a novel method to synthesize a single dynamic input sequence that matches the predicted HDR output. Since the synthesized dynamic input is an exact match for predicted HDR output, they can be used as a supervised training pair.

3. Proposed method

Motivation: Existing CNN-based HDR deghosting methods use only labeled dynamic sequences for training. However, ground truth for diverse dynamic sequences is challenging to acquire. On the contrary, unlabeled dynamic sequences are much easier to capture, as they do not necessitate staticity or tripods to acquire ground truth. Thus, they can consist of diverse scenes with real-world representative motion, without any constraints. In this work, we exploit the hitherto dismissed potential of unlabeled data in enabling few-shot HDR deghosting and achieve performance similar to existing methods trained in a fully supervised setting.

Data distribution: In our approach, we make use of three different types of sets: 1) Labeled ($\mathbf{L}$): K dynamic samples with ground truth HDR, 2) Static ($\mathbf{S}$): Q static samples, 3) Unlabeled ($\mathbf{U}$): M dynamic samples without ground truth. K is assumed to be less than or equal to 5; we keep it minimal as $\mathbf{L}$ sequences are difficult to obtain. Q is fixed at 5. Although easier to acquire, this set does not help to learn HDR deghosting and only guides the initial HDR merging process. For $\mathbf{U}$, we use a pool of unlabeled data available at hand. M is not fixed and is arbitrarily large, as it is easiest to capture and represents real-world content.

Overview: The novelty of our work lies in effectively utilizing the M unlabeled dynamic ($\mathbf{U}$) sequences through two stages of training (see Fig. 2a). In stage one, we train a CNN model $\mathcal{N}$ with $\mathbf{L}$ and $\mathbf{S}$ using a supervised ℓ_2-loss, and $\mathbf{U}$ using a weak self-supervised loss. Then, we use the trained $\mathcal{N}$ model to predict fused HDRs ($\widehat{Y}^P$) for all unlabeled dynamic sequences ($\mathbf{U}$). We note that $\widehat{Y}^P$ is imperfect and possesses minor discoloration and ghosting artifacts, and thus cannot be used as ground truth for $\mathbf{U}$. To resolve this mismatch, we propose a novel method to generate artificial dynamic inputs ($\mathbf{P}$) from $\widehat{Y}^P$. Since $\widehat{Y}^P$ is used to generate $\mathbf{P}$, it is used as ground truth for $\mathbf{P}$. In stage two, we train a new instance of $\mathcal{N}$ using both $\mathbf{L}$ and $\mathbf{P}$ with ℓ_2-loss.

Stage 1 Training: Set $\mathbf{L}$ consists of K dynamic sequences: $\{(X^L, \widehat{Y}^L)_k\}, k = (1, \cdots, K)$. Each X^L sequence has three varying exposure input LDRs: (X_1^L, X_2^L, X_3^L) and the corresponding ground truth HDR,

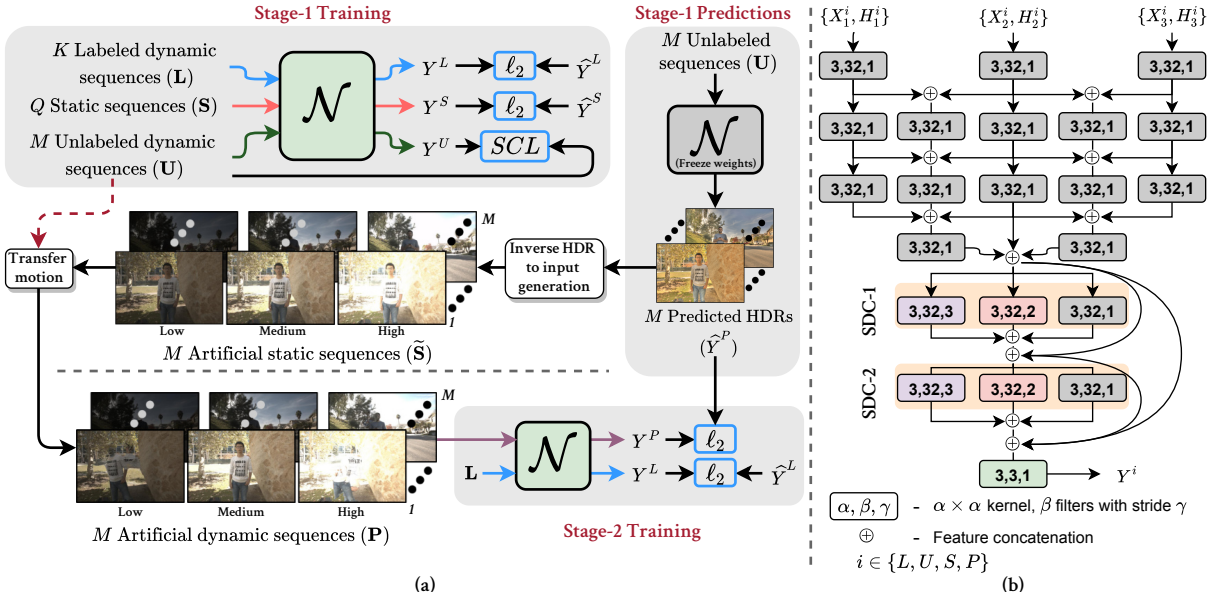

Figure 2. (a) Overview of the proposed method. (b) $\mathcal{N}$ network architecture.

