# HDR video synthesis by a nonlocal regularization variational model

Antoni Buades[1], Onofre Martorell[1,*], Ivan Pereira-Sánchez[1]

*Dept. Mathematics and Computer Science and IAC3, Universitat de les Illes Balears, E-07122 Palma, Spain.*

**Abstract**

High dynamic range (HDR) video synthesis is a very challenging task. Consecutive frames are acquired with alternate expositions, generally two or three different exposure times. Classical methods aim at registering neighbouring frames and fuse them using image HDR techniques. However, the registration often fails to obtain accurate results and the fusion produces ghosting artifacts. Deep learning techniques have recently appeared imitating the structure of existing classical methods. The neural network is intended to estimate the registration function and choose the fusion weights. In this paper, we propose a new method for HDR video synthesis using a variational model. The proposed model uses a nonlocal regularization term to combine pixel information from neighbouring frames. The obtained results are competitive with state-of-the-art. Moreover, the proposed method gives a more reliable and understandable solution than deep-learning based ones.

*Keywords:* HDR video synthesis, variational methods, nonlocal regularization

## 1. Introduction

The fusion of images of the same scene acquired with different exposure

*Corresponding author

*Email addresses:* `toni.buades@uib.es` (Antoni Buades), `o.martorell@uib.cat` (Onofre Martorell), `i.pereira@uib.cat` (Ivan Pereira-Sánchez)

[1]The authors acknowledge the Ministerio de Ciencia, Innovación y Universidades (MCIU), the Agencia Estatal de Investigación (AEI) and the European Regional Development Funds (ERDF) for its support to the project TIN2017-85572-P.

permits to enhance the dynamic range of the image. In order to do so, High Dynamic Range (HDR) imaging methods [1, 2] need to combine the radiance values of the pixels. First, the camera response function (CRF) has to be estimated, which is generally achieved using the method proposed by Devebec and Malik [3]. These radiance values are no longer limited by the 8-bit restriction of general image formats, neither is their combination which permits to merge dark and bright areas. Once the HDR image is built, an extra step is necessary to quantize this output into a fixed number of bits fixed by the visualization, storing device or file format. This last step is known as tonemapping [4, 5].

When the camera is not fixed, e.g hand-held acquisitions, or the objects in the scene move, a direct per pixel combination creates ghosting artefacts. Thus, the sequence of images has to be pre-registered or their combination needs to take into account such a non-static nature [6, 7].

The creation of HDR video is still more challenging than HDR imaging. Although there exists cameras which can record videos with a dynamic range of colors, they require specific hardware that has large costs, which reduces its use [8]. In practice, videos are recorded acquiring frames with alternate exposures (generally 2 or 3). Later, HDR video is synthesized with an offline algorithm.

Despite the multi-image nature of HDR imaging, its extension to video sequences with alternating exposure is not straightforward. Since only two or three different exposures are available, many regions will appear only under or over exposed. For video, we need an HDR version of each frame, which prevents the choice of a middle-exposed image as reference, as it is the case of single-image HDR. There exist several approaches by either using classic or deep learning methods [9, 10]. For all, registration is an important step and their result depends critically on it.

In this paper, we propose a new method for HDR video synthesis by using a variational method. We use a nonlocal regularization term to combine pixel information, either from the same or from neighbouring frames. Motion estimation is used to compensate the search areas of neighbouring frames from which pixels are combined. The patch distance taking part in the weight computation reduces the dependence on the estimated motion. It also reduces the ghosting artefacts.

The use of a variational method allows us to model the HDR video synthesis in a simple and intuitive manner, without the need of deep neural networks and large datasets. There is no extensive and time consuming training or need of retraining whenever the processed data differs from the used for training. Moreover, the results are more reliable and interpretable since the method depends on a few understandable parameters.

The remaining of the paper is organized as follows: in Section 2 we make a review of existing HDR video methods. In Section 3 we present the proposed method. In Section 4 we compare our method with state-of-the-art. Finally, we state the conclusions and future work in Section 5.

## 2. Related work

A common step in all HDR methods is to estimate the radiance of the sequence by inverting the CRF using Devebec and Malik [3] and dividing by the exposure time of each frame. This step permits to compare and combine values from different frames.

There is an extensive literature for HDR imaging, we refer the reader to the comprehensive reviews [11, 12]. In the rest of the section, we focus on video HDR methods. The literature on video HDR imaging is much scarcer.

Due to occlusions, large motion and saturated areas, frame alignment is very challenging. For that reason, ghosting removal, known as deghosting, is the main objective of most proposed methods for video HDR. For the majority, the process applies to a reference frame and makes use of its neighbouring ones. The same procedure is repeated for all frames of the sequence.

**Classical methods.** Kang et al. [13] were the first to propose a HDR method for video. The reference and neighbouring frames are registered using a global transformation and a gradient based optical flow method. The aligned frames are merged using a weighted average that takes into account exposition and errors from optical flow. Many posterior methods share the same general structure but refine the aligning and merging strategies.

Mangiat and Gibson [14] observed that the use of optical flow methods is not sufficient to eliminate all ghosting effects. They proposed to use instead block matching methods for registering neighbouring frames to the reference one. However, the proposed registration introduces blocking effects that need to be corrected a posteriori. These are removed with a cross-bilateral filter at the tonemapped image. In a posterior work [15], the same authors improved the a posteriori correction by locally setting the filtering strength depending on the computed registration function. They increased the filtering magnitude at large motion vectors, which are more likely to be incorrect. However, the strong filtering ends removing many details and texture, making the final result look unnatural. Posterior algorithms try to minimize the dependence on the initial registration, in order to reduce the blurring by Kang et al. [13] or blocking artifacts by [15].

Kalantari et al. [16] proposed an optimization method based on the previous HDR imaging work by Sen et al. [17]. The energy to be optimized uses a temporal similarity measure to get information from forward and backward frames. The computed alignment, global and optical flow, is used only to compensate the search areas for the defined similarity measure. Their algorithm not only creates the HDR version, but the missing exposures for each frame. The proposed measure forces a temporal similarity between the current frame and neighbouring ones for all the reconstructed expositions. In particular, it demands for each patch of a reconstructed exposure that a similar patch is found in the previous and posterior frames. The proposed approach correctly avoids the creation of blocking effects and has a better texture and detail reconstruction than previous approaches, but still, it blurs many details. The same approach was adapted by Gryaditskaya et al. [18] to be applied to sequences where the exposure time is adjusted at each frame in order to reduce motion artifacts.

More recent methods have focused in the improvement of the alignment step [19, 20]. Li et al. [19] first separate the foreground and background areas of the image by a multi-scale regression and rank minimization method. Such a separation is used to estimate the motion between frames and reconstruct the HDR. Van Vo and Lee [20] divide the motion estimation step into two phases: on one hand, they perform optical flow estimation of well-exposed areas in a descriptor domain. On the other hand, they perform a superpixel-based motion estimation on poorly exposed areas. This latter estimation uses the optical flow field of non consecutive frames. That is, the algorithm is only valid for sequences with two alternating exposures.

