# Self-supervised HDR Imaging from Motion and Exposure Cues

Michal Nazarczuk Sibi Catley-Chandar
Ales Leonardis Eduardo Pérez Pellitero

Huawei Noah's Ark Lab

**Abstract.** Recent High Dynamic Range (HDR) techniques extend the capabilities of current cameras where scenes with a wide range of illumination can not be accurately captured with a single low-dynamic-range (LDR) image. This is generally accomplished by capturing several LDR images with varying exposure values whose information is then incorporated into a merged HDR image. While such approaches work well for static scenes, dynamic scenes pose several challenges, mostly related to the difficulty of finding reliable pixel correspondences. Data-driven approaches tackle the problem by learning an end-to-end mapping with paired LDR-HDR training data, but in practice generating such HDR ground-truth labels for dynamic scenes is time-consuming and requires complex procedures that assume control of certain dynamic elements of the scene (*e.g.* actor pose) and repeatable lighting conditions (stop-motion capturing). In this work, we propose a novel self-supervised approach for learnable HDR estimation that alleviates the need for HDR ground-truth labels. We propose to leverage the internal statistics of LDR images to create HDR pseudo-labels. We separately exploit static and well-exposed parts of the input images, which in conjunction with synthetic illumination clipping and motion augmentation provide high quality training examples. Experimental results show that the HDR models trained using our proposed self-supervision approach achieve performance competitive with those trained under full supervision, and are to a large extent superior to previous methods that equally do not require any supervision.

## 1 Introduction

While most uniformly illuminated scenes can be perfectly captured with a conventional camera, it is not uncommon to encounter scenes where the underlying dynamic range, *i.e.* the ratio between the maximum and minimum irradiance, is such that said conventional cameras can not properly capture highlights and shadows simultaneously, and thus suffer from under- and over-exposed pixels. HDR imaging techniques fill in the gap and aim at reconstructing accurate image representations for scenes whose dynamic range is typically beyond three orders of magnitude [21].

The principle behind most HDR capturing strategies rely on obtaining different exposures of the scene that are then incorporated into a single HDR

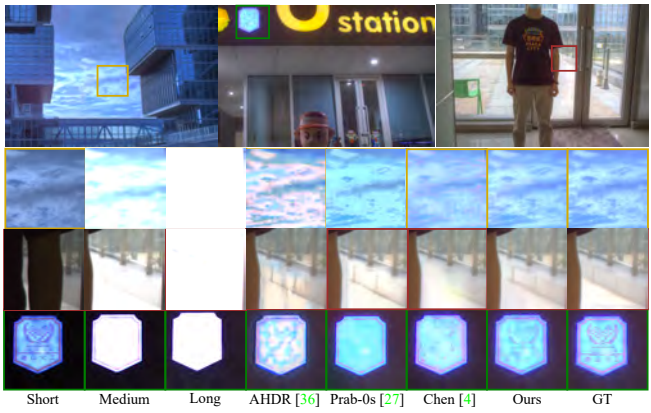



**Fig. 1:** A qualitative comparison of our proposed self-supervised approach against other supervised [36,4] and weakly-supervised [27] methods. We introduce an HDR training strategy that does not require any ground-truth and performs quantitatively better than other unsupervised algorithms and comparable to supervised ones (Tab. 2 and 3).

reconstruction [6]. This can be achieved by multi-camera systems capturing synchronously the scene [23,32,7] but arguably those have been so far less appealing due to *e.g.* increased cost and fragility. Techniques that utilise a single sensor and merge frames captured at different time instants [6] have been proven more pragmatic, and are nowadays widely adopted in consumer cameras of all budgets.

Early frame fusion methods work well under static scenes [6,24], however degrade quickly when there is complex motion (*e.g.* ghosting artefacts). Even though some methods proposed the use of *off-the-shelf* alignment algorithms to find correspondences across frames [34], it is the recent advances of deep learning methods that have established a new *state-of-the-art* by learning an end-to-end mapping between the unaligned LDR input images and the target HDR domain [35,36]. For training such supervised methods, training data where input dynamic scenes are paired to HDR ground-truth images is required.

The seminal work of Kalantari and Ramamoorthi [14] propose for the first time a data-capturing protocol to collect paired dynamic scenes and respective ground-truth labels, which has been adopted in recent datasets [4,28]. Firstly, a subject is asked to stay still and three bracketed exposure images are obtained on a tripod (*static set*) which are then combined to produce the ground truth image. Later, the subject is asked to move and another set of bracketed exposure images is captured (*dynamic set*). The input set is formed by taking the low and high exposure images from this dynamic set and the middle exposure image from the static set. This procedure however comes with substantial drawbacks: it is complex and time-consuming, and most importantly, it assumes control over the dynamic elements of the scene (*e.g.* actor pose) and the repeatability of the scene (*e.g.* composition, illumination) which in essence limit the diversity of scenes and motions that can be captured.

In this work we propose a novel self-supervised HDR methodology that enables training deep models without the need of any ground-truth HDR image.

To the best of our knowledge, no other HDR learning-based previous work has proposed this set-up. In our work, we build on the key observation that LDR input images contain useful information that is transferable to the HDR estimation task when certain conditions are met. In other words, we can systematically study and disentangle the factors of degradation from the HDR to LDR domains, and choose accordingly regions of the LDR images that are not degraded and can thus serve as valuable supervision.

For that purpose, we consider two domains of supervision: (a) the Motion Domain and (b) the Exposure Domain which we then use to generate paired HDR pseudo-labels. The intuition behind (a) is that presence of motion can be automatically and locally determined (*i.e.* as opposed to an image-level rigid *dynamic vs static* label) and therefore pseudo-labels flexibly obtained; and behind (b) that well-exposed regions in the input LDR images are good local approximations of the HDR image, and thus can be used directly as HDR pseudo-labels. These two criteria, in conjunction with further simple synthetic-illumination modelling and motion augmentation enable effective and balanced supervision for challenging dynamic HDR benchmarks [14,4].

In summary, the contributions of this paper are: **(1)** A novel strategy to create HDR pseudo-labels from LDR images based on motion and exposure characteristics, **(2)** a mechanism to transfer and synthetize over-exposed patches via gain mask for well-exposed patches and **(3)** comprehensive experiments and ablation studies to demonstrate the effectiveness of our proposed approach.

## 2 Related work

In this section we provide an overview of relevant multiframe HDR methods that use Convolutional Neural Networks (CNN) and discussion about weakly- and self-supervised approaches. For a more complete review of the HDR SOTA we refer the reader to [33].

**HDR Methods**: Together with their paired dataset, Kalantari and Ramamoorthi proposed an HDR fusion method composed by two stages: alignment and fusion. Firstly, input frames are aligned via optical flow [19] and then a CNN is used to merge aligned frames. Wu *et al.* [35] adopted a similar scheme, however preferring a *simple* global homography rather than a dense optical flow field for the alignment step, and a UNet-like architecture for the fusion of images. These two methods rely on the CNN to suppress errors on alignment at the fusion stage, but do not have an explicit mechanism within their architectures to regulate the contribution of each input frame. Yan *et al.* [36] introduce an attention mechanism that is able to select or suppress features from each respective input frame, which improves performance both for unaligned or pre-aligned input frames. Shortly after, Yan *et al.* [37] explore the non-local correlation in inputs frames in order to reduce ghosting artefacts. In the work of Prabhakar *et al.* [26] 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. Recently,

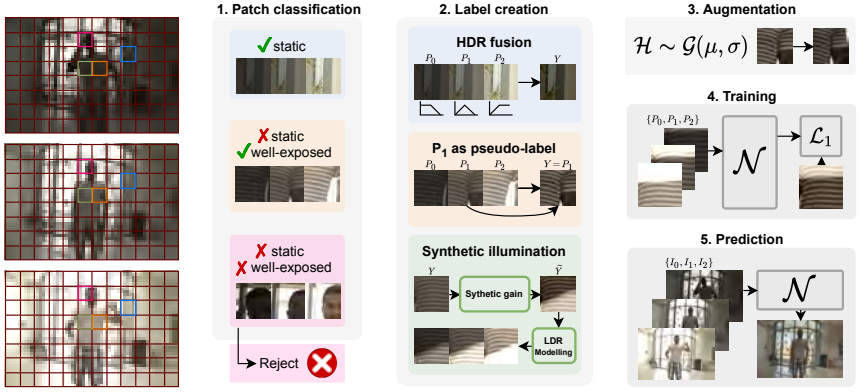



**Fig. 2:** An overview of our method. Input LDR images are considered on a patch level. Image patches are processed to assess their motion and exposure characteristics. HDR fusion is performed for patches that after alignment have no motion. Additionally, we search for dynamic patches, for which the reference frame can serve as an HDR estimation. All well-exposed patches that can be an HDR pseudo-label are applied with synthetic illumination gain, from which an LDR reconstruction is performed. All obtained supervision pairs along with their counterparts augmented with synthetic camera movement are provided as supervision pairs for HDR reconstruction training.