$\widehat{Y}^L$. In our experiments, we vary K between 1 and 5. Following the literature [19, 53, 55], in each X^L sequence we choose the middle image X_2^L as reference. The network's prediction will resemble the reference image in non-saturated regions of the reference image. In saturated regions, details are to be extracted from other images without causing artifacts. The input images are assumed to be in the linear domain; else, they are linearized with CRF [5].

Other images in the sequence (X_1^L, X_3^L) are aligned to the reference image by estimating dense optical flow [27]. Each aligned LDR image is concatenated with their HDR equivalents and fed as input to the model $\mathcal{N}$, which consists of a series of encoder convolution layers, as shown in Fig. 2b. The HDR equivalent of any LDR image I is obtained by $H = I^{2.2}/t_I$, where t_I is the exposure time of I. All encoders' output feature maps are concatenated and fed as input to two Stacked and Dilated Convolution (SDC) blocks [43]. The SDC block outputs are concatenated and passed to a single convolution layer to generate the final predicted HDR, Y^L. We use LeakyRelu [54] activation in all layers except in the last layer, where we use sigmoid activation to predict output within 0 and 1 range.

Additionally, we utilize a small number ($Q = 5$) of static sequences $\mathbf{S} = \{(X^S)_q\}, q = (1, \cdots, Q)$, to guide the network to learn static HDR fusion. Each static sequence X^S, has three varying exposure LDRs: (X_1^S, X_2^S, X_3^S). Static images help the network to stabilize during training. Since static images do not have any motion, the ground truth HDR ($\widehat{Y}^S$) can be obtained by merging the input images using a standard HDR merging technique [5]. In addition to $\mathbf{L}$ and $\mathbf{S}$, we make use of a large pool of unlabeled dynamic sequences ($\mathbf{U}$) during stage one training of $\mathcal{N}$. $\mathbf{U}$ consists of M dynamic sequences $\{(X^U)_m\}, m = (1, \cdots, M)$. Each sequence has three varying exposure images (X_1^U, X_2^U, X_3^U), but no ground truth HDR. The steps for input pre-processing, HDR equivalent concatenation, and model predictions are the same for $\mathbf{L}$, $\mathbf{S}$, and $\mathbf{U}$,

$$Y^i = \mathcal{N}(X^i, H^i), \ \forall i = \{L, U, S\}. \tag{1}$$

For data from $\mathbf{L}$ and $\mathbf{S}$, standard ℓ_2 error between prediction and ground truth is used as loss function (Eq. (2)). As the HDR images are generally displayed after tonemapping, the loss is also computed after tonemapping with μ-law tonemapper. The μ-law tonemapping of any image I is computed by, $T(I) = {}^{log(1+\mu \times I)}/_{log(1+\mu)}$, where μ=5000.

To compute error for $\mathbf{U}$, we use a self-supervised Structural Consistency Loss ($\mathcal{L}_{SCL}$). As the network is trained to predict Y^U with structure similar to the reference image (X_2^U), $\mathcal{L}_{SCL}$ computes ℓ_2 loss between X_2^U and Y^U at the exposure level of X_2^U (Eq. (3)). $\mathcal{L}_{SCL}$ offers ideal supervision in regions where reference image is well-exposed and contains details. In regions where the reference is saturated, it offers only a weak supervision as the predicted HDR loses details in those regions after exposure adjustment and clipping. For example, a model trained with only $\mathcal{L}_{SCL}$ and $\mathbf{U}$, offers highly saturated HDR images (see Fig. 4).