**Learning methods.** There exists an increasing literature in learning based HDR imaging [7, 21, 22, 23, 24, 25] (a wide review can be found on [12]). Niu et al. [7] propose a GAN network. Yan et al. [21] uses an attention module to eliminate the adverse effects of misalignment and saturation. Yan et al. [22] introduces a non-local module to capture global features from the images. Song et al. [23] present a transformer-based medthod. Prabhakar et al. [25] propose a recurrent network.

However, there exist very few learning methods for HDR video. Kalantari and Ramamoorthi [9] were the first ones to propose a learning-based method for HDR video synthesis. They first use a convolutional neural network (CNN) to estimate the optical flow. Afterwards, a second CNN estimates the weights to combine the warped frames. Anand et al. [26], proposed a modification of Kalantari et al., which adds a denoising module at the beginning of the architecture and uses a generative adversarial strategy.

Chen et al. [10] aim at improving Kalantari et al. [9] results by adding another module at the end of the proposed architecture. This new network takes as input the HDR results for 3 consecutive frames and outputs a refined version of the central one. This is the state of the art in video HDR.

Other neural network methods have been proposed for very related problems to HDR video. Kim et al [27] creates a well exposed video from a gradual flash sequence. A video sequence is recorded with a very high frame rate at the same time that the flash is being lighted on and off creating a sequence with 4 different lighting conditions. Cogalan et al [28] shows that video HDR can be achieved by a dual-exposure sensor, which acquires each frame with two different exposures being spatially interleaved in a single image.

**HDR Measures** There exist several measures to evaluate the performance of HDR methods. Although the most of them are intended for HDR imaging, they can be used in HDR video, by sequentially applying them to each frame. The measures can be classified depending on whether they require a ground truth or not. Among the ones using a ground truth, the most relevant and widely used are HDR-VDP-2 [29] and HDR-VQM [30]. The HDR-VDP-2 metric was proposed by Mantiuk et al. [29] and is intended to compute a visual difference image of the ground truth and algorithm's result and calculates a metric that measures the quality of the result based on this difference. Narawaria et al. [30] proposed a measure for video HDR assessment. The video quality is computed based on a spatio-temporal analysis that relates to human eye fixation behavior during video viewing.

Among the ones not requiring any reference, Tursun et al. [31] propose a metric for evaluating ghosting artefacts. It compares the input globally registered images with the HDR result. Karajuzovic-Hadziabdic et al. [32] propose a database for ghosting evaluation and uses the UDQM measure.

## 3. Proposed method

We propose a novel video HDR algorithm under the form of a variational minimization. We use non-local regularization which is commonly used for variational methods in noise removal, deblurring, super-resolution [33, 34, 35], but it is used for the first time in HDR synthesis.

The classical gradient regularization is replaced by a probability distribution which defines the similarity of each pixel with its neighbouring ones. This similarity between two pixels writes as a decreasing function of the distance of patches centred in them.

A fidelity term demands to preserve radiance values for pixels which are well exposed. The regularization term permits to propagate the HDR value of these well exposed pixels to the similar pixels in the same frame and neighbouring ones.

Since registration and alignment is a very challenging task for HDR videos, the estimated transform is used only to compensate the search window in which the probability distribution is defined. The weight similarity depending on patch distance makes the distribution to be robust to optical flow inaccuracies, choosing the correct pixels in these compensated areas. Thus, additionally reducing the ghosting artifacts.

The proposed method presents several novelties that distinguish it from previous works:

- It is the first variational method to use a non local regularization term for HDR synthesis.
- It jointly synthesizes three consecutive HDR frames, increasing the temporal stability of the method.
- It uses patch distances in order to weight the similarity of pixels making the method robust to optical flow errors, thus reducing the ghosting artefacts.
- It obtains results comparable to state of the art deep learning methods, while using a simple and understandable model.
- The regularity term makes the model robust to noise and indeed reduces it notably.

### *3.1. Variational formulation*

The proposed method takes as input a multi-exposure video sequence $\{L_t \mid t = 1, \ldots, N\}$, each frame with exposure $e_t$. In order to compute the HDR of a certain frame $L_t$ we will use its previous and posterior frames, $L_{t-1}$ and $L_{t+1}$ respectively. The current approach can be modified to take into account a temporal neighbourhood of any length.

Let $H_i,\ i \in \{t-1, t, t+1\}$ denote the corresponding radiance frames and $S_i$ the expected HDR frame outputs. Although we will compute the triplet of HDR images, we will only keep as output the one corresponding with the central frame $S_t$.

We propose to minimize a variational energy. In a continuous setting, all the images are defined on $\Omega \subset \mathbb{R}^2$, usually a rectangle, being $\mathbf{x} = (x_1, x_2), \in \Omega$ spatial coordinates on the domain of the image. The proposed variational energy writes as

$$\begin{aligned} J(S_{t-1}, S_t, S_{t+1}) = \sum_{i=t-1}^{t+1} \sum_{c=1}^{3} \int_\Omega ||\nabla_\omega S_i^c(\mathbf{x})||_2 \, \mathbf{dx} \\ + \sum_{i=t-1}^{t+1} \frac{\alpha_i}{2} \sum_{c=1}^{3} \int_\Omega h(L_i^c(\mathbf{x}))\, (S_i^c(\mathbf{x}) - H_i^c(\mathbf{x}))^2 \mathbf{dx}. \end{aligned} \tag{1}$$

where $\alpha_i = \alpha(e_i)$ are the tradeoff parameters between the terms of the functional, $e_i$ indicates the exposure time of frame $i$ and $\alpha(\cdot)$ is an increasing function. The index $c$ indicates the color channel, and $h(L)$ is a weighting function that benefits well-exposed pixels and penalizes under or over-exposed ones,

$$h(L) = \begin{cases} 0 & L \le l_2 \text{ or } L \ge h_2 \\ 1 - \frac{l_1 - L}{l_1 - l_2} & l_2 \le L \le l_1 \\ 1 & l_1 \le L \le h_1 \\ 1 - \frac{L - h_1}{h_2 - h_1} & h_1 \le L \le h_2 \end{cases}. \tag{2}$$

The function $h(L)$ discards pixels for which the color value is below a dark threshold $l_1$ or above a bright threshold $h_2$ (see Figure 1, left). The radiance of well exposed pixels ($h(L) = 1$) will be kept and influence the synthesized HDR value of similar ones.

Each fidelity term, associated to a different image of the temporal neighborhood, has a different weight $\alpha_i = \alpha(e_i)$. This permits to decrease the importance of short exposure frames which have a low signal to noise ratio and quantization effects in the dark parts. The fidelity term applies independently for each channel in order to avoid discarding a particular channel value when the others are saturated. This is crucial in order to recover the value of saturated parts.

The nonlocal gradient $||\nabla_\omega S_i^c(\mathbf{x})||_2$ is defined for a particular frame $i$ and pixel $\mathbf{x} = (x_1, x_2)$ as

$$||\nabla_\omega S_i^c(\mathbf{x})||_2 = \sum_{j=t-1}^{t+1} \sqrt{\int_\Omega \omega_{i,j}(\mathbf{x},\mathbf{y})(|S_i^c(\mathbf{x}) - S_j^c(\mathbf{y})|)^2 \mathbf{dy}}. \tag{3}$$

being $j$ the index of the neighboring images being involved and $\mathbf{y} = (y_1, y_2)$.

