High Dynamic Range and Super-Resolution from Raw Image Bursts

Bruno Lecouat, Thomas Eboli, Jean Ponce, Julien Mairal

High Dynamic Range and Super-Resolution from Raw Image Bursts

BRUNO LECOUAT, Inria and DI/ENS (ENS-PSL, CNRS, Inria), France
THOMAS EBOLI, Université Paris-Saclay, ENS Paris-Saclay, Centre Borelli, France
JEAN PONCE, Inria and DI/ENS (ENS-PSL, CNRS, Inria), France & Center for Data Science, New York University, USA
JULIEN MAIRAL, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France

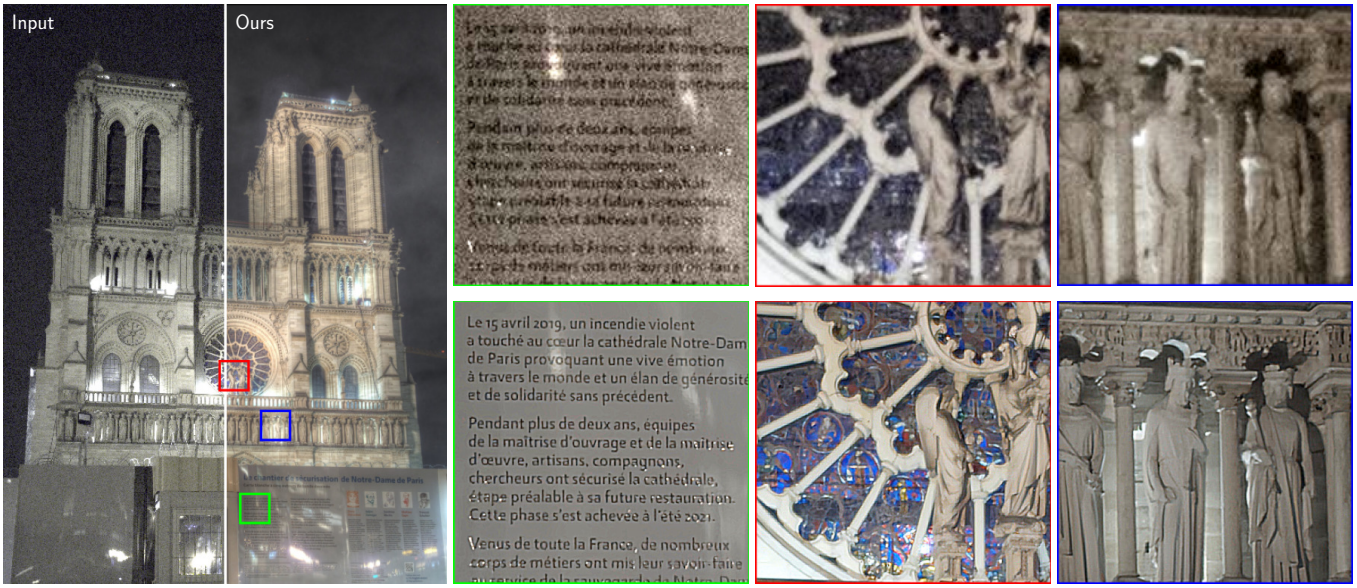


Fig. 1. An example of joint super-resolution (SR) and high-dynamic range (HDR) imaging. **Left:** An 18-photo burst was shot at night from a hand-held Pixel 4a smartphone at 12MP resolution with an exposure time varying from 1/340s to 1/4s. The left half of the central image from the burst is shown along with the right half of the 192MP HDR image reconstructed by our algorithm with a super-resolution factor of ×4 (after tone mapping). **Right:** Three small crops of the two images corresponding to the colored square regions on the left. Crops from the central image of the burst are rendered using Adobe Camera Raw to convert raw files into jpg with highest quality setting. The HDR/SR results are rendered using the PhotoMatix tone mapperhttps://www.hdrsoft.com/. Note that the 192MP HDR image on the left is not reproduced at full resolution because of the corresponding file's size.

Photographs captured by smartphones and mid-range cameras have limited spatial resolution and dynamic range, with noisy response in underexposed regions and color artefacts in saturated areas. This paper introduces the first approach (to the best of our knowledge) to the reconstruction of high-resolution, high-dynamic range color images from raw photographic bursts captured by a handheld camera with exposure bracketing. This method uses a physically-accurate model of image formation to combine an iterative optimization algorithm for solving the corresponding inverse problem with a learned image representation for robust alignment and a learned natural image prior. The proposed algorithm is fast, with low memory requirements compared to state-of-the-art learning-based approaches to image restoration, and features that are learned end to end from synthetic yet realistic data. Extensive experiments demonstrate its excellent performance with super-resolution factors of up to ×4 on real photographs taken in the wild with hand-held cameras, and high robustness to low-light conditions, noise, camera shake, and moderate object motion.

CCS Concepts: • **Computing methodologies** → **Image processing**.

Additional Key Words and Phrases: Computational photography, raw bursts, high-dynamic range imaging, super-resolution

ACM Reference Format:
Bruno Lecouat, Thomas Eboli, Jean Ponce, and Julien Mairal. 2022. High Dynamic Range and Super-Resolution from Raw Image Bursts. *ACM Trans. Graph.* 41, 4, Article 38 (July 2022), 21 pages. https://doi.org/10.1145/3528223.3530180

Authors' addresses: Bruno Lecouat, Inria and DI/ENS (ENS-PSL, CNRS, Inria), Paris, France, bruno.lecouat@inria.fr; Thomas Eboli, Université Paris-Saclay, ENS Paris-Saclay, Centre Borelli , Paris, France, thomas.eboli@ens-paris-saclay.fr; Jean Ponce, Inria and DI/ENS (ENS-PSL, CNRS, Inria), France & Center for Data Science, New York University, New York, USA, jean.ponce@inria.fr; Julien Mairal, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France, julien.marial@inria.fr.

0730-0301/2022/7-ART38
https://doi.org/10.1145/3528223.3530180

1 INTRODUCTION

Key factors limiting the level of detail of photographs captured by digital cameras are their spatial resolution and dynamic range: High resolution is necessary to zoom on small image regions, and

high dynamic range is needed to reveal details hidden in dark areas (*e.g.,* shadows) and avoid color artefacts due to saturation in bright ones (*e.g.,* highlights). For a given sensor size, higher resolution also means smaller pixel size, with less light reaching each photoreceptor, resulting in lower dynamic range and increased noise in dark regions, an effect exacerbated in smartphones by their small sensor size. It is natural, and by now rather common, to use multiple photographs to reconstruct an image with higher spatial resolution, a process known as *super-resolution* (or *SR* for short in this presentation, see, for example [Wronski et al. 2019]), or dynamic range, a process known as *high dynamic range* (or *HDR*) imaging (see, for example [Debevec and Malik 1997]).

We propose in this paper a novel method for *joint* SR and HDR imaging from the *raw* image bursts featuring a range of different exposures that can now be captured by most smartphones and mid-range cameras (Figure 1). A major challenge tackled by our algorithm is the automated alignment with sub-pixel accuracy of the burst elements required to compensate for camera shake and possibly (moderate) object motion, despite the variations in saturation and signal-to-noise ratio due to the different exposures used across the burst. Other notable difficulties include the high contrasts and noise levels encountered in night scenes for example, where a photo might feature both very dark and noisy regions and saturated ones near light sources, as well as the fact that a digital camera only captures one color channel at each pixel according to the corresponding *color filter array* (or *CFA*, often a Bayer pattern). Despite the latter challenge, it now seems clear that it is better to work directly with the raw image data than with the sRGB pictures produced by the *image signal processor* (or *ISP*) of the camera since their construction involves several steps, including white balance, denoising, demosaicking, gamma correction, compression of each color channel content to 8 bits, etc., that result in an unavoidable loss of information in high spatial frequencies and dynamic range.

The approach proposed in the rest of this presentation extends the algorithm for multi-frame super-resolution of [Lecouat et al. 2021] to jointly perform blind denoising, demosaicking, super-resolution and HDR image reconstruction from raw bursts. Its key features can be summarized as follows:

- Our method uses a physically-accurate model of image formation that accounts for the successive transformations applied to the original analog irradiance image, including quantization of the signal, noise, exposure and spatial quantization.

- We combine an iterative optimization algorithm for solving the corresponding inverse problem with a learned image representation for robust alignment and a learned natural image prior. This is the first main technical novelty of our paper, enabling us to address the joint reconstruction of high-resolution, high-dynamic range color images from raw photographics bursts captured by a handheld camera with exposure bracketing.

- The proposed algorithm is fast, with low memory requirements compared to state-of-the-art learning-based approaches to image restoration, and features that are learned end to end from synthetic yet realistic data, generated using again our image formation model.

- We introduce an image alignment method to compensate for camera shake which is robust to (moderate) object motions and an image fusion technique which is itself tolerant to alignment errors. Together, these form the second main technical novelty of our paper, and they are key factors in the robustness of our algorithm in both the SR and HDR imaging tasks with, notably, significant improvement over [Lecouat et al. 2021] in super-resolution.

- Extensive experiments demonstrate the excellent performance of the proposed approach with super-resolution factors of up to $\times 4$ on real photographs taken in the wild with hand-held cameras, and high robustness to low-light conditions, noise, camera shake, and moderate object motion. These results are confirmed by quantitative and qualitative comparisons with the state of the art in super-resolution and HDR imaging tasks on synthetic and real image bursts.

2 BACKGROUND

2.1 High dynamic range imaging

Bracketing techniques. [Debevec and Malik 1997; Granados et al. 2010; Hasinoff et al. 2010; Mann and Picard 1995] construct an HDR image by combining multiple photographs of the same scene with different exposures. The darkest pictures are used to reconstruct areas prone to saturation and the brightest ones are needed for restoring dark regions that are likely to be noisy (we will come back to that point later). They typically work on *linRGB* images, that is, demosaicked images *before* they are transformed by the camera's ISP into *sRGB* images ready for display. A sequence of sRGB input photographs must therefore in general be "linearized" by inverting this mapping, also known as the camera response function (or *CRF*). The HDR image is then reconstructed as a weighted sum of the linearized bracket images, normalized by the corresponding shutter speed. Its pixel values are typically represent as single-precision floating-point numbers, with min and max those of the image bracket. Bracketing-based approaches to HDR imaging face a number of classical issues, including choosing the optimal fusion weights, estimating the CRF [Debevec and Malik 1997], leveraging accurate raw image noise models [Aguerrebere et al. 2014; Granados et al. 2010; Hanji et al. 2020], selecting the best exposure parameters for a fixed number of frames in the bracket [Gallo et al. 2012; Hasinoff et al. 2010], registering images with different exposures [Gallo et al. 2015; Zimmer et al. 2011], which is significantly more challenging than aligning same-exposure images [Ma et al. 2017], and removing ghosting artefacts [Sen et al. 2012; Tursun et al. 2016] due to misalignment.

Using raw bursts with constant shutter speed. Unlike classical exposure bracketing techniques, HDR+ [Hasinoff et al. 2016] takes as input a burst of raw underexposed images captured with the same exposure time. These are mostly free of saturation but noisy in dark regions. A 12-bit, denoised raw image is obtained by aggregating the 10-bit photos of the burst. It is then demosaicked and tone mapped.

Recent updates of HDR+ use a couple of well-exposed frames to achieve better denoising and deghosting [Ernst and Wronski 2021], or leverage the metering technique of Hasinoff et al. [2010] to adapt the original algorithm to low-light situations [Liba et al. 2019].

Using pixelwise ISO sensitivities. Instead of relying on classical imaging devices, Nayar and Mitsunaga [2000] reconstruct a single HDR image from a sensor with spatially-varying pixel exposures. This approach can be further combined with learning-based methods [Martel et al. 2020; Serrano et al. 2016]. Even though our work focuses on standard sensors, we believe it to be flexible enough to be adapted to pixelwise ISO sensitive sensors under simple modification of the image formation model. This is an interesting research direction for future work, but beyond the scope of our paper.

Learning-based methods. for HDR imaging have also been proposed. Kalantari and Ramamoorthi [2017] introduce a convolutional neural network (CNN) to predict the irradiance from three low-dynamic range (LDR) images, with different exposures, camera poses and possibly moving subjects, pre-aligned with an optical flow algorithm. Most recent CNN-based multi-image methods [Niu et al. 2021; Pérez-Pellitero et al. 2021; Wu et al. 2018; Yan et al. 2021, 2020] learn to align and fuse demosaicked images in an end-to-end manner, and they typically operate on image triplets such as those in the dataset of Kalantari and Ramamoorthi [2017]. CNN-based approaches to single-image HDR include [Eilertsen et al. 2017; Endo et al. 2017; Liu et al. 2020; Santos et al. 2020]. They rely on machine learning to recover missing details in the darkest and saturated areas of tone-mapped images.

2.2 Super-resolution

We limit here our discussion to multi-frame super-resolution algorithms. Although single-image learning-based techniques have been used to generate very impressive and highly-detailed images [*e.g. Dahl et al. 2017; Menon et al. 2020]*, their objective is not the same as ours: they aim at generating a high-resolution picture *compatible* with one input photograph, whereas we want to reconstruct the details that are *actually available* in the input burst.