A weighted sum of the supervised ($\mathcal{L}_{sup}$) and self-supervised ($\mathcal{L}_{SCL}$) loss is used to train $\mathcal{N}$,

$$\mathcal{L}_{sup} = \ell_2(T(Y^L), T(\widehat{Y}^L)) + \ell_2(T(Y^S), T(\widehat{Y}^S)), \tag{2}$$

$$\mathcal{L}_{SCL} = \ell_2(clip((Y^U \times t_{X_2^U})^{1/2.2}), X_2^U), \tag{3}$$

$$\mathcal{L}_{S1} = \mathcal{L}_{sup} + \alpha \times \mathcal{L}_{SCL}. \tag{4}$$

Stage one training is carried out for 75 epochs with 5000 batches per epoch using Adam optimizer [23], starting with

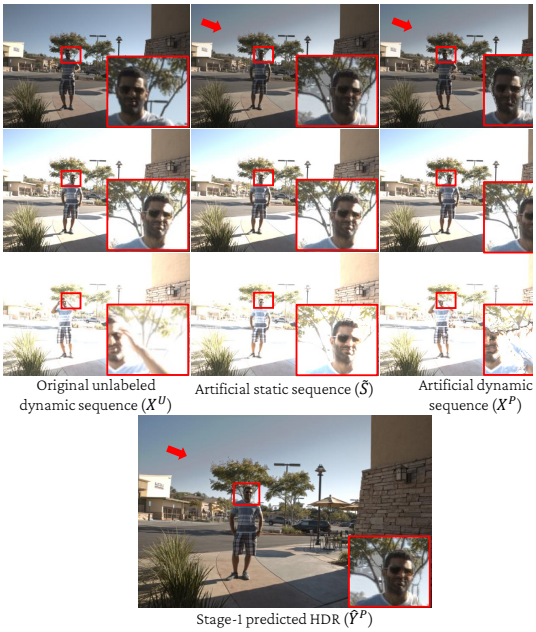

Figure 3. Artificial static ($\widetilde{S}$) and dynamic sequence (X^P) generated from an original unlabeled dynamic sample from X^U. Note the discoloration of sky and artifacts in the predicted HDR ($\widehat{Y}^P$) and its reproduction in artificial sequences.

a learning rate of $1e{-}4$. Each batch consists of 4 random 64×64 patches cropped from an random image in either $\mathbf{L}$ or $\mathbf{S}$, and 4 random patches from $\mathbf{U}$. The learning rate is decayed by 0.75 every 10 epochs. The $\mathcal{L}_{SCL}$ weight factor α starts at 0.5 and is incremented by 0.1 every 10 epochs.

Labeled Artificial Input Synthesis: After stage one training, we use $\mathcal{N}$ to predict M HDR images ($\widehat{Y}^P$) for the M unlabeled dynamic sequences ($\mathbf{U}$), $\widehat{Y}^P = \mathcal{N}(X^U, H^U)$. However, as $\mathcal{N}$ was trained using only limited supervised samples, $\widehat{Y}^P$ contains ghosting artifacts, discoloration and loss of detail in overexposed regions. Thus, although $\widehat{Y}^P$ is generated from $\mathbf{U}$, it is not an ideal or true ground truth and can potentially mislead the network if used as a supervised label for $\mathbf{U}$ in stage two (see Fig. 3).

To overcome this mismatch, our proposed approach inverts the HDR merging pipeline to generate artificial dynamic input sequences ($\mathbf{P}$) that form a proper pair with the respective predicted HDRs ($\widehat{Y}^P$) used to derive them. First, we generate artificial static sequences with different exposure levels from the M predicted HDRs ($\widehat{Y}^P$), using the exposure time of corresponding unlabeled dynamic sequence images ($t_{X_j^U}$). This is followed by gamma correction and clipping of saturated pixels to generate an artificial static input sequence $(\widetilde{S}_1, \widetilde{S}_2, \widetilde{S}_3)_m$,

$$\widetilde{S}_j = clip\left((\widehat{Y}^P \times t_{X_j^U})^{1/2.2}\right), \forall j = (1,2,3). \quad (5)$$

Motion transfer: These static sequences cannot be directly used as input training sequences as they lack dynamic motion and cannot guide HDR deghosting. To induce motion, we make use of the dense backward optical flow F between the original unlabeled dynamic sequence images ($\mathbf{U}$) (Eq. (6)). That is, the flow from X_2^U to X_1^U is calculated as $F_{1,2}$ and used to warp $\widetilde{S}_1$ to obtain X_1^P. Similarly, the flow from X_2^U to X_3^U is calculated as $F_{3,2}$ and used to warp $\widetilde{S}_3$ to obtain X_3^P. The medium exposed reference static image is considered as it is, without any motion, to remain structurally consistent with the predicted ground truth ($\widehat{Y}^P$).

$$X_j^P = \begin{cases} \mathcal{W}(\widetilde{S}_j, F_{j,2}), & j = (1,3), \\ \widetilde{S}_j, & j = 2, \end{cases} \quad (6)$$

where $\mathcal{W}$ denotes the warping function. We perform this operation for all M artificial static sequences to generate M artificial dynamic sequences.

