FlexHDR: Modeling Alignment and Exposure Uncertainties for Flexible HDR Imaging

Sibi Catley-Chandar, Thomas Tanay, Lucas Vandroux, Aleš Leonardis, Gregory Slabaugh, *Senior Member, IEEE*, and Eduardo Pérez-Pellitero

***Abstract*— High dynamic range (HDR) imaging is of fundamental importance in modern digital photography pipelines and used to produce a high-quality photograph with well exposed regions despite varying illumination across the image. This is typically achieved by merging multiple low dynamic range (LDR) images taken at different exposures. However, over-exposed regions and misalignment errors due to poorly compensated motion result in artefacts such as ghosting. In this paper, we present a new HDR imaging technique that specifically models alignment and exposure uncertainties to produce high quality HDR results. We introduce a strategy that learns to jointly align and assess the alignment and exposure reliability using an HDR-aware, uncertainty-driven attention map that robustly merges the frames into a single high quality HDR image. Further, we introduce a progressive, multi-stage image fusion approach that can flexibly merge any number of LDR images in a permutation-invariant manner. Experimental results show our method can produce better quality HDR images with up to 1.1dB PSNR improvement to the state-of-the-art, and subjective improvements in terms of better detail, colours, and fewer artefacts.**

***Index Terms*— High dynamic range imaging, set processing, permutation invariance.**

I. Introduction

DESPITE recent advances in imaging technology, capturing scenes with wide dynamic range still poses several challenges. Current camera sensors suffer from limited or Low Dynamic Range (LDR) due to inherent hardware limitations. The maximum dynamic range a camera can capture is closely related to (a) the sensor's photosite *full well electron capacity* or saturation point, and (b) the black point, which is generally constrained by the uncertainty in the reading due to the dominant presence of noise.

Different solutions have been proposed to overcome these limitations. The principle behind most of them relies on capturing observations of the same scene with different exposure values. This enables a richer coverage of the scene's original dynamic range, but also requires a mechanism to align and unify the different captured observations [1]. Some approaches make use of multi-sensor or multi-camera configurations, e.g. Tocci *et al.* [2], McGuire *et al.* [3], Froehlich *et al.* [4], where a beam splitter enables the light to be captured by multiple sensors. However, such setups are normally expensive, fragile, with bulky and cumbersome rigs, and they may suffer from double contours, light flares, or polarization artefacts [4].

More pragmatic solutions include only a single sensor and obtain multiple exposures by either spatial (i.e. per-pixel varying exposure) [5], [6] or temporal multiplexing (i.e. capturing differently exposed frames) [1]. This simpler hardware setup (and related algorithms) has recently seen widespread adoption, and is now found in cameras ranging from professional DSLR to low-cost smartphones.

Early multi-frame exposure fusion algorithms work remarkably well for almost-static scenes (e.g. tripod, reduced motion) but result in ghosting and other motion-related artefacts for dynamic scenes. Various approaches have achieved success in reducing artefacts such as patch-based methods [7], [8], noise-based reconstruction [9], sparse correspondences [10] and image synthesis [11] but in recent years, Convolutional Neural Networks (CNNs) have greatly advanced the state-of-the-art for HDR reconstruction, especially for complex dynamic scenes [12].

Most HDR CNNs rely on a rigid setup with a fixed, ordered set of LDR input images, which assumes the medium exposure to be the reference image. The most common mechanism for the merging step is image or feature concatenation, and thus for methods where the feature encoder is not shared among input frames [13], there is a dependency between reference frame choice, relative exposure and input image ordering. Optimal exposure parameters [14] or fast object motion might constrain the amount of relevant frames available, and in general, broader flexibility in terms of number of frames and choice of reference is necessary to extend applicability without the burden of model tweaking or retraining.

As for frame registration, previous models largely rely on pre-trained or classical *off-the-shelf* optical flow methods that are rarely designed or optimized for the characteristics of exposure-bracketed LDR images. Recent pixel rejection or attention strategies are disconnected from the alignment stage and mostly ignore uncertainty in exposure or motion.

Manuscript received 22 December 2021; revised 20 May 2022 and 20 July 2022; accepted 16 August 2022. Date of publication 8 September 2022; date of current version 16 September 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Loic Denis. *(Corresponding author: Sibi Catley-Chandar.)*

Sibi Catley-Chandar is with Huawei Noah's Ark Lab, London N1C 4AG, U.K., and also with the School of Electronic Engineering and Computer Science (EECS), Queen Mary University of London, London E1 4NS, U.K. (e-mail: sibi.catley.chandar@huawei.com).

Thomas Tanay, Lucas Vandroux, Aleš Leonardis, and Eduardo Pérez-Pellitero are with Huawei Noah's Ark Lab, London N1C 4AG, U.K.

Gregory Slabaugh is with the School of Electronic Engineering and Computer Science (EECS), Queen Mary University of London, London E1 4NS, U.K.

Digital Object Identifier 10.1109/TIP.2022.3203562

In this paper, we propose a novel algorithm that addresses these limitations in a holistic and unified way. First, we design a HDR-specific optical flow network which can predict accurate optical flow estimates even when the input and target frames are under- or over-exposed. We do this by using symmetric pooling operations to share information between all n input frames, so any missing information in one frame can be borrowed from other frames. Further, we propose models of exposure and alignment uncertainties which are used by our flow and attention networks to regulate contributions from unreliable and misaligned pixels. Finally we propose a flexible architecture that can process any number of input frames provided in any order.

The contributions of this paper are threefold:

1) A lightweight HDR-specific optical flow network which can estimate accurate pixel correspondences between LDR frames, even when improperly exposed, by sharing information between all input frames with symmetric pooling operations, and is trained using an HDR-aware self-supervised loss that incorporates exposure uncertainty.
2) Models of exposure and alignment uncertainty which we use to regulate contributions from unreliable and misaligned pixels and greatly reduce ghosting artefacts.
3) A flexible architecture with a multi-stage fusion mechanism which can estimate an HDR image from an arbitrary set of LDR input images.

II. Related Work

In this section we review the HDR literature with a focus on relevant deep-learning multi-frame exposure fusion methods. For a broader overview we refer the reader to [15], [16].

The seminal work of [12] was the first to introduce a training and testing dataset with dynamic scene content. Their proposed method for learning-based HDR fusion is composed of two stages: first, input LDR images are aligned using a classical optical flow algorithm [17] and then a CNN is trained to both merge images and potentially correct any errors in the alignment. Shortly after, [13] proposed a similar approach that does not perform a dense optical flow estimation, but rather uses an image-wide homography to perform *background* alignment, leaving the more complex non-rigid *foreground* motions to be handled by the CNN. However, this method is highly dependent on the structure of the reference image, and the magnitude and complexity of the motion. Thus, if certain regions are saturated in the reference image, it fails to accurately reconstruct them in the final result. Both [12] and [13] rely on the optimisation of the HDR reconstruction loss to implicitly learn how to correct ghosting and handle the information coming from different frames. However, neither provides an explicit mechanism to prevent incorrect information (e.g. overexposed regions) from influencing the final HDR estimation. Despite the noteworthy performance improvement over existing methods at the time, these approaches still suffer from ghosting, especially for fast moving objects and saturated or near-saturated regions.

Yan *et al.* [18] address some limitations of its predecessors by establishing an attention mechanism to suppress undesired information before the merging stage, e.g. misalignments, overexposed regions, and focus instead on desirable details of non-reference frames that might be missing in the reference frame. In the work of Prabhakar *et al.* [19] parts of the computation, including the optical flow estimation, are performed in a lower resolution and later upscaled back to full resolution using a guide image generated with a simple weight map, thus saving some computation. [20] propose the first end-to-end deeep learning based video HDR algorithm which drastically improved inference speeds compared to classical methods.

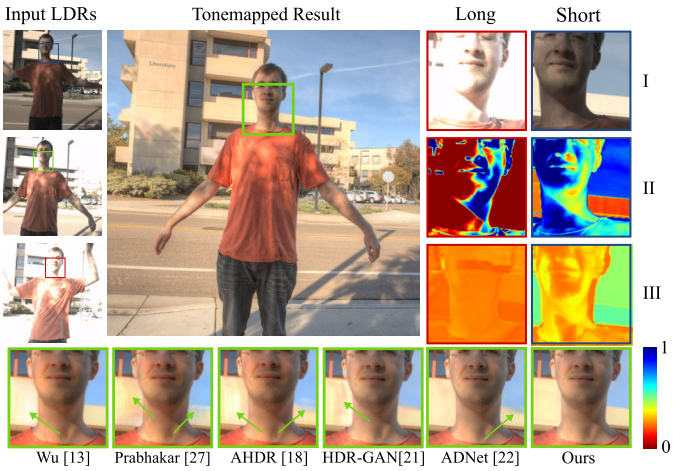

Fig. 1. Our method, intermediate results, and comparisons. LDR input images are shown on the left and our tone mapped result is in the centre. The visualisations on the right are: **I** Input Image, **II** Exposure Map, **III** Attention Map. The bottom row compares to several state-of-the-art methods. Our uncertainty modelling more effectively handles regions of overexposure and motion between input frames.