Chen *et al.* [4] estimate HDR video streams, and provide a new test set and related benchmark that retains similarities with [14] in the capturing procedure.

**Self-supervised Learning**: All the discussed methods above require training with full ground-truth supervision. As discussed in Section 1, capturing HDR labels is inconvenient, time-consuming and inherently limits the scene and motion variability. Self-supervised learning (SSL) has attracted much interest during the last years in part for its clear success on representation learning [5,1,9,13], but also due to its general appeal for problems where capturing large amount of data is still an open problem. Literature has mainly focused on high-level tasks, however there have been a number of works that explore self-supervision for *e.g.* image denoising [18,17], super resolution [33]. These works however are not directly applicable to the HDR estimation task, as some of the necessary assumptions on the estimated residual distribution do not hold, or still require some form of supervision. The recent work of Prabhakar *et al.* [27] is the first attempt at training-data efficiency applicable to HDR imaging. They propose a novel weakly-supervised training strategy, where they use *easy-to-capture* static scenes in combination with unlabelled dynamic sequences and only a few *costly* dynamic HDR labels, and achieve results that are competitive to fully supervised methods. Their strategy is divided into two stages: Firstly they train their model with the few labelled dynamic scenes, and use that model to predict HDR pseudo-labels from the pool of unlabelled dynamic data. In stage two the model is further fine-tuned with all the available original data and the generated pseudo-labels. Despite the important contributions of this work, the method does still require some manual labelling effort, *i.e.* whole image static scenes need to be

captured with a tripod, and ideally a number of dynamic scenes captured in the stop-motion fashion.

## 3 Method

Existing data-driven approaches for HDR reconstruction use either full, or weak supervision from ground truth HDR images. Those HDR labels are very difficult to capture and require a specific setup (static camera and controllable elements of the scene, *e.g.* moving actor). We propose a method that alleviates the need for ground truth labels and is applicable to various learning-based approaches for HDR reconstruction. We introduce a self-supervision strategy that leverages various LDR image characteristics to produce supervision signals that do not require access to ground truth HDR image. Hence, this allows for the use of any varying exposure LDR sequence captured with no constraints on the procedure.

Given the set of LDR images with varying exposure $(I_0, I_1, I_2)$, where $I_1$ refers to an arbitrarily chosen reference image, our goal is to produce a set of HDR pseudo-labels in the form of sets of patches $\{(P_{0i}, P_{1i}, P_{2i}), Y_i\}$ suitable for use as a supervision signal in data-driven approaches.

We observe that there are two main conditions for a straightforward capture and reconstruction of HDR patches: (1) the scene, or at least a part of it, is static and allows thus to reliably perform a linear combination of input LDR images with varying exposure; or (2) the LDR reference image is well-exposed on its own, *i.e.* does not contain over- or under-exposed regions or any form of degradation. We argue that sequences of LDR images often contain regions that satisfy one of the given conditions, and by using patches rather than images we can flexibly select them. Static parts of the image can be used to reconstruct HDR via direct linear fusion. Well-exposed image regions represent areas where input images provide a good approximation of the underlying luminosity and can thus be used as HDR pseudo-labels. We introduce further synthetic illumination changes on those pseudo-labels such that the respective LDR patches contain information loss due to over-exposure, and thus create valuable LDR to HDR supervision on dynamic sets with over-exposed regions. In summary, we propose a method that exploits static and well-exposed regions in the image separately to produce HDR image patches for self-supervision. Figure 2 presents an overview of the proposed method. A set of input images is divided into square patches $\{P_{0i}, P_{1i}, P_{2i}\}$ and parsed by characteristics classification and label creation modules. Sets of patches and created pseudo-labels provide supervision LDR-HDR pairs that can be used by any learning method. Additionally, we propose the use of simple motion augmentation on all patches to provide additional supervision for dynamic, misaligned patches.

### 3.1 Motion domain

In the proposed approach, we suggest to obtain a set of pseudo-labels for HDR supervision based on motion characteristics of the image. The intention is to

extract image regions suitable to undergo static HDR fusion. We perform classification based on the optical flow estimation of the unlabelled input data. A high level overview of the the method is shown in Algorithm 1. Firstly we estimate optical flow between each image and the reference image, in both directions (*i.e.* $I_1I_0$, $I_0I_1$, $I_1I_2$, $I_2I_1$). We use the recent GMA [12] pretrained on the Sintel dataset [3]. In order to provide an HDR estimation, images have to be fully aligned. We assume that very small camera motions can be easily corrected. We create a histogram of optical flow magnitudes and consider only images with dominant optical flow magnitude below a given threshold $t_f$. The camera movement is corrected by estimating a homography transformation $H_{i1}$ between all images with respect to the reference frame. Further, we apply warping to all non-reference images ($I_0$, $I_2$) to the reference image coordinate system obtaining thus an aligned set of images ($I_{0W}, I_1, I_{2W}$). For optical flow visualisations and details on alignment see Supplementary Material.

**Algorithm 1** An algorithm for patch classification.

**Input:** LDR: $(I_0, I_1, I_2)$
  $\{P_{0i}, P_{1i}, P_{2i}\} \leftarrow patches(I_0, I_1, I_2)$
  *optical flow (OF)* $\leftarrow f(I_j, I_k)$
  **if** $mode(OF\{I_1, I_2, I_3\}) < t_f$ **then**
    $H_{01}, H_{21} \leftarrow \texttt{RANSAC}(I_0, I_1), \texttt{RANSAC}(I_2, I_1)$
    $I_{0W}, I_{2W} \leftarrow warp(I_0, H_{01}), warp(I_2, H_{21})$
    $\{P_{0Wi}, P_{2Wi}\} \leftarrow patches(I_{0W}, I_{2W})$
    **for** $i = 0...N$ **do**
      **if** $(P_{0i}, P_{1i}, P_{2i})$ *static* **then**
        $Y_i \leftarrow \Lambda_0 P_{0Wi} + \Lambda_1 P_{1i} + \Lambda_2 P_{2Wi}$
        *Save:* $((P_{0i}, P_{1i}, P_{2i}),\ Y_i)$
      **else**
        **if** $P_{1i}$ *well-exposed* **then**
          $Y_i \leftarrow P_{1i}$
          *Save:* $((P_{0i}, P_{1i}, P_{2i}),\ Y_i)$
  **else**
    **for** $i = 0...N$ **do**
      **if** $P_{1i}$ *well-exposed* **then**
        $Y_i \leftarrow P_{1i}$
        *Save:* $((P_{0i}, P_{1i}, P_{2i}),\ Y_i)$

Thereafter, we extract patches from the warped images $\{P_{0Wi}, P_{2Wi}\}$ corresponding directly to patches from original images. Further we consider each set of patches separately and classify them as *static* or *dynamic*. We measure the difference of the optical flow magnitude with respect to its median and compare to a given threshold (we set a threshold $t_s$ as a function of the median $m$: $T = max(min(m, 2), 0.5)$ to allow for slightly more lenience with bigger movements). The condition has to be satisfied by all the computed optical flows across all aforementioned pairs of images to consider the patch as static.

If the patch is considered *static*, we perform HDR fusion based on the warped set of patches $(P_{0Wi}, P_{1i}, P_{2Wi})$. Merging is done by computing a weighted combination of patches. We use a triangular weighting scheme similar to Debevec

*et al*. [6] and Kalantari *et al*. [14]. The fusion is done in the linear image domain and we assume a gamma shaped curve for conversion. Finally, we perform a last consistency check to ensure the HDR reconstruction has gone well: we reject static ground truth patches that have a low PSNRs when compared to the well-exposed regions of the input LDR image (see Subsection 3.2) in order to avoid misalignment and ghosting artefacts in the pseudo-labels. The resulting HDR patches alongside the respective input LDR patches constitute LDR-HDR supervision pair $((P_{0i}, P_{1i}, P_{2i}),\ Y_i)$.

For all *dynamic* patches (including patches from highly misaligned images), we check if the reference frame is well-exposed (see Subsection 3.2). If the patch is considered reliably exposed, we use the reference patch directly as the HDR pseudo-label and create a supervision pair $((P_{0i}, P_{1i}, P_{2i}),\ Y_i)$ with real dynamic motion. These examples provide useful real dynamic training pairs that can guide methods on how to align LDR inputs to the HDR reference frame. We show in Figure 3 (left) examples of our proposed HDR pseudo-labels alongside corresponding LDR patches for both static and dynamic stacks of input patches.