The generated M artificial dynamic sequences, $X^P = (X_1^P, X_2^P, X_3^P)$, agrees with the M predicted HDRs ($\widehat{Y}^P$), and forms a proper labeled training set, $\mathbf{P} = \{(X^P, \widehat{Y}^P)_m\}, m = (1, \cdots, M)$. The motion transferred from real-world motion in $\mathbf{U}$, ensures that input dynamic motion stays meaningful and diverse. The artifacts present in $\widehat{Y}^P$ restrict it from functioning as a ground truth for $\mathbf{U}$. However, all those artifacts present in $\widehat{Y}^P$ becomes part of $\mathbf{P}$ and hence $\widehat{Y}^P$ is a clean ground truth for $\mathbf{P}$.

Stage 2 training: In this stage, we train a new instance of $\mathcal{N}$ with the K labeled dynamic samples from $\mathbf{L}$, along with M artificially generated labeled samples from $\mathbf{P}$ in a supervised fashion. Since both $\mathbf{L}$ and $\mathbf{P}$ have ground truths, $\mathcal{N}$ can be trained using the standard ℓ_2 loss function. We do not use $\mathbf{S}$ in stage two training, as it does not help learn HDR deghosting due to lack of motion. We also ignore $\mathbf{U}$, as the weak self-supervision ($\mathcal{L}_{SCL}$) does not support convergence in the presence of $\mathbf{P}$'s supervision. As in stage 1, the loss is computed between the tonemapped variants of predicted and ground truth images.

$$Y^i = \mathcal{N}(X^i, H^i), \forall i = \{L, P\}. \quad (7)$$

$$\mathcal{L}_{S2} = \ell_2(T(Y^L), T(\widehat{Y}^L)) + \ell_2(T(Y^P), T(\widehat{Y}^P)). \quad (8)$$

In stage two, the model is trained for 75 epochs with 5000 batches per epoch. Each batch consists of 4 random 64×64 patches cropped from a randomly picked image in $\mathbf{P}$ and $\mathbf{L}$. Adam optimizer [23] is used with an initial learning rate of $1e{-}4$, which is decayed by a factor of 0.75 after every 10 epochs.

Stage 2 refinement (S2R): The stage two trained models can predict better quality HDRs for $\mathbf{U}$ than the stage one models. Thus, we use second stage predicted HDRs to generate a new set of artificial dynamic inputs and refine the model. We show in Fig. 5 (d, e) that there is only a minor improvement in PSNRs. The scores reported in Tables 1

Table 1. Quantitative comparison against five existing CNN-based methods under constrained few-shot scenario. Refer Section 4 for details. The best score is highlighted in **bold**.

	Kalantari *et al.* [19]				Prabhakar *et al.* [35]			
	1-shot 5-way		5-shot 5-way		1-shot 5-way		5-shot 5-way	
	P_L	P_μ	P_L	P_μ	P_L	P_μ	P_L	P_μ
Kalantari *et al.* [19]	39.32 ±0.45	37.40 ±1.28	40.08 ±0.10	39.96 ±0.16	33.63 ±0.44	34.78 ±0.72	34.32 ±0.22	36.19 ±0.18
Wu *et al.* [53]	37.03 ±0.71	36.44 ±1.63	38.62 ±0.27	38.03 ±0.29	31.19 ±0.98	33.09 ±1.17	32.66 ±0.65	35.28 ±0.20
Prabhakar *et al.* [35]	37.88 ±0.12	35.54 ±0.44	38.24 ±0.06	36.14 ±0.36	31.26 ±0.83	31.52 ±1.47	31.85 ±0.47	33.65 ±0.43
Yan *et al.* [55]	37.81 ±0.38	36.96 ±1.06	39.37 ±0.16	38.86 ±0.34	31.73 ±0.81	33.27 ±1.24	33.14 ±0.67	34.95 ±0.48
Prabhakar *et al.* [34]	39.82 ±0.41	36.92 ±1.09	40.54 ±0.12	38.66 ±0.34	32.34 ±0.99	33.89 ±1.18	32.95 ±0.97	35.45 ±0.29
Ours	**41.04** ±0.11	**41.13** ±0.07	**41.39** ±0.12	**41.40** ±0.11	**35.74** ±0.13	**36.47** ±0.16	**35.86** ±0.12	**36.61** ±0.10

and 2 correspond to stage 2 refined models, and inference for any test sequence is done only with this model.

4. Experiments and Results

We perform extensive experiments to show our proposed few-shot approach's effectiveness compared to existing methods. Since there are no other few-shot HDR deghosting approaches, we compare our approach against five existing CNN-based methods [19, 34, 35, 53, 55], under the constrained few-shot scenario. For comparison, we make use of Kalantari *et al.* [19] dataset with 74 training and 15 validation sequences, and Prabhakar *et al.* [35] dataset with 466 training and 116 validation sequences. We report results in Table 1 with $K \in \{1, 5\}$ and $Q = 5$ settings. For each value of K, we report the average and 95% margin of variation across 5 runs (denoted as 5-way in Table 1). In each run, we randomly choose K different random labeled dynamic sequences and 5 different static sequences. The rest of the dataset sequences are used as unlabeled data (**U**) without ground truth. It should be noted that all three sets are maintained disjoint.

