Digital Gimbal: End-to-end Deep Image Stabilization with Learnable Exposure Times

Omer Dahary¹, Matan Jacoby¹, and Alex M. Bronstein¹

¹Technion - Israel Institute of Technology
{omerd,matanj,bron}@cs.technion.ac.il

Abstract

Mechanical image stabilization using actuated gimbals enables capturing long-exposure shots without suffering from blur due to camera motion. These devices, however, are often physically cumbersome and expensive, limiting their widespread use. In this work, we propose to digitally emulate a mechanically stabilized system from the input of a fast unstabilized camera. To exploit the trade-off between motion blur at long exposures and low SNR at short exposures, we train a CNN that estimates a sharp high-SNR image by aggregating a burst of noisy short-exposure frames, related by unknown motion. We further suggest learning the burst’s exposure times in an end-to-end manner, thus balancing the noise and blur across the frames. We demonstrate this method’s advantage over the traditional approach of deblurring a single image or denoising a fixed-exposure burst on both synthetic and real data.

1. Introduction

Through advances in imaging technology, optical systems have become lighter and more portable than ever. These improvements, however, are sometimes negated by the need for stabilization. Due to the natural tremor of handheld cameras, or the movement of mounted vehicles, camera motion is often unavoidable. As a result, images captured this way may suffer from motion blur, which is especially evident during long exposures or with long focal lengths. Image stabilization, the task of mitigating this phenomenon, has been mostly explored in the optical and mechanical domains. In the optical regime, stabilization emerges from using high-cost, uniquely tailored lenses or cheaper but less effective shift sensors. On the other hand, mechanical stabilization devices, such as actuated gimbals, can be attached to any camera with no need for complicated optical equipment. Such systems are highly popular but limited in their ability to compensate for blur due to scene motion or atmospheric turbulence. However, the most severe limitation of mechanical image stabilization is its cost, form factor, weight, and power requirements, which might be prohibitive in many scenarios.

Figure 1. A qualitative evaluation of our approach on images captured by a vibrating camera with a long focal length. (See the supplement for a description of our setup.) The fully-exposed image (a) suffers from considerable motion blur, which is hard to mitigate using image deblurring methods (b). Employing the same time budget to capture a fixed-exposure burst and merging it via a burst denoising algorithm (c) results in over-smoothing and ghosting artifacts. Our approach (d) of learning a non-uniform exposure regime enables utilizing the sharpness of short-exposed frames and the high SNR of long-exposed ones to yield a superior reconstruction.
In recent years, deep learning approaches have been successfully utilized for solving traditional image processing tasks. The field of burst and multi-frame imaging, in particular, has seen rising attention, with a variety of works on denoising [19] and deblurring [1, 29]. These methods exploit the information entangled in different degraded realizations of the same scene to reassemble them into a single high-quality image. Relying on this concept, together with the well-known trade-off between strong blur at long exposures and low SNR at short exposures, we propose solving the image stabilization task using a burst of images, each captured consecutively over a short period. Thus we reduce the problem to joint burst denoising and deblurring, as the two are already well-established. While most burst methods assume the data across the burst are complementary, we further suggest explicitly intervening in the acquisition pipeline and forcing diversity in noise and blur levels. We achieve this by designing an end-to-end learned system that includes both a deep burst image processing network and the camera’s exposure configuration. The latter is implemented using a novel network layer that numerically models the frame acquisition process, and can potentially be further tuned for other imaging tasks. To the best of our knowledge, this approach has never been attempted before for still image reconstruction from SNR-limited bursts in the presence of camera motion.

2. Related work

Burst imaging aims to compensate for non-optimal optics or imaging conditions by fusing frames captured sequentially over a short period. This ongoing research field has been excessively studied for denoising and deblurring, with most works focusing on a single kind of degradation.


