Lucas-Kanade Reloaded: End-to-End Super-Resolution from Raw Image Bursts

Bruno Lecouat, Jean Ponce, Julien Mairal

Lucas-Kanade Reloaded: End-to-End Super-Resolution from Raw Image Bursts

Bruno Lecouat[1,2] Jean Ponce[1,3] Julien Mairal[2]

[1]Inria and DIENS (ENS-PSL, CNRS, Inria), Paris, France
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[3]Center for Data Science, New York University, New York, USA
firstname.lastname@inria.fr

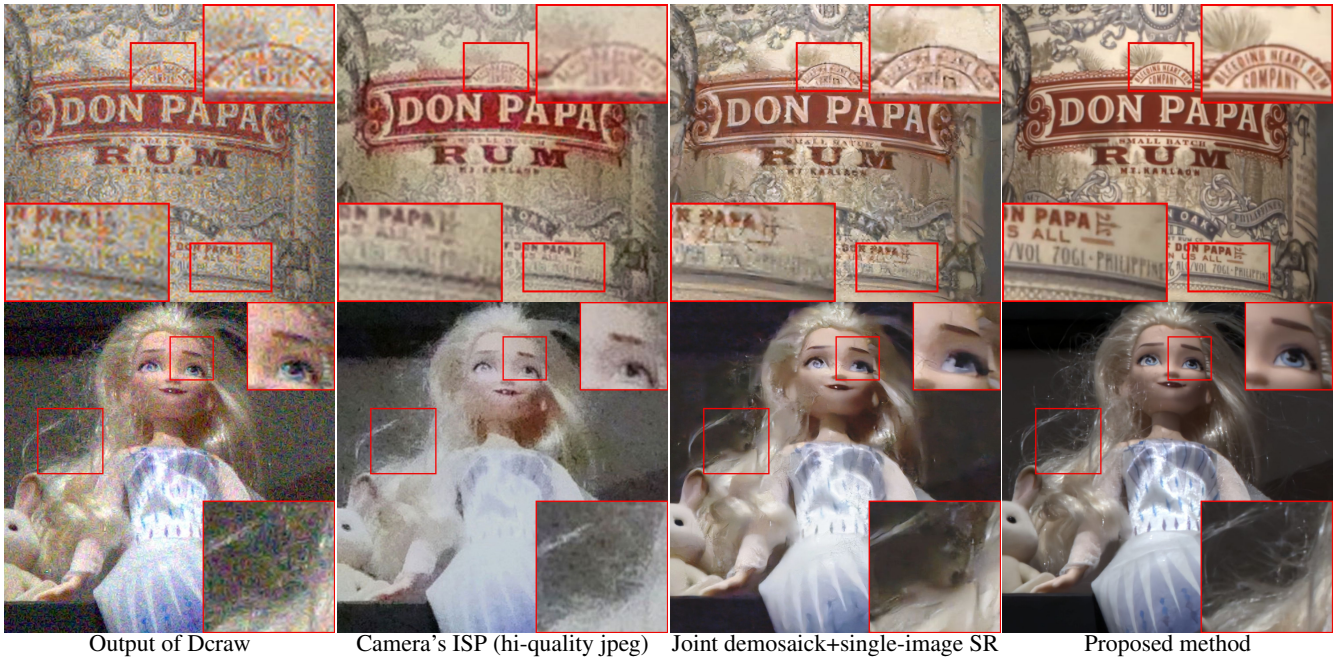

Figure 1: $\times 4$ super-resolution results obtained from a burst of 30 raw images acquired with a handheld Panasonic Lumix GX9 camera at 12800 ISO for the top image and 25600 for the bottom image. Dcraw performs basic demosaicking.

Abstract

This presentation addresses the problem of reconstructing a high-resolution image from multiple lower-resolution snapshots captured from slightly different viewpoints in space and time. Key challenges for solving this super-resolution *problem include (i) aligning the input pictures with sub-pixel accuracy, (ii) handling raw (noisy) images for maximal faithfulness to native camera data, and (iii) designing/learning an image prior (regularizer) well suited to the task. We address these three challenges with a hybrid algorithm building on the insight from [45] that aliasing is an ally in this setting, with parameters that can be learned end to end, while retaining the interpretability of classical approaches to inverse problems. The effectiveness of our approach is demonstrated on synthetic and real image bursts, setting a new state of the art on several benchmarks and delivering excellent qualitative results on real raw bursts captured by smartphones and prosumer cameras. Our code is available at* https://github.com/bruno-31/lkburst.git.