More recently, the state of the art in HDR imaging has been pushed to new highs. [21] propose the first GAN-based approach to HDR reconstruction which is able to synthesize missing details in areas with disocclusions. Liu *et al.* [22] introduce a method which uses deformable convolutions as an alignment mechanism instead of optical flow and was the winning submission to the 2021 NTIRE HDR Challenge [23]. Contemporary work has explored and pioneered new training paradigms, such as the weakly supervised training strategy proposed by [24].

Extending these methods to an arbitrary number of images requires changes to the model definition and re-training. Set-processing neural networks [25] can naturally handle those requirements. In [26], a permutation invariant CNN is used to deblur a burst of frames which present only rigid, 0-mean translations with no explicit motion registration. For the HDR task, [27] proposed a method that uses symmetric pooling aggregation to fuse any number of images, but requires pre-alignment [28] and artefact correction by networks which only work on image pairs.

III. Proposed Method

Given a set of n LDR images with different exposure values $\{I_1, I_2, \ldots, I_n\}$ our aim is to reconstruct a single HDR image H which is aligned to a reference frame I_r. To simplify

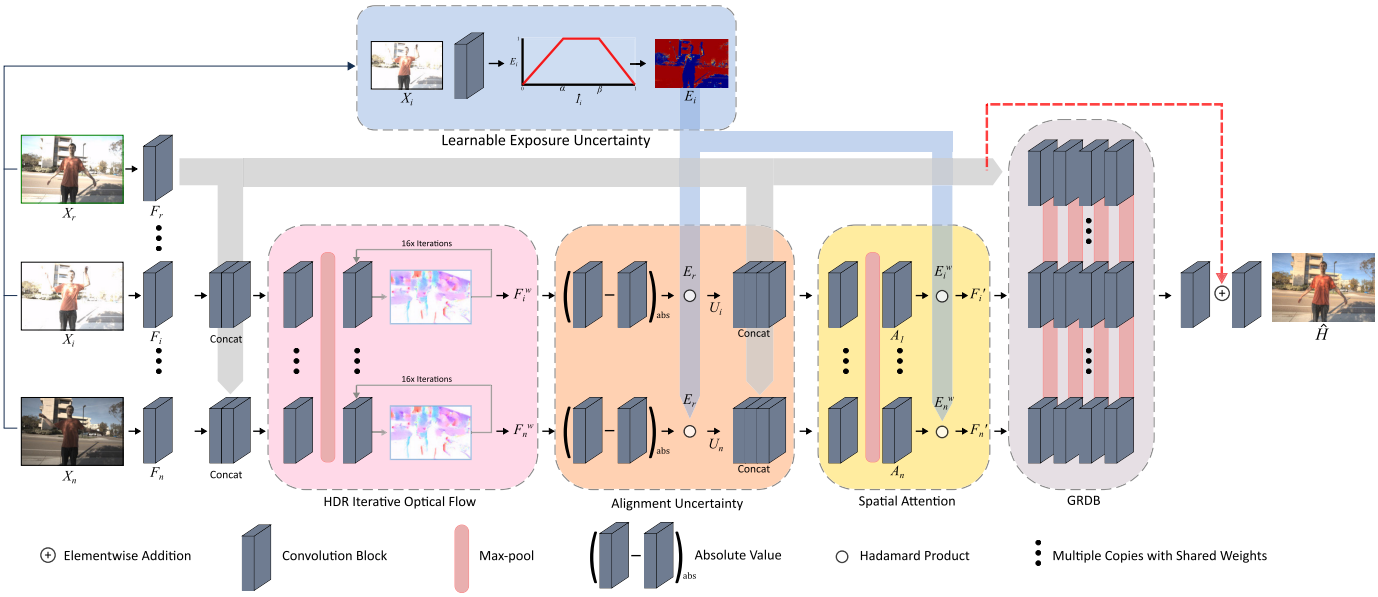

Fig. 2. Model architecture. Our model accepts any number of LDR images as input and aligns them with a *HDR flow network* which shares information between frames with pooling operations. We then model exposure and alignment uncertainties which are used by our *attention network* to suppress untrustworthy regions. Finally, the *merging network* consists of a grouped residual dense block with multi-stage max-pooling operations for gradual merging of input frames.

notation, we denote $I_r = I_1$, but any input frame can be chosen as the reference frame. To generate the inputs to our model, we follow the work of [12], [13], [18] and form a linearized image L_i for each I_i as follows:

$$L_i = I_i^{\gamma}/t_i, \tag{1}$$

where t_i is the exposure time of image I_i with power-law non-linearity γ. Setting $\gamma = 2.2$ inverts the CRF, while dividing by the exposure time adjusts all the images to have consistent brightness. We concatenate I_i and L_i in the channel dimension to form a 6 channel input image $X_i = [I_i, L_i]$. Given a set of n inputs $\{X_1, X_2, \ldots, X_n\}$ our proposed network estimates the HDR image $\hat{H}$ by:

$$\hat{H} = h(\{X_i\} : \theta), \tag{2}$$

where $h(\cdot)$ denotes our network, θ the learned weights of the network and $\hat{H}$ is the predicted radiance map in the linear domain. Our network accepts any number of frames n and is invariant to the order of the non-reference inputs. This is different from the work of [12], [13], [18] where the value of n is fixed to 3 and the order of inputs is fixed, and the work of [27] where only the fusion stage is performed on n inputs, but frame alignment and attention are performed on image pairs only. Our method performs alignment, regulates the contribution of each frame based on related alignment and exposure uncertainties and flexibly fuses any number of input frames in a permutation-invariant manner. Our network is also trained end-to-end and θ is learned entirely during the HDR training.

A. Architecture Overview

Our architecture is composed of: *Learnable Exposure Uncertainty* (Sec. III-C), *HDR Iterative Optical Flow* (Sec. III-D), *Alignment Uncertainty and Attention* (Sec. III-E), and *Merging Network* (Sec.III-F). An overview of the architecture can be seen in Figure 2. Our architecture makes use of max-pooling operations to share information between frames and to fuse frames together (Sec. III-B). This improves the accuracy of our flow and attention networks and gives us the advantage of an architecture that is flexible enough to accept an arbitrary number of images. The flow network and the attention network work together to align non-reference frames to the reference frame and suppress artefacts from misaligned and over-exposed regions. The merging network then combines the aligned features to predict a single HDR image. By explicitly modelling the two most common sources of error, motion and exposure, we create a network that is aware of uncertainty and is able to greatly reduce artefacts compared to state-of-the-art methods, as shown in Figure 1.

B. Flexible Set Processing

Many state-of-the-art CNN HDR reconstruction methods require a fixed number of inputs in fixed order of exposure [12], [13], [18]. To overcome this limitation, we design a set-processing network that can naturally deal with any number of input images. Related concepts have previously shown strong benefits for problems such as deblurring [26] and we here propose to leverage set-processing and permutation invariance tools for HDR fusion.

Given n input images, our network uses n identical copies of itself with shared weights to process each image separately in its own *stream*. We use a multi-stage fusion mechanism, where features F_i^k of each individual stream i at an arbitrary point k within the network can share information with each other as follows:

$$F_i^{\max} = \text{conv}(\left[F_i^k, \max(F_1^k, \ldots, F_n^k)\right]), \tag{3}$$

where max(·) denotes a max-pooling operation, [·] denotes concatenation and conv(·) denotes a convolutional layer (see Fig. 5). This operation is repeated at multiple points in the network. Finally, the outputs of each stream are then pooled together into a single stream with a global max-pooling operation $F^{\max}_{\text{global}} = \max(F^k_1, \ldots, F^k_n)$. This result is processed further in the final layers of our network to obtain the HDR prediction. This allows the network to process any number of frames in a permutation invariant manner while still being informed by the other frames.