The task of image restoration from bursts with non-uniform exposure times was first explored using traditional image processing tools. Ben- Ezra and Nayar [2] introduce a hybrid imaging system that records camera motion to predict the point spread function of a blurry image. Zhang et al. [33] use a sparse prior to adaptively combine information from noisy and blurry inputs. Yuan et al. [31] employ a noisy/blurry image pair of the same scene to estimate the blur kernel and reduce ringing artifact. More recently, researchers have been adopting the idea of using short- and long-exposure pairs as the input to deep learning-based restoration algorithms. These include tasks such as high dynamic range (HDR) imaging [13], deblurring [34], and low-light restoration [4]. However, these works rely on fixed and pre-selected timing regimes, with two separately-tuned data generation pipelines: one for the short-exposure frames, and another one for the long-exposure ones.

In parallel, recent works by Google Research have shown that burst imaging can mitigate smartphone cameras’ limitations by increasing their dynamic range [8], spatial resolution [30], and performance in low light conditions [18]. These improvements are achieved by an alternative image signal processing (ISP) pipeline that robustly aligns and merges frames and applies post-processing effects. Furthermore, it introduces dynamic exposure selection: either by interpolating a hand-crafted database, or by aggregating motion prediction with inertial sensor data.

3. Approach

We assume the scenario in which a latent scene is captured by a moving camera over a temporal interval \([0, T]\). With some abuse, we denote the irradiance image at time \(t\) as a latent transformation \(\tau_t\) of a latent irradiance \(E\), i.e., \(E_t = \tau_t(E)\). The irradiance images are not directly observable; instead, the camera acquires a discrete set of \(n\) frames \(Y_1, \ldots, Y_n\). Each frame is captured during its integration interval \([t_i, t_i + \Delta t_i]\) \(\subset [0, T]\), where \(t_i\) denotes the opening time of the shutter, and \(\Delta t_i\) the frame exposure. Each frame is related to the latent image through the sensor forward model \(F:\)

\[
Y_i = F(\tau_{t_i \in [t_i, t_i + \Delta t_i]}(E)). \tag{1}
\]

Assuming the exposure times \(\Delta t_i\) are relatively low (around a few milliseconds), each frame in the burst suffers from low SNR due to imaging noise. Furthermore, we expect the frames to be slightly blurred due to the camera movement. The trade-off between these two sources of degradation is controlled by the exposure time \(\Delta t_i\), with longer exposures corresponding to better SNR and stronger blur, and vice versa.

Our aim is to utilize the information entangled in these measurements of the scene, by combining the frames into a single sharp high-SNR estimation of the latent irradiance image,

\[
\hat{E} = I(Y_1, \ldots, Y_n), \tag{2}
\]

where \(I\) is the reconstruction (inverse) operator.

We propose to learn the latter reconstruction operator concurrently with the user-controlled parameters of the camera, which in our case is the shutter schedule \(\{t_i, \Delta t_i\}\).
This leads to a learning problem of the form
\[
\min_{\Delta, I} \mathbb{E} f \left( I \left( \{ F \left( \tau \in [t_i, t_i + \Delta t_i] (E) \} \} \right)_{i=1}^n \right), E. \tag{3}
\]
Here \( \mathbb{E} f \) denotes the expectation of a loss function measuring the discrepancy between the latent \( E \) and its estimated version \( \hat{E} \), and \( \Delta \) describes the burst parameters detailed in the sequel. In what follows, we describe the construction of differentiable forward and inverse models and their learning scheme.

4. Sensor forward model

This section describes our numerical forward model (denoted as \( F \) above) of the imaging process, starting with the scene irradiance and ending with a fully formed raw image. We generally follow the noise estimation of Konnik and Welsh [16], while paying particular attention to time-dependent properties and visibly dominant factors at short exposures.

4.1. Image formation

Our model describes a moving camera capturing short-exposure images of a static scene. For simplicity, we assume that both the camera and the light source are monochromatic in relation to the same wavelength \( \lambda \). We further assume device-specific parameters are known, as they are customarily supplied by manufacturers [10].

From photons to electrons. Let \( E_t (x) \) be the irradiance of a pixel \( x \) at time \( t \). We can express it in terms of the photon flux
\[
\gamma_t (x) = \frac{\lambda A E_t (x)}{h c}, \tag{4}
\]
where \( \lambda \) denotes the wavelength, \( A \) is the effective pixel area, \( h \) is the Planck constant, and \( c \) is the speed of light.