### 3.2. Nonlocal weights

For a fixed pixel $\mathbf{x}$ belonging to a frame $i$, the family of weights $\omega_{i,\cdot}(\mathbf{x},\cdot)$ favours those neighbouring pixels having a similar radiance image neighborhood, that is, a weight $\omega_{i,j}(\mathbf{x},\mathbf{y})$ will be higher when the patches centered at pixels $\mathbf{x}$ in radiance image $i$ and $\mathbf{y}$ in radiance image $j$, respectively, look similar. Also, the family of weighs takes into account the motion between images and the saturation of pixel colors.

In order to compute the optical flow between images $L_i$ and $L_j$, as in [36], we first photometrically calibrate the color values of the darker image to look alike the brighter one and then both are converted to grayscale. This calibration is carried out by using the method proposed by Martorell et al. [6]. The proposed photometric calibration computes a joint histogram between the globally registered images using [37] and looks for the color increasing transformation curve passing near the peaks of that histogram. The flow is then computed with the optical flow algorithm from Brox et al. [38] with weights $\gamma = 0.3$ and $\alpha = 9$.

For a pixel $\mathbf{x}$ from frame $i$, let $P$ be the squared window centered at $\mathbf{x}$ and radius $r$. We compare $P$ with other patches of the same size located in a spatial neighbourhood of $(2R + 1) \times (2R + 1)$ pixels at frame $i$ and the motion compensated neighbourhoods in the other two frames.

For each frame $j \in \{t-1, t, t+1\}$, we compute the set of neighbouring pixels as

$$\mathcal{N}_{\mathbf{x}}^{j} = \{\mathbf{y} = \mathbf{x} + \mathbf{u}_{i,j}(\mathbf{x}) + (s_1, s_2) \mid \; -R \leq s_1, s_2 \leq R\} \tag{4}$$

being $u_{i,j}$ the estimated flow between frames $i$ and $j$. The corresponding set of neighbouring patches in frame $j$ is

$$\mathcal{N}_{P}^{j} = \{Q_{\mathbf{y}} \text{ centered at } \mathbf{y} \mid \mathbf{y} \in \mathcal{N}_{\mathbf{x}}^{j}\}. \tag{5}$$

Then, the set of neighbouring pixels and corresponding patches are defined is

$$\mathcal{N}_{\mathbf{x}} = \bigcup_{j=t-1}^{t+1} \mathcal{N}_{\mathbf{x}}^{j}, \qquad \mathcal{N}_{P} = \bigcup_{j=t-1}^{t+1} \mathcal{N}_{P}^{j}. \tag{6}$$

We can compute the weight for each pixel $\mathbf{y}$ from frame $j$ centered at $Q_y \in \mathcal{N}_P$, named $\omega_{i,j}(\mathbf{x}, \mathbf{y})$. This weight is formed by the product of three terms. The first one is based on the patch color difference between neighbourhoods $P$ and $Q_{\mathbf{y}}$

$$\lambda_{i,j}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||H_i(P) - H_j(Q_{\mathbf{y}})||_2}{\kappa(L_i(\mathbf{x}))^2}\right), \tag{7}$$

being $\kappa(L_i(\mathbf{x}))$ an adaptive bandwidth, whose objective is to minimize the importance of patch distance in the overall weight when $x$ belongs to a white saturated part and thus its radiance is not reliable. Therefore the function $\kappa$ takes large values whenever $L_i(\mathbf{x})$ gets close to saturation (see Figure 1, right, and Equation (16)).

The second one is based on the distance between the pixel $\mathbf{y}$ and the center of the search area in frame $j$

$$\beta_{i,j}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||\mathbf{x} + \mathbf{u}_{i,j}(\mathbf{x}) - \mathbf{y}||_2}{\theta}\right). \tag{8}$$

where $\theta$ is a bandwidth related to the radius $R$ of the search window. We favor pixels being close the center of the compensated neighborhood, since these are more likely, to belong to the same object.

Finally, the third one is based on the saturation of pixel $\mathbf{y}$

$$\eta_j(\mathbf{y}) = \frac{1}{3} \sum_{c=1}^{3} h(L_i^c(\mathbf{y})) \tag{9}$$

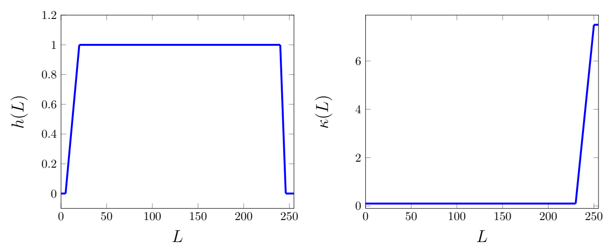



Figure 1: Plot of the weighting function $h(L)$ in (2) and parameter $\kappa(L)$ in (16).

with $h(L)$ being the weighting function in (2). The term $\eta_j(\mathbf{y})$ indicates whether the pixel $y$ is near dark or bright saturation. Brightly saturated pixels will lead to incorrect radiance values, while pixels close to dark saturation might be noisier and more quantized.

Finally, combining the three terms we get

$$\omega_{i,j}(\mathbf{x},\mathbf{y}) = \frac{1}{\mathcal{C}_{\mathcal{P}}} \cdot \lambda_{i,j}(\mathbf{x},\mathbf{y}) \cdot \beta_{i,j}(\mathbf{x},\mathbf{y}) \cdot \eta_j(\mathbf{y}) \tag{10}$$

with $\mathcal{C}_{\mathcal{P}}$ being the normalization factor

$$\mathcal{C}_{\mathcal{P}} = \sum_{y \in \mathcal{N}_{\mathbf{x}}} \lambda_{i,j}(\mathbf{x},\mathbf{y}) \cdot \beta_{i,j}(\mathbf{x},\mathbf{y}) \cdot \eta_j(\mathbf{y}). \tag{11}$$

The weight for all other pixels outside of $\mathcal{N}_{\mathbf{x}}$ is set to zero.

### *3.3. Minimization*

We propose to minimize

$$\arg \min_{S_{t-1}, S_t, S_{t+1}} J(S_{t-1}, S_t, S_{t+1}), \tag{12}$$

where $J(S_{t-1}, S_t, S_{t+1})$ is the variational energy shown in Eq. (1). This energy (1) is convex but non-smooth. To find a global optimal solution we use the primal-dual algorithm proposed by Chambolle and Pock [39]. This algorithm reformulates the minimization into a saddle point problem by introducing a dual variable $q$. Then, the saddle optimal point is computed with an iterative scheme consisting of a descending step with the primal variable, an ascending step in


Figure 2: HDR results on three consecutive frames of *Skateboarder* sequence from Kalantari et al. dataset [16].

the dual variable followed by an overrelaxation. In our model, this leads to the following iterative scheme.