C. Modelling Exposure Uncertainty

A key limitation of LDR images is that any pixel values above the sensor saturation point results in information loss. Values approaching the saturation level are also unreliable due to negative post-saturation noise [14]. When reconstructing an HDR image from multiple LDR images, values at or close to the saturation point can produce artefacts in the final output. Furthermore, underexposed values close to zero are also unreliable due to dark current and low signal-to-noise ratio. We seek to regulate the contribution of such values by modelling our confidence in a given pixel value being correct. For a given input image I_i, we propose the following piecewise linear function where α and β are predicted by the network for each image:

$$E_i = \begin{cases} \frac{1}{\alpha}\hat{I}_i & \hat{I}_i < \alpha \\ 1 & \alpha \leqslant \hat{I}_i \leqslant \beta \\ \frac{1}{(1-\beta)}(1-\hat{I}_i) & \beta < \hat{I}_i \end{cases}. \tag{4}$$

Here $\hat{I}_i$ denotes the mean value across the three RGB channels and E_i is the predicted exposure map which represents our estimated confidence in a given pixel. This function is plotted in Figure 4. We learn from data how to predict α and β by means of a shallow network, i.e. a convolution acting on the concatenation of $[X_i, \hat{I}_i]$ followed by a spatial average pooling. We constrain α and β such that $0 < \alpha < 0.5 < \beta < 1$. As $\hat{I}_i$ approaches 0 or 1, the pixel becomes increasingly unreliable and the value of the exposure mask approaches zero. The slope with which E_i approaches zero is determined by α and β. As shown in Figure 1 this allows us to regulate the contribution that improperly exposed regions in an image can have on our result.

D. HDR Specific Efficient Iterative Optical Flow Network

Recent learning based optical flow methods [29] typically do not work well for HDR. Successive frames can have large amounts of missing information due to overexposure, which makes aligning frames difficult. This is especially true if the reference and non-reference frames are both overexposed. We solve this issue by using max-pooling operations to share information between all n input frames in our flow network's encoder, as described in Eq. (3). This lets the network *fill in* missing information from any of the n available input frames and predict more accurate flows.

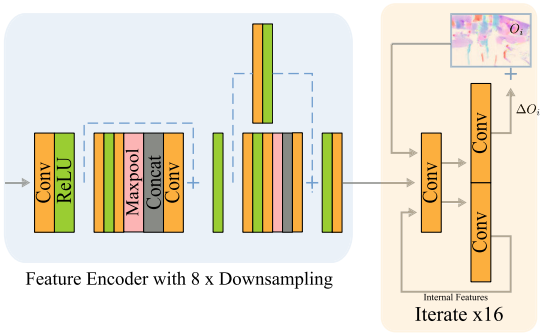

Fig. 3. Architecture of our HDR Iterative Optical Flow Network. The feature encoder first downsamples the input features by 8x before the recurrent convolutions iteratively refine the estimated optical flow field.

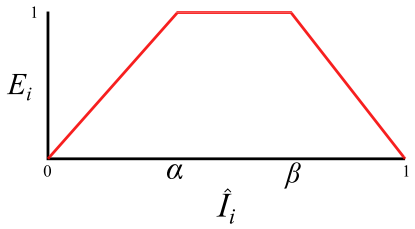

Fig. 4. We model exposure uncertainty as a piecewise linear function where α and β are predicted by the network. Given the mean pixel values of an image, $\hat{I}_i$, our model predicts an image specific response of exposure confidence, E_i.

The architecture of our proposed flow network is inspired by RAFT [29], however we design the network to be lightweight and efficient. We do not use a context encoder, a correlation layer or a convolutional gated recurrent unit, instead using only simple convolutional layers to predict our optical flow field.

Given an input X_i and an exposure mask E_i, we use a convolutional layer to extract features F_i from X_i. The inputs into the flow network are then concatenated as follows: $[F_i, F_r, E_i]$, where F_r corresponds to the features extracted from the reference image. The flow network is informed by E_i so that our predictions are aware of the exposure uncertainty in the image. As recurrent convolutions can be computationally expensive at full resolution, the flow network first downsamples the input features by 8× using strided convolutions. It then iteratively refines the predicted flow over 16 iterations, with a flow initialized to zero, to obtain the optical flow field O_i via:

$$O_i = f([F_i, F_r, E_i]), \tag{5}$$

where $f(\cdot)$ denotes our optical flow network. The optical flow field is resized to the original resolution with bilinear upsampling and used to warp our features F_i:

$$F^w_i = w(F_i, O_i), \tag{6}$$

where F^w_i are the warped features and $w(\cdot)$ denotes the function of warping an image with an optical flow field. The architecture of our flow network can be seen in Figure 3.

Unlike other methods which use fixed alignment [13], [27], our flow network is trained in a fully self-supervised manner. As ground truth optical flows for our datasets are unavailable, we use the self-supervised photometric loss between the reference features F_r and the warped features F^w_i as supervision to

guide the learning of the flow network. We multiply the loss by E_r so that the reference frame is only used as supervision in regions where it is well exposed. We also apply the optical flow field to the exposure mask, so it remains spatially aligned with the warped features:

$$E_i^w = w(E_i, O_i), \tag{7}$$

where E_i^w is the warped exposure mask.

E. Alignment Uncertainty and Attention

Our attention network is informed by two measures of uncertainty: exposure and alignment. To model the alignment uncertainty we compute an uncertainty map U_i as:

$$U_i = \texttt{abs}(F_i^w - F_r) \circ E_r, \tag{8}$$

where $\texttt{abs}(\cdot)$ denotes the elementwise absolute value and $\circ$ denotes element-wise multiplication. This map captures the difference in motion between the reference frame and the warped frame and helps inform our attention network of any inconsistencies in the alignment. We multiply by E_r so that only the well exposed regions of the reference frame are used to calculate misalignments. By regulating the contributions from misaligned areas of the image, our network can significantly reduce ghosting in the final output. The exposure uncertainty is given by the warped exposure map E_i^w. The inputs to the attention network are then concatenated as follows $[F_i^w, F_r, U_i, E_i^w]$. The attention network predicts a 64 channel attention map A_i as follows:

$$A_i = a([F_i^w, F_r, U_i, E_i^w]), \tag{9}$$

where $a(\cdot)$ denotes our attention network. As in our flow network, we use max-pooling to share information between all n input frames. We then obtain our regulated features by multiplying the warped features F_i^w by the attention map and the exposure map:

$$F_i' = F_i^w \circ A_i \circ E_i^w, \tag{10}$$

where $\circ$ denotes element-wise multiplication and F_i' denotes the regulated features. Multiplication by the exposure map enforces a strict constraint on our network and prevents unreliable information leaking into our output. Our HDR-aware attention effectively regulates the contribution of each frame, taking into account both alignment and exposure uncertainty.

F. Merging Network

Our merging network takes the regulated features obtained from Equation 10 and merges them into a single HDR image. The merging network is based on a Grouped Residual Dense Block (GRDB) [30], which consists of three Residual Dense Blocks (RDBs) [31]. We modify the GRDB so that each stream can share information with the other streams for a multi-stage fusion of features. An overview of the fusion mechanism can be seen in Figure 5. Specifically we add a max-pooling operation after each RDB which follow the formulation described in Equation 3. This allows the network to progressively merge features from different streams, instead of merging them together in a single concatenation step where information might be lost. This is followed by a final global max-pooling operation which collapses the n streams into one. The merging network then processes this result further with a global residual connection and refinement convolutions.

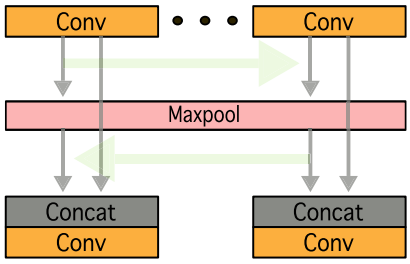

Fig. 5. An overview of our multi-stage fusion mechanism. The green arrows show the direction of information flow between the different streams. By sharing information between streams at multiple points, the network is able to produce clear, detailed images.

G. Loss Function

As HDR images are not viewed in the linear domain, we follow previous work and use the μ-law to map from the linear HDR image to the tonemapped image:

$$\mathcal{T}(H) = \frac{log(1+\mu H)}{log(1+\mu)}, \tag{11}$$

where H is the linear HDR image, $\mathcal{T}(H)$ is the tonemapped image and $\mu = 5000$. We then estimate the ℓ_1-norm between the prediction and the ground truth to construct a tone mapped loss as follows:

$$\mathcal{L}_{tm} = \left\| \mathcal{T}(\hat{H}) - \mathcal{T}(H) \right\|_1. \tag{12}$$

To improve the quality of reconstructed textures we also use the perceptual loss as in [32]. We pass the tonemapped images through a pre-trained VGG-19 [33] and extract features from three intermediate layers. We reduce the ℓ_1-norm between the features of the ground truth and our prediction:

$$\mathcal{L}_{vgg} = \left\| \phi(\mathcal{T}(\hat{H})) - \phi(\mathcal{T}(H)) \right\|_1, \tag{13}$$

where ϕ is a pre-trained VGG-19 network. Finally, to provide supervision for our optical flow network, we calculate a simple photometric loss between the warped features F_i^w and the reference features F_r and multiply by E_r to limit supervision to well exposed regions in the reference frame:

$$\mathcal{L}_{phot} = \left\| (F_i^w - F_r) \circ E_r \right\|_1, \tag{14}$$

where $\texttt{abs}$ is the elementwise absolute value. Our total loss function can be expressed as:

$$\mathcal{L}_{tot} = \mathcal{L}_{tm} + \mathcal{L}_{phot} + 10^{-3}\mathcal{L}_{vgg}. \tag{15}$$

H. Implementation Details

During training, we take a random crop of size 256×256 from the input image. We perform random horizontal and vertical flipping and random rotation by 0°, 90°, 180° or 270° degrees to further augment the training data. We train

using a batch size of 16 and a learning rate of 0.0001 with the Adam optimizer. During test time, we run inference on the full test image of size 1500×1000 for the Kalantari *et al.* dataset, and 1536×813 for the Chen *et al.* dataset. We implement the model in PyTorch, and train the model on 4 Nvidia V100 GPUs for approximately 2 days.