Energy-based methods. High-frequency information present in low-resolution (LR) photos with aliasing artefacts is useful for reconstructing a high-resolution (HR) image from multiple LR frames [Farsiu et al. 2006]. Unfortunately, this information is typically lost during the denoising and demosaicking steps performed by the camera ISP pipeline to produce sRGB images. Farsiu et al. [2006] estimate an HR demosaicked image from a sequence of raw photographs by minimizing a penalized energy—that is, they solve an inverse problem via optimization. Wronski et al. [2019] adapt the kernel method of Takeda et al. [2007] and exploit natural hand tremor to jointly demosaick and super-resolve a raw image burst with magnification factors up to ×3 in a fraction of a second on a handheld smartphone.

Learning-based techniques. Bhat et al. [2021a] learn a CNN with attention module to align, demosaick and super-resolve a burst of raw images. In a follow-up work, Bhat et al. [2021b] minimize a penalized energy including a data term comparing the sum of parameterized features residuals. Lecouat et al. [2021], learn instead a hybrid neural network alternating between aligning the images with the Lucas-Kanade algorithm [Lucas and Kanade 1981], predicting an HR image by solving a model-based least-squares problem and evaluating a learned prior function. Luo et al. [2021] propose a neural network architecture that aligns an input burst of images while performing super-resolution with a non-local fusion module.

2.3 Joint HDR imaging and super-resolution

The algorithms proposed by Choi et al. [2009]; Gunturk and Gevrekci [2006] address joint SR and HDR imaging with an existing SR energy-based solver. To tackle the multi-exposure setting, they introduce weights inspired by bracketing techniques in the least-squares term. More generally, this joint image restoration problem has been addressed in a two-stage fashion: (i) image registration with an algorithm robust to varying exposures and (ii) solving a least-squares problem including operators modelling both SR and HDR. For instance, Rad et al. [2007] propose an exposure-invariant transform before applying the FFT-based registration technique of Vandewalle et al. [2006]. The image is then obtained by solving a penalized least-squares problem. Zimmer et al. [2011] use an optical flow approach with normalized gradients for robustness to changes of exposure, and the HR/HDR image is found by solving again a penalized least-squares problem. Traonmilin and Aguerrebere [2014] adapt a backprojection algorithm to the multi-exposure setting and simply solve a weighted least-squares problem without prior, with comparable performance but lower computational cost. Vasu et al. [2018] explore the case where the LDR SR images are also blurred with camera shake or motion blur. Similar to the HDR case, CNNs have also been proposed for single-image joint SR and HDR, *e.g.,* [Kim et al. 2019], while Deng et al. [2021] address instead joint SR, HDR and tone mapping by merging a pair of previously aligned over- and under-exposed images with a two-stream CNN. In contrast with these techniques, we use trainable image features to adapt the raw image registration module of Lecouat et al. [2021] to the varying-exposure setting in a robust manner, and jointly learn these features and a parametric image prior in an end-to-end manner.

Figure 2 shows examples of the input data these methods use and samples of the the predicted high-resolution HDR images we predict with the proposed approach.

3 IMAGE FORMATION MODEL

We now describe the process generating a burst of low-dynamic low-resolution raw images from a high-resolution HDR image. This process yields a natural inverse problem formulation, which we will leverage later to build a trainable architecture.

3.1 Dynamic range

After analog-to-digital conversion, a camera sensor outputs a black-and-white mosaicked image whose pixel values are integers obtained by quantizing the number of photons collected by each photosite on a linear q-bit scale [Clark 2006], where q is called the *bit depth* of the sensor. We denote by P_q the set of the discrete values a pixel may take, as measured in *data numbers* (or *DNs* [Clark 2006; Martinec 2008]), from 0 to $2^q - 1$.

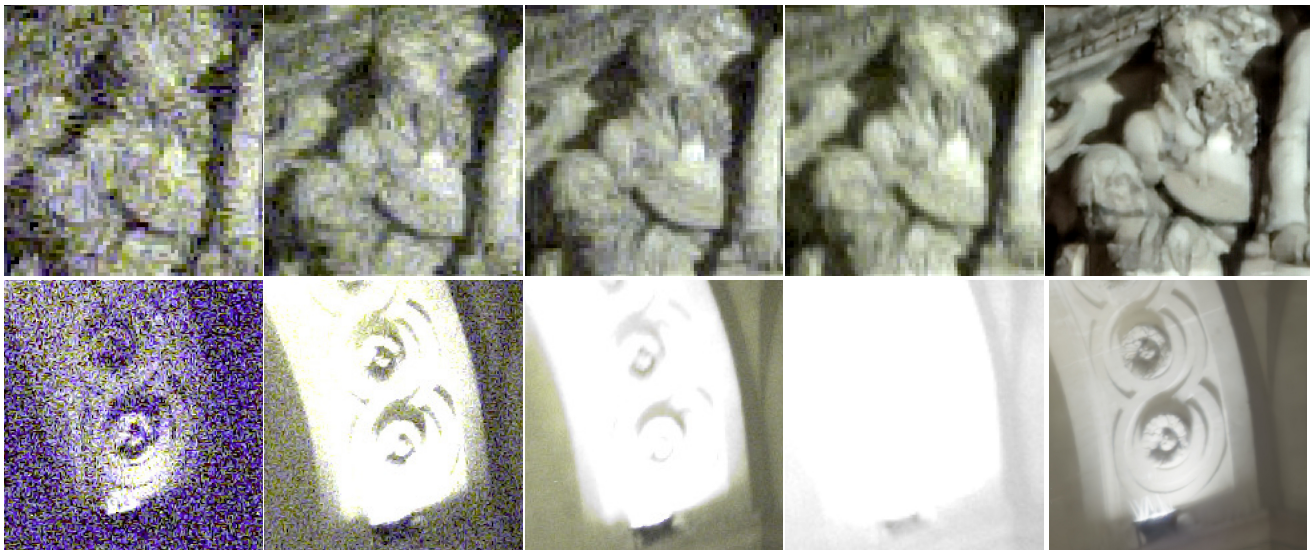


Fig. 2. Exposure bracketing: **Left:** Three high dynamic range, high-resolution images obtained by our method from 18-image bursts taken by a handheld Pixel 4a smartphone with a ×4 super-resolution factor. We show post-processed sRGB pictures for the sake of presentation. **Right:** Small crops from sample photos in the burst and our reconstruction. Note the high level of noise in the short-exposure images, in particular in the second row, and the saturated regions in the long-exposure ones. As shown by the last column of the figure, our algorithm recovers details in saturated areas and remove noise in the darkest regions. The reader is invited to zoom in on a computer screen.

The dynamic range $R(u)$ for a pixel u is defined as the ratio of the largest to the smallest values this pixel may take: the larger the bit depth of the sensor, the greater is its maximal value in P_q. The ratio is usually given in photographic *stops*, where each stop corresponds to a multiple of 2. In practice, the largest value u can take is limited either by the bit depth q or the white level c set by the camera, to prevent color artefacts in highlights [Luijk 2007], whereas the lowest value is actually limited by the noise $\varepsilon(u)$ and by the camera black level b [Foi et al. 2008]. Note that even in the absence of light, $\varepsilon(u)$ is never 0 since any digital camera suffers from various sources of electronic noise [Hasinoff et al. 2010]. This also shows that increasing dynamic range is strongly related to denoising, as discussed later in this section.

3.2 Exposure

As mentioned above, raw pixel values depend linearly on the number of photons captured by each photosite (ignoring quantization effects) and thus on exposure time. In photography, this effect is quantified by the *exposure value* (or *EV*): Increasing it by +1EV (resp. decreasing by -1EV) corresponds to doubling (resp. halving) the raw pixel values. The EV depends on the *ISO gain*, aperture size and exposure time. In this work, we will only control the exposure time Δt, keeping it small enough to (mostly) avoid motion blur, and keep the other two quantities constant since modifying the ISO gain may change the noise distribution [Hasinoff et al. 2010] and adjusting the aperture size changes the blur of out-of-focus regions [Levin et al. 2007].

The raw value $y(u)$ in P_q recorded at some pixel u is thus related to the irradiance $x(u)$ in $\mathbb{R}^+$ at the same location by

$$y(u) = S(\Delta t x(u)), \tag{1}$$

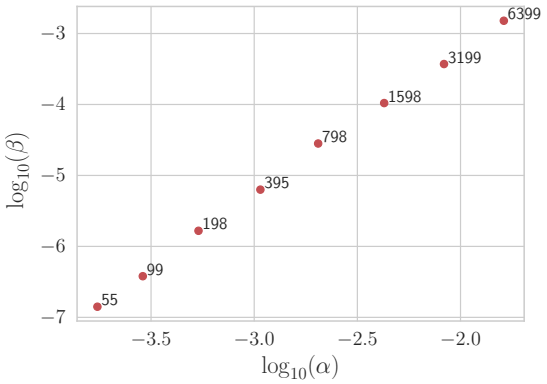


Fig. 3. Empirical measurements of the shot and read noise levels α and β from the metadata of raw images taken with the Google Pixel3a smartphone. The numbers next to the markers are the corresponding empirical ISO levels. As observed by Brooks et al. [2019], there exists a linear relationship between $\log_{10}(\alpha)$ and $\log_{10}(\beta)$ that we leverage to train our models.

where S is the function mapping pixel values from $\mathbb{R}^+$ to P_q. This equation is only valid when $S(\Delta t x(u)) < 2^q - 1$, with saturation occurring for higher values. Using short exposure times limits saturation, but, as shown in the next section, leads to a poor *signal-to-noise ratio* (or *SNR*).

3.3 Noise and SNR

The raw image noise $\varepsilon(u)$ at each pixel comes from the physics of light and the electronics of the camera. The former is called *shot*

noise, and it can be modelled with a Poisson distribution [Foi et al. 2008]. The latter is often referred to as *read* noise and corresponds to random signal fluctuations caused by the electronics and quantization effects. It is usually modelled with a zero-mean Gaussian distribution [Foi et al. 2008]. The combination of shot and read noise can be modelled by a single random variable $\varepsilon(u)$ following a zero-mean Gaussian distribution with pixel-dependent standard deviation, defined for any pixel value $y(u)$ as [Brooks et al. 2019; Foi et al. 2008; Plötz and Roth 2017]:

$$s(u) = \sqrt{\alpha y(u) + \beta}, \tag{2}$$

where α and β are respectively the variances of the shot and read noise. Figure 3 shows the distribution of α (shot noise level) and β (read noise level) for the Google Pixel3a camera. We have obtained theses values from the EXIF metadata of raw images taken with the smartphone. Each marker corresponds to a couple $(\log_{10}(\alpha), \log_{10}(\beta))$ for an ISO level. In dark regions, read noise dominates shot noise, and limits the total dynamic range.

For the Poissonian-Gaussian noise model of Eq. (2), the SNR is:

$$\mathrm{SNR}(u) = \frac{m(u)y(u)}{s(u)} = \frac{m(u)y(u)}{\sqrt{\alpha y(u) + \beta}}, \tag{3}$$

where m is a binary mask excluding the saturated pixels. It is a monotonically increasing function of the pixel value $y(u)$, essentially linear in dark regions (*e.g.,* shadows) where read noise dominates shot noise, and essentially proportional to $\sqrt{y(u)}$ in bright regions (*e.g.,* highlights) where the opposite occurs [Granados et al. 2010]. As already discussed in the previous section, noise removal is essential for generating images with high dynamic range, and Equation (3) shows that high raw pixel values lead to better SNR and thus better dynamic range in both dark and bright image regions. But high pixels values everywhere in an image can typically only be achieved at the cost of saturating the brightest areas. Exposure bracketing avoids this problem by using the longest exposures to eliminate read noise from dark regions and the shortest ones to avoid saturation in bright spots.

3.4 Overall image formation model

The original analog image cannot be recovered on a computer and we instead focus on estimating a discrete HR/HDR $sh \times sw \times 3$ photograph x with pixel values in $\mathbb{R}_+$ from a burst of K raw LR and LDR images y_k ($k = 1, \dots, K$) of size $h \times w$ with entries in P_q. The integer s is the super-resolution factor. Following [Lecouat et al. 2021], let us introduce the warp operator W_k associated with the kth photo in the burst and accounting for camera shake, the blur operator B taking into account the integration of the signal over the pixel area is modeled by a convolution, the decimation operator D_s associated with the super-resolution factor s, and the C operator is a binary mask modeling the sensor CFA. Putting them together and taking into account the exposure time Δt_k, the analog low-resolution image associated with the irradiance image x is $a_k = CD_sBW_k(\Delta t_k x)$, which can be rewritten as $a_k = A_k x$, where $A_k = \Delta t_k CD_sBW_k$ (the factor Δt_k commutes with the operators since it only scales the image values).

Combining this model with Eq. (2) and (1) yields, for all $k = 1, \dots, K$:

$$y_k = S\left(A_k x + \varepsilon_k\right), \tag{4}$$

where we abuse the notation so S operates on a whole image instead of a scalar, and ε_k is a zero-mean Gaussian noise with pixel-dependent variance $\alpha A_k x + \beta$ according to Eq. (2). The operator CD_sB impacts the spatial resolution, while S and the noise variance limit the dynamic range of each image y_k.