Each collected photon deposits its energy in the form of electric charge, generating photoelectrons inside the pixel. Their mean amount is given by
\[
\bar{e} (x) = \eta_{\lambda} \int_t^{t + \Delta t} \gamma_t (x) \, dt, \tag{5}
\]
where \( \Delta t \) is the time interval on which the shutter was open, and \( \eta_{\lambda} \) is the quantum efficiency of the pixel at wavelength \( \lambda \).

Noise generation. The photons are not the sole source of charge in the pixel. Some amount of current, named the dark current, thermally deposits charge in it, even if the scene is completely dark. The mean number of electrons it generates grows linearly with the exposure time:
\[
\bar{e}_0 = \frac{I_0 \Delta t}{q_e}, \tag{6}
\]
where \( I_0 \) is the average dark current, and \( q_e \) is the elementary charge.

The measured number of electrons fluctuates randomly, obeying Poisson statistics due their discrete nature. This phenomenon, known as shot noise, is often approximated in the image processing domain by a Gaussian distribution [4, 13, 19], which is a valid assumption in high-light regimes. However, since we are dealing with short exposures, at which shot noise dominates other components, we model photoelectron production as a realization of a Poisson variable,
\[
e_q (x) \sim \text{Poiss} (\bar{e} (x) + \bar{e}_0). \tag{7}
\]

Another major source of noise, the readout noise, is independent of the exposure time. It mostly originates in thermal fluctuations in the analog-to-digital converter (ADC) and follows a zero-mean Gaussian distribution with some device-specific standard deviation, \( e_{\text{ro}} (x) \sim \mathcal{N} (0, \sigma_{\text{ro}}^2) \).

Therefore, we can express the total number of collected electrons as
\[
e (x) = e_q (x) + e_{\text{ro}} (x). \tag{8}
\]

From electrons to digital numbers. The pixel circuits convert the collected electrons into voltage, which is then amplified and translated into digital numbers (DNs). Following the EMVA 1288 standard [10], we assume this process to be described by an almost linear curve, whose sensitivity dampens towards the full-well capacity (FWC) of the
In order to make the sensor forward model amenable to learning, we assume the total time budget $T$ and the number of frames $n$ in the burst to be fixed, and parametrize the individual frame exposure parameters as

$$
\Delta t_i = \Delta t_{\text{min}} + \alpha_i (T - n (\Delta t_{\text{min}} + \Delta t_{\text{ro}})).
$$

Here $\Delta t_{\text{min}}$ is the minimum exposure time, and $\Delta t_{\text{ro}}$ is the frame readout time (including any additional blank time needed between two consecutive frames). The parameters $\alpha_i$, representing the relative frame exposures, are, in turn, parametrized as

$$
\alpha_i = \sigma(\Delta)_i = \frac{e^{\Delta_i}}{\sum_{j=1}^{m} e^{\Delta_j}},
$$

where $\sigma$ denotes softmax. The vector $\Delta$, serving as the trainable camera parameters, can be set to be either $n$-dimensional, in which case the burst consumes the entire available time budget, or $(n + 1)$-dimensional, allowing not to utilize all the available time (Fig. 3).

4.3. Numerical approximation

The learning of the optimal burst exposure parameters $\Delta$ mandates a differentiable calculation of the forward model described by equations (5-11). While most of these computations are straightforward, special consideration should be taken in steps (5,7,10).

The forward pass of (5) calculates a temporal integral of a continuous irradiance function $E_t(x)$, scaled according to (4). At training, we approximate this integral via the trapezoidal method from a sequence of $N \gg n$ uniformly-sampled irradiance values $\{E_t(x)\}_{i=1}^{N}$ obtained from simulated scene flow.

We discuss the backpropagation through equations (7,10) in the supplement.