$$\begin{cases} q_{i,j}^{k+1}(\mathbf{x},\mathbf{y}) = \dfrac{(q^k + \sigma \nabla_\omega \bar{S}^k)_{i,j}(\mathbf{x},\mathbf{y})}{\max\{1, |(q^k + \sigma \nabla_\omega \bar{S}^k)|_{i,j}(\mathbf{x},:)\}} \\ S_i^{k+1}(\mathbf{x}) = \dfrac{S_i^k(\mathbf{x}) + \tau(\mathrm{div}_\omega q^{k+1})_i(\mathbf{x}) + \tau\alpha_i h_i(\mathbf{x}) H_i(\mathbf{x})}{1 + \tau\alpha_i h_i(\mathbf{x})} \\ \bar{S}_i^{k+1}(\mathbf{x}) = 2S_i^{k+1}(\mathbf{x}) - S_i^k(\mathbf{x}) \end{cases} \tag{13}$$

where $q_{i,j}^k(\mathbf{x},\mathbf{y})$ is the value of the dual variable at pixels $\mathbf{x}$ and $\mathbf{y}$ and frames $i$ and $j$ in the iteration $k$. The norm $|\cdot|_{i,j}(\mathbf{x},:)$ is defined for a given variable $p$


Figure 3: Excerpt of the results shown in Figure 2. Kalantari13 tends to blur moving details (e.g foot of the girl on the right and trousers of the skater) and has some ghosting artifacts (e.g left shoulder of the skater). Kang also presents some ghosting artifacts (e.g right shoulder and left foot of the skater). Chen and our does not have apparent ghosting but Chen result is noisier.

as

$$|p|_{i,j}(\mathbf{x},:) = \sqrt{\int_{\{\mathbf{y}\in\Omega | w_{i,j}(\mathbf{x},\mathbf{y})\neq 0\}} \left(p_{i,j}(\mathbf{x},\mathbf{y})\right)^2 \, d\mathbf{y}} \tag{14}$$

and $(\mathrm{div}_\omega q^k)_i$ is the nonlocal divergence of the dual variables defined as

$$\begin{aligned}&(\mathrm{div}_\omega q^k)_i(x) = \\ &\sum_{j=t-1}^{t+1} \int_\Omega \sqrt{\omega_{i,j}(\mathbf{x},\mathbf{y})} q^k_{i,j}(\mathbf{x},\mathbf{y}) - \sqrt{\omega_{j,i}(\mathbf{y},\mathbf{x})} q^k_{j,i}(\mathbf{y},\mathbf{x}) \, d\mathbf{y}\end{aligned} \tag{15}$$

The parameters $\tau$ and $\sigma$ of the Chambolle-Pock minimization method are set to 0.2 the default values commonly used.

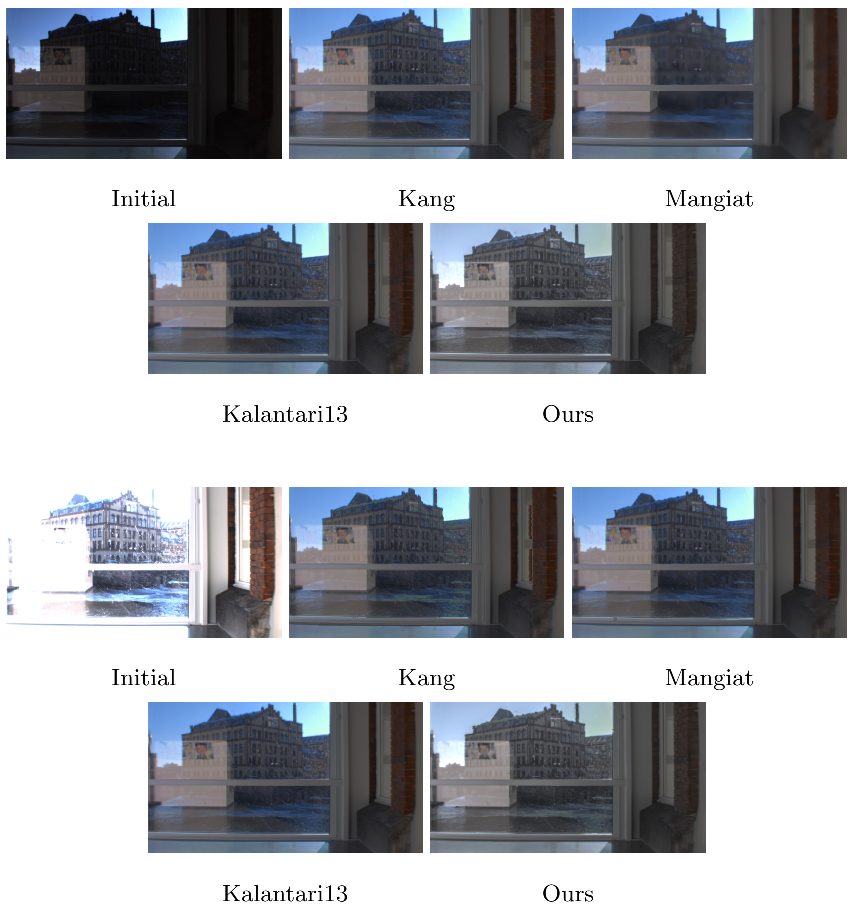



Figure 4: HDR results on two consecutive frames of *Hallway* sequence provided by Li et al. [19].


Figure 5: Excerpt of the results shown in Figure 4. Mangiat removes many details and Kalantari13 deforms the roof of the building. Kang and our method have no noticeable artifacts.

Since the energy can be decoupled per channel, the iterative scheme is applied separately at each channel. The iterative scheme (13) is repeated a fixed number of iterations and is early stopped if the difference between consecutive iterations is smaller than a threshold.

### *3.4. Parameter setting and implementation details*

The use of patch comparison is quite common in both image and video processing algorithms. The size of the patch depends on the particular processing task, the number of color or spectral channels of the image and the transformed used to combine patch values. Generally, if patches are combined by simple statistical tools, the patch size is set as small as possible being robust to noise. Since the processed sequences have three channels and a high signal to noise ratio, a $3 \times 3$ patch has shown to be sufficient for comparison. In addition, since radiance values at saturated zones are not correct, the use of a small window minimizes the effect that these radiances have when processing spatially close but non saturated pixels. As a consequence, the radius $r$ of patch $P$ is set to 1.

**Algorithm 1** Energy minimization

**Input:** Input images: $L_i, \quad i \in \{t-1, t, t+1\}$

**Output:** HDR frames $\{S_{t-1}, S_t, S_{t+1}\}$

$H_i \leftarrow \text{radiance}(L_i), \quad i \in \{t-1, t, t+1\}$

$\omega_{i,j}(\mathbf{x}, \mathbf{y}) \leftarrow \text{nonlocal\_weights\_computation} \quad i, j \in \{t-1, t, t+1\}, \mathbf{x}, \mathbf{y} \in \Omega$

$k \leftarrow 0$

**while** $k < K$ or error $< \varepsilon$ **do**

$\quad q_{i,j}^{k+1}(\mathbf{x}, \mathbf{y}) \leftarrow \frac{(q^k + \sigma \nabla_\omega \bar{S}^k)_{i,j}(\mathbf{x}, \mathbf{y})}{\max\{1, |(q^k + \sigma \nabla_\omega \bar{S}^k)|_{i,j}(\mathbf{x}, :)\}}, \quad i, j \in \{t-1, t, t+1\}, \mathbf{x}, \mathbf{y} \in \Omega$