Note that our model assumes that the scene is static during burst acquisition, which may result in ghosting artefacts in the presence of scene motion, when using this model within an inverse problem formulation. We will, however, introduce in the next section simple weighting strategies to make our approach robust to moderate scene motion.

4 PROPOSED APPROACH

The goal of this work is to design a function F_θ with learnable parameter θ which, given K raw images $Y = \{y_1, \dots, y_K\}$ and corresponding exposure times $\Delta = \{\Delta t_1, \dots, \Delta t_K\}$, predicts a single-precision floating-point estimate $\hat{x}$ of the the HR $sh \times sw \times 3$ irradiance map:

$$\hat{x} = F_\theta(Y, \Delta). \tag{5}$$

As explained later in this section, all images of the burst are automatically aligned on a reference frame y_{k_0} (typically the central one that has in general a reasonable exposure).

4.1 Formulation of the problem

Inverse problem. Our image formation model (4) suggests using an inverse problem formulation to the design of F_θ and the recovery of $\widehat{x}$. We first convert the discrete raw pixel values from y_k in P_q into 32-bits real values in $[0, 1]$, and construct the binary mask $m(y_k)$ representing saturated pixels containing non-informative values. With an abuse of notation, we keep the notation y_k for the floating-point burst images in the rest of this presentation, and formulate the solution of our inverse problem as the joint recovery of the warp operators $W_1, \dots, W_K$ (parameterized with a piecewise-affine model, as detailed later), and the irradiance image x:

$$\min_{x, W_1, \dots, W_K} \frac{1}{2} \sum_{k=1}^{K} \left\| w_k \odot \left(y_k - A_k x\right) \right\|_F^2 + \lambda \Omega(x), \tag{6}$$

where A_k is the image formation operator defined in the previous section, $\odot$ denotes pointwise multiplication, and the function Ω is a regularizer, and it will be discussed in details later. The $h \times w$ maps w_k store pixel-wise weights used to control the relative contribution of each frame to the reconstruction of each pixel, a key factor for robustness in bracketing methods [Aguerrebere et al. 2014; Granados et al. 2010].

A robust weighting strategy. We write

$$w_k = \frac{\Delta t_k m(y_k)}{\sum_{j=1}^{K} \Delta t_j m(y_j)} \odot g\left(y_k, W_k y_1\right), \tag{7}$$

where $m(y_k)$ is the binary with zero values at saturated pixels (this formulation assumes the existence of non-saturated pixels at corresponding locations in the burst; when all pixels are saturated, we use

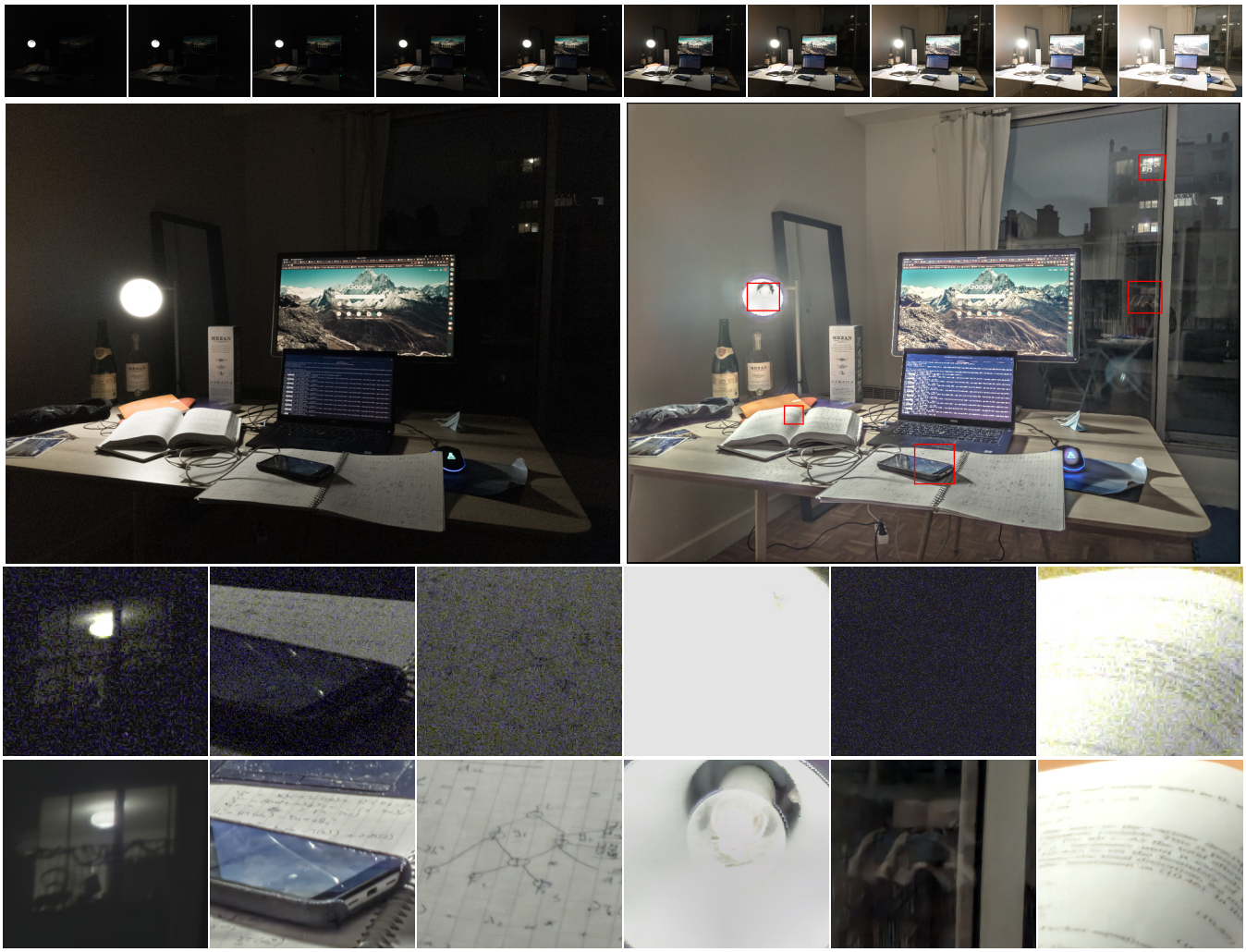


Fig. 4. Joint HDR imaging and super-resolution ×4 with a burst taken with a hand-held Pixel4a at night, facing a spotlight. **Top:** The original burst. **Middle:** The central image in the burst (left) and the reconstructed HDR/SR image after tone mapping (right). **Bottom:** Six crops showing details of the original and HDR/SR images, presented respectively in the first and second rows.

uniform weights instead). Here, the function g is a confidence factor, often used in HDR imaging to weight down images incorrectly aligned [Tursun et al. 2016] and avoid ghosting effects. It can be handcrafted from classical image features and/or priors, but we will instead follow a plug-and-play strategy (detailed in the next section) to directly learn a parametric function g from supervisory data. Our overall weighting strategy is useful for HDR since it provides larger weights to frames obtained with longer exposure time that are less noisy, but it also accounts for registration errors through the learned function g, which turns out to be critical for robustness to moderate scene motion.

Warp parameterization. We align images with piecewise-affine warps $W_k = W_{p_k}$, where W_{k_0} is the identity and $p = \{p_1, \ldots, p_K\}$ is the set of warp parameters. This is implemented by tiling the images into small (*e.g.*, 200×200) crops, that are aligned independently with affine transformations with 6 parameters.

Regularizer. Many classical regularizers can be used in the formulation of inverse problems in image processing applications, for example sparse total-variation priors [Choi et al. 2009] or combinations of penalty functions computed from pixel or histogram values [Debevec and Malik 1997; Heide et al. 2014; Rad et al. 2007]. We instead follow the same plug-and-play strategy as for the confidence function g, and learn a CNN in place of the proximal operator [Parikh and Boyd 2014] of the penalty function Ω. We detail its implementation in Sec. 4.3.

4.2 Optimization strategy

We solve our optimization problem with half-quadratic splitting (or HQS) [Geman and Yang 1995] by introducing an auxiliary variable

z and minimize

$$\min_{x,z,p} \frac{1}{2} \sum_{k=1}^{K} \|w_k \odot (y_k - A_k z)\|_F^2 + \frac{\eta}{2} \|x - z\|_F^2 + \lambda \Omega(x). \quad (8)$$

The parameter η is usually increased at each iteration according to some preset schedule, which guarantees that, as η grows, the solution of this relaxed problem converges to that of the original one (6) [Geman and Yang 1995]. As detailed in Sec. 4.3, we choose instead to learn this parameter from training data, which improves performance in practice. Note that we now find the warp operators by minimizing the energy with respect to the warp parameters p, and that all operators involved are implemented efficiently by exploiting the image structure (*e.g.,* convolutions instead of large sparse operators, etc.). The optimization is carried out by first initializing z and p, then, in an alternate fashion, repeating T times ($T = 3$ in our implementation) an HQS stage consisting of the three steps detailed below. The motivation for this strategy is that it allows us to gracefully convert our optimization method into a trainable architecture, as discussed in Sec. 4.3, thanks to automatic differentiation tools [Baydin et al. 2018] implemented in modern deep learning frameworks.

Updating z. The auxiliary image z is updated by a few steps of a simple gradient descent (GD) algorithm:

$$z \leftarrow z - \delta \left(\eta(z - x) + \sum_{k=1}^{K} A_k^\top \left(w_k^2 \odot (A_k z - y_k) \right) \right), \quad (9)$$

where δ is a step size (which will be learned automatically by the procedure presented in the next section), and of course A_k depends on the current warping parameters p.

Updating x. Minimizing (8) with respect to the image x while keeping the other variables fixed amounts to compute the so-called proximal operator G of Ω [Parikh and Boyd 2014]:

$$x = G(z, \lambda/\eta) = \arg\min_x \frac{1}{2} \|x - z\|_{\mathrm{F}}^2 + \frac{\lambda}{\eta} \Omega(x). \quad (10)$$

We will detail in the next section how we implement G.

Updating p. Lecouat et al. [2021] estimate the warp parameters p_k ($k \neq k_0$) on 200×200 tiles in a 4-scale Gaussian image pyramid, running three stages of the Lucas-Kanade algorithm [Lucas and Kanade 1981] at each stage. We will show in Sec. 4.3 how to do significantly better, both quantitatively and qualitatively, by using a similar approach to align *learned* features instead.

Initialization of p and z. A fast and coarse initialization of the warp parameters p is obtained using a sub-pixel variant of the FFT-based algorithm of [Anuta 1970] with the features of [Ward 2003]. After having estimated p for the first time with the Lucas-Kanade algorithm and before the first z-update stage, we initialize z as follows: we demosaick each frame y_k with bilinear interpolation, align them with the warping operators W_k, average them with the normalized weights $\Delta_k / \sum_{j=1}^{K} \Delta_j$, and finally upscale the resulting image by a factor s with bilinear interpolation. This procedure yields a fast and coarse estimate of the HR and HDR image to start the GD algorithm in Eq. (9).

4.3 Learnable architecture

The optimization procedure described in the previous section is implemented as a function F_θ that produces an estimate $\hat{x}$ from a burst Y and exposure times Δ, according to Eq. (5). By writing this function as a finite sequence of operations that are differentiable with respect to the model parameters θ, it is then possible to leverage training data—that is, pairs of HR/HDR images x associated to LR/LDR bursts—to *learn* these parameters for the reconstruction task. This of course raises questions about data collection and generation, which are discussed later, but it also opens up many possibilities for further improvements. In particular, as described in the rest of this section, this allows us to learn implicitly the regularization function Ω by taking advantage of deep learning principles, as well as learning appropriate weighting strategies, and robust features to improve image alignment.

Learnable proximal operator G. Following the *plug-and-play* strategy [Venkatakrishnan et al. 2013] which has proven powerful in the signal processing literature, we replace the proximal operator G above by a function G_ω represented by a CNN and parameterized by ω, such that the update (10) becomes

$$x = G_\omega(z, \gamma), \quad (11)$$

where γ is also a trainable parameter. The CNN has a residual U-net architecture, which is a smaller variant of the network of Zhang et al. [2020] for single-image super resolution. This network has four scales with respectively 32,64,128,128 channels per scale. We also run experiments with an even smaller version of the network with 32 features per channel (dubbed *small*) and 16 features per channel (dubbed *tiny*). Note that for our problem, the first layer has 4 input channels: three for the predicted RGB auxiliary variable z and one for the scalar γ.

Learnable confidence function g. Similarly, since designing the function g by hand is difficult, we choose to learn instead a CNN g_ρ, and the fusion weights w_k become for all $k \neq k_0$:

$$w_k = \frac{\Delta t_k m(y_k, c)}{\sum_{j=1}^{K} \Delta t_j m(y_j, c)} \odot g_\rho(y_k, W_k y_{k_0}), \quad (12)$$

The function g_ρ is implemented with the tiny variant of the U-Net architecture used above. The network takes as input the concatenation along the channel dimension of RGB versions of the images y_k and $W_k y_{k_0}$ obtained by bilinear interpolation.