5. Reconstruction network

The following section describes our inverse model (denoted as $I$ in Section 3) responsible for estimating the clean irradiance image $I$ given the burst frames $Y_1, \ldots, Y_n$. We base our reconstruction model on the recent work in kernel prediction networks (KPN), which were first introduced by Jia et al. [11], and have shown promising results in various image processing tasks, including video interpolation [21, 22], super-resolution [12], and deblurring [24], and burst denoising [19]. Its success in the latter two problems encourages its use in this work, as our goal is to merge different noisy and blurry realizations of the same scene. To do so, we predict temporally- and spatially-variant kernels and apply them to merge the captured burst into a clean and sharp image.

Since fixed-size kernels are limited in their ability to align frames, we pre-warp the burst according to a reference one, as customarily applied in video deblurring works [15, 25]. Gast and Roth [7] proposed using pre-trained flow networks as the most efficient option for the latter operation. However, this approach is impractical in our case, as the amount of noise and blur in the input frames is expected to change while the exposure times are learned. Therefore, we opted for an end-to-end training methodology, in which the flow network parameters are co-learned, not specifically for perfect alignment, but to compensate for the kernels’ narrow receptive field.

A schematic representation of our network is summarized in Fig. 2.

5.1. Flow network

The goal of the flow network is to align the frames $\{Y_i\}_{i=1}^{n}$ according to a pre-selected reference one $Y_{i_0}$. Since the entire burst is captured during continuous motion, we choose $Y_{i_0}$ to be the middle frame, as we expect it to overlap the others significantly, making alignment easier. Aiming to deal with non-rigid motion and parallax, we
Figure 4. Our flow network obeys the architecture proposed by Kalantari and Ramamoorthi [13]. Our sub-CNN comprises four \(5 \times 5\) convolutions intertwined with ReLUs, with the number of channels being \(n, 100, 50, 25, 2 (n - 1)\). In every scale, we feed the sub-CNN the appropriate pyramid level of the normalized burst, each frame in its own channel. We interpret the output as \(XY\) displacement fields for each non-reference frame.

use flow networks [6, 9, 23, 26]. However, these methods align individual images and therefore are prone to fail on degraded input. Kalantari and Ramamoorthi [13] handle this concern by jointly aligning the reference’s neighboring frames, thus complementing missing information. We follow the same architecture and simultaneously align the entire burst.

The network utilizes a hierarchical approach [13, 23, 27], which develops and refines a flow from a Gaussian pyramid of the input. At each level, from coarse to fine, we apply the same sub-CNN. At the coarsest level, we treat the sub-CNN’s output as a displacement field, which we then upsample and apply to the next level before feeding it to the sub-CNN again. Then, we sum the output flow with the upsampled one to produce the next flow in the chain and repeat. This architecture, depicted in Fig. 4, has the benefit of producing large-scale and high-precision pixel displacements.

Note that since the exposure times are not stable, the burst’s brightness varies from frame to frame and during training. Therefore, we normalize the frames according to their portion of the exposure time budget before feeding them into the flow network:

\[
Y^n_i = \frac{T}{\Delta t_i} Y^n_i. \tag{14}
\]

5.2. Kernel prediction network

Fig. 5 depicts our kernel prediction architecture, which follows the noise-blind version of the network suggested by Mildenhall et al. [19]. It is designed as an encoder-decoder with skip connections, which predicts adaptive kernels for each input pixel. The kernels are then applied to merge all given frames into the resulting image.

The use of pixel-adaptive kernels has a few key benefits. They reduce noise by pixel averaging without risking crossing edges and can also fix non-uniform blur. The latter is especially evident in our problem setting, where long focal lengths are commonly applied [28]. Moreover, by picking the most reliable frames for each area in the final image, the kernels allow us to exploit the variability across the burst (see Fig. 6). This choice is affected by each frame’s exposure time and content but can also compensate for alignment artifacts by the flow network. Thus, similarly to Kalantari and Ramamoorthi [13], we merge the original and registered burst to allow the former to be picked over the other.