IV. Results

We conduct several experiments both comparing against well-known state-of-the-art algorithms and also individually validating the contributions in an extensive ablation study. The experimental setup is described below.

1) Datasets: We use the dynamic training and testing datasets provided by Kalantari and Ramamoorthi [12] which includes 89 scenes in total. Each of these scenes include three differently exposed input LDR images (with EV of -2.00, 0.00, $+2.00$ or -3.00, 0.00, $+3.00$) which contain dynamic elements (e.g. camera motion, non-rigid movements) and a ground-truth image aligned with the medium frame captured via static exposure fusion. Additionally we use the dynamic testing dataset provided by Chen *et al.* [34] for further evaluation. As this dataset does not have a corresponding training set, all methods are trained on the Kalantari dataset and evaluated on the Chen dataset. We test on the 3-Exposure setting which has the ground truth aligned to the middle exposure. To keep it consistent with training, we restrict the number of input frames to three with EVs of -2.00, 0.00, $+2.00$. For purely qualitative evaluation of our method, we include testing sequences from the Tursun [35] dataset.

2) Metrics: We include seven different objective metrics in our quantitative evaluation. First, we compute the PSNR-L, which is a fidelity metric computed directly on the linear HDR estimations. HDR linear images are normally tonemapped for visualization, and thus we include PSNR-μ, which evaluates PSNR on images tonemapped using the μ-law, as defined in Eq. (11), which is a simple canonical tonemapper. We also calculate PSNR-PU, which uses the perceptual uniform encoding (PU21) for HDR images introduced by [36] which aims to improve the correlation between standard metrics and subjective scores on HDR images. For each of the three image domains (linear, μ-tonemapped, PU21), we also calculate the SSIM (Structural Similarity Index) metric introduced by [37] which aims to evaluate perceived changes in the underlying structure of the image. This gives us three further metrics, namely SSIM-L, SSIM-μ and SSIM-PU. Lastly, we also compute the HDR-VDP 2.2 [38], which estimates both visibility and quality differences between image pairs. For each metric, we also report a confidence interval calculated using a t-test at the 95% significance level. We compute the confidence intervals per image and report the mean across the test set.

A. Ablation Studies

We evaluate the contribution of the different parts of our model architecture on the Kalantari dataset.

In Table I, we evaluate the quantitative impact of using our multi-stage fusion mechanism as well as the performance gain from our proposed flow network and our uncertainty modelling. Our baseline model uses the same architecture as our proposed method, but with the flow network, uncertainty modelling and multi-stage max-pooling removed, instead using concatenation as the fusion mechanism, and the attention mechanism from [18]. We also qualitatively evaluate the impact of our contributions in Figures 6, 7 and 8.

1) Fusion Mechanism: We show in Table I that using our multi-stage fusion mechanism (MSM) outperforms concatenation (Baseline Model) by 0.39dB PSNR-L and 0.27dB PSNR-μ. The progressive sharing of information between streams allows the network to retain more information and produce sharper, more detailed images.

TABLE I

An Ablation Study Showing the Contributions of Our Proposed Fusion Mechanism, HDR Iterative Flow, Learnable Exposure Modelling and Alignment Uncertainty

Model	PSNR-$\mu \pm t_{0.95}$	PSNR-L $\pm t_{0.95}$
Baseline Model	43.91 $\pm$ 0.031	41.39 $\pm$ 0.156
+ Multi Stage Max-pooling (MSM)	44.18 $\pm$ 0.032	41.78 $\pm$ 0.172
+ MSM + Flow	44.23 $\pm$ 0.031	42.49 $\pm$ 0.178
+ MSM + Flow + Fixed Exposure	44.21 $\pm$ 0.032	42.20 $\pm$ 0.168
+ MSM + Flow + Exposure Uncertainty	44.28 $\pm$ 0.033	42.28 $\pm$ 0.194
+ MSM + Flow + Fixed Exposure + Alignment Uncertainty	44.28 $\pm$ 0.032	42.51 $\pm$ 0.155
+ MSM + Flow + Exposure Uncertainty + Alignment Uncertainty	**44.35** $\pm$ 0.033	**42.60** $\pm$ 0.165

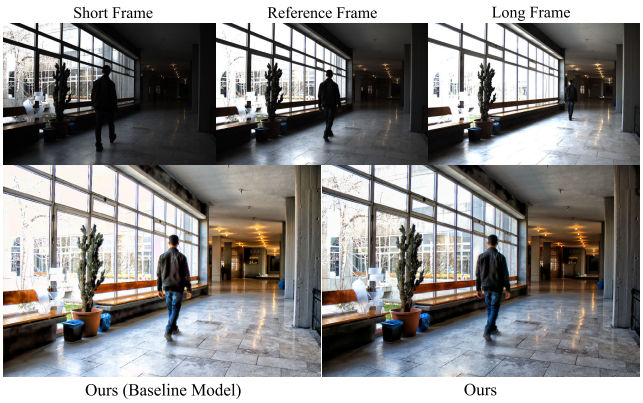

Fig. 6. Qualitative evaluation of our model on a test scene from the Tursun dataset with large foreground motion and severe over/under-exposure. Our exposure uncertainty is able to regulate the contributions of overexposed pixels in the window panes while our max-pooling mechanism restores details which are not visible in the reference frame.

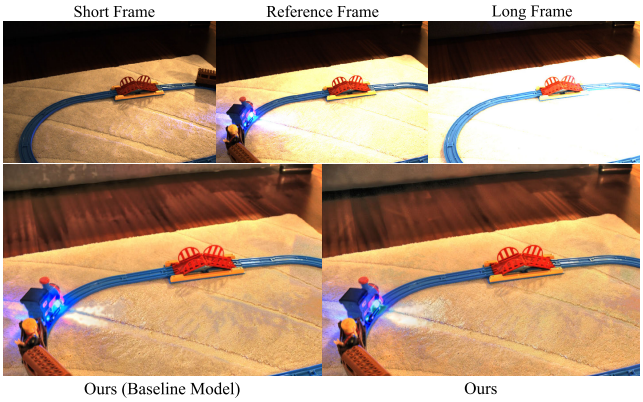

Fig. 7. Qualitative evaluation of our model on a test scene from the Tursun dataset with a fast moving object. Our efficient flow network reduces ghosting artefacts while our exposure uncertainty reduces exposure artefacts.

TABLE II
QUANTITATIVE RESULTS ON THE KALANTARI *et al.* [12] DATASET. BEST PERFORMER DENOTED IN BOLD AND RUNNER-UP IN UNDERSCORED TEXT. † VALUES AS REPORTED BY AUTHORS