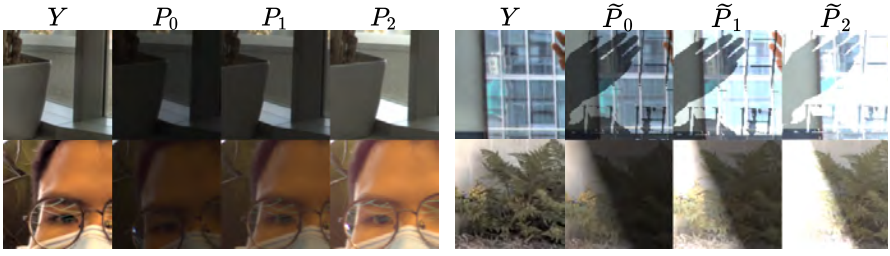



**Fig. 3:** Example of HDR pseudo-label with short, medium, long LDR patches. **Left:** HDR pseudo-label from *static* patch (top), and well-exposed reference patch used as pseudo-label (bottom). **Right:** Original pseudo-label and patches obtained with synthetic illumination and LDR modelling for transfer (top) and synthetic (bottom) masks.

### 3.2 Exposure domain

Intuitively, HDR static fusion weights [6,14] provide higher confidence in areas of the image that are well-exposed given each exposure value. Similarly, we locally select regions of the LDR images that are well exposed and use those images as HDR pseudo-labels. We then synthetically extend their dynamic range so that they provide valuable supervision on the exposure domain. Firstly, for each patch we assess whether it is well-exposed. We perform that via comparison to lower and upper threshold of the fusion triangular functions (*i.e*. pixel illumination values) [14] and consider a patch well-exposed if the majority of pixels lay within the given range. In such a way we reject patches containing over- and under-exposed regions. We consider such well-exposed patches as a reliable estimation for the corresponding HDR patch. Further, we apply a synthetic illumination change to the patch, providing an HDR pseudo-label whose range of values goes beyond that of the original LDR patch, *i.e*. it has a higher range of illumination.

We apply this illumination change as a gain mask over the patch. We consider two alternatives for creating the mask:

1. **transfer** – saturation masks obtained from non-well-exposed patches (selected randomly),
2. **synthetic** – randomly generating a line going through the patch and applying mask increasing progressively towards the patch boundary (resembling the sun falling onto the planar surface).

The value of the gain applied for the masked pixels is adjusted such that a given percentage of pixels within the patch become saturated. These synthetically generated HDR pseudo-labels are re-exposed (and clipped) according to the exposure values of the original LDR input following an image formation model [8,29]. Examples of original well-exposed patches with corresponding sets of generated LDR patches are presented in Figure 3 (right). For more examples of generated supervision pairs (both motion and exposure domains) we refer the reader to the Supplementary Material.

### 3.3 Movement augmentation

To complement and further balance the dynamic and static pseudo-labels we introduce simple motion augmentation. We consider two modes of handheld photo shooting: *pseudo-static*, similar to *e.g.* the camera shake induced by a person taking a picture while trying to keep the camera steady; and *free-moving* when the camera is held freely with no consideration of the movement. We introduce global camera motion as a form of augmenting the generated supervision pairs. For the movement generation we randomly choose values of horizontal and vertical displacement from a Gaussian distribution (mean 0 and given standard deviation for pseudo-static movement, higher mean and deviation for the *free-moving* mode). Pseudo-static movement is applied only to synthetically generated supervision pairs from well-exposed patches as it is assumed to be naturally present in the pairs estimated from static patches. Due to the nature of our approach, most of the high-movement patches are filtered out, and thus we apply larger displacement values to both sets of generated supervision pairs.

### 3.4 Model

Our self-supervision method provides a supervision of LDR image sets paired with respective HDR pseudo-label. Therefore, it can be applied to any existing data-driven approach, *e.g.* any HDR estimation network. In our experiments we have chosen to focus on a simple UNet-based architecture[30], as it is a well-explored backbone, and has proven effective in several image-to-image translation problems [2,11,20]. Our selected UNet architecture contains 4 down- and upsampling modules and is used to predict an HDR residual signal from the stack of channel-wise concatenated inputs. Additionally, a reference input with over- and underexposed regions masked out is passed though a shallow convolutional module to produce a vanilla attention map which is later used to guide

the merging of reference image and residual signal. We use a weighted sum of mean absolute errors for linear and $\mu$-law tonemapped images ($\mathcal{T}(I) = \frac{log(1+\mu I)}{log(1+\mu)}$ where $\mu = 5000$) as a loss function (Equation 1). More details on the network architecture are provided in the Supplementary Material.

$$l(I, GT) = \mathcal{L}_1(I, GT) + \alpha\mathcal{L}_1(\mathcal{T}(I), \mathcal{T}(GT)) \tag{1}$$

## 4 Experiments

We test our proposed approach on the Kalantari *et al.* dataset [14] and the recent datasets introduced and used by Chen *et al.* [4]: **D** - dynamic dataset with ground truth, **DnGT** - dynamic dataset without ground truth, **S** - static dataset, **SRM** - static dataset with random synthetic movement, **HdM2** - 2 sequences from HdM-HDR [7]. We consider video sequences as separate triplets of images with alternating exposures.

All sets of LDR images were processed as described in Section 3. We used patches of size $(128, 128)$, extracted with stride 64. The threshold for the well-exposed values was set to $T_L = 0.125$, $T_H = 0.75$ for under- and overexposure respectively (for images considered in sRGB domain). Additionally, saturation masks from non-well-exposed patches were used only if covered at least 10% of the patch surface. In the motion domain module, an image was considered *globally static* if the dominant value of optical flow magnitude was smaller than $t_f = 15px$. A minimal value of PSNR calculated within well-exposed mask for the *static* patch to be considered as supervision was set to $45dB$. Values of thresholds for considering a patch as well-exposed were set the same as in the exposure domain module. A *pseudo-static* movement augmentation was set to be drawn from Gaussian distribution: $\mathcal{N}(0, 4)$, while for larger movements: $\pm\mathcal{N}(20, 3)$. Thereafter, we obtained 4 subsets from each processed dataset:

1. exposure domain *pseudo-static* - ED,
2. exposure domain with bigger movement - EDM,
3. motion domain *pseudo-static* - MD,
4. motion domain with bigger movement - MDM.

We show in Table 1 a breakdown of the number of supervision pairs generated for each dataset, considering each domain split.

**Table 1:** A summary of the number of supervision pairs generated for different domain splits for various datasets. Presented as a percentage of the total number (in *italic*) of pairs in the given dataset. In **bold** - subsets used for testing.

| Subset | KTr | **KTe** | Sum | **D** | DnGT | S | **SRM** | HdM2 | Sum |
|---|---|---|---|---|---|---|---|---|---|
| ED | 24.11 | 4.47 | 28.58 | 5.49 | 7.27 | 5.14 | 4.57 | 0.54 | 23.02 |
| EDM | 23.18 | 4.16 | 27.34 | 5.61 | 7.15 | 5.00 | 4.39 | 0.52 | 22.67 |
| MD | 18.76 | 3.74 | 22.50 | 3.61 | 9.57 | 6.51 | 6.23 | 0.27 | 26.19 |
| MDM | 17.88 | 3.70 | 21.57 | 3.50 | 11.81 | 6.38 | 6.18 | 0.27 | 28.12 |
| Total | 83.93 | 16.07 | *39382* | 18.21 | 35.80 | 23.03 | 21.37 | 1.59 | *106916* |

### 4.1 Results

We report in Table 2 and 3 the results of our approach compared against other HDR estimation approaches, including weakly-, and unsupervised methods for Kalantari [14] and Chen *et al.* [4] datasets respectively.

**Table 2:** Quantitative comparison of our method against existing approaches on Kalantari *et al.* dataset. Table is split based on the level of supervision in the respective method: left (**S**) - fully-supervised, top right (**WS**) - weakly-supervised, bottom right (**US**) - unsupervised. The best unsupervised score is highlighted in **bold**, the best score overall - underlined. †Values as reported in [33].

| | | Kalantari - KTe | | |
|---|---|---|---|---|
| | | $P_L$ | $P_\mu$ | $HV2$ |
| **S** | AHDR[36] | 41.16 | 43.57 | 64.83 |
| | Kalantari[14] | 41.23 | 42.70 | 64.63 |
| | Wu[35] | 41.62 | 42.01 | 65.78 |
| | Prabhakar [28] | 40.31 | 42.79 | 62.95 |
| | Prab-SV[27] | 41.79 | 41.94 | 65.30 |
| | Ours (Supervised) | 40.83 | 42.39 | 64.20 |
| **WS** | Prab-5s[27] | 41.28 | 41.67 | 65.15 |
| | Prab-1s[27] | 41.03 | 41.22 | 64.61 |
| **US** | Prab-0s[27] | **40.90** | 41.14 | **64.89** |
| | Hu†[10] | 30.84 | 32.19 | 55.25 |
| | Oh†[25] | 27.11 | 27.35 | 46.83 |
| | Sen [31] | 38.38 | 40.98 | 60.54 |
| | Ours | 40.54 | **42.15** | 63.99 |

We test the proposed method on Kalantari *et al.* [14] Test split (KTe), while training on self-supervision pairs from the Training and Test splits (KTr+KTe). Additionally, we provide results on Chen dynamic dataset (D) and static dataset with random synthetic movement (SRM), trained on full data introduced by the authors (D+DnGT+S+SRM+HdM2). We provide values of PSNR, PSNR for $\mu$-law tonemapped images, and HDR-VDP2 [22] averaged across all the images in the dataset. All our results were obtained by training the aforementioned UNet-based model for 300 epochs, setting $\alpha = 0.2$, using Adam optimiser [16] with learning rate $1e-4$, and multi-step scheduler decreasing learning rate by 90% in epochs 210 and 285.