We train all five existing methods and our multi-stage approach with the same set of images for a fair and uniform comparison. We train the existing methods using **L** and **S** sequences but not **U**, as the existing methods are fully supervised. We report average PSNR across all validation sequences in the linear (P_L) and μ-law tonemapped domains (P_μ). It is clear that all existing methods perform poorly when trained with only a few labeled samples. However, our proposed multi-stage training approach outperforms all of them across both metrics by a significant margin with only 1 labeled dynamic sample.

Table 2. Quantitative comparison against nine methods from three different categories. Refer Section 4 for details. The best score is highlighted in red and the second best in blue.

		Kalantari *et al.* [19]					Prabhakar *et al.* [35]				
		P_L	P_μ	S_L	S_μ	HV2	P_L	P_μ	S_L	S_μ	HV2
C1	Sen [44]	38.57	40.94	0.971	0.978	64.74	32.93	33.43	0.972	0.964	65.47
	Hu [16]	31.25	35.75	0.941	0.963	62.07	29.47	32.58	0.954	0.949	63.50
	Endo [8]	8.846	21.33	0.622	0.715	54.00	9.760	8.890	0.641	0.675	55.76
	Eilertsen [7]	14.21	14.13	0.350	0.882	57.95	14.19	15.66	0.442	0.869	58.74
	Ours (K=0)	40.97	41.11	0.989	0.987	67.08	35.06	36.25	0.976	0.946	67.42
C2	Ours (K=1)	41.04	41.13	0.988	0.987	67.19	35.74	36.47	0.978	0.947	67.53
	Ours (K=5)	41.39	41.40	0.990	0.989	67.25	35.86	36.61	0.979	0.948	67.56
C3	Kalantari [19]	41.27	42.74	0.981	0.987	66.10	32.50	35.63	0.969	0.961	65.40
	Wu [53]	40.91	41.65	0.986	0.986	67.44	34.40	38.03	0.977	0.971	66.59
	Prabhakar [35]	39.68	40.47	0.980	0.975	66.50	32.74	36.08	0.967	0.959	66.10
	Yan [55]	41.08	41.21	0.989	0.989	67.53	35.28	38.65	0.963	0.961	66.88
	Prabhakar [34]	41.33	42.82	0.986	0.989	67.15	34.98	38.30	0.978	0.970	66.25
	Ours	41.79	41.92	0.990	0.990	67.70	35.57	38.63	0.980	0.974	67.60

In Table 2, we exhaustively compare our approach against all major HDR deghosting techniques. We use PSNR and SSIM [50] in linear and μ-law tonemapped domains, and HDR-VDP-2 (HV2) [28] as comparison metrics. We show comparison on three categories depending upon the amount of labeled dynamic data used: C1 - zero-shot, C2 - few-shot, and C3 - fully supervised. C1 consists of two non-learning HDR deghosting methods [16, 44], two single-image HDR reconstruction techniques [7, 8]. It also contains our approach trained in a zero-shot setting with $K=0$, averaged across 5 different **S** sets while the rest of dataset is considered as **U** without using the ground truth. C2 contains our proposed few-shot approach with $K \in \{1, 5\}$, each averaged across 5 different $\{\mathbf{L}, \mathbf{S}, \mathbf{U}\}$ sets. C3 contains 5 state-of-the-art CNN-based approaches [19, 34, 35, 53, 55] trained in fully-supervised setting with 100% of labeled dataset ($K = 74$ for [19] and $K = 466$ for [35]). We also include our approach trained in a fully-supervised setting for one stage, which is using the entire dataset as **L**, without **S** and **U**.

Our zero-shot approach outperforms other non-learning and single-image HDR approaches by a large margin. It outperforms some fully supervised methods like Wu *et al.* [53] and Prabhakar *et al.* [35] in terms of P_L and SSIM. Also, it outperforms Kalantari *et al.* [19] and [35] in HDR-VDP-2 metric. Our few-shot approach can match the performance of many state-of-the-art fully-supervised techniques while using less than or equal to 5 labeled dynamic data. For instance, our few-shot approach with K=5 outperforms all other fully supervised methods in P_L and S_L metrics, and ties in first position with [55] and [34] for S_μ. Comparatively, our fully supervised method outperforms all existing methods in 3 out of 5 metrics among both datasets.