Method	PSNR-$\mu \pm t_{0.95}$	PSNR-PU $\pm t_{0.95}$	PSNR-L $\pm t_{0.95}$	SSIM-$\mu \pm t_{0.95}$	SSIM-PU $\pm t_{0.95}$	SSIM-L $\pm t_{0.95}$	HDR-VDP-2 $\pm t_{0.95}$
Sen [7]	40.98 ± 0.031	33.27 ± 0.038	38.38 ± 0.140	0.9880 ± 4.58×10^{-5}	0.9782 ± 7.24×10^{-5}	0.9758 ± 8.37×10^{-5}	60.54 ± 1.17
Kalantari [12]	42.70 ± 0.030	33.86 ± 0.037	41.23 ± 0.146	0.9915 ± 3.48×10^{-5}	0.9832 ± 5.66×10^{-5}	0.9858 ± 5.85×10^{-5}	64.63 ± 1.23
Wu [13]	42.01 ± 0.024	30.82 ± 0.015	41.62 ± 0.145	0.9898 ± 3.13×10^{-5}	0.9805 ± 5.17×10^{-5}	0.9872 ± 5.20×10^{-5}	65.78 ± 1.15
AHDR [18]	43.57 ± 0.031	33.46 ± 0.023	41.16 ± 0.181	0.9922 ± 2.98×10^{-5}	0.9843 ± 5.23×10^{-5}	0.9871 ± 5.82×10^{-5}	64.83 ± 1.19
Prabhakar19 [27]	42.79 ± 0.027	29.06 ± 0.017	40.31 ± 0.211	0.9912 ± 3.08×10^{-5}	0.9762 ± 5.51×10^{-5}	0.9874 ±5.44×10^{-5}	62.95 ± 1.24
Pu [39]†	43.85	-	41.65	0.9906	-	0.9870	-
Prabhakar20 [19]†	43.08	-	41.68	-	-	-	-
NHDRR [40]†	42.41	-	-	0.9887	-	-	-
Prabhakar21 [24]	41.94 ± 0.027	32.18 ± 0.022	41.80 ± 0.141	0.9901 ± 3.20×10^{-5}	0.9813 ± 5.26×10^{-5}	0.9892 ± 4.87×10^{-5}	65.30 ± 1.15
ADNet [22]	43.87 ± 0.031	30.68 ± 0.014	41.69 ± 0.156	0.9925 ± 2.88×10^{-5}	0.9845 ± 4.96×10^{-5}	0.9885 ± 5.34×10^{-5}	65.56 ± 1.14
HDR-GAN [21]	43.96 ± 0.032	34.04 ± 0.032	41.76 ± 0.164	0.9926 ± 2.90×10^{-5}	0.9853 ± 5.00×10^{-5}	0.9884 ± 5.43×10^{-5}	65.07 ± 1.14
Ours	**44.35** ± 0.033	**35.13** ± 0.030	**42.60** ± 0.165	**0.9931** ± 2.72×10^{-5}	**0.9865** ± 4.57×10^{-5}	**0.9902** ± 4.61×10^{-5}	**66.56** ± 1.18

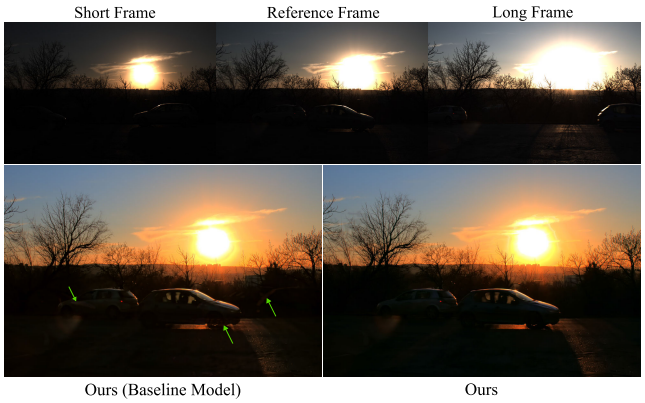

Fig. 8. Qualitative evaluation of our model on a test scene from the Tursun dataset with multiple fast moving objects which are present in all input frames. The baseline model struggles with ghosting artefacts, indicated by the green arrows. Our method is able to eliminate these artefacts entirely.

2) Motion Alignment and Modelling Uncertainty: We look at the performance of our proposed flow network and uncertainty modelling in Table I. Our flow network (MSM + Flow) improves PSNR-L by a large 0.7dB, and PSNR-μ by 0.05dB. compared to using just MSM. We validate the contribution of our learnable model of exposure uncertainty by comparing it to the non-learnable fixed exposure model used by [12]. We fix the values of α and β to match the triangle functions used to generate the ground truths of the Kalantari dataset, which are in essence the oracle α and β parameters. Our learnable exposure modelling (MSM + Flow + Exposure Uncertainty) shows an improvement in PSNR-μ of 0.07dB and PSNR-L of 0.08dB compared to the fixed exposure model (MSM + Flow + Fixed Exposure) In this case, the gain from learning exposure values is small as it is possible to easily fix the values to their optima due to prior knowledge of how the dataset was created. However, this is not always possible, especially in scenarios with different numbers of input images and exposure levels, where the underlying ideal weights are not known. Our decision to learn the weights allows our model to handle any number of input frames without any manual tuning. We also validate the contribution of our alignment uncertainty (MSM + Flow + Fixed Exposure + Alignment Uncertinaty), which gives an improvement in PSNR-μ of 0.07dB and PSNR-L of 0.32dB when compared to using only exposure uncertainty.

B. Performance Evaluation

We evaluate the performance of our proposed method for the HDR estimation task and compare it to other state-of-the-art methods both quantitatively and qualitatively. The methods included in our benchmark cover a broad range of approaches, namely: the patch-based method of Sen *et al.* [7], methods which use traditional alignment followed by CNNs to correct dense and global alignment, [12], [13], [24], the flexible aggregation approach of [27] that also uses dense alignment, methods which rely on attention or feature selection followed by a CNN to deghost and merge images [18], [40], a GAN-based approach which can synthesize missing details in areas with disocclusions [22] and a method which uses deformable convolutions as an alignment mechanism [22]. For the Chen *et al.* test set, we also compare against the HDR video method proposed by [34], which uses a coarse to fine architecture to align and reconstruct input frames. As this method requires seven input frames for the three exposure setting, we do not re-train this on the Kalantari dataset and instead use the pre-trained weights provided by the authors. We show in Table II the quantitative evaluation on the Kalantari test set. The differences in PSNR between our method and the runners up on the Kalantari dataset are large (i.e. +1.1dB PSNR-PU, +0.4db PSNR-μ, +0.8dB PSNR-L). Similarly, the HDR-VDP-2 score obtained by our method outperforms all others by a wide margin (i.e. 0.8). Furthermore, we outperform all methods on all seven metrics, showing our method is consistently better across different evaluation criteria. To further evaluate the consistency of our improvement over other methods, we look at the distribution of PSNRs per image across the Kalantari test set. In Figure 12, we show the improvement achieved in PSNR-μ of our method and other top performers over AHDR [18], which is a well known and strong baseline method in HDR imaging. Our method demonstrates a consistent and significant improvement over AHDR, being the only method which achieves an improvement in performance on every single test image. Apart from a few exceptions, our method also outperforms the existing state-of-the-art methods for most images. We observe similar performance on the

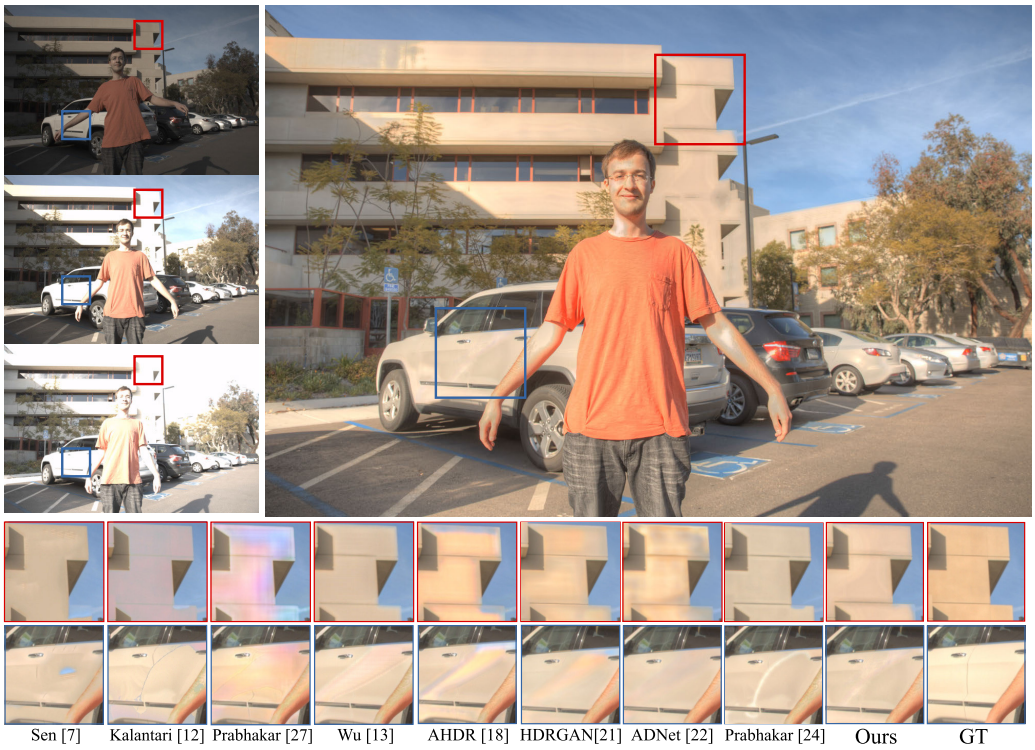

Fig. 9. Qualitative evaluation of our method for an image from the Kalantari test set. Our method obtains images with noticeably less ghosting artefacts, together with sharper and finer details. Best viewed zoomed-in.