We compare our method against recent self-supervised and unsupervised approaches, *i.e.* Sen [31], Hu [10], Oh [25], Prabhakar [27] in their proposed zero-shot setting. Additionally, we show the results of weakly-supervised method - Prabhakar [27] with the supervision from 5 static, and 1 or 5 dynamic scenes with ground truth labels. Additionally, we compare ourselves to Prabhakar [27] supervised, Chen [4], AHDR [36], Kalantari [14], Wu [35], Prabhakar [28].

We test the proposed method on Kalantari *et al.* [14] Test split (KTe), while training on self-supervision pairs from the Training and Test splits (KTr+KTe). Additionally, we provide results on Chen dynamic dataset (D) and static dataset with random synthetic movement (SRM), trained on full data introduced by the authors (D+DnGT+S+SRM+HdM2). We provide values of PSNR, PSNR for $\mu$-law tonemapped images, and HDR-VDP2 [22] averaged across all the images in the dataset. All our results were obtained by training the UNet-based model

**Table 3:** Quantitative comparison of our method against existing approaches on Chen *et al.* datasets. Table is split based on the level of supervision in the respective method: top - fully-supervised, middle - weakly-supervised, bottom - unsupervised. The best unsupervised score is highlighted in **bold**, the best score overall - underlined.

| | Chen - D | | | Chen - SRM | | |
|---|---|---|---|---|---|---|
| | $P_L$ | $P_\mu$ | $HV2$ | $P_L$ | $P_\mu$ | $HV2$ |
| AHDR[36] | 35.09 | 39.56 | 63.78 | 35.49 | 35.48 | 61.31 |
| Chen[4] | 42.33 | 41.47 | 71.87 | 41.02 | 35.89 | 68.25 |
| Prab-SV[27] | 38.19 | 40.89 | 65.48 | 40.27 | 37.61 | 67.46 |
| Prab-5s[27] | 38.90 | 40.02 | 67.51 | 40.42 | 35.86 | 67.31 |
| Prab-1s[27] | 39.29 | 39.65 | 67.05 | 39.74 | 35.08 | 66.93 |
| Prab-0s[27] | 39.01 | 39.51 | 69.12 | 40.52 | 35.04 | 69.97 |
| Sen[31] | 39.58 | 40.79 | 68.83 | 40.51 | 36.83 | 66.01 |
| Ours | **42.80** | **42.05** | **71.55** | **45.90** | **40.71** | **71.72** |

for 300 epochs, with $\alpha = 0.2$, using Adam optimiser [16] with learning rate $1e{-}4$, and multi-step scheduler decreasing learning rate by 90% in epochs 210 and 285.

We compare our method against recent self-supervised and unsupervised approaches, *i.e.* Sen [31], Hu [10], Oh [25], Prabhakar [27] in their proposed zero-shot setting. Additionally, we show the results of weakly-supervised method - Prabhakar [27] with the supervision from 5 static, and 1 or 5 dynamic scenes with ground truth labels. Additionally, we compare ourselves to Prabhakar [27] supervised, Chen [4], AHDR [36], Kalantari [14], Wu [35], Prabhakar [28].

Our method outperforms other methods that do not require supervision. Note that zero-shot experiment of Prabhakar *et al.* [27] still requires providing 5 completely static scenes. Our approach does not require any assumption on any number of data. Additionally, we achieve a significant improvement over the non-learning approach by Sen *et al.* [31]. Further, our algorithm provides results comparable to those of fully supervised methods. For Chen *et al.* dataset, we do outperform other fully-supervised methods trained on the mentioned dataset ($0.47dB$ improvement on dynamic dataset, and $4.88dB$ on static with random motion over the runner-up). We attribute such a good performance to the ability of our self-supervised method to facilitate in-domain data (D and SRM) which are not presented to the network in supervised methods. We show that our approach provides an HDR estimation of quality superior to other unsupervised methods, and competitive with supervised ones, even though we do not use any annotated data and use a simple, lightweight vanilla architecture. Additionally, we would like to highlight that our proposed self-supervised strategy obtains very close performance to the same architecture trained in a fully supervised setting (Table 3 Ours (Supervised) vs Ours), which further validates the effectiveness of our HDR pseudolabels.

Figure 4 presents a qualitative results of our method on an image from Kalantari *et al.* and an image from Chen *et al.*- Dynamic compared against main methods reported in Table 2. In the red patch of the left image, note the dis-

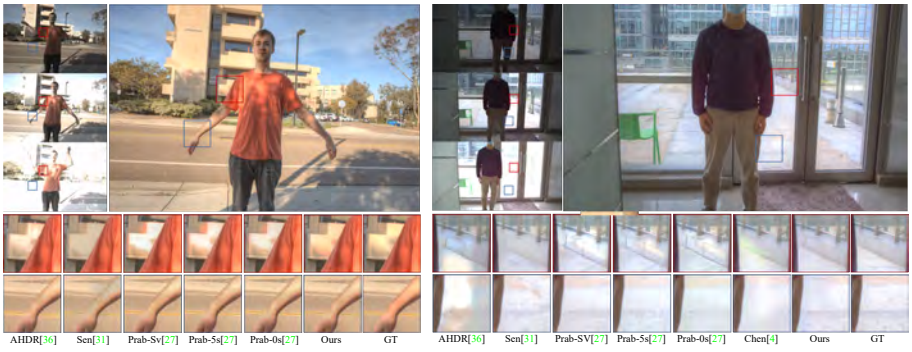



**Fig. 4:** Qualitative comparison of our method against fully-, weakly-, and unsupervised approaches. Left image - example from Kalantari *et al*., note artefacts in colour reconstruction in red, hand discolouration in blue. Right image - example from Chen *et al*. Dynamic, note artefacts on the wall in red, pavement ghosting in blue.

colouration of the building observable in various methods, while reconstructed well in ours - we attribute this to supervision from images with synthetically increased illumination. Similarly, blue-marked patch present skin-tone colour reconstruction other methods are struggling with, leaking the yellow colour from road marks. In the right image, red patch focuses on various artefacts on the highly illuminated in the reference image, wall. Even though some slight ghosting is visible, our method provides a good reconstruction of that region, keeping its colour correct, and uniform. Blue patch accounts for ghosting effects present in most of the methods in that region. We argue that our method and Sen [31] provide the most visually satisfying reconstruction of the pavement plane, with the latter, however, introducing new lines not present in the ground truth.

### 4.2 Ablation study

In Table 4 we emphasise the importance of each module in creating a self-supervision signal for network training. We perform an ablation study by removing a single block from the pipeline in each experiment, and training only on non-augmented versions of motion and exposure domains. It is worth to notice that training in augmented motion domain (MD+MDM) only achieved a very good performance on both Chen *et al*. datasets, whereas training in exposure domain (ED+EDM) performed significantly worse. On the other hand, exposure domain supervision performed almost as good as the whole supervision set for Kalantari *et al*. dataset (both ED, and ED+EDM), while motion domain supervision performance was noticeably degraded. This, in conjunction with full supervision signal achieving the highest results proves the importance of exploiting both domains for self-supervision. Additionally, motion augmentation always improves the results over single ED or MD on dynamic datasets showing that non-augmented splits may not have enough motion samples. Qualitative comparison is also present in Figure 5. We can observe the network struggling

**Table 4:** Ablation results obtained by training with different domain splits of datasets for Kalantari *et al.* and Chen *et al.*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | MD | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ |
| | MDM | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ |
| | ED | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ |
| | EDM | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| KTe | $P_L$ | 36.44 | 39.55 | 38.80 | 37.65 | 40.40 | 40.54 |
| | $P_\mu$ | 36.73 | 41.16 | 38.50 | 37.76 | 41.68 | 42.15 |
| D | $P_L$ | 38.72 | 38.38 | 40.09 | 41.34 | 40.22 | 42.80 |
| | $P_\mu$ | 39.40 | 39.15 | 39.25 | 41.62 | 39.70 | 42.05 |
| SRM | $P_L$ | 45.44 | 42.50 | 46.01 | 43.99 | 41.93 | 45.90 |
| | $P_\mu$ | 41.45 | 35.04 | 40.10 | 41.48 | 34.98 | 40.71 |

with ghosting artefacts when not trained with any exposure domain split. With lack of motion domain split, we notice a decreased capability in fusing the HDR images. We observe a non-motion-augmented training to be the most effective for non-dynamic dataset (Chen *et al.*- static with random movement). However, presenting the network with higher values of misalignment in the images is shown to be important for improving the results on real, dynamic data.