In Table 3, we compare the effectiveness of each component of $\{\mathbf{L}, \mathbf{S}, \mathbf{U}\}$ in guiding convergence during stage

Figure 4. Qualitative results of stage one training using different combinations of $\{\mathbf{L}, \mathbf{S}, \mathbf{U}\}$.

one training. We perform this ablation on [19]'s dataset for 3 sets of $\{\mathbf{L}, \mathbf{S}, \mathbf{U}\}$. Using only $\mathbf{S}$, the model learns only static-HDR merging. Thus, resultant images have ghosting artifacts, which is evident by the low P_μ value (see Fig. 4). Using both $\mathbf{S}$ and $\mathbf{U}$ is essentially the stage one training of zero-shot approach, which performs much better than using only weakly self-supervised $\mathbf{U}$. This shows that although $\mathbf{S}$ does not help HDR deghosting, it guides the HDR merging process in case of low dynamic labeled data.

For the remaining runs, we average 3 different sets, for both $K \in \{1, 5\}$, and show that using both $\mathbf{S}$ and $\mathbf{U}$ helps improve performance over direct few-shot training by a large margin. Without $\mathbf{S}$, the HDR merging process struggles in low-data regime. Without $\mathcal{L}_{SCL}$ and $\mathbf{U}$, there are only 1 or 5 dynamic sequences in the training set, which leads to poor deghosting performance. In Fig. 4, we see that using all 3 sets $\{\mathbf{L}, \mathbf{S}, \mathbf{U}\}$ achieves best deghosting performance with minimal artifacts.

Qualitative Results: In Figure 1, we present results produced by major HDR deghosting methods on a test sample from the Kalantari *et al.* [19] dataset. Even though the CNN-based methods were trained using 100% of labeled dataset, our 5-shot model produces comparable, if not better results in challenging regions. In Fig. 6, we present results of CNN-based HDR deghosting methods under the constrained few-shot scenario, on 2 test samples from Kalantari *et al.* [19]'s dataset. Trained with only 5 labeled dynamic samples, existing methods produce lot of ghosting and discoloration artifacts. However, our approach is capable of reconstructing details in heavily saturated dynamic regions.

5. Discussion

Number of Samples: We ran multiple ablations to determine the number of samples required from each set ($\mathbf{S}$, $\mathbf{L}$, $\mathbf{U}$). We perform all the ablation experiments and report results on Kalantari *et al.* [19] dataset. Firstly, we fixed $\mathbf{U}$ at $M = 20$, $\mathbf{L}$ at $K = 1$, and varied the number of static samples ($\mathbf{S}$) from $Q \in \{0, 5, 10, 20, 40, 74\}$ in stage one training. As seen in Fig. 5 (a), just 5 static samples are enough to guide the HDR merging process. Using many static samples does not offer any noticeable improvement in metrics. So we use $Q = 5$ static samples for all our experiments.

Table 3. Ablation analysis to determine importance of each set during stage one training. The numbers are reported on Kalantari *et al.* [19]'s test set averaged over three different runs.

S	L	U	P_L	P_μ	HV2
✓	✗	✗	38.33	34.88	61.27
✗	✗	✓	33.51	37.37	61.79
✓	✗	✓	39.84	39.92	65.68

			1-shot			5-shot		
S	L	U	P_L	P_μ	HV2	P_L	P_μ	HV2
✗	✓	✗	36.40	32.57	64.70	40.56	40.14	66.36
✓	✓	✗	39.54	38.61	65.08	40.43	39.40	65.42
✗	✓	✓	35.85	38.26	64.35	39.90	40.72	66.30
✓	✓	✓	**40.15**	**40.63**	**65.89**	**40.64**	**41.08**	**66.39**

Secondly, we fixed $\mathbf{S}$ at $Q = 20$, $\mathbf{L}$ at $K = 1$, and varied the number of unlabeled dynamic sequences ($\mathbf{U}$) from $M \in \{0, 5, 10, 20, 40, 74\}$ in stage one training. Although we usually use all unlabeled samples available at hand, Fig. 5 (b) suggests that a handful of $M = 20$ unlabeled dynamic sequences is enough to achieve reported stage one results.

Finally we vary $|\mathbf{L}|$ from $K \in \{1, 2, 3, 4, 5, 7, 18, 37\}$, while all remaining samples are used as $\mathbf{U}$. We kept $\mathbf{S}$ fixed at $Q = 74$, whose quantity, as we have already established, does not affect performance. For each K value, we average across 3 different sets of $\mathbf{L}$, whose results are shown in Fig. 5 (c). We also show direct few-shot training scores, where the model is trained only on $\mathbf{L}$. With just $K = 1$, our stage one training achieves over 4 to 8dB improvement in PSNRs and almost matches fully supervised performance.