Fig. 10. Qualitative evaluation of our method for an image from the Kalantari test set. Our method obtains images with noticeably less ghosting artefacts, together with sharper and finer details. Best viewed zoomed-in.

Chen *et al.* dynamic test set, demonstrating the generalization ability of our model on out of domain data. Our method is best or second best in six out of seven metrics, with a significant improvement in PSNR-μ (i.e 0.5dB) over the runner up Chen *et al.* [34]. We outperform Chen *et al.* on several key metrics such as PSNR-PU, PSNR-μ, SSIM-PU and SSIM-μ despite the fact their method is trained on video data and has an in-domain advantage.

Fig. 11. Qualitative evaluation of our model on different numbers of input frames of varying exposures. The quality of the reconstruction clearly increases with the number of input frames used. Our model can utilize information from all 9 input frames, despite having only seen 3 input frames during training.

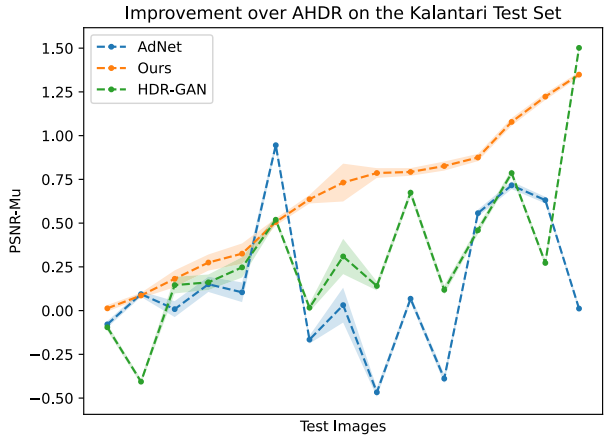

Fig. 12. We show performance per image on the Kalantari test set of the best performing methods and associated confidence intervals as shaded areas.

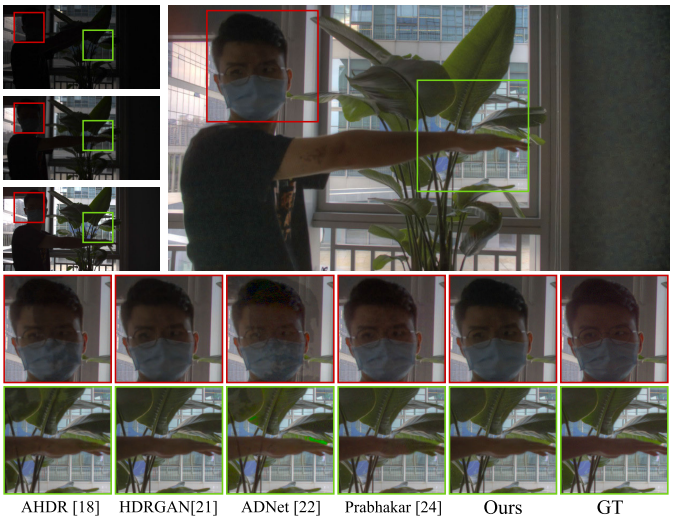

Fig. 13. Qualitative evaluation of our method for an image from the Chen test set. Our method performs well even in low light conditions while other deep learning methods suffer from ghosting and artefacts. Best viewed zoomed in.

We also quantitatively evaluate the performance of our optical flow network compared to previous optical flow methods in Table V. We compare against the traditional method introduced by Liu *et al.* [17] which is used by [12] and [24] in their HDR pipelines, as well as the deep learning based approach introduced by Sun *et al.* [28] which is used by [27] to pre-align input frames. We substitute our optical flow network with the comparison methods but we keep the rest of our proposed architecture the same. We show that our flow network improves on the runner up by 0.55dBs in PSNR-μ and 0.12dBs in PSNR-L, while having the additional advantages of being end-to-end trainable and requiring no pre-training with ground truth optical flows.

In Figures 9, 10, and 13 we show some visualizations of our algorithm compared with the benchmarked methods for qualitative, subjective evaluation. All other methods present traces of ghosting artefact around the edges near a moving object, especially where disocclusions happen and one or more frames have overexposed values in those locations (e.g. moving head, moving arm). Our method tackles such challenges effectively thanks to the exposure confidence awareness, and strongly suppresses the ghosting artefact. Additionally, our method also demonstrates better performance when it comes to edges and

TABLE III

QUANTITATIVE RESULTS ON THE CHEN *et al.* [34] DYNAMIC DATASET. BEST PERFORMER DENOTED IN BOLD AND RUNNER-UP IN UNDERSCORED TEXT. *CHEN IS TRAINED ON A SYNTHETIC VIDEO DATASET WHILE ALL OTHER METHODS ARE TRAINED ON KALANTARI

Method	PSNR-$\mu \pm t_{0.95}$	PSNR-PU $\pm t_{0.95}$	PSNR-L $\pm t_{0.95}$	SSIM-$\mu \pm t_{0.95}$	SSIM-PU $\pm t_{0.95}$	SSIM-L $\pm t_{0.95}$	HDR-VDP-2 $\pm t_{0.95}$
Sen [7]	40.79 $\pm$ 0.011	**34.33** $\pm$ 0.012	39.58 $\pm$ 0.043	0.9862 $\pm$ 2.35 $\times 10^{-5}$	0.9617 $\pm$ 6.41 $\times 10^{-5}$	0.9912 $\pm$ 5.86 $\times 10^{-5}$	68.83 $\pm$ 0.77
AHDR [18]	39.56 $\pm$ 0.019	26.71 $\pm$ 0.006	35.09 $\pm$ 0.058	0.9814 $\pm$ 3.84 $\times 10^{-5}$	0.9632 $\pm$ 7.08 $\times 10^{-5}$	0.9900 $\pm$ 7.24 $\times 10^{-5}$	63.78 $\pm$ 0.75
Prabhakar21 [24]	40.89 $\pm$ 0.015	29.06 $\pm$ 0.008	38.19 $\pm$ 0.052	0.9857 $\pm$ 3.17 $\times 10^{-5}$	0.9670 $\pm$ 6.16 $\times 10^{-5}$	0.9899 $\pm$ 6.27 $\times 10^{-5}$	65.48 $\pm$ 0.69
ADNet [22]	38.14 $\pm$ 0.042	30.45 $\pm$ 0.029	40.97 $\pm$ 0.042	0.9797 $\pm$ 4.30 $\times 10^{-5}$	0.9622 $\pm$ 7.14 $\times 10^{-5}$	0.9935 $\pm$ 4.15 $\times 10^{-5}$	70.00 $\pm$ 0.66
HDR-GAN [21]	40.80 $\pm$ 0.018	27.37 $\pm$ 0.007	36.19 $\pm$ 0.048	**0.9904** $\pm$ 2.06 $\times 10^{-5}$	0.9730 $\pm$ 5.28 $\times 10^{-5}$	0.9907 $\pm$ 6.37 $\times 10^{-5}$	66.33 $\pm$ 0.72
Chen* [34]	41.47 $\pm$ 0.014	33.37 $\pm$ 0.011	**42.33** $\pm$ 0.038	0.9883 $\pm$ 2.65 $\times 10^{-5}$	0.9698 $\pm$ 5.35 $\times 10^{-5}$	**0.9945** $\pm$ 3.40 $\times 10^{-5}$	**71.87** $\pm$ 0.82
Ours	**41.98** $\pm$ 0.014	33.67 $\pm$ 0.011	41.32 $\pm$ 0.053	0.9884 $\pm$ 2.41 $\times 10^{-5}$	**0.9732** $\pm$ 4.84 $\times 10^{-5}$	0.9935 $\pm$ 4.08 $\times 10^{-5}$	69.42 $\pm$ 0.78

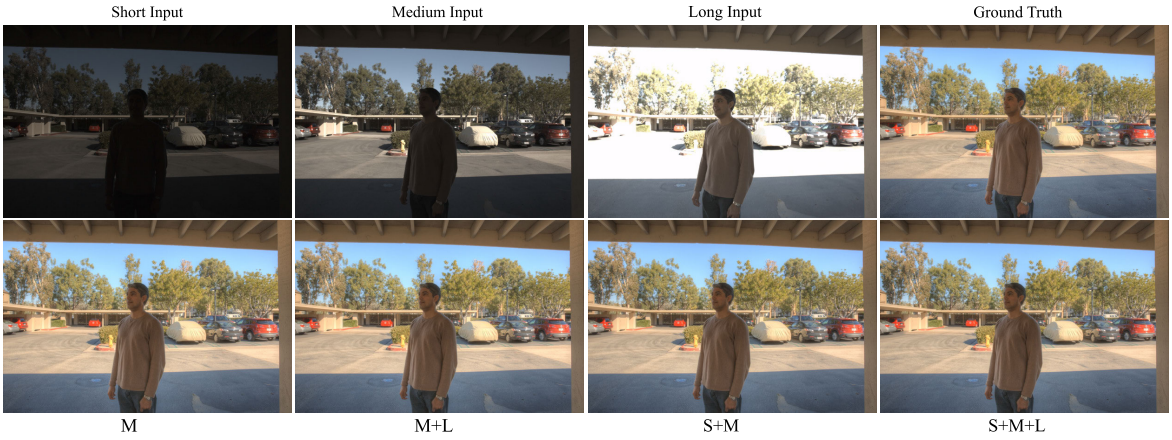

Fig. 14. Qualitative comparison of our model trained on all permutations with different input frames. When the medium frame is well exposed, our model can attain a high quality prediction with just one frame. There is no noticeable increase in image quality when more inputs are provided.