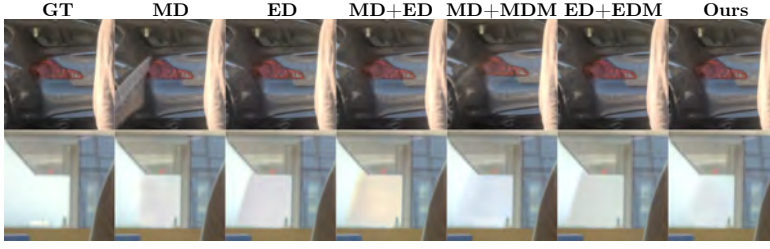



**Fig. 5:** Qualitative results for domain split ablation. **Top:** Note the ghosting artifacts from high illumination long photo, when trained only with motion domain. **Bottom:** Observe inaccuracies of HDR fusion when trained only in exposure domain.

Table 5 focuses on understanding the behaviour of our strategy when using various datasets for training. We can observe a good performance on both Chen *et al.* datasets when training the network only with patches extracted from the respective set of images, or from both Chen *et al.* test datasets combined. The results for Kalantari show, however, a significant drop in performance when trained only on test set patches. We hypothesise that such a worsening of the results might be caused by the reduced size of the dataset, and thus the reduced amount of supervision patches extracted from Kalantari *et al.* Test (KTe, see Table 1). This suggests a high capability of our method when used only within the given set of data, provided a sufficient number of input images. Additionally, the results of training our method with non-target dataset, and in particular full sets of data prove its usefulness in incorporating additional, not expensive to

obtain, unlabelled series of images into the training pipeline. Similarly, performance of the models trained on Chen *et al.* datasets is reported to be higher on Kalantari *et al.* Test than the model trained on the respective dataset. We attribute this behaviour to the larger number of patches extracted from Chen *et al.* datasets. Additionally, this proves usefulness of our method by presenting the transferability between various image domains.

Finally, in Table 6 we present a comparison of results of training with various architectures. We substitute the backbone of our approach (UNet [30]) with (i) a ResNet backbone, as described in Zhang *et al.* [38] (with no dense connections); and (ii) a Grouped Residual Dense Blocks architecture as described in Kim *et al.* [15] (GRDN). We show that the method we propose is suitable to be used with various neural networks. Additionally, we note that increasing the capacity of the network (GDRN characterises with high complexity) yields increase in performance when provided with enough data (number of patches generated from Chen *et al.* is significantly greater than number of patches obtained from Kalantari *et al.* dataset).

**Table 5:** Ablation results obtained by networks trained on different datasets. In **bold** patches only from dataset used for testing.

| Test\Train | | Ours | KTe | D | SRM | D+SRM |
|---|---|---|---|---|---|---|
| KTe | $P_L$ | 40.54 | **37.09** | 38.86 | 38.48 | 38.72 |
| | $P_\mu$ | 42.15 | **37.76** | 41.49 | 41.31 | 40.30 |
| D | $P_L$ | 42.80 | 35.48 | **41.11** | 39.01 | 41.19 |
| | $P_\mu$ | 42.05 | 38.74 | **40.03** | 41.54 | 40.95 |
| SRM | $P_L$ | 45.90 | 36.58 | 39.71 | **44.67** | 45.48 |
| | $P_\mu$ | 40.71 | 34.64 | 35.09 | **40.65** | 39.48 |

**Table 6:** Comparison of quantitative results on Chen[4] D and SRM for various backbone architectures.

| | | UNet | ResNet | GRDN |
|---|---|---|---|---|
| D | $P_L$ | 42.80 | 41.64 | 44.36 |
| | $P_\mu$ | 42.06 | 42.20 | 43.30 |
| | $HV2$ | 71.55 | 70.76 | 73.77 |
| SRM | $P_L$ | 45.90 | 43.34 | 45.97 |
| | $P_\mu$ | 40.71 | 39.93 | 42.39 |
| | $HV2$ | 71.72 | 68.88 | 70.46 |

## 5 Conclusions

In this work we propose a self-supervised approach for HDR Imaging that exploits LDR patches based on its motion and exposure characteristics. Our method provides supervision pairs for data-driven algorithms obtained only from sets of LDR images captured without any constrains. We have shown that the supervision provided by our approach enables training networks with performance superior to other weakly- and unsupervised approaches, and comparable to supervised ones. Our strategy can leverage any image sequence captured with varying exposure to increase the accuracy of HDR estimation, saving a lot of time and effort on capturing training data. Additionally we provide an effective way to incorporate unlabelled in-domain data into the training process.

Our work provides a simple yet effective model for applying synthetic illumination gain to the image. A further step to develop supervision in the exposure domain would be to study the possibility of learning that degradation model, *e.g.* via training a GAN network with a set of unpaired saturated patches.

# Supplementary material

## A Generalisation Experiments

In order to show generalisability of our approach, we performed a comparison of various network architectures used for the HDR reconstruction training and inference (Table 6 and 7). In Table 8 we present the comparison of the number of parameters and number of floating point operations per second (GFLOPs) for a single test image of size 1000x1500 px. The ResNet architecture presents a lower number of parameters, however slightly higher computational complexity than UNet. GRDN is a much more complex architecture than UNet in terms of both number of parameters and complexity.

In Table 7 we report the results of various backbones experiments for Kalantari *et al.* [14] Test as well as both Chen *et al.* [4] splits. In the experiment we can observe a similar performance for all backbone architectures. We believe that these results show the generalisation capability of our framework with respect to various backbone architectures with different number of parameters and complexity. We suggest that an increase of the gap in training performance on Chen *et al.* [4] for GRDN indicates the sensitivity of complex architectures to the cardinality of a dataset (Chen *et al.* yielded significantly more patches than Kalantari *et al.* dataset).

**Table 7:** Comparison of quantitative results on Kalantari [14] and Chen [4] for various backbone architectures.

| | Kalantari - Te | | | Chen - D | | | Chen - SRM | | |
|---|---|---|---|---|---|---|---|---|---|
| | $P_L$ | $P_\mu$ | $HV2$ | $P_L$ | $P_\mu$ | $HV2$ | $P_L$ | $P_\mu$ | $HV2$ |
| Unet | 40.54 | 42.15 | 63.99 | 42.80 | 42.06 | 71.55 | 45.90 | 40.71 | 71.72 |
| ResNet | 40.45 | 42.18 | 63.40 | 41.64 | 42.20 | 70.76 | 43.34 | 39.93 | 68.88 |
| GRDN | 40.81 | 42.31 | 64.36 | 44.36 | 43.30 | 73.77 | 45.97 | 42.39 | 70.46 |

**Table 8:** Number of parameters and GFLOPs comparison for various backbone architectures.

| Architecture | Parameters | GFLOPs |
|---|---|---|
| UNet | 17.3M | 1856 |
| ResNet | 0.8M | 2240 |
| GRDN | 21.9M | 6570 |

**Fig. 6:** Examples of patches generated with synthetic gain applied progressively from a randomly generated line. Top to bottom: short, medium, long, HDR pseudo-label.