$\quad S_i^{k+1}(\mathbf{x}) = \frac{S_i^k(\mathbf{x}) + \tau(\text{div}_\omega q^{k+1})_i(\mathbf{x}) + \tau \alpha_i h_i(\mathbf{x}) H_i(\mathbf{x})}{1 + \tau \alpha_i h_i(\mathbf{x})}, \quad i \in \{t-1, t, t+1\}, \mathbf{x}, \mathbf{y} \in \Omega$

$\quad \bar{S}_i^{k+1}(\mathbf{x}) = 2S_i^{k+1}(\mathbf{x}) - S_i^k(\mathbf{x}), \quad i \in \{t-1, t, t+1\}, \mathbf{x}, \mathbf{y} \in \Omega$

$\quad$ error $\leftarrow ||S^{k+1} - S^k||_2$

$\quad k \leftarrow k + 1$

**end while**

Since we use motion compensation, the size $R$ of the spatial neighbourhood does not need to take into account large displacements. It just needs to take into account optical flow imprecisions, which are supposed to be at most of two or three pixels. If the flow fails completely, patch distance is able to discard such matches, reducing ghosting effects. Therefore, $R$, the radius of the spatial neighbourhood, is set to 3. Additionally, we favor pixels being close the center of the compensated neighborhood, since these are more likely, to belong to the same object. This is reflected in the weight term $\beta_{i,j}(\mathbf{x}, \mathbf{y})$ in (8) and the parameter $\theta$ which has been estimated empirically to 1.

The tradeoff parameters $\alpha_i = \alpha(e_i)$ balance the weight of the regularity and fidelity terms. Since short exposures are noisy and quantized in very dark parts, we privilege long exposures, setting $\alpha(\cdot)$ to be an increasing function. In practice, the sequences used in the experimentation section have exposure times in $\{2^{-3}, 2^{-2}, 1, 2^2, 2^3\}$. We have experimentally set the corresponding $\alpha$ values to $\{7, 8, 10, 12, 13\}$, espectively.

The function $h(L)$ is designed to avoid the use of saturated values in both the regularization and fidelity terms. We set $l_2 = 5$ as the dark saturation value and

$h_2 = 246$ as the bright one. Since most sequences were originally compressed, values between 0 and 10 are highly quantified and have a poor signal to noise ratio and artifacts. Taking a $l_2 = 5$ is a good compromise to avoid enhancing artifacts. The $h_2$ is the white saturation value we observed for all the examples used in the experimentation section. As displayed in Figure 1 in order to avoid a drastic zero to one change, the values $l_1 = 20$ and $h_1 = 240$ permit a smoother change. These values are not critical for the performance of the method.

The function $\kappa(L_i(\mathbf{x}))$ sets the kernel bandwidth in an adaptive manner, depending on the exposure of the reference point: pixels that are saturated provide a wrong radiance value, hence, we cannot rely on patch distances when computing weights $\omega_{i,\cdot}(\mathbf{x}, \cdot)$. In such a case, setting a high value of $\kappa$ we diminish the importance of patch distance in the overall weight computation. We define a continuous variation between over-exposed and well-exposed pixels for defining $\kappa(L)$

$$\kappa(L) = \begin{cases} \delta & L \leq h_1 \\ \delta + \frac{L-h_1}{h_2-h_1}(\delta_\infty - \delta) & h_1 < L \leq h_2 \\ \delta_\infty & L > h_2 \end{cases} . \tag{16}$$

We experimentally set $\delta = 0.1$, while $\delta_\infty$ is set to any large enough value, so any weight $\lambda_{i,j}(\mathbf{x}, \mathbf{y})$ equals one.

For image regions saturated in all exposures, the value of the normalization term $\mathcal{C}_\mathcal{P}$ will be very small or even zero, and thus the weight distribution will not be reliable. In such cases, being $\mathcal{C}_\mathcal{P}$ too small, if $L_i(\mathbf{x})$ is over $h_2$, we set to 1 all weights from the image with lowest exposure time and renormalize the weight distribution. In the case $L_i(\mathbf{x})$ is below $l_2$, the same procedure is applied to the weights of the image with highest exposure.

## 4. Results

In this section we present the results of the proposed model for HDR video synthesis, as well as a comparison with state-of-the-art methods. We compare qualitatively and numerically the proposed method with sequences from the datasets provided by Kalantari et al. [16] and Chen et al. [10].

On one hand, we evaluate our method on sequences *Waving Hands*, *Skateboarder*, *Dog* and *Ninja* from the dataset provided by Kalantari et al. [16]. Unfortunately for this dataset, there is no true HDR image that can be used as reference to compare with. We compare them with the precomputed results by Mangiat et al. [14], Kalantari et al. [16] (Kalantari13), Kang et al. [13], Kalantari et al. [9] (Kalantari19) and Chen et al. [10]. The results of the firsts three methods were available at Kalantari's website [2]; and the last two where kindly provided by Chen et al. [10]. Since, not all methods published their result for each sequence, we display in each case the available ones. The tonemap in our experiments is performed by using the tonemapping algorithm from Reinhard et al. [40].

On the other hand, we compare our method with the sequences with 2 exposures from the dataset by Chen et al. [10]. In this case, we compare with Kalantari et al. [16], Chen et al. [10] and the provided ground truth.

### *4.1. Numerical comparison*

Table 1 displays the quantitive results in terms of the UDQM measure [31]. The UDQM is an objective metric that takes into account the most common deghosting artefacts that appear in the output of widely used algorithms and does not require any reference ground truth. The evaluations have been performed on the sequences for which do not dispose of a reference: *Dog, Skateboarder, Hands* and *Ninja*. It has been evaluated for one particular frame for each exposure. We have highlighted for each sequence the method giving the best score. Our method gives the highest score in most sequences which is in accordance with the visual quality analysis performed below.

Table 2 shows the PSNR-$\mu$ computed on tonemapped images for the first four sequences of the dataset provided by Chen et al. [10] with two different

[2] `https://web.ece.ucsb.edu/~psen/PaperPages/HDRVideo/`

| Sequence | *Dog* | | | *Skateboarder* | | | *Hands* | | *Ninja* | |
|---|---|---|---|---|---|---|---|---|---|---|
| Exposure | low | mid | high | low | mid | high | low | high | low | high |
| Kalantari13 | 0.36 | 0.36 | 0.36 | 0.34 | 0.33 | 0.34 | 0.33 | 0.34 | 0.36 | 0.36 |
| Kang | 0.36 | 0.36 | 0.36 | 0.34 | 0.34 | 0.34 | 0.33 | 0.34 | | |
| Chen | **0.40** | **0.39** | **0.38** | 0.35 | 0.37 | 0.35 | **0.42** | **0.42** | 0.36 | 0.40 |
| Kalantari19 | 0.37 | 0.37 | 0.37 | | | | | | 0.36 | 0.36 |
| Mangiat | | | | | | | 0.40 | 0.40 | | |
| Ginger | | | | | | | 0.33 | 0.34 | | |
| Ours | 0.38 | 0.37 | **0.38** | **0.37** | **0.38** | **0.37** | **0.42** | **0.42** | **0.39** | **0.41** |

Table 1: Evaluation of the results using the UDQM metric [31]. Higher values of this index represents less ghosting artefacts. The highest values are highlighted in bold. The empty cells come from not having the result of that method in that sequence. The methods giving the lowest values are Kalantari13 and Kang. Our method gives the highest score in most sequences. These values are supported by the visual analysis on the HDR results.

exposures (one of these sequences is displayed in Figure 11). As in [10] and [9], we apply the following transformation to the images

$$T_i = \frac{\log(1+\mu S_i)}{\log(1+\mu)}, \tag{17}$$

with $\mu = 5000$ to compare them.