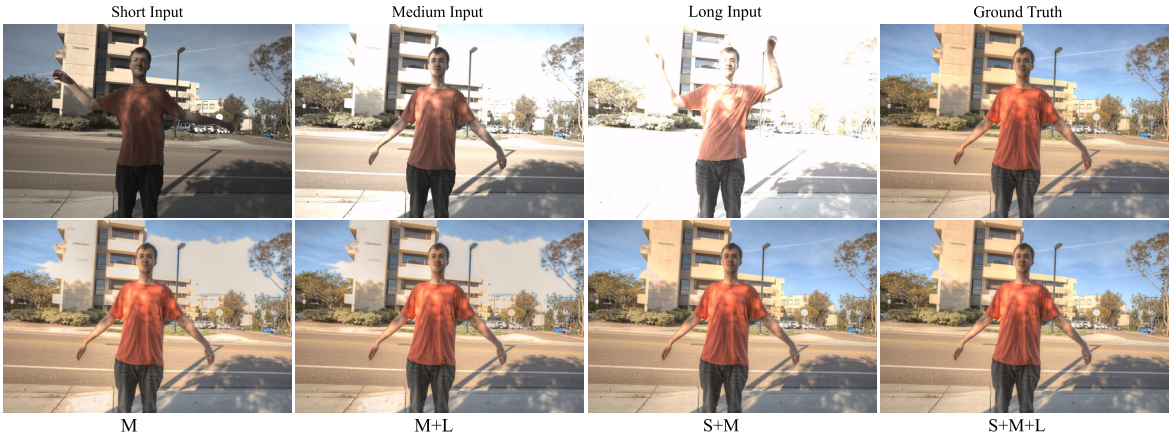

Fig. 15. Qualitative comparison of our model trained on all permutations with different input frames. When the medium frame is overexposed, our model is not able to completely hallucinate details in large overexposed regions. The short frame is essential to reconstruct the missing details. There is no noticeable improvement in image quality from including the extremely overexposed long frame.

textures (e.g. building facade), as well as out of domain low-light performance.

C. Flexible Imaging

We show that our model is flexible enough to accept an arbitrary number of images without the need for re-training. In Table IV we evaluate the performance of our proposed model when trained and tested on different numbers of images with different exposures. We use the following input frame configurations for training and testing: Short + Medium + Long (S + M + L), Short + Medium (S + M), Medium + Long (M + L), Medium (M). The reference frame in all settings is the medium frame, which is spatially aligned to the ground truth. As expected, performance is best when the testing configuration is seen during training. Our model trained on all permutations achieves competitive cross-setting performance, obtaining the best results for the S + M and M + L settings. It is also competitive with our best model for the M and S + M + L settings, without needing any extra training time, and is capable of accepting a range of different input configurations without the need for re-training. In fact, we show that our model using only two frames (S + M) can obtain results outperforming current state-of-the-art methods using all three frames in PSNR-μ [21] and

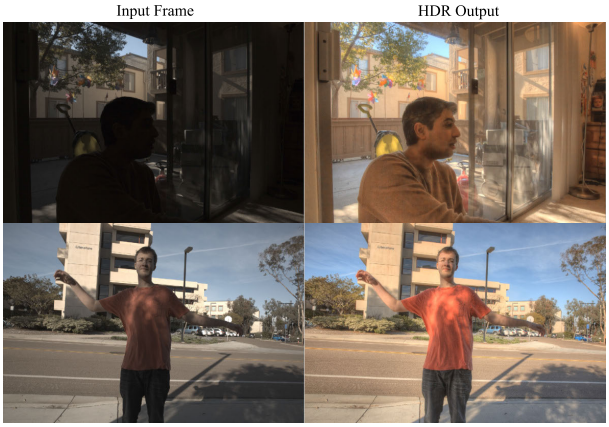

Fig. 16. Qualitative evaluation of our model using just the short frame as input. When the input frame is severely underexposed (top), we see quantization and noise related artefacts. However when the input frame is well exposed (bottom), the output is free of artefacts.

TABLE IV

COMPARISON OF DIFFERENT TRAINING REGIMES ON THE KALANTARI TEST SET USING OUR PROPOSED MODEL. * BATCHES SAMPLED UNIFORMLY FROM THE TWO OPTIONS. ** BATCHES SAMPLED UNIFORMLY FROM ALL FOUR OPTIONS. THE FOUR OPTIONS ARE: S + M + L, S + M, M + L, M

Training Frames	Test Frames	PSNR-μ $\pm$ $t_{0.95}$	PSNR-L $\pm$ $t_{0.95}$
S + M + L	S + M + L	**44.35** $\pm$ 0.033	**42.60** $\pm$ 0.165
S + M / M + L*	S + M + L	43.44 $\pm$ 0.031	40.58 $\pm$ 0.144
M	S + M + L	20.86 $\pm$ 0.006	24.66 $\pm$ 0.030
All Permutations **	S + M + L	44.24 $\pm$ 0.031	42.30 $\pm$ 0.170
S + M + L	S + M	40.24 $\pm$ 0.028	41.15 $\pm$ 0.158
S + M / M + L*	S + M	44.11 $\pm$ 0.032	42.17 $\pm$ 0.167
M	S + M	22.68 $\pm$ 0.005	25.38 $\pm$ 0.037
All Permutations **	S + M	**44.18** $\pm$ 0.031	**42.29** $\pm$ 0.168
S + M + L	M + L	41.85 $\pm$ 0.025	37.60 $\pm$ 0.148
S + M / M + L*	M + L	42.58 $\pm$ 0.031	38.32 $\pm$ 0.234
M	M + L	22.03 $\pm$ 0.006	25.00 $\pm$ 0.032
All Permutations **	M + L	**42.74** $\pm$ 0.030	**38.38** $\pm$ 0.222
S + M + L	M	33.10 $\pm$ 0.010	32.84 $\pm$ 0.032
S + M / M + L*	M	28.26 $\pm$ 0.005	22.44 $\pm$ 0.019
M	M	**42.86** $\pm$ 0.031	**38.79** $\pm$ 0.226
All Permutations **	M	42.67 $\pm$ 0.030	38.40 $\pm$ 0.216

TABLE V

QUANTITATIVE COMPARISON OF DIFFERENT OPTICAL FLOW METHODS ON THE KALANTARI TEST SET

Method	PSNR-μ $\pm$ $t_{0.95}$	PSNR-L $\pm$ $t_{0.95}$
Liu et al. [17]	43.68 $\pm$ 0.030	42.48 $\pm$ 0.157
PWC-Net. [28]	43.80 $\pm$ 0.031	42.36 $\pm$ 0.143
Ours	**44.35** $\pm$ 0.033	**42.60** $\pm$ 0.165

PSNR-L [24]. We qualitatively show the performance gains from our flexible architecture in Figure 11. The figure shows that our method works with 3, 5, 7, or 9 input frames with varying exposures from -4.00 to $+4.00$, without the need to re-train. As the number of frames increases, the reconstruction quality also clearly increases (i.e. the colour and detail of the flames improves as we go from 3 to 9 frames). Our proposed method is capable of utilizing the information provided in all 9 input frames. In Figure 14 we show that our model is also capable of producing high quality HDR outputs with fewer than 3 input frames. Furthermore we show in Figure 17

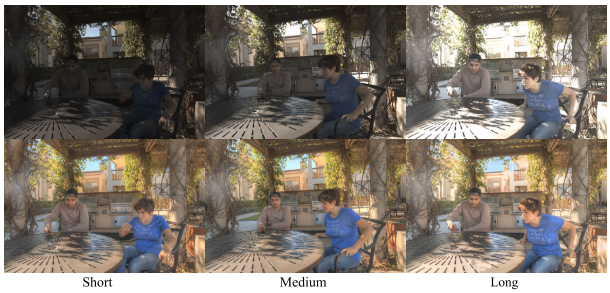

Fig. 17. Our method is flexible enough to accept any frame as the reference frame without re-training, providing superior choice to the user.