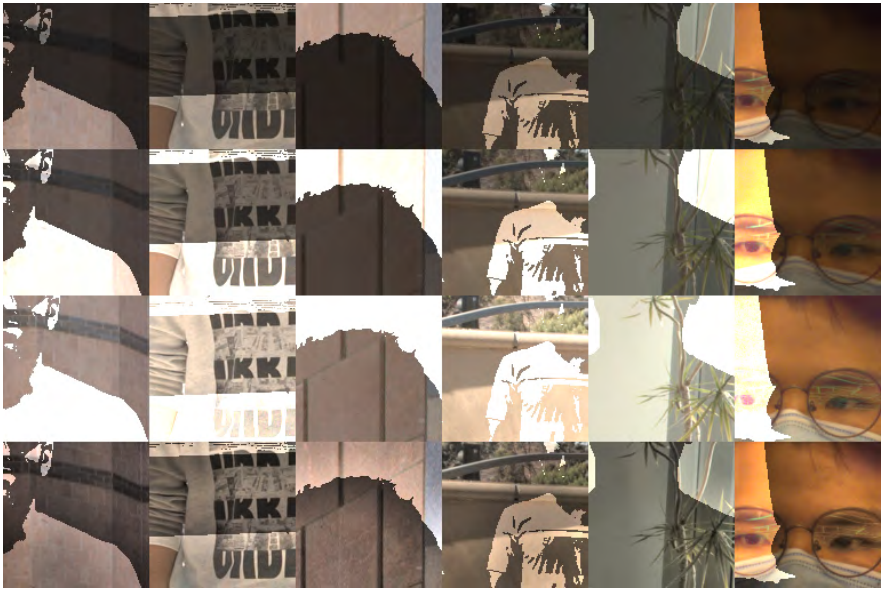


**Fig. 7:** Examples of patches generated with synthetic gain applied by transferring a saturation mask from overexposed patches. Top to bottom: short, medium, long, HDR pseudo-label.

**Fig. 8:** Examples of patches generated as a fusion of static patches. Top to bottom: short, medium, long, HDR pseudo-label.

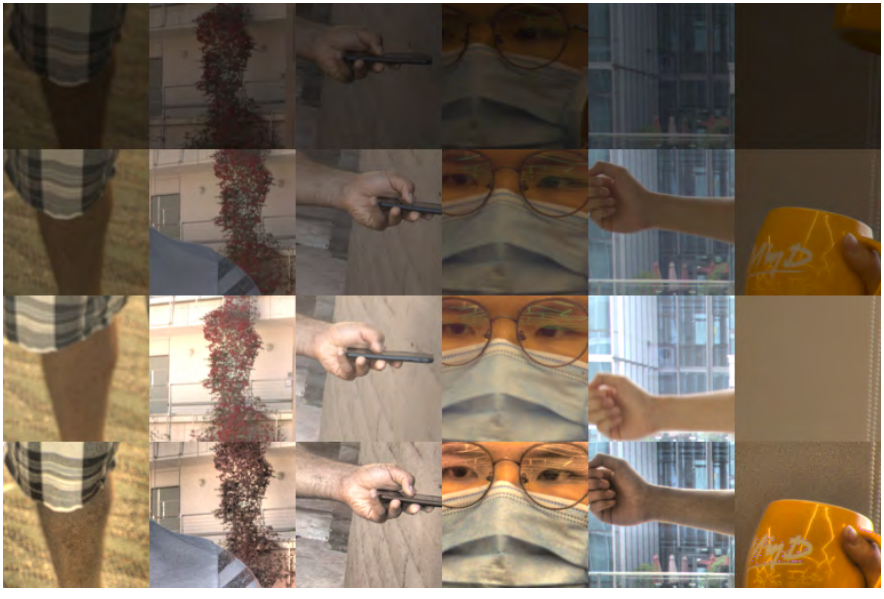


**Fig. 9:** Examples of patches where reference frame is reused as HDR. Top to bottom: short, medium, long, HDR pseudo-label.

## B Extended Visualisations

We provide additional examples of patches generated with our approach. Figure 6 presents a set of images generated with synthetic gain mask applied progressively from a randomly generated line. Figure 7 corresponds to supervision pairs in which the synthetic gain mask is transferred from saturated patches. Figure 8 shows examples of patches generated via the fusion of static patches. Figure 9 presents patches with reference frame reused as HDR supervision. Note that all HDR pseudo-labels were tonemapped for visualisation.

Figure 10 presents additional qualitative results of our method, including examples obtained in generalisation experiment (GRDN), unsupervised approach from Prabhakar [27] (zero-shot setting), and supervised methods: Chen *et al.* [4], and AHDR [36].

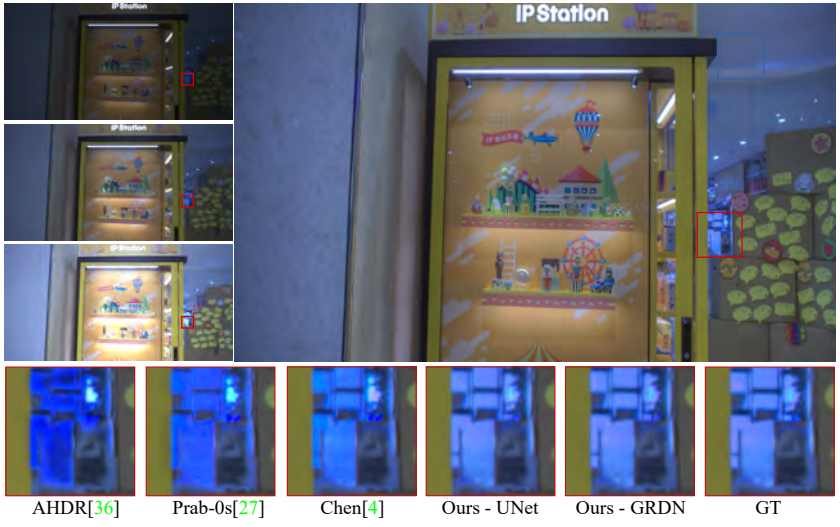
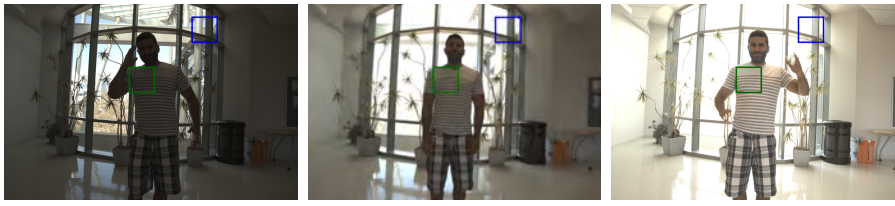



**Fig. 10:** Qualitative comparison of our method against fully-, and unsupervised approaches, including generalisation experiment. Example from Chen *et al.*

## C Implementation details

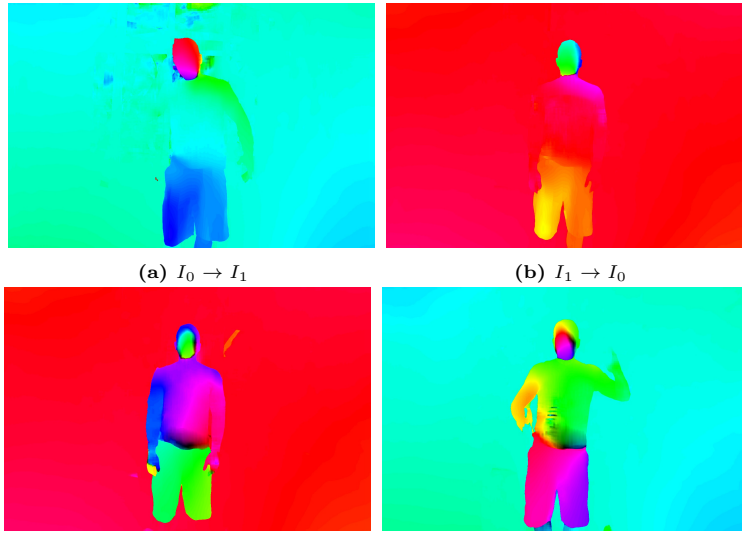
### C.1 Motion Domain

In this section we present a detailed explanation of patch classification with respect to motion cues. Firstly, we consider a set of images captured with varying exposure times - $(I_0, I_1, I_2)$, where $I_1$ is the reference frame. An example of such sequence is presented in Figure 11 (note patch colouring, consistent across this section). We calculate optical flow values for all combinations of images that include a reference frame, *i.e.* $I_1 \rightarrow I_0$, $I_1 \rightarrow I_2$, $I_0 \rightarrow I_1$, $I_2 \rightarrow I_1$. A visualisation of optical flow maps is presented in Figure 12. In the next step, we



**(a)** Short. **(b)** Medium. **(c)** Long.

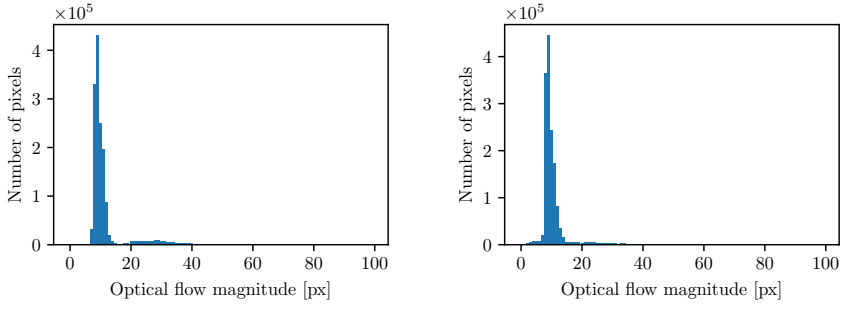
**Fig. 11:** Example of image sequence used for detailed description of patch classification based on motion cues.