Learnable features for alignment. A classical approach to the registration of frame captured with different exposure times is to use MTB features [Ward 2003]. Here, we construct instead a single-channel feature map for each raw image using again the tiny CNN with U-net architecture, then perform the multi-scale Lucas Kanade algorithm for a fixed number of iterations (3 iteration per scale of the pyramid) *directly on the feature map.* Our implementation of the forward additive version of the Lucas Kanade algorithm is fully differentiable. Therefore we can learn the parameters of the feature map jointly with all the trainable parameters of our model, following a strategy similar to [Chang et al. 2017]. As shown in the experimental section this significantly improves registration performance.

4.4 Learning the model parameters θ

We denote here by θ all the learnable parameters of our methods, including those of the CNNs and the scalar parameters involved in the HQS optimization procedure introduced above (*e.g.*, , δ, η, ...). We use triplets of the form $(x^{(i)}, Y^{(i)}, \Delta^{(i)})$ $(i = 1, \ldots, n)$ of training data to supervise the learning procedure. In our setting where ground-truth HDR/HR images are normally not available for real image bursts, the training data is necessarily semi-synthetic, that is, obtained by applying various transformations to real images. Obtaining robust inference with real raw bursts is thus challenging. The hybrid nature of our algorithm, which exploits both a learning-free inverse problem formulation and data-driven priors, appears to be a key to achieving good generalization on real raw data acquired in various conditions that do not necessarily occur in the training dataset.

Dataset generation. Given a collection of *sRGB* images, we construct bursts of LDR/LR raw images and HDR/HR RGB targets using the ISP inversion method of [Brooks et al. 2019] and our image formation pipeline, adjusting the gain to simulate different exposure times. The noise levels are sampled following the empirical model of Figure 3.

Training loss. With this training data in hand, we supervise our model using the ℓ_1 distance between the target irradiance images $x^{(i)}$ and the predicted ones $F_\theta(Y^{(i)}, \Delta^{(i)})$, and minimize the cost function:

$$\min_\theta \sum_{i=1}^{n} \left\| x^{(i)} - F_\theta(Y^{(i)}, \Delta^{(i)}) \right\|_1 . \tag{13}$$

We have also tried to use the so-called μ-law [Kalantari and Ramamoorthi 2017] to include some kind of tone mapping in the supervision but it only marginally improved the visual quality of the images predicted by our model.

Optimizer. We minimize Eq. (13) using Adam optimizer with learning rate set to 10^{-4} for 400k iterations. We decrease the learning rate by 0.5 every 100k iterations. The weights of the CNNs are randomly initialized with the default setting of the PyTorch library.

5 RESULTS

We first show in Section 5.1 several qualitative results illustrating the performance of our method for joint HDR imaging, super-resolution, demosaicking and denoising from real raw image bursts. Qualitative and quantitative comparisons with existing methods for super-resolution, HDR imaging, and registrations are presented in Section 5.2, Section 5.3 and Section 5.4 respectively. The effect of the choice of prior and the robustness of our method for real images are discussed in Section 5.5. Additional results, ablations studies, and discussions of its limitations can be found in the appendix.

Note that all the HDR images are rendered using *Photomatix*[1] for tone mapping, which is itself a challenging task [Reinhard et al. 2002] beyond the scope of this paper. For baselines operating on RGB images instead of raw photographs, we first process raw files with Adobe Camera Raw to generate RGB images with the highest quality possible.

[1] https://www.hdrsoft.com/

5.1 Joint SR and HDR on raw image bursts

To the best of our knowledge, we are the first to address jointly HDR, super resolution, demosaicking, and denoising on bursts of raw images. Therefore, we will mostly present here qualitative results, and will defer quantitative comparisons to the following sections that evaluate the performance of our algorithm on separate HDR or SR tasks.

We consider bursts acquired in different settings by a Pixel 3a or 4a camera, by using an Android application to shoot bursts of 11 to 18 raw images. We choose an EV step of 1/3 to 2/3 between each shot. This is particularly important for night scenes to avoid motion blur in the longest-exposure frames. Our method successfully restores finer details and extends the dynamic range of the original shot by denoising dark areas and restoring clipped signals. More precisely:

- Figures 4 and 6 show night-time photos with large dynamics, similar to Figure 1, with both under- and over-exposed areas in the low-resolution central frame. Both the dynamics and the resolution are significantly improved by our algorithm.
- An outdoor day-time photograph is shown in Figure 5, with a particularly large dynamic range. The scene contains both under-exposed, noisy areas in the shadows and large bright saturated areas. Note also that the scene contains patterns which are smaller than the resolution of the native image which is a particularly hard setting for demosaicking. Our approach handles such situations well.
- A night scene with both very dark building parts and light bulbs, resulting in a very large dynamic range (Figure 6). Our approach, unlike our competitors, can recover details in both the dark and saturated areas.

5.2 Pure super-resolution

We now move to pure super-resolution from raw image bursts, and compare our approach with Lecouat et al. [2021], using examples from their paper. The bursts in this section all have the same exposure, making the alignment simpler compared to the previous section. We first perform a quantitative evaluation on the semi-synthetic benchmark of Bhat et al. [2021a], following their experimental setup and using their dataset. Table 1 presents a comparison of our approach for SR, which can be seen as an improved variant of Lecouat et al. [2021]. All methods in the comparison are designed to process raw image bursts. We first note that our improvements in the image registration module yields +1dB over [Lecouat et al. 2021] for similar network capacities. The geometric error, measuring alignment discrepancies, is also four times smaller than that of [Lecouat et al. 2021], which further suggests the usefulness of our modified Lucas-Kanade module. Since the other methods of the panel do not explicitly predict any motion vector, we cannot compute the corresponding geometric errors. We also have a PSNR gain of about 0.5 to 1dB over three of the recent competitors and fall only behind Luo et al. [2021] by less than 1dB but with 13 times fewer learnable parameters. Therefore, the proposed approach is also a compact and competitive algorithm for SR alone. A speed comparison, presented later in Section A.4 also shows that our method is faster at inference time.

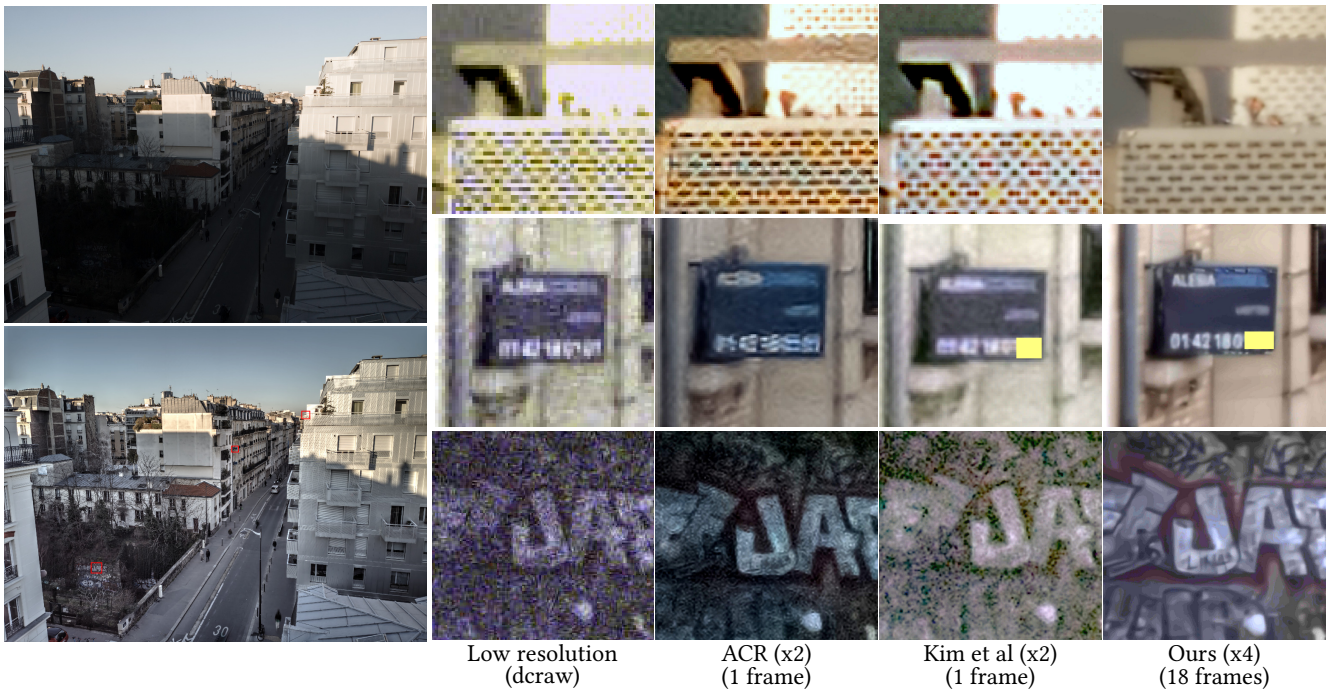


Fig. 5. Day-time comparisons of joint HDR imaging and super-resolution algorithms with bursts acquired by a Pixel4a. **Left:** The central image in the burst (top) and our reconstruction (bottom). **Right:** Comparison of close-ups of the reconstructions obtained by the CNN-based Adobe Camera Raw single-image algorithm for ×2 super-resolution and demosaicking, the CNN-based ×2 super-resolution method of [Kim et al. 2019], and our method. (Note: part of the phone number legible in our case is masked for privacy reasons.)

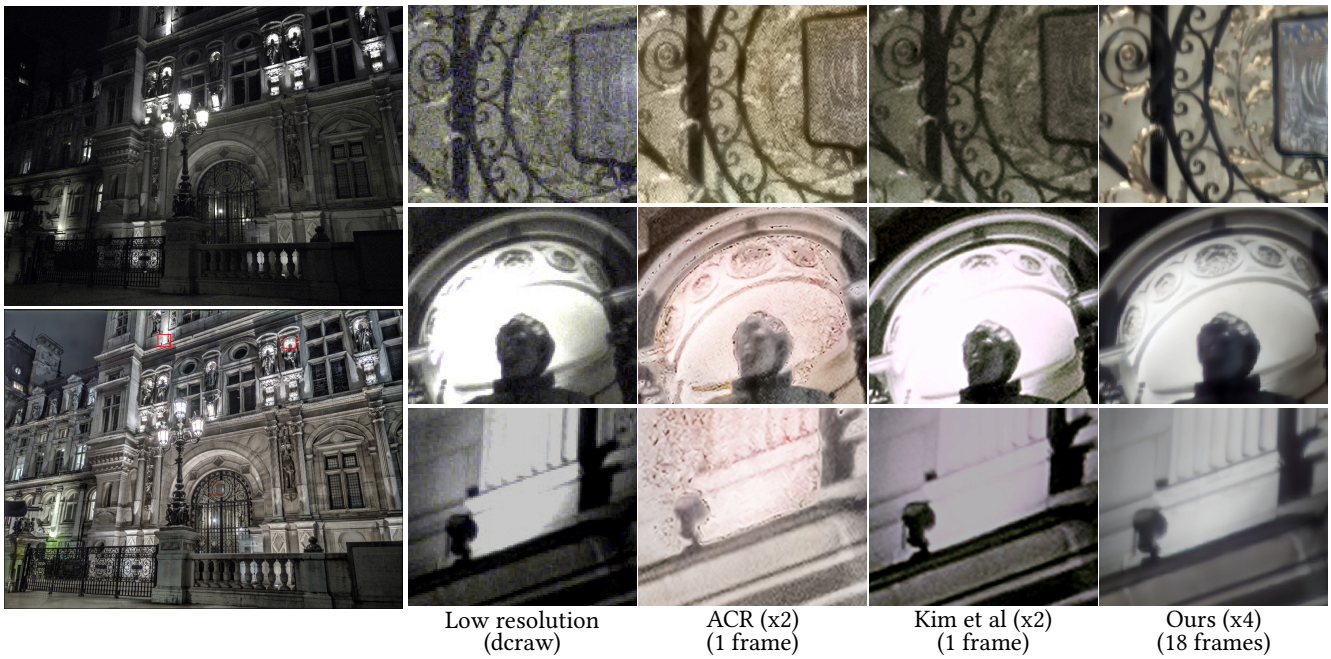


Fig. 6. Night-time comparisons of joint HDR imaging and super-resolution algorithms with bursts acquired by a Pixel4a. **Left:** The central image in the burst (top) and our reconstruction (bottom). **Right:** Comparison of close-ups of the reconstructions obtained by the CNN-based Adobe Camera Raw single-image algorithm for ×2 super-resolution and demosaicking, the CNN-based ×2 super-resolution method of [Kim et al. 2019], and our method.

Fig. 7. Visual comparison for super-resolution only on real same-exposure raw bursts, of respectively $K = 20$ and $K = 30$ frames, with state-of-the-art competitors. We *do not* present HDR results in this figure. Our approach limits Moiré artefacts in the first row and reveals in general more high frequency image details in both rows. The last row shows an example requiring deghosting. The ghosted LR image on the left is obtained by averaging the whole burst to show the pedestrian's motion. Bhat et al. [2021a] and our method effectively handle small object motions. The reader is invited to zoom in.