Chen et al. algorithm, being the state of the art on HDR video, is a complex and computationally demanding neural network. It applies an initial network in which a first HDR result is obtained by combining three consecutive frames. Afterwards, a second net combines three consecutive first HDR in order to obtain the final result. As a consequence, five frames are used in order to compute the HDR of a particular frame. In our case, we use only three frames and a single pass algorithm.

We also compare with Kalantari et al. [16], which is the closest algorithm to ours, since it uses a patch similarity measure and minimizes an energy containing this similarity and a fidelity term for well exposed pixels. We perform better than both methods on low exposure frames and we perform similarly to Chen on high exposure ones. On average, Chen and the proposed method obtain a similar PSNR, while Kalantari13 has a poorer performance.

### *4.2. Visual comparison*

Figure 2 shows the results on three consecutive frames of sequence *Skateboarder*, each of them taking as reference a frame with a different exposure time. Figure 3 displays an excerpt on the results of the previous figure. It is noticeable that Kang's result has ghosting artifacts on the left feet and right shoulder of the skater and on the right feet of the girl. Kalantari13 result looks blurry on the trousers of the skater and it has ghosting artifacts in its left shoulder. On the second crop, it blurs the right foot of the right side girl. Chen and our does not have apparent ghosting but Chen result seems noisier than ours.

Figure 4 shows the results on two consecutive frames of sequence *Hallway* provided by Li et al. [19]. Figure 5 displays an excerpt on the results of the previous figure. Mangiat excessively filters the details of the image. Kalantari13, because of a bad registration, deforms the roof of the building. Kang and our method do not have apparent ghosting or excessive filtering.

Figure 6 shows the results on two consecutive frames of sequence *Waving hands*. Kang and Magiat results present ghosting artifacts and Kalantari13 result looks blurry.

Figure 7 shows results on two frames of *Dog* sequence. Figure 8 displays an excerpt of it, centered at the head of the dog. Kang and Kalantari19 results have ghosting artifacts near the nose and mouth of the dog and the Kalantari13 result looks blurry.

Figure 9 shows the results on the *Ninja* sequence. Taking a closer look on the excerpt on Figure 10, we can see that Kalantari13 and Kalantari19 produce blurry results and Chen result looks noisier than ours.

Finally, Figures 11 and 12 display the results with Sequence 1 and 4 from Chen et al. dataset [10]. Kalantari13 presents several ghosting artefacts, while at a first look Chen and our method perform similarly better. For Figure 11, a close look into the shirt, reveals that Chen over-smooths the texture and has a ghosting artifact on the left side. For Figure 12, a close look into the object being hold, reveals that Chen over-smooths the vertical detail.

| Sequence | 0 | | 1 | | 2 | | 3 | | mean |
|---|---|---|---|---|---|---|---|---|---|
| Exposure | high | low | high | low | high | low | high | low | |
| Kalantari13 | 27.17 | 28.59 | 28.50 | 29.39 | 31.00 | 30.36 | 29.23 | 30.35 | 29.32 |
| Chen | **49.86** | 43.78 | **50.14** | 43.99 | **48.87** | 44.55 | **49.79** | 45.55 | **47.07** |
| Ours | 48.08 | **44.51** | 49.53 | **44.97** | 48.46 | **45.23** | 49.37 | **45.91** | **47.01** |

Table 2: Numerical comparison on two exposure sequences from the dataset of Chen et al. [10]. We display the PSNR-$\mu$ metric for particular low and high exposure frames. Each sequence number corresponds to the index on the dataset. One of these sequences is displayed in Figure 11. We compare with Kalantari et al. [16] and Chen et al. [10]. Our algorithm gives the highest PSNR for low exposure frames and performs similarly to Chen on high exposure ones. On average, Chen and the proposed method obtain identical PSNR, while Kalantari13 has a poorer performance.

### *4.3. Computational complexity*

We run our method in a computer with a processor 2.8GHz Intel Core i7-1165G7. With the images from the Kalantari dataset [16] having 1280x720 pixels, our method takes 150 seconds. This time corresponds to all the steps of the proposed algorithm, including optical flow. With the same image, [16] being the closest method to ours, takes 400 seconds.

According to the work by Chen et al. [10], the application of their convolutional network takes 0.51 s. However, the network runs on a GPU and requires an extensive and computationally intensive training, while our method runs on a CPU and does not require a training stage. Therefore, the computational times are not comparable.

## 5. Conclusions

We have presented a new variational model for HDR video synthesis. It writes as an energy minimization using nonlocal regularization across the neighbouring frames and a fidelity term for well exposed pixels. The proposed method obtains competitive results compared with state-of-the-art deep learning techniques Kalantari et al. [9] and Chen et al. [10]. Compared to these methods, our approach depends on a few understandable parameters, making the results more reliable and interpretable.

As a future work, we plan to study the incorporation of unrolling methods [41, 42, 43]. These hybrid methods combine energy minimization strategies and deep learning techniques. The advantages of replacing the proximal operator with a neural network include gains in representation power and direct learning of algorithm parameters from real data.

Constraining the network structure to particular mathematical expressions reduces its complexity, improves its generalization capabilities and increases the interpretability.

## References


[1] T. Mitsunaga, S. K. Nayar, Radiometric self calibration, in: Proceedings IEEE computer society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Vol. 1, IEEE, 1999, pp. 374–380.

[2] S. Mann, R. Picard, Beingundigital'with digital cameras, MIT Media Lab Perceptual 1 (1994) 2.

[3] P. E. Debevec, J. Malik, Recovering high dynamic range radiance maps from photographs, in: ACM SIGGRAPH 2008 classes, ACM, 2008, p. 31.

[4] J. Tumblin, H. Rushmeier, Tone reproduction for realistic images, IEEE Computer Graphics and Applications 13 (6) (1993) 42–48.

[5] G. Ward, A contrast-based scalefactor for luminance display., Graphics Gems 4 (1994) 415–21.

[6] O. Martorell, C. Sbert, A. Buades, Ghosting-free dct based multi-exposure image fusion, Signal Processing: Image Communication 78 (2019) 409–425.

[7] Y. Niu, J. Wu, W. Liu, W. Guo, R. W. Lau, Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions, IEEE Transactions on Image Processing 30 (2021) 3885–3896.

[8] M. D. Tocci, C. Kiser, N. Tocci, P. Sen, A versatile hdr video production system, ACM Transactions on Graphics 30 (4) (2011) 1–10.