TABLE VI

A COMPARISON OF RUNTIMES OF OUR METHOD ON DIFFERENT INPUT RESOLUTIONS AND NUMBER OF INPUT FRAMES, COMPUTED ON AN NVIDIA V100 GPU

Input Resolution	# Input Frames	Runtime (s)
1024x682	1	0.24
1024x682	2	0.48
1024x682	3	0.70
1280x720	1	0.32
1280x720	2	0.58
1280x720	3	0.92
1500x1000	1	0.51
1500x1000	2	0.97
1500x1000	3	1.55

TABLE VII

BREAKDOWN OF OUR MODEL PARAMETERS BY MODEL SUB-COMPONENTS

Model	# Parameters (millions)
Flow Network	0.87
Attention Network	0.33
Merging Network	0.92
Ours Total	2.12

that our method can use any frame as the reference frame without re-training, providing superior flexibility and choice when compared to state-of-the-art methods.

D. Parameters and Runtime

We provide a breakdown of our model parameters by sub-components in Table VII and provide a comparison of our flow network with state-of-the-art optical flow methods in Table VIII. Our flow network is an order of magnitude smaller than [28], which is used by [27] to align images, and 6$\times$ smaller than RAFT [29]. We also explore how the runtime of our model varies depending on both the input resolution and the number of input frames in Table VI. The model runtime grows approximately linearly with the number of input frames, and quadratically with the input resolution.

E. Limitations

One limitation of our method is that it is not able to hallucinate details in large over-exposed regions, as seen in Figure 15. The missing information needs to be provided in one of the input frames for our model to be able to accurately reconstruct the HDR image. This is not surprising given that

TABLE VIII
A COMPARISON OF THE NUMBER OF PARAMETERS IN OUR EFFICIENT FLOW NETWORK AND OTHER STATE-OF-THE-ART FLOW NETWORKS

Model	# Parameters (millions)
PWC-Net [28]	8.75
PWC-Net-Small [28]	4.08
RAFT [29]	5.30
Our Flow Network	0.87

the loss functions used during training have a focus on HDR reconstruction. There is potential to improve our single image HDR performance by exploring a training strategy similar to those used for inpainting. Similarly our model can not fully denoise extremely underexposed regions, as shown in Figure 16. We also note that for the datasets used in our work, the CRF is a simple fixed gamma curve with $\gamma = 2.2$, as defined by the dataset authors. However in a general case, the gamma correction we use is only a coarse approximation of the CRF and we rely on the network to implicitly perform finer adjustments. We expect accurate and explicit estimation of the CRF [41], [42] to positively impact the reconstruction performance of our model, especially for cases where the CRF is non-trivial. Finally, although our model can theoretically accept any number of input frames, the amount of activation memory required increases linearly with the number of input frames. On a single Nvidia V100 GPU, we can process up to nine full sized input frames from the Tursun dataset as shown in Figure 11.

V. CONCLUSION

In this paper we explored modelling exposure and alignment uncertainties to improve HDR imaging performance. We presented (1) an HDR-specific optical flow network which is capable of accurate flow estimations, even with improperly exposed input frames, by sharing information between input images with a symmetric pooling operation. (2) We also presented models of exposure and alignment uncertainty which we use to regulate contributions from unreliable and misaligned pixels and greatly reduce ghosting artefacts. (3) Lastly a flexible architecture which uses a multi-stage fusion to estimate an HDR image from an arbitrary set of LDR input images. We conducted extensive ablation studies where we validate individually each of our contributions. We compared our method to other state-of-the-art algorithms obtaining significant improvements for all the measured metrics and noticeably improved visual results.

REFERENCES

[1] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. ACM SIGGRAPH*, 2008, pp. 1–10.

[2] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," in *Proc. ACM SIGGRAPH Papers*, 2011, pp. 1–10.

[3] M. McGuire, W. Matusik, H. Pfister, B. Chen, J. F. Hughes, and S. K. Nayar, "Optical splitting trees for high-precision monocular imaging," *IEEE Comput. Graph. Appl.*, vol. 27, no. 2, pp. 32–42, Mar. 2007.

[4] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," *Proc. SPIE*, vol. 9023, pp. 279–288, Mar. 2014.

[5] F. Heide *et al.*, "FlexISP: A flexible camera image processing framework," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 1–13, Nov. 2014.

[6] S. Hajisharif, J. Kronander, and J. Unger, "Adaptive dualISO HDR reconstruction," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, pp. 1–13, Dec. 2015.

[7] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–203, 2012.

[8] J. Zheng, Z. Li, Z. Zhu, S. Wu, and S. Rahardja, "Hybrid patching for a sequence of differently exposed images with moving objects," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5190–5201, Dec. 2013.

[9] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free HDR reconstruction," *ACM Trans. Graph.*, vol. 32, no. 6, p. 201, 2013, doi: 10.1145/2508363.2508410.

[10] O. Gallo, A. J. Troccoli, J. Hu, K. Pulli, and J. Kautz, "Locally non-rigid registration for mobile HDR photography," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 48–55, doi: 10.1109CVPRW.2015.7301366.

[11] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 1163–1170, doi: 10.1109/CVPR.2013.154.

[12] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Jul. 2017.

[13] S. Wu, J. Xu, Y. Tai, and C. Tang, "End-to-end deep HDR imaging with large foreground motions," in *Proc. Eur. Conf. Comput. Vis.*, vol. 2, 2018, pp. 120–135.

[14] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 553–560.

[15] A. Artusi, T. Richter, and R. K. Mantiuk, "High dynamic range imaging technology," *IEEE Signal Process. Mag.*, vol. 34, no. 5, pp. 165–172, Sep. 2017.

[16] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High Dynamic Range Imaging*, 2nd ed. Burlington, MA, USA: Morgan Kaufmann, 2010.

[17] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[18] Q. Yan *et al.*, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1751–1760.

[19] K. R. Prabhakar, S. Agrawal, D. K. Singh, B. Ashwath, and R. V. Babu, "Towards practical and efficient high-resolution HDR deghosting with CNN," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 12366. Cham, Switzerland: Springer, 2020, pp. 497–513, doi: 10.1007/978-3-030-58589-1_30.

[20] N. K. Kalantari and R. Ramamoorthi, "Deep HDR video from sequences with alternating exposures," *Comput. Graph. Forum*, vol. 38, no. 2, pp. 193–205, 2019.

[21] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. H. Lau, "HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions," *IEEE Trans. Image Process.*, vol. 30, pp. 3885–3896, 2021.

[22] Z. Liu *et al.*, "ADNet: attention-guided deformable convolutional network for high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 463–470.

[23] E. Pérez-Pellitero *et al.*, "NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 691–700.

[24] K. R. Prabhakar, G. Senthil, S. Agrawal, R. V. Babu, and R. K. S. S. Gorthi, "Labeled from unlabeled: Exploiting unlabeled data for few-shot deep HDR deghosting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4873–4883.

[25] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[26] M. Aittala and F. Durand, "Burst image deblurring using permutation invariant convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 11212. Cham, Switzerland: Springer, 2018, pp. 748–764, doi: 10.1007/978-3-030-01237-3_45.

[27] K. R. Prabhakar, R. Arora, A. Swaminathan, K. P. Singh, and R. V. Babu, "A fast, scalable, and reliable deghosting method for extreme exposure fusion," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2019, pp. 1–8, doi: 10.1109/ICCPHOT.2019.8747329.

[28] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[29] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

[30] D.-W. Kim, J. R. Chung, and S.-W. Jung, "GRDN: Grouped residual dense network for real image denoising and GAN-based real-world noise modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.

[31] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[34] J. Chen, Z. Yang, T. N. Chan, H. Li, J. Hou, and L.-P. Chau, "Attention-guided progressive neural texture fusion for high dynamic range image restoration," *IEEE Trans. Image Process.*, vol. 31, pp. 2661–2672, 2022.

[35] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "An objective deghosting quality metric for HDR images," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 139–152, May 2016.

[36] R. Mantiuk and M. Azimi, "PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[38] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: A calibrated method for objective quality prediction of high-dynamic range and standard images," *J. Electron. Imag.*, vol. 24, no. 1, 2015, Art. no. 010501.

[39] Z. Pu, P. Guo, M. S. Asif, and Z. Ma, "Robust high dynamic range (HDR) imaging with complex motion and parallax," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 1–16.

[40] Q. Yan *et al.*, "Deep HDR imaging via a non-local network," *IEEE Trans. Image Process.*, vol. 29, pp. 4308–4322, 2020.

[41] M. D. Grossberg and S. K. Nayar, "Modeling the space of camera response functions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1272–1282, Oct. 2004.

[42] J.-Y. Lee, Y. Matsushita, B. Shi, I. S. Kweon, and K. Ikeuchi, "Radiometric calibration by rank minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 144–156, Jan. 2013.