Low resolution [Bhat et al. 2021a] [Lecouat et al. 2021] [Luo et al. 2021] Ours

The previous comparison is conducted on semi-synthetic data, both for training the models and for testing, which makes its conclusions difficult to generalize to the real world of raw bursts from handheld cameras. Nevertheless, it remains the best existing quantitative experimental setup, to our knowledge, since it is not possible to acquire reliable HR ground-truth data along with LR raw bursts. Figure 7 shows two challenging real-world examples on which we compare qualitatively the approaches of Bhat et al. [2021a], Lecouat et al. [2021] and Luo et al. [2021] to ours, for ×4 super-resolution factor. We display in the first row the results for a burst of $K = 20$ raw frames of a textured surface. Moiré artefacts and aliasing can respectively be noticed in the results from Bhat et al. [2021a] and Luo et al. [2021]. Such artefacts are not visible in our reconstruction and that of of Lecouat et al. [2021]. The second row shows the results for a burst of $K = 30$ raw images from [Lecouat et al. 2021]. Amongst the four methods in the panel, ours returns the sharpest image, with for instance easier-to-read characters than competitors. We point out that we *have not* used any sharpening algorithm on any of these images.

As remarked by Wronski et al. [2019], there is a physical limit to the maximum frequency one can reconstruct with aliasing, due to the sensor pitch or the lens point-spread function. We verify this property in Figure 8 where we show two crops from the same image, with ×2 and ×4 resolution factors. The first row shows details of a balcony clearly benefiting from a ×4 gain in resolution compared to its ×2 counterpart. The second row shows however that sometimes, as predicted by Wronski et al. [2019], ×4 upsampling factor may not reveal finer details that its ×2 counterpart.

Table 1. Super-resolution (×4) comparison with a selected panel of recent methods with average PSNR and geometric error when it can be computed. We *do not* perform HDR generation in this experiment. Our method falls behind that Luo et al. [2021] within a margin of less than 1dB but with 13 times fewer parameters. We gain 1dB compared to [Lecouat et al. 2021] with a similar number of parameters by upgrading the registration module.

Model	# parameters	PSNR	Geom (avg)
[Bhat et al. 2021a]	13M	40.76	N/A
[Lecouat et al. 2021]	3M	41.45	<u>2.56</u>
[Bhat et al. 2021b]	-	41.56	N/A
[Dudhane et al. 2021]	6.6M	41.93	N/A
[Luo et al. 2021]	26M	**43.35**	N/A
Ours	3M	<u>42.42</u>	**0.80**

5.3 Pure HDR imaging

We evaluate the ability of our approach to align and merge raw images into HDR image at the same resolution as the input.

We compare our approach with a bracketing technique, implemented with the weights of [Hasinoff et al. 2010], two state-of-the-art CNNs [Wu et al. 2018; Yan et al. 2021] trained to predict a 32-bit image from only three LDR images with -2, 0 and +2EV or -3, 0 and +3EV, and recent single-image HDR CNNs [Liu et al. 2020; Santos

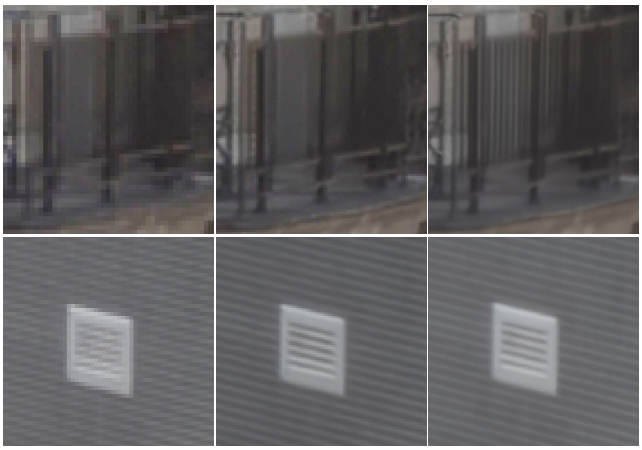


Fig. 8. Visualizing super-resolution limit at resolutions increased by ×2 and ×4 with our model. The first image in the first row benefits from the ×4 improvements whereas the one in the second row (from the same photograph) is not further enhanced after ×2. See the discussion in the text.

et al. 2020]. We generate 266 raw bursts with 32-bit ground-truth images, each burst containing 11 synthetic raw images with small random shifts and rotations and Poissonian-Gaussian noise with parameters α and β selected according the distribution in Figure 3. More details about data generation can be in found in Section B of the appendix. To evaluate the CNNs trained on RGB images, we first pick the three raw frames corresponding to {-2.4,0,+2.4} EV in the burst and demosaick them with the approach of Malvar et al. [2004]. We also demosaick the frames before merging the HDR images with the bracketing technique. If the raw frames are not aligned, after demosaicking, we align the frames either with the phase correlation algorithm [Tursun et al. 2016] on the MTB features Ward [2003] or with our Lucas-Kanade-based registration technique. For fairness with the CNNs, we compare our approach when there are only three frames in the bracket (the same as for the CNNs) and with the whole burst.

We present in Table 2 the results of our comparison. We evaluate the PSNR and the SSIM metrics on both the output of each algorithm and after evaluating the irradiance maps with μ-law, playing the role of a tone mapping algorithm [Kalantari and Ramamoorthi 2017]. However these typical image processing metrics may not be adapted to HDR imaging [Aydin et al. 2008; Eilertsen et al. 2021] we thus also report the HDRVDP2 perceptual quality score of Narwaria et al. [2015] (version 2.2.2). Note that Wu et al. [2018] and Yan et al. [2021] use RGB images for training, while our method leverages more information by directly processing raw frames. We also compare our method to the single-image methods of Santos et al. [2020] and Liu et al. [2020] running on the central frame of the burst.

Our algorithm using 11 frames achieves the best results as expected, with HDRVDP2 margins ranging from +4 to +9 over recent CNN-based methods and of +4 over the bracketing technique of Hasinoff et al. [2010] using 11 frames too. The gap with CNNs comes from our ability to restore the darker areas in raw photographs containing large read noise whereas these networks are trained on RGB images only. Figure 10 shows qualitative comparisons with the baselines in Table 2 for bursts of 21 images taking during day time and night time. Our method achieves the best visual results in both dark and saturated areas. Note that the CNN baselines considered here have been designed to handle 1 or 3 images only, which is not sufficient to achieve effective denoising through image fusion in challenging settings.

We also compare our approach with a public-domain implementation [Monod et al. 2021] of Google's HDR+ [Hasinoff et al. 2016] that addresses HDR imaging by fusing images with the same exposure. In this setting, HDR essentially boils down to burst denoising, which is effectively handled by our approach. Figure 9 shows a qualitative comparison of HDR+ with our technique. We achieve better denoising, especially in the darkest areas, while also increasing spatial resolution.

5.4 Multi-exposure registration

We evaluate the performance of our registration module based on learnable features. We measure the geometric alignment error between the ground-truth motion and predicted one [Sanchez 2016] computing the Euclidean distance between the aligned image corners with that of the ground-truth ones and is counted in number of pixels in the HR image.

We report the mean and the median over 266 validation bursts (containing 11 images per burst) synthesized with the same protocol as for generating the training data. We compare a typical multi-exposure registration scheme consisting in combining MTB features and phase correlation [Tursun et al. 2016] (used for prealigning the images in our model), with the 3 iterations of the pyramid Lucas-Kanade (PLK) algorithm over plain pixels and deep features learnt in an end-to-end manner. The three methods are run on the mosaicked and possibly noisy images, prior to any ISP processing. We evaluate this panel over three scenarios: (i) HDR generation without SR from noise-free raw bursts, (ii) HDR generation without SR from raw bursts with noise and (iii) joint HDR and SR with factor ×4 from raw bursts with noise.

Table 3 shows that, in all cases, our approach achieves the best quantitative results, with a margin ranging from 0.5px for the (unrealistic) noise-free benchmark to more than 1px for the more challenging ones featuring noise. Interestingly, using more iterations does not always mean a better alignment. A plausible explanation is that our model is trained for using three iterations of the LK algorithm, and may be sub-optimal for more iterations.

We have also empirically observed that the errors in this table are always greater than that reported by Lecouat et al. [2021] in their work for aligning frames with the same exposure. This gap is caused in practice by the darkest and brightest frames, much harder to align because of the noise in dark regions and large saturated areas.

Figure 11 compares the advantage of running the Lucas-Kanade algorithm with deep features and plain pixels in a real situation. Note the purple zipping artefacts caused by faulty alignment before

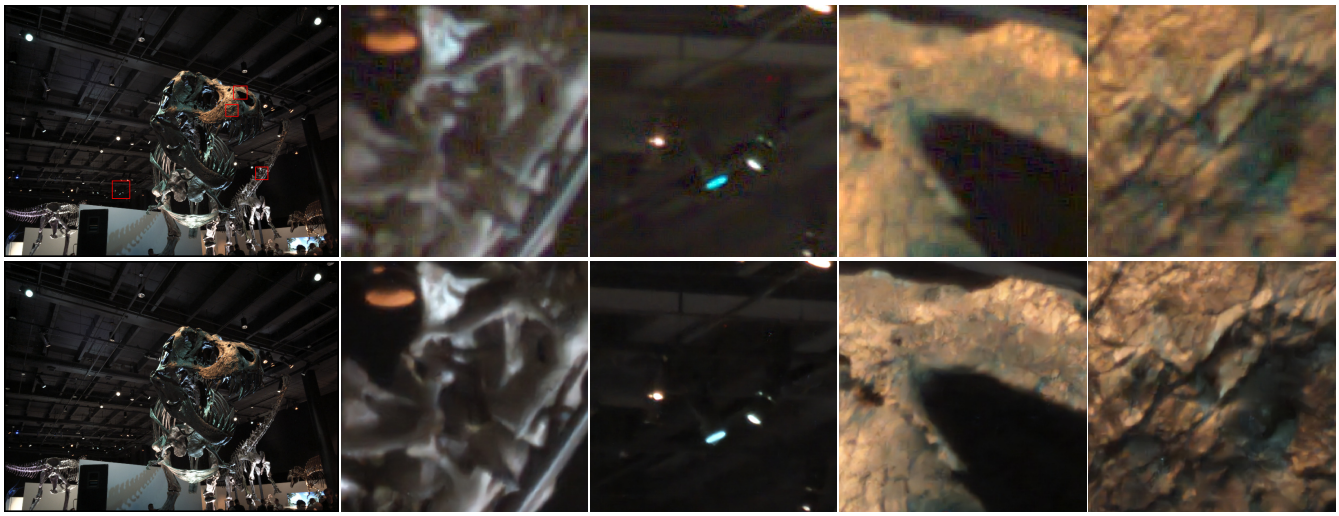


Fig. 9. Comparison between the public-domain implementation [Monod et al. 2021] dubbed here HDR+Ipol of Google's HDR+ [Hasinoff et al. 2016] (top) with our HDR/SR×4 method. **Left:** The images reconstructed by HDR+Ipol (top) and our method (bottom) from a burst of 8 same-exposure images acquired by a Nexus 5. Note that they are barely distinghuisable at this resolution. **Right:** Crops showing that our algorithm reveals finers details while effectively suppressing noise in dark areas.

Table 2. Quantitative comparison of various algorithms for HDR imaging – we do not perform super-resolution in this experiment – on a synthetic dataset consisting of bracketed raw bursts simulated with our pipeline. Our method directly takes raw frames as an input. The other methods process RGB frames obtained here with VNG demosaicking. Our algorithm quantitatively outperforms the other HDR methods on this dataset, which is not surprising as it is trained leverage the information lost in the raw to rgb conversion.

Method	PSNR (dB)	μ-PSNR (dB)	SSIM (%)	μ-SSIM (%)	HDR-VDP2 (Q)
K=1 frames					
[Liu et al. 2020]	20.11	24.42	0.611	0.690	57.32
[Santos et al. 2020]	**22.14**	**25.85**	**0.641**	**0.702**	**62.94**
K=3 frames					
[Hasinoff et al. 2010] + MTB	28.08	29.46	0.819	0.847	61.13
[Hasinoff et al. 2010] + PLK	27.25	28.69	0.814	0.836	60.82
[Wu et al. 2018]	26.47	27.61	0.771	0.782	61.80
[Yan et al. 2021]	26.31	27.11	0.761	0.774	61.14
Ours	**33.75**	**34.39**	**0.942**	**0.943**	**63.24**
K=11 frames					
[Hasinoff et al. 2010] + MTB	29.54	30.96	0.862	0.892	62.07
[Hasinoff et al. 2010] + PLK	28.80	30.21	0.862	0.888	61.95
Ours	**37.83**	**39.22**	**0.964**	**0.971**	**65.44**

image fusion in the left image obtained with the plain-pixel Lucas-Kanade algorithm. These artefacts vanish in the image on the right using deep features.

5.5 Discussion

Choice of the prior function. An important component of our approach is the image proximal operator G_ω. Figure 12 shows a qualitative comparison of a prior-free version, solely aligning and merging the frames, using the image gradients soft-thresholding function derived from the classical TV-ℓ_1 prior, and our approach with a learnable module. The TV-based version is significantly sharper than that the one without prior. The parametric prior returns a better zoomed-in image, *e.g.,* next to the head and the dress of the statue.