**(a)** $I_0 \to I_1$ **(b)** $I_1 \to I_0$

**(c)** $I_1 \to I_2$ **(d)** $I_2 \to I_1$

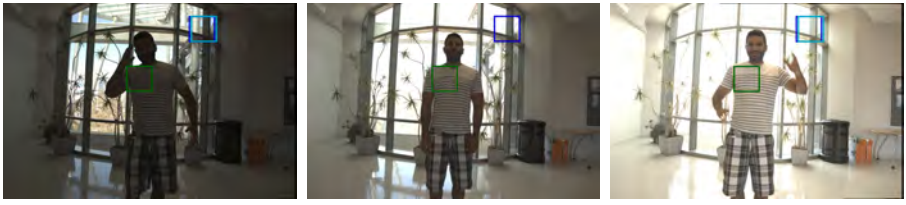
**Fig. 12:** Optical flow map visualisation with *uv* coordinates mapped to angle as value and magnitude as colour in HSV space.




**(a)** $I_1 \to I_0$ **(b)** $I_1 \to I_2$

**Fig. 13:** Histograms of optical flow magnitude with respect to the reference frame for the considered example.

calculate histograms with $1px$ bin width of the magnitude of optical flow for the optical flows calculated with respect to reference frame ($I_1 \rightarrow I_0$, $I_1 \rightarrow I_2$). Such histograms are used to consider whether the images can be aligned based of the threshold discussed in D.1. Histograms obtained for the optical flows related to the given example are presented in Figure 13.



**(a)** Short. **(b)** Medium. **(c)** Long.

**Fig. 14:** Image sequence after applying the warping. Pale blue patch corresponds to the warping of dark blue patch.

If the dominant value in the histogram is higher that the threshold $t_f$, then image is considered to have high misalignment which is not corrected, and its patches are considered as in section 3.1 of main manuscript (*dynamic* patches). For other images, we consider the optical flow estimated displacements to correct the perspective. However, we are not trying to correct dynamic behaviour of the objects in the scene, only camera movement. Therefore, we consider only pixel displacements of dominant and neighbouring bins, and use `RANSAC` to estimate homography matrices between $I_0$ and $I_1$ – $H_{01}$, and $I_2$ and $I_1$ – $H_{21}$. Further we apply the homographies to non-reference images to obtain aligned images - $I_{0W}$ and $I_{2W}$. Figure 14 presents given example after warping - note that corresponding patches in warped images are aligned, whereas, in original input, alignment is not guaranteed (in fact, extremely rare). Further, we consider images on the patch level - a set of patches $\{P_{0i}, P_{1i}, P_{2i}, P_{0Wi}, P_{2Wi}\}$ coming from $I_0, I_1, I_2, I_{0W}, I_{2W}$ respectively. For each patch, we test whether its content is static. Given the patch set, we consider the optical flow magnitude in the corresponding region. The magnitude of the optical flow has to not deviate from its median more than a threshold across the whole patch. The threshold is based on the value of the median $m$: $T = max(min(m, 2), 0.5)$ - it allows for a bigger deviation with bigger displacements. The condition for the optical flow magnitude has to be satisfied for the optical flow calculated between all aforementioned pairs of images to consider patch as *static*.

*Static* patches are processed as described in Section 3.1 of the main manuscript. Figure 15 presents a set of warped input patches with resulting HDR pseudo-label for the example given in Figures 11 and 14. Similarly, supervision pair corresponding to the same set of patches is shown in Figure 16. A supervision pair for a *dynamic* patch for the given example is presented in Figure 17

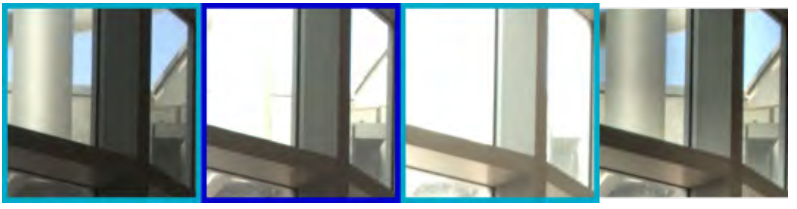


**Fig. 15:** HDR estimation for the *static* patch example. Left to right: short warped, medium, long warped, HDR estimation - tonemapped.

**Fig. 16:** Supervision pair from *static* patch. Left to right: short, medium, long, HDR estimation - tonemapped. Note: input supervision patches are not warped.

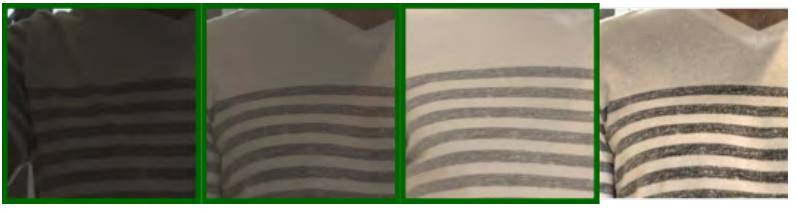


**Fig. 17:** Supervision pair from *dynamic* patch. Left to right: short, medium, long, HDR estimation - tonemapped.

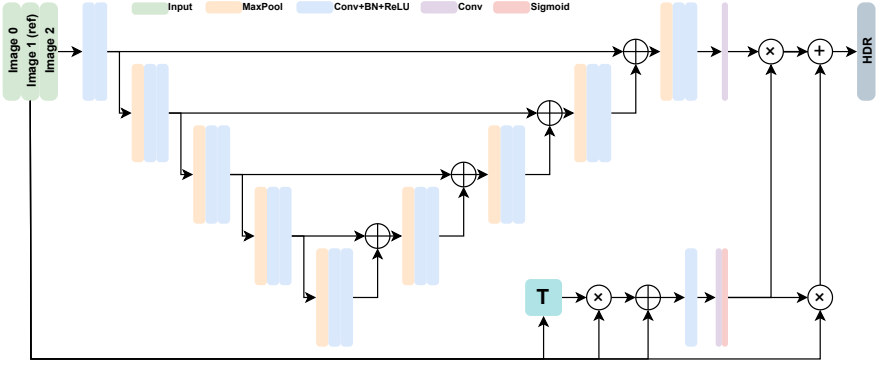



**Fig. 18:** Detailed overview of used UNet architecture.

### C.2 UNet Architecture

In our experiments we used a UNet [30] based model. A detailed overview is presented in Figure 18. We consider a triplet of LDR images as 9 channel input to a regular UNet network that predicts a residual signal to incorporate higher dynamic range in the reference frame. Additionally, we extract a well-exposed mask from a reference image, and input a concatenation of the mask, and masked reference frame (4 channels) to a shallow convolutional branch that predicts blending weights. Finally, HDR estimation is obtained by fusing reference frame with residual signal according to predicted weights.

### C.3 Source Code

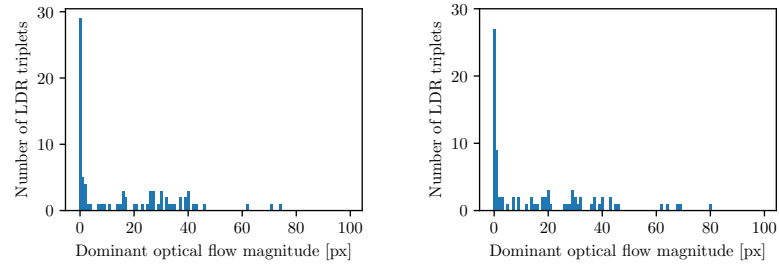
We intend to release the source code together with trained network weights to reproduce the presented results subject to an ongoing internal review and approval.

## D Hyper-parameters

### D.1 Static camera threshold

In our experiments we have chosen 15 pixels to be the threshold for considering the image alignment viable for correction ($t_f$). The threshold is computed based on the dominant value of optical flow magnitude for the given triplet (considering optical flows with respect to the reference frame - see Figure 13). The motivation for choosing a threshold value is to eliminate big movements of camera, for which estimating the homography transformation may lead to imprecise HDR fusion. To provide more insight, we report histograms of dominant value of optical flow magnitude for datasets used in our experiments - see Figure 19. We observe close to unimodal distribution for Chen [4] datasets suggesting that they were collected in *pseudo-static* manner. With such a distribution it is desirable to correct the alignment of most of the LDR samples. On the other hand, Kalantari [14] dataset characterises with close to bimodal distribution of camera movements with a strong peak in close to 0 values and a small increase in values from 20 to 40. In this case, we want to set the threshold to categorise most of the second mode as not suitable for alignment. Therefore, we have chosen to consider 15 pixels as a good threshold that separates modes of Kalantari distribution, whereas preserves almost all Chen LDR images.