[9] N. K. Kalantari, R. Ramamoorthi, Deep hdr video from sequences with alternating exposures, Computer Graphics Forum 38 (2) (2019) 193–205. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13630, doi:https://doi.org/10.1111/cgf.13630.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13630

[10] G. Chen, C. Chen, S. Guo, Z. Liang, K.-Y. K. Wong, L. Zhang, Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 2502–2511.

[11] O. T. Tursun, A. O. Akyüz, A. Erdem, E. Erdem, The state of the art in hdr deghosting: A survey and evaluation, Computer Graphics Forum 34 (2) (2015) 683–707. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12593, doi:https://doi.org/10.1111/cgf.12593.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12593

[12] L. Wang, K.-J. Yoon, Deep learning for hdr imaging: State-of-the-art and future trends, IEEE transactions on Pattern Analysis and Machine Intelligence (2021) 1–1.

[13] S. B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High dynamic range video, ACM Transactions on Graphics 22 (3) (2003) 319–325. doi:10.1145/882262.882270.
URL https://doi.org/10.1145/882262.882270

[14] S. Mangiat, J. Gibson, High dynamic range video with ghost removal, in: Applications of Digital Image Processing XXXIII, Vol. 7798, International Society for Optics and Photonics, 2010, pp. 307 – 314.

[15] S. Mangiat, J. Gibson, Spatially adaptive filtering for registration artifact

removal in hdr video, in: 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 1317–1320.

[16] N. K. Kalantari, E. Shechtman, C. Barnes, S. Darabi, D. B. Goldman, P. Sen, Patch-based high dynamic range video., ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2013) 32 (6) (2013).

[17] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, E. Shechtman, Robust patch-based hdr reconstruction of dynamic scenes., ACM Transactions on Graphics 31 (6) (2012) 1–11.

[18] Y. Gryaditskaya, T. Pouli, E. Reinhard, K. Myszkowski, H.-P. Seidel, Motion aware exposure bracketing for hdr video, Computer Graphics Forum 34 (4) (2015) 119–130.

[19] Y. Li, C. Lee, V. Monga, A maximum a posteriori estimation framework for robust high dynamic range video synthesis, IEEE Transactions on Image Processing 26 (3) (2016) 1143–1157.

[20] T. Van Vo, C. Lee, High dynamic range video synthesis using superpixel-based illuminance-invariant motion estimation, IEEE Access 8 (2020) 24576–24587.

[21] Q. Yan, D. Gong, Q. Shi, A. v. d. Hengel, C. Shen, I. Reid, Y. Zhang, Attention-guided network for ghost-free high dynamic range imaging, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1751–1760.

[22] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, Y. Zhang, Deep hdr imaging via a non-local network, IEEE Transactions on Image Processing 29 (2020) 4308–4322.

[23] J. W. Song, Y.-I. Park, K. Kong, J. Kwak, S.-J. Kang, Selective transhdr: Transformer-based selective hdr imaging using ghost region mask, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv,

Israel, October 23–27, 2022, Proceedings, Part XVII, Springer, 2022, pp. 288–304.

[24] Q. Yan, B. Wang, P. Li, X. Li, A. Zhang, Q. Shi, Z. You, Y. Zhu, J. Sun, Y. Zhang, Ghost removal via channel attention in exposure fusion, Computer Vision and Image Understanding 201 (2020) 103079.

[25] K. R. Prabhakar, S. Agrawal, R. V. Babu, Self-gated memory recurrent network for efficient scalable hdr deghosting, IEEE Transactions on Computational Imaging 7 (2021) 1228–1239.

[26] M. Anand, N. Harilal, C. Kumar, S. Raman, Hdrvideo-gan: deep generative hdr video reconstruction, in: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, 2021, pp. 1–9.

[27] J.-W. Kim, J.-H. Ryu, J.-O. Kim, Deep gradual flash fusion for low-light enhancement, Journal of Visual Communication and Image Representation 72 (2020) 102903.

[28] U. Cogalan, M. Bemana, K. Myszkowski, H.-P. Seidel, T. Ritschel, Learning hdr video reconstruction for dual-exposure sensors with temporally-alternating exposures, Computers & Graphics 105 (2022) 57–72.

[29] R. Mantiuk, K. J. Kim, A. G. Rempel, W. Heidrich, Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions, ACM Transactions on graphics (TOG) 30 (4) (2011) 1–14.

[30] M. Narwaria, M. P. Da Silva, P. Le Callet, Hdr-vqm: An objective quality measure for high dynamic range video, Signal Processing: Image Communication 35 (2015) 46–60.

[31] O. T. Tursun, A. O. Akyüz, A. Erdem, E. Erdem, An objective deghosting quality metric for hdr images, in: Computer Graphics Forum, Vol. 35, Wiley Online Library, 2016, pp. 139–152.

[32] K. Karajuzovic-Hadziabdic, J. H. Telalovic, R. K. Mantiuk, Assessment of multi-exposure hdr image deghosting methods, Computers & Graphics 63 (2017) 1–17.

[33] J. Duran, J. Navarro, A. Buades, Variational densification and refinement of registration maps, SIAM Journal on Imaging Sciences 14 (3) (2021) 879–912.

[34] G. Gilboa, S. Osher, Nonlocal operators with applications to image processing, Multiscale Modeling & Simulation 7 (3) (2009) 1005–1028.

[35] M. Jung, L. A. Vese, Nonlocal variational image deblurring models in the presence of gaussian or impulse noise, in: International Conference on Scale Space and Variational Methods in Computer Vision, Springer, 2009, pp. 401–412.

[36] O. Martorell, A. Buades, Variational temporal optical flow for multi-exposure video, in: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, INSTICC, SciTePress, 2022, pp. 666–673. doi:10.5220/0010908300003124.

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

[38] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: European Conference on Computer Vision, Springer, 2004, pp. 25–36.

[39] A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, Journal of Mathematical Imaging and Vision 40 (1) (2011) 120–145.

[40] E. Reinhard, K. Devlin, Dynamic range reduction inspired by photoreceptor physiology, IEEE Transactions on Visualization and Computer Graphics 11 (1) (2005) 13–24. doi:10.1109/TVCG.2005.9.

[41] Y. Chen, W. Yu, T. Pock, On learning optimized reaction diffusion processes for effective image restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5261–5269.

[42] E. Kobler, T. Klatzer, K. Hammernik, T. Pock, Variational networks: connecting variational methods and deep learning, in: German Conference on Pattern Recognition, Springer, 2017, pp. 281–293.

[43] U. Schmidt, S. Roth, Shrinkage fields for effective image restoration, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 2774–2781.