To generate the output, we first take the pixel-wise mean of the resulting frames:

\[
\hat{E} = \Gamma \left( \frac{1}{2n - 1} \sum_{i=1}^{n} Y^n_i \odot k^n_i + \sum_{i \neq i_0} Y^n_i \odot k^n_i \right) \tag{15}
\]

Here \(\{Y^n_i\}_{i=1,i \neq i_0}^{n}\) are the aligned frames, \(\{k^n_i\}_{i=1,i \neq i_0}^{n}\) are the predicted location-dependent kernels, and \(\odot\) denotes the application of the kernel. Furthermore, to improve perceptual quality, we apply the differentiable gamma correcting function \(\Gamma\) [19].

6. Training

6.1. Loss function

Given the ground-truth image \(E\), we define our basic loss function as:

\[
\ell_{\text{basic}} \left( \hat{E}, E \right) = \left\| \hat{E} - E \right\|_1 + \mu \left\| \nabla \hat{E} - \nabla E \right\|_1, \tag{16}
\]

where \(\nabla\) are the combined horizontal and vertical Sobel filters, and \(\mu\) is a fixed constant. We choose the \(\ell_1\) loss as it was shown to promote sharp outputs, while the second term is applied as a regularization aiming to suppress patterned artifacts [34].

In our experiments, we found that while this loss yielded good results, the predicted kernels ignored \(\{Y^n_i\}_{i \neq i_0}^{n}\), as the flow network fails to properly align the burst. Thus, similarly to Mildenhall et al. [19], we propose using an annealed
where $\alpha$ and $\beta$ are fixed constant, and $\ell_{\text{basic}}$ is the pixel-wise sum over all $E$.

\[ \ell \left( \hat{E}, E, E \right) = \ell_{\text{basic}} \left( \hat{E}, E \right) + \beta \alpha^t \ell_{\text{basic}} \left( \hat{E}^a, E \right), \]

where $\beta, \alpha \in (0, 1)$ are fixed constant, and $t$ is the current iteration number. Hence, the loss in the first phase of training will produce gradients that compensate for the kernels’ narrow receptive field by improving the flow network’s performance.

### 6.2. Synthetic dataset

While standard burst and video denoising and deblurring algorithms operate on a few neighboring frames, our method must be evaluated on burst frames matching the irregular exposure regime dictated by our training on dense irradiance maps. Therefore, in the absence of any existing dataset that allows this flexibility, we generated our own.

Since our network is proposed as an alternative to mechanical gimbals, we create our data using 720p scenery drone videos from YouTube. We divide each video into short clips of 31 frames while skipping 500 frames between each clip. This process yielded around 10,000 different clips. Since undersampling may cause discontinuous artifacts in the motion trajectory, we follow other blur simulating works [3, 20] and increase the frame rate using a CNN for frame interpolation [22]. Although deep interpolation algorithms may create unreliable individual frames, they can still handle nonlinear motion and produce naturally looking blur when averaged. We feed each clip to the CNN eight times, resulting in 241 frames. Since the middle frame is not artificially generated, we use it as the ground-truth.

At each training step, we randomly choose a $512 \times 512$ patch from the input clip. Although it already depicts camera motion, which is imperative for producing parallax and occlusions, we further augment the set by adding a controlled amount of motion using homographies. Since drone videos are usually filmed using long focal lengths, their primary source of blur is ego-rotations [28]. We therefore apply randomly aggregated 3D rotations from frame to frame, which are computed using a pre-selected camera intrinsic matrix. Choosing the range of angles to rotate by affects the amount of blur in the final burst. However, these transformations may damage the original video quality and introduce black margins. To mitigate this degradation, we crop...
Table 1. Average PSNR and SSIM for our synthetic test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMPHN [32]</td>
<td>17.99</td>
<td>0.6391</td>
</tr>
<tr>
<td>Analysis-synthesis network pair [14]</td>
<td>18.6</td>
<td>0.7308</td>
</tr>
<tr>
<td>DeblurGAN-v2 [17]</td>
<td>23.12</td>
<td>0.7230</td>
</tr>
<tr>
<td>KPN [19]</td>
<td>28.46</td>
<td>0.8296</td>
</tr>
<tr>
<td>Ours</td>
<td>31.42</td>
<td>0.8948</td>
</tr>
</tbody>
</table>

7. Experiments

7.1. Baseline

We analyze the performance of our method against two traditional approaches: burst denoising and single image deblurring. While our algorithm incorporates non-uniformly-exposed bursts, which may suffer from both noise and blur, these alternatives often handle a single kind of degradation. Therefore, to allow a fair comparison, we apply burst denoising to a uniformly-exposed burst with the same overall budget and image deblurring to a single image captured using the budget in its entirety.