1. Introduction

The problem of reconstructing high-resolution (HR) images from lower-resolution (LR) ones comes in multiple flavors, that may significantly differ from each other in both technical detail and overall objectives. When a single LR image is available, the corresponding inverse problem is severely ill-posed, requiring very strong priors about the type of picture under consideration [18, 47]. For natural images, data-driven methods based on convolutional neural networks (CNNs) have proven to be very effective [26, 44]. Generative adversarial networks (GANs) have also been used to synthesize impressive HR images that may, how-

ever, contain "hallucinated" high-frequency details [9, 28].

In the true *super-resolution* setting [31, 39, 47],[1] where multiple LR frames are available, HR details *are* present in the data, but they are spread among multiple misaligned images, with technical challenges such as recovering subpixel registration, but also the promise of recovering veridical information in applications ranging from amateur photography to astronomy, biological and medical imaging, microscopy imaging, and remote sensing.

Videos are of course a rich source of multiple, closely-related pictures of the same scene, with several recent approaches to super-resolution in this domain, often combining data-driven priors from CNNs with self-similarities between frames [21, 27, 43]. However, most digital videos are produced by a complex pipeline mapping raw sensor data to possibly compressed, lower-resolution frames, resulting in a loss of high-frequency details and spatially-correlated noise that may be very difficult to invert [12]. With the ability of modern smartphone and prosumer cameras to record raw image bursts, on the other hand, there is a new opportunity to restore the corresponding frames *before* the image signal processor (ISP) of the camera produces irremediable damage [4, 45]. This is the problem addressed in this presentation, and it is challenging for several reasons: (i) images typically contain unknown motions due to hand tremor,[2] making subpixel alignment difficult; (ii) converting noisy raw sensor data to full-color images is in itself a difficult problem known as *demosaicking* [22, 25]; and (iii) effective image priors are often data driven, thus requiring a differentiable estimation procedure for end-to-end learning.

In this paper, we jointly address these issues and propose a new approach that retains the interpretability of classical inverse problem formulations while allowing end-to-end learning of models parameters. This may be seen as a bridge between the "old world" of signal processing and the "brave new one" of data-driven black boxes, without sacrificing interpretability: On the one hand, we address an inverse problem with a model-based optimization procedure alternating motion and HR image estimation steps, directly building on classical work from the 1980s [1, 29] and 1990s [16]. On the other hand, we also fully exploit modern technology in the form of a *plug-and-play* prior [6, 42] that gracefully mixes deep neural networks with variational approaches. In turn, unrolling the optimization procedure [7, 25, 48] allows us to learn the model parameters end to end by using training data with synthetic motions [4].

Since aliasing produces low-frequency artefacts associated with undersampled high-frequency components of the original signal, it is typically considered a nuisance, motivating camera manufacturers to add anti-aliasing (optical) filters in front of the sensor.[3] Yet, aliased images carry high-frequency information, which may be recovered from multiple shifted measurements. Perhaps surprisingly, aliasing is thus an ally in the context of super-resolution, a fact already noted in earlier references, see [41]. As shown in the rest of this presentation, our approach to raw burst super-resolution also exploits this insight, and it achieves a new state of the art on several standard benchmarks that use synthetic motion for ground truth. It also gives excellent qualitative results on real data obtained with smartphone and prosumer cameras. Interestingly, as illustrated by Figure 1, our method has turned out to be surprisingly robust to noise given the particularly challenging setting of raw image super-resolution, which involves simultaneous blind denoising, demosaicking, registration, and upsampling.

Summary of contributions.

- To the best of our knowledge, we propose the first model-based architecture learnable end to end for joint image alignment and super-resolution from raw image bursts.
- We introduce a new differentiable image registration module that can be applied to images of different resolutions, is readily integrable in neural architectures, and may find other uses beyond super-resolution.
- We show that our approach gives excellent results on both real image bursts (with up to $\times 4$ upsampling for raw images) and synthetic ones (up to $\times 16$ for RGB images).

2. Related Work

Classical multiframe super-resolution. Tsai and Huang wrote the seminal paper in this setting [39], with a restoration model in the frequency domain assuming known translations between frames. Most latter approaches have focused on the spatial domain, and they generally fall into two main categories [23]: In interpolation-based methods, LR snapshots aligned with sub-pixel precision are jointly interpolated into an HR image [15, 38]. Impressive results have recently been obtained for hand-held cameras using the variant of this method proposed by Wronski *et al.* [45], whose insight of exploiting aliasing effects has been one of the inspirations of our work. However, due to the sequential nature of their algorithm, errors may propagate from one stage to the next, leading to sub-optimal reconstructions [34]. In contrast, iterative spatial domain techniques iteratively refine an estimate for the super-resolved image so as to best explain the observed LR frames under some image formation model. Variants of this approach include the early iterated backprojection algorithm of Irani *et al.* [20], the maximum likelihood technique of Elad and Feuer [11],

[1] "Single-image super-resolution" has become a popular nickname for single-image upsampling under strong priors; here, we use the classical definition of super-resolution from multiple LR snapshots [39, 47].