Robustness on real images. A key advantage of our approach is the accuracy of its registration module, as detailed on Table 3 and illustrated in Figure 11. We have remarked that this module is particularly efficient for aligning raw frames with the same exposure, as illustrated by Figure 7 in the context of SR with factor ×4. Given

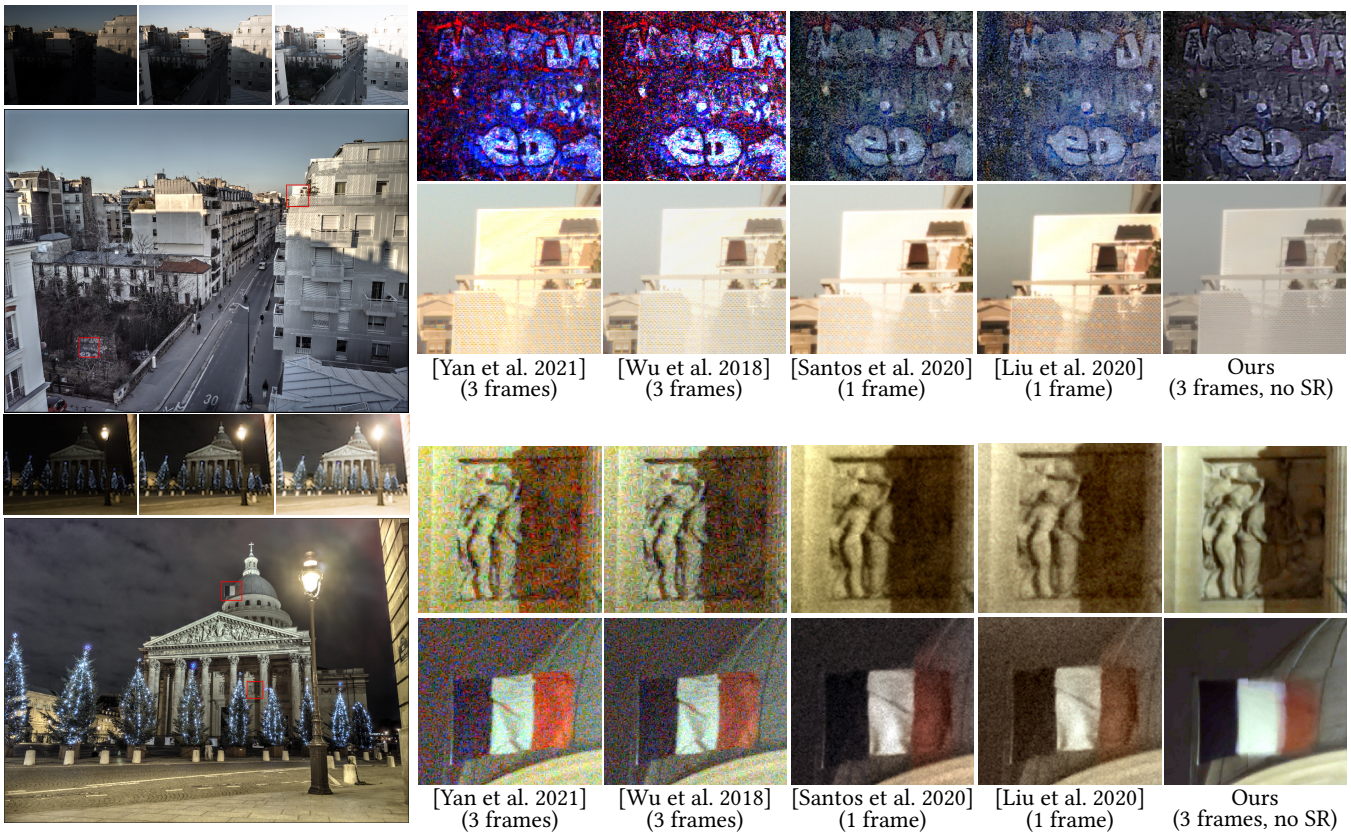


Fig. 10. Comparison with CNN-based HDR methods processing one to three input frames. **Left:** A sequence of three input frames, followed by our result after tone mapping, for two scenes. **Right:** Small crops from the scenes obtained by various methods. To be fair, we compare them with a version of our model that does not perform super-resolution (×1 upscaling factor) and only processes a burst of 3 images, in the EV range [-2.4,0,2.4]. We observe that, in well-exposed regions, the reconstruction performances of the three methods are similar. Our method appears to be more robust to noise, but more sensitive to non-rigid motion as shown in the case of the flag.

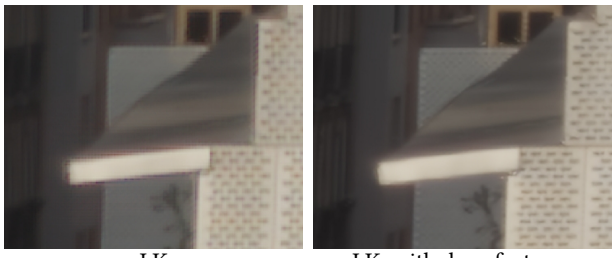


Fig. 11. Qualitative comparison of the reconstructed image with a pyramid of Lucas-Kanade run on plain pixels or deep features. Note the zipping artefacts along the edges of the large white rectangle. Our learnable variant is faster and leads to more accurate results. The reader is invited to zoom in.

a burst of raw photographs including moving objects with reasonable motion during exposure, *e.g.*, the pedestrian in the figure, we can predict high-quality HR image well-aligned with the reference frame whereas the competitors may introduce ghosting or colored artefacts. Notwithstanding, we have also noted that non-rigid motions in the raw frame burst may lead to blur in the final predicted image. For instance, Figure 10 compares the restoration results from Wu et al. [2018], Yan et al. [2021] and our model for a crop featuring a waving flag, *i.e.*, a non-rigid motion. We select $K = 3$ images with EV values of {-2.4,0,2.4} EV for the CNNs and for our model. The CNNs trained to remove ghosting artefacts accurately align the flag with the reference frame whereas our prediction is blurry in the red section of the flag. This may stem from the fact that multi-exposure image registration is a very challenging problem and that we have not such non-rigid motions in our training data. For a better deghosting, the injection of non-rigid motion in the training data, similarly to the dataset introduced in [Kalantari and Ramamoorthi 2017], is an interesting future research direction.

6 CONCLUSION

We have introduced an effective algorithm for the reconstruction of high-resolution, high-dynamic range color images from raw photographic bursts with exposure bracketing. We have demonstrated its excellent performance with super-resolution factors of up to

Table 3. Quantitative comparison of registration methods on synthetic data with average and median geometric errors [Sanchez 2016]. We compare MTB features combined with sub-pixelic phase correlation [Tursun et al. 2016], the pyramid Lucas-Kanade (PLK) algorithm and our variant of PLK using deep features. The three algorithms are run on the mosaicked images. On each benchmark, we outperform both vanilla PLK and the MTB-based approach.

Model	Geom (avg.)	Geom (med.)
×1 - No noise - 11 raw frames		
MTB + phase correlation	2.93	2.61
3 PLK iterations	1.32	0.97
3 PLK iteration +deep features (ours)	0.91	**0.60**
5 PLK iterations	1.47	1.10
5 PLK iterations + deep features (ours)	**0.88**	0.61
×1 - Noise - 11 raw frames		
MTB + Phase correlation	3.58	2.99
3 PLK iterations	2.77	2.40
3 PLK iteration +deep features (ours)	**1.25**	**0.95**
5 PLK iterations	2.76	2.10
5 PLK iterations + deep features (ours)	1.40	1.00
×4 (aliasing) - Noise - 11 raw frames		
MTB + Phase correlation	5.93	4.67
3 PLK iterations	3.82	3.58
3 PLK iteration +deep features (ours)	**2.04**	**2.03**
5 PLK iterations	3.87	3.50
5 PLK iterations + deep features (ours)	2.62	2.17

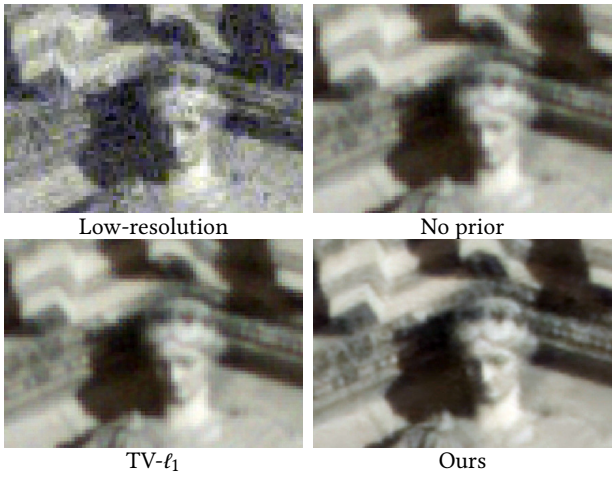


Fig. 12. Visual comparison of the impact of the prior for joint HDR and ×4 SR. We fuse $K = 20$ images in this example. We note for the three methods effectively suppress the noise present in the original LR frames. However, our learnable prior (here with 300k parameters) yields a higher quality image. The reader is invited to zoom in.

×4 on real photographs taken in the wild with hand-held cameras, and high robustness to low-light conditions, noise, camera shake, and moderate object motion. We have also shown that it compares favorably to the state of the art in both the HDR imaging and super-resolution domains. The keys to the success of this algorithm are the powerful underlying image formation model, a hybrid approach for solving the corresponding inverse problem that combines the interpretability and generalization power of model-based techniques with the flexibility and robustness of (deep) machine learning technology, and new twists on classical alignment techniques that make them robust by design. The proposed approach also has limitations: In particular its performance in saturated regions is far from perfect, and it is only robust to relatively small scene motions. Future work will include improving these aspects, but also adapting the method to large camera displacements (wide-baseline setting) and scene motions, which in turn will lead to 3D capture applications. A missing piece is of course handling motion and defocus blur, and we also plan to adapt our method to a new range of applications such as focus stacking, where image alignment requires comparing images with different focuses that may be even more different than those obtained under changing exposures. Finally, we are also interested in scientific applications in astronomy, microscopy, and remote sensing.

ACKNOWLEDGMENTS

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). JM and BL were supported by the ERC grant number 714381 (SOLARIS project) and by ANR 3IA MIAI@Grenoble Alpes (ANR-19-P3IA-0003). JP was supported in part by the Louis Vuitton/ENS chair in artificial intelligence and the Inria/NYU collaboration. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011011252R2 made by GENCI. This work was done while TE was a PhD student at Inria.

REFERENCES

Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. 2014. Best Algorithms for HDR Image Generation. A Study of Performance Bounds. *SIAM Journal on Imaging Science* 7, 1 (2014), 1–34.

Paul E. Anuta. 1970. Spatial Registration of Multispectral and Multitemporal Digital Imagery Using Fast Fourier Transform Techniques. *IEEE Transactions on Geoscience eletronics* 8, 4 (1970), 353–368.

Tunç Ozan Aydin, Rafal Mantiuk, and Hans-Peter Seidel. 2008. Extending quality metrics to full luminance range images. In *Proceedings of Human Vision and Electronic Imaging (SPIE Proceedings)*, Bernice E. Rogowitz and Thrasyvoulos N. Pappas (Eds.), Vol. 6806. SPIE, 68060B.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. 2018. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research (JMLR)* 18 (2018), 1–43.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. 2021a. Deep Burst Super-Resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 9209–9218.

Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. 2021b. Deep Reparametrization of Multi-Frame Super-Resolution and Denoising. (2021), 2460–2470.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. 2019. Unprocessing Images for Learned Raw Denoising. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 11036–11045.

Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. 2017. CLKN: Cascaded lucas-kanade networks for image alignment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 2213–2221.

Jongseong Choi, Min Kyu Park, and Moon Gi Kang. 2009. High Dynamic Range Image Reconstruction with Spatial Resolution Enhancement. *Computer Journal* 52, 1 (2009), 114–125.

Roger N. Clark. 2006. Digital Camera Reviews and Sensor Performance Summary. "https://clarkvision.com/articles/digital.sensor.performance.summary/".

Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. 2017. Pixel Recursive Super Resolution. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Paul E. Debevec and Jitendra Malik. 1997. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH*. ACM, 369–378.

Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. 2021. Deep Coupled Feedback Network for Joint Exposure Fusion and Image Super-Resolution. *IEEE Transactions on Image Processing (TIP)* 30 (2021), 3098–3112.

Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Khan, and Ming-Hsuan Yang. 2021. Burst Image Restoration and Enhancement. *arXiv preprint arXiv:2110.03680* (2021).

Gabriel Eilertsen, Saghi Hajisharif, Param Hanji, Apostolia Tsirikoglou, Rafal K. Mantiuk, and Jonas Unger. 2021. How to cheat with metrics in single-image HDR reconstruction. In *Proceedings of the workshops of the International Conference on Computer Vision (ICCVW)*. 3981–3990.

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafal K. Mantiuk, and Jonas Unger. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics* 36, 6 (2017), 178:1–178:15.

Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. 2017. Deep reverse tone mapping. *ACM Transactions on Graphics (ToG)* 36, 6 (2017), 177:1–177:10.

Manfred Ernst and Bartlomiej Wronski. 2021. HDR+ with Bracketing on Pixel Phones. "https://ai.googleblog.com/2021/04/hdr-with-bracketing-on-pixel-phones.html".

Sina Farsiu, Michael Elad, and Peyman Milanfar. 2006. Multiframe demosaicing and super-resolution of color images. *IEEE Transactions on Image Processing (TIP)* 15, 1 (2006), 141–159.

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen O. Egiazarian. 2008. Practical Poissonian-Gaussian Noise Modeling and Fitting for Single-Image Raw-Data. *IEEE Transactions on Image Processing (TIP)* 17, 10 (2008), 1737–1754.

Orazio Gallo, Marius Tico, Roberto Manduchi, Natasha Gelfand, and Kari Pulli. 2012. Metering for Exposure Stacks. *Computer Graphics Forum* 31, 2 (2012), 479–488.

Orazio Gallo, Alejandro J. Troccoli, Jun Hu, Kari Pulli, and Jan Kautz. 2015. Locally non-rigid registration for mobile HDR photography. In *(CVPRW)*. IEEE Computer Society, 48–55.

Donald Geman and Chengda Yang. 1995. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing (TIP)* 5, 7 (1995), 932–946.

Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik P. A. Lensch. 2010. Optimal HDR reconstruction with linear digital cameras. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 215–222.

Bahadir K. Gunturk and Murat Gevrekci. 2006. High-resolution image reconstruction from multiple differently exposed images. *IEEE Signal Processing Letters* 13, 4 (2006), 197–200.

Param Hanji, Fangcheng Zhong, and Rafal K. Mantiuk. 2020. Noise-Aware Merging of High Dynamic Range Image Stacks Without Camera Calibration. In *Proceedings of the workshops of the European Conference on Computer Vision (ECCVW)*. 376–391.

Samuel W. Hasinoff, Frédo Durand, and William T. Freeman. 2010. Noise-optimal capture for high dynamic range photography. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 553–560.

Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)* 35, 6 (2016), 192:1–192:12.

Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu abd Wolfgang Heidrich, Karen Egiazarian, Jan Kautz, and Kari Pulli. 2014. FlexISP: A flexible camera image processing framework. *ACM Transactions on Graphics (ToG)* 33, 6 (2014), 231:1–231:13.

Andrey Ignatov, Luc Van Gool, and Radu Timofte. 2020. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 536–537.

Nima Khademi Kalantari and Ravi Ramamoorthi. 2017. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 144:1–144:12.

Soo Ye Kim, Jihyong Oh, and Munchurl Kim. 2019. Deep SR-ITM: Joint Learning of Super-Resolution and Inverse Tone-Mapping for 4K UHD HDR Applications. In *Proceedings of the International Conference on Computer Vision (ICCV)*. 3116–3125.

Bruno Lecouat, Jean Ponce, and Julien Mairal. 2021. Lucas-Kanade Reloaded: End-to-End Super-Resolution from Raw Image Bursts. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Anat Levin, Robert Fergus, Frédo Durand, and William T. Freeman. 2007. Image and depth from a conventional camera with a coded aperture. *ACM Transactions on Graphics (ToG)* 26, 3 (2007), 70.

Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T. Barron, Dillon Sharlet, Ryan Geiss, Samuel W. Hasinoff, Yael Pritch, and Marc Levoy. 2019. Handheld mobile photography in very low light. *ACM Transactions on Graphics (ToG)* 38, 6 (2019), 164:1–164:16.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2020. Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 1648–1657.

Bruce D. Lucas and Takeo Kanade. 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 674–679.

Guillermo Luijk. 2007. Dcraw tutorial. "http://guillermoluijk.com/tutorial/dcraw/index_en.htm".

Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. 2021. EBSR: Feature enhanced burst super-resolution with deformable alignment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*. 471–478.

Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. 2017. Robust Multi-Exposure Image Fusion: A Structural Patch Decomposition Approach. *IEEE Transactions on Image Processing (TIP)* 26, 5 (2017), 2519–2532.

Henrique S. Malvar, Li-wei He, and Ross Cutler. 2004. High-quality linear interpolation for demosaicing of Bayer-patterned color images. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 485–488.

Steve Mann and Rosalind W. Picard. 1995. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proceedings of Is&T*. 442–448.

Julien NP Martel, Lorenz K Mueller, Stephen J Carey, Piotr Dudek, and Gordon Wetzstein. 2020. Neural sensors: Learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 7 (2020), 1642–1653.

Emil Martinec. 2008. Noise, Dynamic Range and Bit Depth in Digital SLRs. "https://photonstophotos.net/EmilMartinec/noise.html".

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul Srinivasan, and Jonathan T Barron. 2021. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. *arXiv preprint arXiv:2111.13679* (2021).

Antoine Monod, Julie Delon, and Thomas Veit. 2021. An Analysis and Implementation of the HDR+ Burst Denoising Method. *Image Processing On Line* 11 (2021), 142–169.

Manish Narwaria, Rafal K. Mantiuk, Matthieu Perreira Da Silva, and Patrick Le Callet. 2015. HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal on Electronic Imaging* 24, 1 (2015), 010501.

Shree K Nayar and Tomoo Mitsunaga. 2000. High dynamic range imaging: Spatially varying pixel exposures. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE, 472–479.

Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson W. H. Lau. 2021. HDR-GAN: HDR Image Reconstruction From Multi-Exposed LDR Images With Large Motions. *IEEE Transactions on Image Processing (TIP)* 30 (2021), 3885–3896.

Neal Parikh and Stephen P. Boyd. 2014. Proximal Algorithms. *Foundations and Trends in Optimization* 1, 3 (2014), 127–239.

Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. 2021. NTIRE 2021 Challenge on High Dynamic Range Imaging: Dataset, Methods and Results. In *CVPR Workshops*. 691–700.

Tobias Plötz and Stefan Roth. 2017. Benchmarking Denoising Algorithms with Real Photographs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 2750–2759.

Ali Ajdari Rad, Laurence Meylan, Patrick Vandewalle, and Sabine Süsstrunk. 2007. Multidimensional image enhancement from a set of unregistered and differently exposed images. In *Computational Imaging (SPIE Proceedings)*, Vol. 6498. SPIE, 649808.

Erik Reinhard, Michael M. Stark, Peter Shirley, and James A. Ferwerda. 2002. Photographic tone reproduction for digital images. *ACM Transactions on Graphics (ToG)* 21, 3 (2002), 267–276.

Javier Sanchez. 2016. The inverse compositional algorithm for parametric registration. *Image Processing On Line* (2016).

Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. 2020. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *ACM Transactions on Graphics (ToG)* 39, 4 (2020), 80.

Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. 2012. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Transactions on Graphics (ToG)* 31, 6 (2012), 203:1–203:11.

Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. 2016. Convolutional sparse coding for high dynamic range imaging. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 153–163.

Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. 2007. Kernel Regression for Image Processing and Reconstruction. *IEEE Transactions on Image Processing (TIP)* 16, 2 (2007), 349–366.

Yann Traonmilin and Cecilia Aguerrebere. 2014. Simultaneous High Dynamic Range and Superresolution Imaging without Regularization. *SIAM Journal on Imaging Science* 7, 3 (2014), 1624–1644.

Okan Tarhan Tursun, Ahmet Oguz Akyüz, Aykut Erdem, and Erkut Erdem. 2016. An Objective Deghosting Quality Metric for HDR Images. *Computer Graphics Forum* 35, 2 (2016), 139–152.

Patrick Vandewalle, Sabine Süsstrunk, and Martin Vetterli. 2006. A Frequency Domain Approach to Registration of Aliased Images with Application to Super-resolution. *EURASIP Journal on Advances in Signal Processing* 2006 (2006).

Subeesh Vasu, Abhijeet Shenoi, and A. N. Rajagopalan. 2018. Joint HDR and Super-Resolution Imaging in Motion Blur. In *Proceedings of the International Conference on Image Processing (ICIP)*. 2885–2889.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. 2013. Plug-and-play priors for model based reconstruction. In *Proceedings of the Global Conference on Signal and Information Processing*. 945–948.

Greg Ward. 2003. Fast, Robust Image Registration for Compositing High Dynamic Range Photographs from Hand-Held Exposures. *Journal on Graphics, GPU, & Game Tools* 8, 2 (2003), 17–30.

Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. 2019. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (ToG)* 38, 4 (2019), 28:1–28:18.

Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. 2018. Deep High Dynamic Range Imaging with Large Foreground Motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 120–135.

Qingsen Yan, Dong Gong, Javen Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. 2021. Dual-attention-guided network for ghost-free high dynamic range imaging. *International Journal of Computer Vision (IJCV)* (2021), 1–19.

Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. 2020. Deep HDR Imaging via A Non-Local Network. *IEEE Transactions on Image Processing (TIP)* 29 (2020), 4308–4322.

Kai Zhang, Luc Van Gool, and Radu Timofte. 2020. Deep unfolding network for image super-resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 3217–3226.

Henning Zimmer, Andrés Bruhn, and Joachim Weickert. 2011. Freehand HDR Imaging of Moving Scenes with Simultaneous Resolution Enhancement. *Computer Graphics Forum* 30, 2 (2011), 405–414.

A ABLATION STUDIES

In this section, we provide additional experiments to better understand the impact of our method's components.

A.1 Learning the image prior

We quantitatively validate the choice of a parametric proximal function in Eq. (11) to address joint HDR and SR by a factor of 4. We compare the proposed CNN-based implementation with the the classical total-variation ℓ_1 (TV-ℓ_1), for instance used by Choi et al. [2009], and a simple weighted least-squares problem, *i.e.,* without prior. We generate 266 bursts of $K = 9$ images with exposure values in $[-3, 3]$EV. Table 4 shows average PSNRs for methods embedding no prior (simple weighted least-squares), TV-ℓ_1 prior and three variants of the CNN G with parameter ω of sizes 30K, 300K and 3M. The variants with the parametric priors are trained according to the protocol described in the previous section. This table shows, as expected, a clear advantage of learnable penalty functions over handcrafted ones, with a margin of more than 4dB for the shallowest network and more than 7dB for the deepest one over the TV-ℓ_1 variant. Note that the prior-free version is only 0.4dB below its TV-ℓ_1-based counterpart, suggesting that machine learning is important to design an efficient prior function to address joint HDR and SR restoration.

Table 4. Quantitative comparison of the choice of the prior over the total performance of the method. Average PSNR on predicted linear HDR images jointly super-resolved by a factor of 4 for typical handcrafted image priors and variants of the proposed parametric one with several parameter sizes. The learnable ones achieve the best scores overall by an important margin. The more parameters yields the best PSNRs.

Prior	PSNR
No prior	26.15
TV-ℓ_1	26.51
Tiny (30k)	30.71
Small Prior (300k)	<u>32.56</u>
Large Prior (3M)	**34.18**

A.2 Alignement sub-components evaluation

We quantify the impact of each components of the alignment module in Table 5 by measuring the mean PSNR of predicted HDR images with resolution enhanced by ×4. We generate 266 raw bursts with 11 frames for each burst with the same protocol than the other experiments. We decompose it into three bricks: using bracketed images, the confidence function g and running the pyramid Lucas-Kanade (PLK) algorithm on deep features. Adding each component one-by-one gradually increases the mean PSNR, the maximum value being naturally reached when the three components are gathered. Note that the PLK algorithm run on deep features brings an improvement of about +2dB, which alone is a better contribution than the total of +1.3dB by combining bracketed images and the confidence function g. We also give an upper-bound to this performance by running a version of our model where we give the ground-truth motion to align the images. Such an oracle model achieves an average PSNR of 34.18dB compared to the 31.42dB of the best setting where the motion is estimated instead. It suggests that there is room to improvement but each sub-components in the alignment actually helps to further narrow the gap with the oracle model.

A.3 Performance with the number of frames

We compare the performance of our approach with respect to the number of frames in the burst for joint SR and HDR on synthetic data. Figure 13 shows the mean PSNR taken over 3 seeds for bursts of length ranging from 3 to 30. Our model greatly benefits from additional

Table 5. Ablation study for the alignment module. We report mean PSNR on HDR images with ×4 super-resolution. The first configuration (#1) uses a burst but no bracketing (constant exposure). The fourth configuration (#4) is the setting we use in practice, with bracketed exposures, the confidence function g and the pyramid Lucas-Kanade algorithm run on learnt features. Adding these three components one by one gradually improves the mean PSNR, showcasing the importance of each module. The fifth (#5) configuration is an upper bound where we use the ground-truth motion (and thus do not need LK with deep deatures).