Stage 2 Ablations: Conventional pseudo-labeling [4, 14, 26] self-supervision suggests that the predicted HDRs ($\hat{Y}^P$) can be used as a label for $\mathbf{U}$ during stage two. However, a model trained using only $\{\mathbf{U}, \mathbf{L}\}$ with $\{\hat{Y}^P, \hat{Y}^L\}$ as labels gave P_L of only 40.65dB compared to our proposed approach improving up to 41.30dB in stage two. This indicates that naively pseudo-labeling is flawed for HDR deghosting. The model gets incorrectly penalized for the artifacts present in predicted HDRs that are not a part of the input sequence. In contrast, our artificial sequences stay true to the predicted HDRs and form valid training pairs. In addition, including the $\mathcal{L}_{SCL}$ self-supervision loss in stage two, gave even worse P_L of 40.45dB. This is due to the conflict between $\mathcal{L}_{SCL}$ and $\mathcal{L}_2$ losses in saturated regions which misleads network convergence.

We also tried including $\mathbf{S}$ during stage two training, but it decreased P_μ by over 0.4dB. In stage two, when the model is fine-tuning its deghosting performance, including static sequences proves detrimental. Finally, while we align the input sequences in $\mathbf{L}$ to their respective reference images using optical flow [27], we do not perform the same for the generated dynamic sequences ($\mathbf{P}$). We found that doing

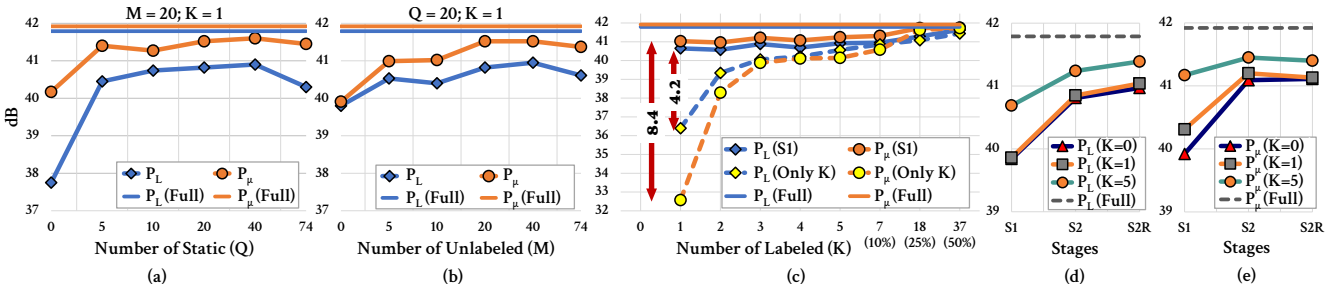

Figure 5. Ablation on number of samples and stages. Please refer to Section 5 for details.

Figure 6. Qualitative comparison of existing CNN-based methods against proposed 5-shot approach on two challenging examples from Kalantari *et al.* [19]'s validation set. The results for existing methods are generated after training them with 5 labeled dynamic examples (refer Section 4).

so resulted in aggravated warping artifacts, which rendered learning difficult and lead to a drop of 0.3dB P_L. However, for real dynamic images from **L** and **U**, we always perform optical flow alignment before merging.

Few Shot HDR Video: We manually annotate HDR frames for the first 4 frames of two LDR videos provided by [20], and use them as labeled sequences. The remaining unlabeled frames are used as unlabeled dynamic sequences. This way, we extend our few-shot two-stage approach to HDR video and generate 2 HDR videos. The resultant videos and training details are included in supplementary.

6. Conclusion

In this paper, we propose a novel few-shot HDR deghosting method using unlabeled data through self-supervision. In the first stage, we train a model with limited dynamic labeled data and boost performance using unlabeled data with a self-supervision loss. In stage two, we found that performance can be further improved with the help of our proposed novel approach to generate labeled data from unlabeled samples. Such an approach brings up a promising paradigm shift and can be extended to many other image enhancement and photography applications. We have shown that as low as 5 labeled dynamic samples and a pool of unlabeled samples is sufficient to achieve deghosting performance comparable to a model trained with fully supervised data. Finally, our work eliminates the necessity to capture ground-truth for all sequences, and utilizes few labeled dynamic data and unlabeled data to achieve similar, if not better, results. Our few-shot approach saves much tedious effort, time, and manual scrutiny in collecting accurate ground truth for large-scale HDR deghosting dataset.