[2] Image bursts acquired on a tripod may also present subpixel misalignments in practice due to floor vibrations, as observed in our experiments.

[3] There is, however, a trend today toward removing these filters, as in the prosumer camera used in some of our experiments with real images.

and the model regularized by bilateral total variation of Farsiu *et al.* [13]. The image formation parameters are either be assumed to be known a priori through calibration, or estimated jointly with the HR image. In general, inter-frame motion can either be estimated separately, or be treated as an integral part of the super-resolution problem [2, 16], thus avoiding motion estimation between LR frames, whose accuracy may be affected by undersampling [40]. The method proposed in the rest of this paper combines the best of both worlds since it performs joint estimation while aligning the LR frames with the reconstructed HR image.

Learning-based approaches. In this context, the multi-frame case has received less attention than its single-image counterpart, for which several loss functions and architectures have been proposed [9, 28, 48]. Most multi-frame algorithms focus on video super-resolution. Model-based techniques learn non-uniform interpolation or motion compensation using convolutional neural networks [37] but the most successful approaches so far are model free, leveraging instead diversity with 3D convolutions or attention mechanisms [21, 43]. Learning-based methods have also been used in remote sensing applications, using 3D convolutions [32] or joint registration/fusion architectures [8] for example. Finally, and closer to our work, Bhat et al. [4] have recently proposed a network architecture for raw burst super-resolution, together with a very interesting dataset featuring both synthetic and real images for training and testing. It is important to note that learning-based approaches to super-resolution are typically trained on synthetically generated LR images [30], a strategy that may not generalize well to real photographs unless great care is taken in modeling the image corruption process [5]. Learning super-resolution models from real LR/HR image pairs is quite challenging since it requires in general using separate cameras with different lenses and spatial resolution, with inevitable spatial and spectral misalignments. As shown by our experiments, our method, although trained from synthetic LR images, gives excellent results with real bursts taken from different smartphones and cameras. Leveraging real images at training time is, for now, left for future work.

3. Proposed Approach

This section presents the three main components of our approach: its image formation model, an optimization procedure for solving the corresponding inverse problem, and its unrolled implementation in a feedforward architecture whose parameters can be learned end to end.

3.1. Image formation model

Image acquisition in a digital camera starts from an instantaneous irradiance function $f_{\gamma,t} : [0,1]^2 \to \mathbb{R}^+$ defined on a continuous retinal domain with nonnegative values,

Figure 2: Proof of concept for extreme $\times 16$ upsampling. The right image is obtained by processing a burst of 20 LR images presented on the left obtained with synthetic random affine movements and bilinear downsampling.

such that $f_{\gamma,t}(\mathbf{u})$ is the spectral irradiance value at point $\mathbf{u}$, time t, and wavelength γ, accounting for blur due to optics, atmospheric effects, etc. The camera sensor integrates $f_{\gamma,t}$ in the spatial, time, and spectral domains to construct a *raw* digital image $\mathbf{y} : [1,\dots,n]^2 \to \mathbb{R}^+$, where each pixel's spectral response is typically dictated by the 2×2 RGGB Bayer pattern, with twice as many measurements for the green channel than for the red and blue ones [22]. Modern cameras turn the raw image $\mathbf{y}$ into a full blown, three-channel *RGB* image $\mathbf{x}$ with the same spatial resolution through an interpolation process called *demosaicking*.

In practice, we do not have access to $f_{\gamma,t}$ to use as ground truth for learning an image restoration process, even when an accurate model of the $f_{\gamma,t} \mapsto \mathbf{x}$ map is available. Thus, we model instead the process $\mathbf{x} \mapsto \mathbf{y}_k$, where $\mathbf{x}$ is a latent high-resolution (HR) image we wish to recover, and the low-resolution (LR) images $\mathbf{y}_k$ $(k = 1,\dots,K)$ have been observed in a burst of length K. We assume that $\mathbf{x}$ is sharp, without any blur, and noiseless. The burst images are obtained through the following forward model (Figure 3):

$$\mathbf{y}_k = DBW_{\mathbf{p}_k}\,\mathbf{x} + \boldsymbol{\varepsilon}_k \text{ for } k = 1,\dots,K, \qquad (1)$$

where $\boldsymbol{\varepsilon}_k$ is some additive noise. Here, both the HR image $\mathbf{x}$ and the frames $\mathbf{y}_k$ of the burst are flattened into vector form. The operator $W_{\mathbf{p}_k}$ parameterized by $\mathbf{p}_k$ warps $\mathbf{x}$ to compensate for misalignments between $\mathbf{x}$ and $\mathbf{y}_k$ caused by camera or scene motion between frames, assumed here to be a 6-parameter affine transformation of the image plane, then resamples the warped image to align its pixel grid with that of $\mathbf{y}_k$. Finally, the corresponding HR image is blurred to account for integration over both space (the LR pixel area, using either simple averaging or, as in the figure, a Gaussian filter) and time (accounting for camera and/or scene motion during exposure), and it is finally downsampled in both the spatial and spectral domains by the operator D, with an (a priori) arbitrary choice of *where* to pick the sample from (pixel corner or center for example), the spectral part corresponding to selecting one of the three RGB values to assemble the raw image. It will prove convenient in the sequel to