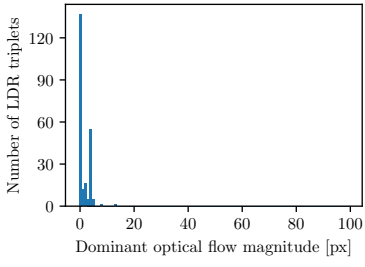
Similar reasoning was used in choosing the amount of movement augmentation - drawn from the following Gaussian distributions: *pseudo-static* - $\mathcal{N}(0, 4)$, bigger movements: $\pm\mathcal{N}(20, 3)$. Note that *pseudo-static* augmentation ensures the augmentation within alignment threshold and corresponds to the main mode of movement magnitudes across the images (Fig. 19). Values corresponding to bigger movement lie outside the alignment threshold and try to account for patches that may be discarded in the process of generation within motion domain.



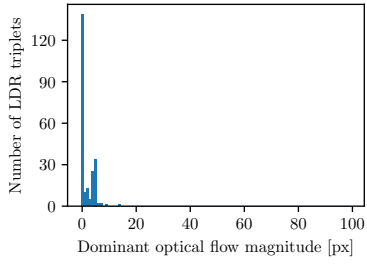

**(a)** $I_1 \rightarrow I_0$ - Kalantari [14]




**(b)** $I_1 \rightarrow I_2$ - Kalantari [14]




**(c)** $I_1 \rightarrow I_0$ - Chen [4]




**(d)** $I_1 \rightarrow I_2$ - Chen [4]

**Fig. 19:** Histograms of dominant optical flow values in Kalantari [14] dataset and Chen [4] - dynamic, and static with random movement datasets.

### D.2 Generalisability

We note that the choice of hyper-parameters is not fine-tuned to any particular dataset but rather inspired by intuition from the data collection process and previous literature. For the motion domain we consider optical flow threshold similar to values observed for a static, hand-held camera. In the exposure domain the range of illumination for well-exposed regions is inspired by the triangular function weights used in early model-based approaches whose underlying principle is an increasing confidence for pixel values closer to 0.5 [6, 14]. Our hyper-parameters are not fine-tuned but rather fixed for all sub-datasets (Kalantari, D, SRM, HdM2, DnGT, S) used in the paper, hinting robustness and little data-dependency.

## E Future Work

In this section we extend further the main manuscript's discussion with respect to limitations of our proposed approach and possible future work avenues which we could not fit within the page limit.

**Camera Response Function**: In our approach we assume that a gamma transformation can approximate accurately enough the Camera Response Function (CRF) necessary to bring images into *exposure alignment* in the linear

domain. This holds true for the most commonly used multi-frame HDR datasets included in the paper (the Kalantari *et al*. [14] and Chen *et al*. [4] datasets). For the case where CRF is arbitrarily complex and unknown, we would need to estimate the CRF, similarly to *e.g*. [39].

**Noise Modelling**: We note that carefully collected dataset presented by Chen *et al*. [4] and Kalantari *et al*. [14] contain images charaterised with negligible amount of noise. Therefore, we do not include a noise model in the process of patch classification and LDR modelling. When working with highly noisy data, patch classification process would have to be adjusted to match the noise level between generated pseudo-labels and source images. Additionally, LDR modelling process would account for noise parameters as in [8].

**Motion Modelling**: In our work we introduce the use of motion augmentation in order to further increase and balance the number of training examples where there is misalignment among input LDR frames, as this is crucial to learn LDR-to-HDR mappings for dynamic scenes. We use for this purpose translation-only motion, as despite its simplicity it has proven to be effective. Other more sophisticated transformations could be explored, such as a perspective transform (where the transformation parameters are fitted to represent the target motion distribution), or motion transfer as recently done in [27].

**Weak supervision and pre-training**: In our manuscript we explore the full self-supervision scenario as we believe this is the most novel and interesting aspect of our work, however we argue that our approach could be easily used as a self-supervised "pre-training" step that can be later on fine-tuned on available annotated data (*i.e*. weakly supervised), and will explore in the future the impact of size on both unlabelled and labelled datasets to have a deeper understanding of the dynamics between weakly and self-supervised set-ups within our proposed strategy.

## References


1. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: Int. Conf. Learn. Represent. (2020) 4
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017) 8
3. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Eur. Conf. Comput. Vis. (2012) 6
4. Chen, G., Chen, C., Guo, S., Liang, Z., Wong, K.Y.K., Zhang, L.: HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In: Int. Conf. Comput. Vis. (2021) 2, 3, 4, 9, 10, 11, 14, 15, 18, 22, 23, 24
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (2020) 4
6. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Proc. Conf. on Computer Graphics and Interactive Techniques. SIGGRAPH (1997) 2, 7
7. Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., Brendel, H.: Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In: Proc. of SPIE (2014) 2, 9
8. Hasinoff, S.W., Durand, F., Freeman, W.T.: Noise-optimal capture for high dynamic range photography. In: IEEE Conf. Comput. Vis. Pattern Recog. (2010) 8, 24
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020) 4
10. Hu, J., Gallo, O., Pulli, K., Sun, X.: Hdr deghosting: How to deal with saturation? In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1163–1170 (2013) 10, 11
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. IEEE Conf. Comput. Vis. Pattern Recog. (2017) 8
12. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.I.: Learning to estimate hidden motions with global motion aggregation. In: Int. Conf. Comput. Vis. (2021) 6
13. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **43**(11) (2021) 4
14. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. on Graphics (Proc. of SIGGRAPH) **36**(4) (2017) 2, 3, 4, 7, 9, 10, 11, 15, 22, 23, 24
15. Kim, D.W., Chung, J., Jung, S.W.: GRDN: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 2086–2094 (2019) 14
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR (2015) 10, 11
17. Laine, S., Karras, T., Lehtinen, J., Aila, T.: High-quality self-supervised deep image denoising. Adv. Neural Inform. Process. Syst. **32** (2019) 4
18. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T.: Noise2noise: Learning image restoration without clean data. In: International Conference on Machine Learning (2018) 4

19. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009) 3
20. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Eur. Conf. Comput. Vis. (2018) 8
21. Mann, S., Ali, M.: Chapter 1. In: Dufaux, F., Le Callet, P., Mantiuk, R.K., Mrak, M. (eds.) High Dynamic Range Video. Academic Press (2016) 1
22. Mantiuk, R., Kim, K., Rempel, A., Heidrich, W.: HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph. **30**(4) (2011) 10
23. McGuire, M., Matusik, W., Pfister, H., Chen, B., Hughes, J.F., Nayar, S.K.: Optical splitting trees for high-precision monocular imaging. IEEE Computer Graphics and Applications **27**(2) (2007) 2
24. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: Pacific Conference on Computer Graphics and Applications (2007) 2
25. Oh, T.H., Lee, J.Y., Tai, Y.W., Kweon, I.S.: Robust high dynamic range imaging by rank minimization. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1219–1232 (2015) 10, 11
26. Prabhakar, K.R., Agrawal, S., Singh, D.K., Ashwath, B., Babu, R.V.: Towards practical and efficient high-resolution HDR deghosting with CNN. In: Eur. Conf. Comput. Vis. (2020) 3
27. Prabhakar, K.R., Senthil, G., Agrawal, S., Babu, R.V., Gorthi, R.K.S.: Labeled from unlabeled: Exploiting unlabeled data for few-shot deep HDR deghosting. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021) 2, 4, 10, 11, 18, 24
28. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: IEEE International Conference on Computational Photography (2019) 2, 10, 11
29. Pérez-Pellitero, E., Catley-Chandar, S., Leonardis, A., Timofte, R., et. al.: NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (2021) 8
30. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI (2015) 8, 14, 22
31. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based HDR reconstruction of dynamic scenes. ACM Transactions on Graphics (TOG) **31**, 1 – 11 (2012) 10, 11, 12
32. Tocci, M.D., Kiser, C., Tocci, N., Sen, P.: A versatile HDR video production system. In: SIGGRAPH (2011) 2
33. Wang, L., Yoon, K.J.: Deep learning for HDR imaging: State-of-the-art and future trends. IEEE Trans. Pattern Anal. Mach. Intell. (2021) 3, 4, 10
34. Ward, G.: Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. Journal of Graphics Tools **8**(2) (2003) 2
35. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Eur. Conf. Comput. Vis. (2018) 2, 3, 10, 11
36. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) 2, 3, 10, 11, 18
37. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Trans. Image Process. **29** (2020) 3
38. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018) 14

39. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image HDR reconstruction by learning to reverse the camera pipeline. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)