In our comparison, we use recent state-of-the-art deep learning algorithms. For image deblurring, we apply DMPHN [32], DeblurGAN-v2 [17], and the analysis-synthesis network pair [14]. To conduct a fair comparison, we re-train one representative method on million full-exposure images generated from our forward model. As the latter method does not have a publicly available training code, we re-train DeblurGAN-v2, which showed better results than DMPHN. As for burst denoising, we re-train KPN [19] on our dataset.

7.2. Synthetic images

We first evaluate our method on a test set of 276 clips obtained using the same procedure as our training set, although originating from different videos. We present a quantitative comparison to our baseline in Table 1. Our model outperforms traditional methods, obtaining an improvement of almost 3dB over KPN.

Fig. 7 presents an example result. As we can see, all three single image deblurring algorithms fail to deblur the image: DMPHN’s output is distorted, while the other two exhibit ghosting artifacts. This is expected, as the full-exposure image demonstrates an evident amount of blur, which is hard to mitigate without any complementary information. Furthermore, this image also suffers from noise, raising the task difficulty even more.

On the other hand, multi-frame methods show better performance. KPN manages to clean the fixed-exposure bursts’ noise and moderately improve the resulting images’ sharpness. Nevertheless, due to the lack of frame diversity, it fails in capturing the fine details of the scene. Our approach of learning the burst exposures is the best alternative, producing a clean and sharp output.

7.3. Real images

We further evaluate our method on real images acquired using two different cameras at two different scenes: an indoor one with controlled light conditions and an outdoor one. (See the supplement for a full description of our setup.) Fig. 1 and 8 present a qualitative comparison, showing that the methods’ perceptual quality appears to be generally consistent with the synthetic assessment, with our learned-exposure model significantly outperforming other approaches. However, this time, the analysis-synthesis network pair achieves the best results among the single image deblurring methods, despite not being re-trained for the task, while the re-trained DeblurGAN-v2 yields poor results. This deviation can be explained by the training SNR not precisely matching the de-facto one, as evidenced by the burst methods’ brighter output. Nevertheless, this proves our method’s advantage even when the training conditions are not met. Finally, we note that as the camera we used for the indoor experiment suffers from harsh fixed-pattern noise at low-exposures, traces of such patterns can be found in the final output. However, this phenomenon was not replicated by the other camera, and we expect that incorporating non-uniformities in the forward model should mitigate it. We leave this revision for future work.

7.4. Impact of noise and blur levels

The main factors that determine the trade-off between deblurring and denoising are illumination and camera motion. Thus, they have a substantial effect on our method’s learned exposures and reconstruction quality. Fig. 9 explores these factors’ impact by displaying the method’s performance on the same synthetic scene, altered with varying noise and blur levels. To achieve optimal results, we re-train the model for each working point on suitable ranges of SNR and blur while keeping the rest of the hyper-parameters fixed, including the predicted kernel size, which we set to $5 \times 5$. (We list the full configuration in the supplement.)

Interestingly, the middle frame’s exposure, which acts as the reference frame in the flow network, corresponds
8. Conclusions

In this work, we presented a new approach for image stabilization via joint burst denoising and deblurring for fast unstabilized cameras. We further proposed an end-to-end learning scheme for optimizing the camera’s exposure regime along with a reconstruction model, made possible via a novel differentiable layer simulating the camera sensor. This method’s key benefit is its ability to exploit the trade-off between high SNR and strong blur at long exposure, and vice versa. Synthetic and real results suggest that this approach significantly improves current deep state-of-the-art methods, both perceptually and quantitatively.
References