Acknowledgements: This work was supported by a project grant from MeitY (No.4(16)/2019-ITEA), Govt. of India and WIRIN.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 3

[2] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Advances in neural information processing systems*, pages 523–531, 2016. 3

[3] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proceedings. 15th International Conference on Pattern Recognition.*, 2000. 2, 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3, 7

[5] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, page 31. ACM, 2008. 4

[6] Ashley Eden, Matthew Uyttendaele, and Richard Szeliski. Seamless image stitching of scenes with large motions and exposure differences. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2006. 2

[7] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017. 3, 6

[8] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177–1, 2017. 3, 6

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017. 3

[10] Orazio Gallo, Natasha Gelfandz, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. In *ICCP*, pages 1–7. IEEE, 2009. 2

[11] Orazio Gallo, Alejandro Troccoli, Jun Hu, Kari Pulli, and Jan Kautz. Locally non-rigid registration for mobile HDR photography. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015. 2

[12] Orazio Gallo, Alejandro Troccoli, Jun Hu, Kari Pulli, and Jan Kautz. Locally non-rigid registration for mobile HDR photography. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 49–56, 2015. 3

[13] Thorsten Grosch. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling and Visualization, RWTH Aachen*, pages 277–284, 2006. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3, 7

[15] Yong Seok Heo, Kyoung Mu Lee, Sang Uk Lee, Youngsu Moon, and Joonhyuk Cha. Ghost-free high dynamic range imaging. In *Asian Conference on Computer Vision*, pages 486–500. Springer, 2010. 2

[16] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2, 3, 6

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

[18] Takao Jinno and Masahiro Okuda. Motion blur free hdr image acquisition using multiple exposures. In *2008 15th IEEE International Conference on Image Processing*, pages 1304–1307. IEEE, 2008. 3

[19] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, 36(4), 2017. 1, 2, 3, 4, 6, 7, 8

[20] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Trans. Graph.*, 32(6):202–1, 2013. 8

[21] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 319–325. ACM, 2003. 2

[22] Erum Arif Khan, AO Akyiiz, and Erik Reinhard. Ghost removal in high dynamic range images. In *IEEE International Conference on Image Processing*, 2006. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5

[24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[25] 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. 2

[26] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013. 3, 7

[27] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 4, 7

[28] 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. *ACM Transactions on graphics (TOG)*, 30(4):40, 2011. 6

[29] Tae-Hong Min, Rae-Hong Park, and SoonKeun Chang. Histogram based ghost removal in high dynamic range images. In *2009 IEEE International Conference on Multimedia and Expo*, pages 530–533. IEEE, 2009. 2

[30] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 3

[31] 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, 2014. 2

[32] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[33] Fabrizio Pece and Jan Kautz. Bitmap movement detection: HDR for dynamic scenes. In *2010 Conference on Visual Media Production*, pages 1–8. IEEE, 2010. 2

[34] K Ram Prabhakar, Susmit Agrawal, Durgesh Singh, Balraj Ashwath, and R Venkatesh Babu. Towards practical and efficient high-resolution HDR deghosting with CNN. In *2020 European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6

[35] 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. 2, 3, 6

[36] K Ram Prabhakar and R Venkatesh Babu. Ghosting free hdr for dynamic scenes via shift-maps. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8, 2016. 2

[37] K Ram Prabhakar and R Venkatesh Babu. Ghosting-free multi-exposure image fusion in gradient domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016. 3

[38] Shanmuganathan Raman and Subhasis Chaudhuri. Bilateral filter based compositing for variable exposure photography. In *Eurographics (short papers)*, pages 1–4, 2009. 2

[39] Shanmuganathan Raman and Subhasis Chaudhuri. Reconstruction of high contrast images for dynamic scenes. *The Visual Computer*, 27(12):1099–1114, 2011. 2

[40] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 3

[41] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. 2

[42] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016. 3

[43] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2556–2565, 2019. 4

[44] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203, 2012. 2, 3, 6

[45] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 3

[46] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? *arXiv preprint arXiv:1910.03560*, 2019. 3

[47] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 3

[48] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multiexposure high dynamic range image acquisition. In *Proceedings of the International Conference on Computer Graphics, Visualization and Computer Vision*, 2007. 2, 3

[49] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 3

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[51] Greg Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of Graphics Tools*, 8(2):17–30, 2003. 2, 3

[52] Shiqian Wu, Shoulie Xie, Susanto Rahardja, and Zhengguo Li. A robust and fast anti-ghosting algorithm for high dynamic range imaging. In *2010 IEEE International Conference on Image Processing*, pages 397–400. IEEE, 2010. 2

[53] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *European Conference on Computer Vision*, pages 120–135, 2018. 2, 3, 4, 6

[54] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 4

[55] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 3, 4, 6

[56] Qingsen Yan, Dong Gong, Pingping Zhang, Qinfeng Shi, Jinqiu Sun, Ian Reid, and Yanning Zhang. Multi-scale dense networks for deep high dynamic range imaging. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 41–50. IEEE, 2019. 2, 3

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

[58] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 3

[59] Wei Zhang and Wai-Kuen Cham. Reference-guided exposure fusion in dynamic scenes. *Journal of Visual Communication and Image Representation*, 23(3):467–475, 2012. 2

[60] Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, volume 30, pages 405–414. Wiley Online Library, 2011. 2