Settings	#1	#2	#3	#4	#5
Bracketing		✓	✓	✓	✓
Confidence function g			✓	✓	✓
LK with deep features				✓	
Oracle motion					✓
PSNR	28.50	29.28	29.77	31.42	34.18

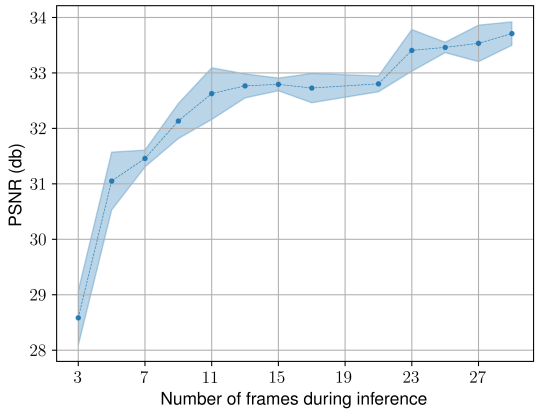


Fig. 13. Average PSNR on predicted HDR images with spatial resolution increased by (×4), from a varying number of frames in the burst (from 3 to 30). Our approach benefits from any additional input frame, especially for less than $K = 11$ images. The average PSNR is evaluated from 3 seeds.

frames for burst sizes smaller than 11; Starting from 3 images and a PSNR score of 28.6dB, we gain up to 4dB when accumulating 11 frames. Beyond this number, we gain an extra decibel by accumulating more than 20 frames. It is consistent with typical bracketing techniques, *e.g.,* [Granados et al. 2010; Hasinoff et al. 2010], for which more images means better noise removal in the dark regions. Thanks to the learnt robust registration algorithm and prior, the performance of our approach hardly falls down when accumulating more and more images, unlike typical multi-image algorithms that may accumulate registration, *e.g.,* as noted by Wronski et al. [2019].

A.4 Computational speed

Our algorithm leverages optimization and machine learning techniques, which leads to a dramatically smaller number of parameters than state-of-the-art CNNs for tasks such as super-resolution. We evaluate the computational speed and memory consumption of our model embedding three variants of the learnable operator G with varying number of parameters. We compare our three versions of the proposed network with that of [Bhat et al. 2021a] and [Luo et al. 2021], the best performers for ×4 SR in Table 1. We run this five-way comparison within the same python environment, *e.g.,* same version of Pytorch, and on the same GPU (Nvidia Titan RTX) for fairness. We show in Table 6 that our hybrid method exceeding the SR state of the art in the previous paragraphs, is also the lightest in the panel. The method of [Luo et al. 2021] has 26 million parameters and that of [Bhat et al. 2021a] about 13 million parameters whereas our deepest model has 3 million of them, *i.e.,* four time less than [Bhat et al. 2021a]. This gap in size of parameter is due to the building blocks in the competitors' architectures. Indeed, they heavily rely on memory-greedy attention modules, whereas our implementation of G is based on the fully convolutional U-net architecture of [Lecouat et al. 2021; Zhang et al. 2020]. This table also shows that, for resolution factor of ×4, our approach is much faster than the state of the art, while coping with them according to Table 1. In this table, our "Small" model is less than a decibel below Luo et al. [2021]'s model but with an inference time forty times smaller on the same GPU. We are also four times faster than Bhat et al. [2021a]. Likewise, our models require three to four times less GPU memory than our selected competitors, which is an important designing point for deploying such a technology in commercial software running on consumer-grade devices. We also report in Table 6 information about the ×2 case since in many situations pushing the resolution further brings little improvement, *e.g.,* Figure 8 and the analysis of Wronski et al. [2019]. In this configuration, our model requires even less memory to process 400×400 tiles and may run on modest GPUs.

A.5 Limitations

Albeit our approach favorably addresses HDR, SR and joint HDR and SR against the state of the art, we have noted throughout our experiments a few limitations in certain cases that may degrade the performance of our model.

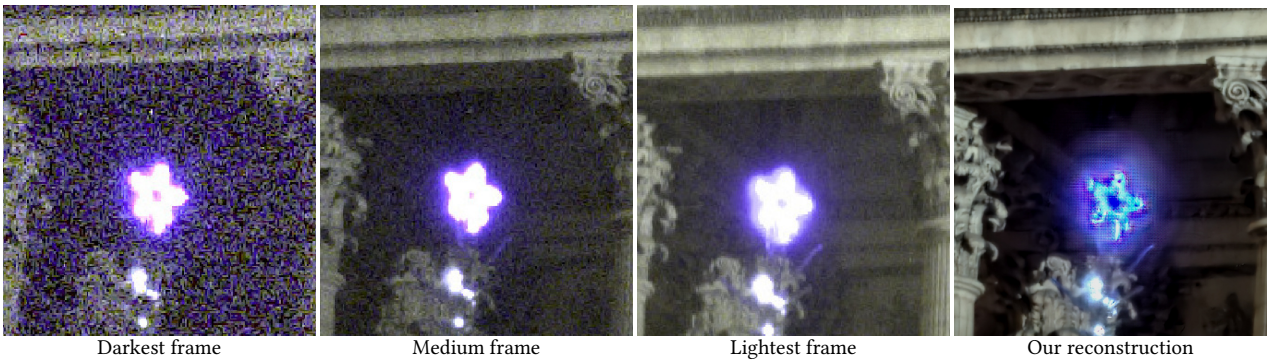
Lack of robustness to non-rigide motion for joint HDR and SR.. We have observed that our model which performs joint HDR and super-resolution are less robust to non-rigid motion than our models performing only burst super-resolution. An example of artefacts that we typically get is shown in Figure 10 in the case of the moving flag.

Saturated areas. In the pictures shot with a smartphone, we have sometimes noticed color halos next to saturated areas (Figure 14). They may be caused by the fact that the corresponding very high level of contrast is hard to simulate in our synthetic data.

Hot pixels. The method is not trained to correct hot pixels, that may locally alter HDR imaging techniques. We assume that these pixels are corrected upstream in the camera pipeline, which is a classical assumption in the field.

Table 6. Comparison of inference speed for different models for burst super-resolution. We have benchmarked the inference speed of different models for processing a burst of 14 12Mpixel raw images (Pixel 4a) on a single Titan RTX GPU. We have used the official implementations released by the authors without any modification. We have not optimized inference speed (yet) using with standard tools such as mixed precision and/or model compression.

Model	# parameters	Runtime	Memory (200x200)	Memory (400x400)
Competitors (×4)				
[Bhat et al. 2021a]	13,000k	40.0sec	3.5Gb	11.5Gb
[Luo et al. 2021]	26,000k	9.5min	3.5Gb	12Gb
Ours (×4)				
Very Small	60k	13.4sec	1.2Gb	2.8Gb
Small	250k	20.0sec	1.2Gb	2.8Gb
Large	3,000k	38.2sec	1.3Gb	3.1Gb
Ours (×2)				
Very Small	60k	4.7sec	800Mb	1.2Gb
Small	250k	6.2sec	820Mb	1.2Gb
Large	3,000k	10.7sec	860Mb	1.3Gb



Darkest frame | Medium frame | Lightest frame | Our reconstruction

Fig. 14. Color halos in very bright areas and saturated regions. Because all input views are saturated in the neighborhood of the star, key info required to correctly reconstruct the image there is missing, leading to severe artefacts there.

A.6 Supervision with various loss functions

In this subsection, we propose an ablation study to assess the effectivness of different supervision loss instead of the basic L1 loss. We use the same loss function as the one described in [Mildenhall et al. 2021], which gives a tone curve $\psi(x) = \log(x + \epsilon)$ which more strongly penalizes errors in dark regions. Results of the ablation study are presented in Table 7. Our experiment that the log loss gives better result in term of μ-psnr. The selection of the right supervision loss for the training of our model is an interesting direction for future research.

loss	psnr (dB)	μ-psnr (%)	μ-ssim (%)
L_1	36.52	38.08	0.9682
$\log(\epsilon = 10^{-1})$	37.04	37.86	**0.964**
$\log(\epsilon = 10^{-2})$	**37.71**	**38.70**	0.967
$\log(\epsilon = 10^{-3})$	13.74	8.07	0.199

Table 7. Super resolution factor x1, ablation study with different training loss.

A.7 Ablation HDR with no motion

In addition to Table 2, we provide in Table 8 below a HDR fusion evaluation of the same methods, but on a variant of the synthetic test set where we have not simulated motion between frames. In this new table, our method still achieves the best HDR-VDP scores for brackets of both $K = 3$ and $K = 11$ images, but with a margin of only 1 point over our implementation of [Hasinoff et al. 2010]. We can conclude that both

Table 8. Quantitative comparison of various algorithms for HDR imaging – we do not perform super-resolution in this experiment – on a synthetic dataset consisting of bracketed raw bursts simulated with our pipeline. Our method directly takes raw frames as an input. The other methods process RGB frames obtained here with VNG demosaicking. Our algorithm quantitatively outperforms the other HDR methods on this dataset, which is not surprising as it is trained leverage the information lost in the raw to rgb conversion

Method	psnr (dB)	μ-psnr (dB)	ssim	μ-ssim (dB)	HDR-VDP2(Q)
K=1 frame					
Liu *et al.* [Liu et al. 2020]	19.98	24.25	0.608	0.687	56.51
Santos *et al.* [Santos et al. 2020]	22.05	25.80	0.635	0.699	61.69
K=3 frames					
Wu *et al.* [Wu et al. 2018]	26.42	27.51	0.765	0.774	61.04
Yan *et al.* [Yan et al. 2021]	26.22	27.01	0.752	0.768	60.37
Hassinof *et al.* [Hasinoff et al. 2010]	30.55	31.26	0.874	0.878	67.77
Ours	**34.29**	**34.31**	**0.945**	**0.934**	**68.63**
K=11 frames					
Hassinof *et al.* [Hasinoff et al. 2010]	33.80	33.43	0.917	0.927	68.86
Ours	**38.73**	**38.56**	**0.973**	**0.969**	**70.05**

approaches achieve similar results. However, when we compare these margins that of 3 points in Table 2 between our approach over [Hasinoff et al. 2010], it suggests that our technique is more robust to alignment failures than the pure bracketing technique of Hasinoff et al. [2010].

B IMPLEMENTATION DETAILS

We include below details about our datasets and implementation for reproducibility purposes. See also Table 6 for the number of parameters used in different variants of our method.

Data Generation. Given a collection of sRGB images, we construct bursts of LDR low-resolution raw images and HDR/high-resolutions RGB targets. For the generation of realistic raw data from sRGB images, we follow the approach described in [Bhat et al. 2021a], using the author's publicly available code on the training split of the Zurich raw to RGB dataset [Ignatov et al. 2020]. The approach consists of applying the inverse RGB to raw pipeline introduced in [Brooks et al. 2019]. For the training of our model, we generate bursts of 11 frames of size 256x256 with random motions. Displacements are randomly generated, applying random translations of ±6 pixels and random rotations of $\pm 1^{\circ}$. Frames are downscaled with bilinear interpolation in order to simulate LR frames containing aliasing.

We then apply two different random gains in order to simulate frames with varying exposure. First, a random gain in the range [-5ev,5ev] is applied to the ground-truth image, in order to simulate 32-bit ground-trugh images with a large dynamic. We then apply a different gains in the range [-3ev, 3ev] for each image of the burst in order to simulate images with different exposure times. This results in synthetic bursts with different saturated areas and signal to noise ratios. Synthetic noise is added to the frames. The noise levels are sampled following the empirical model of Figure 3. Finally, color values are discarded according to the Bayer pattern.

Validation Split. In order to perform further comparison and conduct the ablation study, we build a validation set by randomly extracting 266 images from the Zurich raw to RGB dataset [Ignatov et al. 2020].

Model. We summarize in Figure 15 our proposed pipeline. In all our experiments, we unroll 3 iterations of the HQS algorithm.

Deep Prior. We give more details about the architecture of the deep prior used in our experiments. For all our experiments, we use a smaller variant of the ResUNet architecture introduced in [Zhang et al. 2020] for single-image super resolution. This architecture involves four scales, each of which has an identity skip connection between downscaling and upscaling operations. Downscaling operations are implemented using 2x2 strided convolution while upscaling are implemented with pixel-shuffling. Each residual block is made of two 3x3 convolution layers and ReLU activation combined with an identity skip. For each scale we apply a cascade of 2 residual blocks. The network has respectively 32,64,128,128 channels for each convolution per scale.

Model variants. We also run experiments with an even smaller version of the network with 32 features per channel (dubbed small) and 16 features per channel (dubbed tiny).

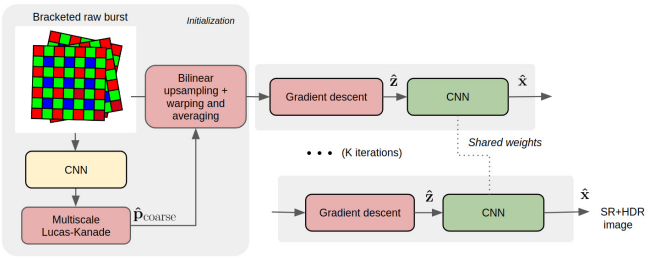


Fig. 15. A diagramatic view of our model.

Training procedure. We minimize Eq. 13 using Adam optimizer with learning rate set to $10^{-}5$ for 400k iterations. We decrease the learning rate by a factor 2 every 100k iterations. The weights of the CNNs are randomly initialized with the default setting of the PyTorch library. Our approach is implemented in Pytorch and takes approximately 2 days to train on a Nvidia Titan RTX GPU.