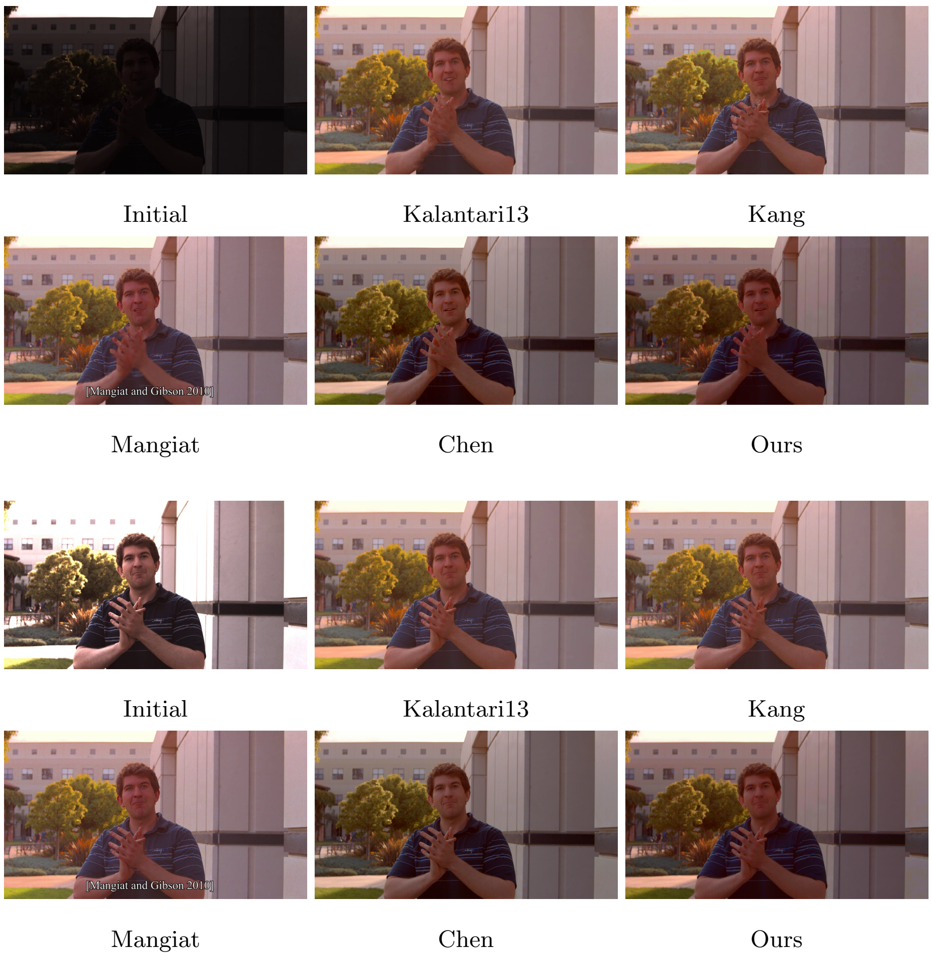
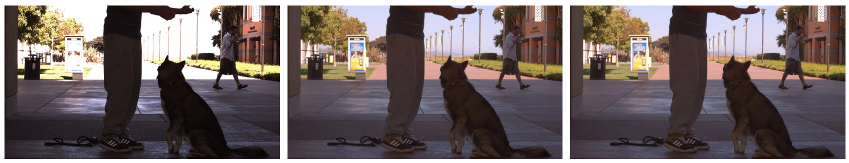
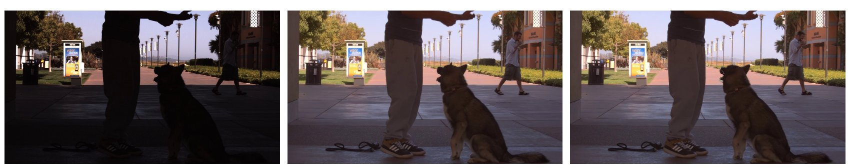



Figure 6: HDR results on two consecutive frames of *Waving hands* sequence from Kalantari et al. dataset [16]. Kang and Mangiat results present ghosting artifacts, Kalantari13 results look blurry and Chen and our method obtain pleasant results without noticable artifacts.



Initial Kalantari13 Kang

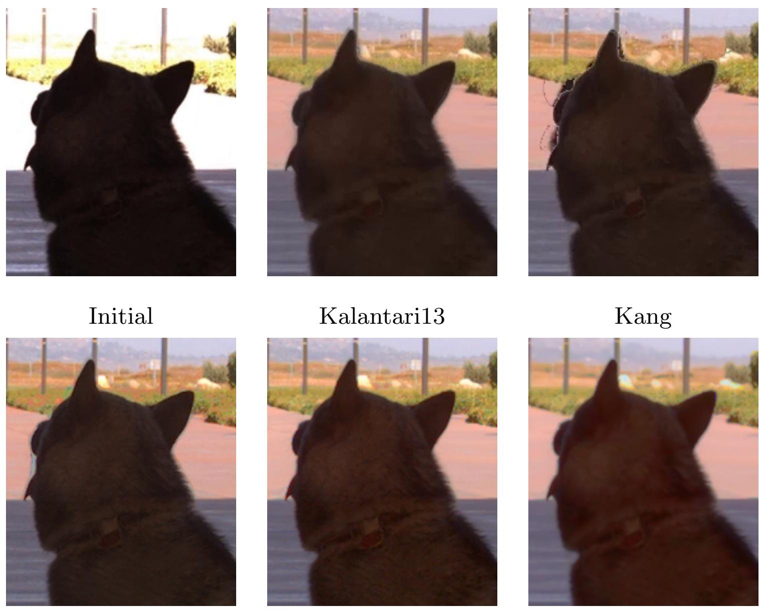
Kalantari19 Chen Ours



Initial Kalantari13 Kang

Kalantari19 Chen Ours

Figure 7: HDR results on two consecutive frames of *Dog* sequence from Kalantari et al. dataset [16]. Kalantari13, Kang and Kalantari19 have problems at registering the head of the dog. Chen and our method both produce high quality results (see details on Figure 8).



Initial | Kalantari13 | Kang

Kalantari19 | Chen | Ours

Figure 8: Excerpt of the results shown in Figure 7. Kalantari13, Kang and Kalantari19 present ghosting artifacts.


Figure 9: HDR results on two consecutive frames of *Ninja* sequence from Kalantari et al. dataset [16].


Figure 10: Excerpt of the results shown in Figure 9. The result from Kalantari13 looks blurry and results from Kalantari19 and Chen look noisier than ours.

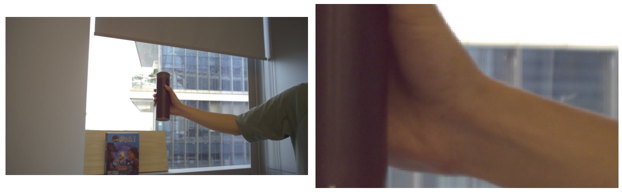
Input frame

Kalantari13 | Chen | Ours

Input frame

Kalantari13 | Chen | Ours

Figure 11: Full frame and excerpt of the input frame and HDR results on sequence 1 of Chen et al's dataset. Kalantari13 presents several ghosting artefacts, while at a first look Chen and our method perform similarly better. A close look into the shirt, reveals that Chen over-smooths the texture and has a ghosting artifact on the left side when centred at the shorter exposure.



Input frame

Kalantari13 Chen Ours

Input frame

Kalantari13 Chen Ours

Figure 12: Full frame and excerpt of the input frame and HDR results on sequence 4 of Chen et al's dataset. Kalantari13 presents several ghosting artefacts, while at a first look Chen and our method perform similarly better. A close look into the object being hold, reveals that Chen over-smooths the vertical detail and has a slight ghosting artifact on the finger.