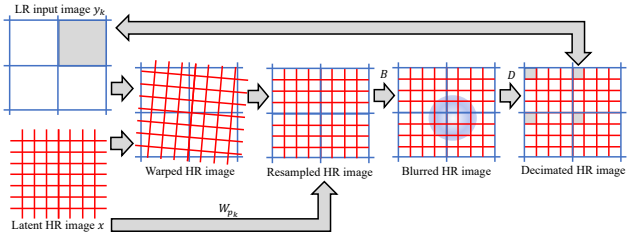

Figure 3: Image formation: The HR image $\mathbf{x}$ is warped then resampled to align it with the LR image $\mathbf{y}$ using the operateur $W_{\mathbf{p}_k}$. It is then blurred by the operator B to account for integration over LR pixels and finally downsampled in the spatial and spectral domains by the operator D (the spectral downsampling from RGB to R, G or B is not illustrated here for simplicity).

rewrite (1) as $\mathbf{y} = U_{\mathbf{p}}\mathbf{x} + \boldsymbol{\varepsilon}$, where

$$U_{\mathbf{p}} = \begin{bmatrix} DBW_{\mathbf{p}_1} \\ \vdots \\ DBW_{\mathbf{p}_K} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_K \end{bmatrix}, \mathbf{p} = \begin{bmatrix} p_1 \\ \vdots \\ p_K \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_K \end{bmatrix}. \quad (2)$$

Before closing this section, let us note that simpler motion models with two (translation) or three (rigid motion) parameters, or (much) more complex piecewise-affine or elastic models could be considered depending on the application. We focus here on the scenario where a user wishes to zoom in on a relatively small crop (say, between 100×100 to 800×800 pixels) of a multi-megapixel image, and the affine model has proven effective with real handheld cameras in this setting. This implicitly corresponds to a globally piecewise-affine motion model.

3.2. Inverse problem and optimization

Given the image formation model of Eqs. (1)–(2), recovering the HR image $\mathbf{x}$ from the K LR frames $\mathbf{y}_k$ in the burst can be formulated as finding the values of $\mathbf{x}$ and $\mathbf{p}$ that minimize

$$\frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x}), \quad (3)$$

where ϕ_θ is a parameterized regularizer, to be detailed later, and λ is a parameter balancing the data-fidelity and regularization terms. Many methods are of course available for minimizing this function. Like others (e.g., [24]), and mainly for simplicity, we choose here a quadratic penalty method [33, Sec. 17.1] often called half-quadratic splitting (or HQS) [14]: the original objective is replaced by

$$E_\mu(\mathbf{x},\mathbf{z},\mathbf{p}) = \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{z}\|^2 + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x}), \quad (4)$$

where $\mathbf{z}$ is an auxiliary variable, and μ is a parameter increasing at each iteration, such that, as $\mu \to +\infty$, the minimization of (4) with respect to $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{p}$ becomes equivalent to that of (3) with respect to $\mathbf{x}$ and $\mathbf{p}$ alone. Each iteration of HQS can be viewed as one step of a block-coordinate descent procedure for minimizing E, changing one of variables $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{p}$ at a time while keeping the others fixed, with the value of μ increasing after each iteration. Convergence guarantees for quadratic penalty methods require an approximate minimization of Eq. (4) with increasing precision over time [33]. Following common practice in computer vision (e.g. [24]), we use HQS without formally checking that its precision indeed increases with iterations. This very simple procedure turns out to work well in practice. Its steps are detailed in the next three paragraphs, the exponent t being used to designate the value of the variables at iteration t. The sequence of weights $(\mu^t)_{t\geq 0}$ is learned end-to-end as explained in Section 3.3.

Updating z. Several strategies are possible for minimizing Eq. (4) with respect to $\mathbf{z}$. Given the dimension of the problem, one may choose for instance a fast iterative minimization procedure such as conjugate gradient descent. Since an approximate minimization is sufficient for our needs, we have chosen to use instead a single step of plain gradient descent, which converges more slowly in theory, but is also simpler and more easily amenable to the unrolled optimization strategy for end-to-end learning that will be presented next. The update at iteration t is given by

$$\mathbf{z}^t \leftarrow \mathbf{z}^{t-1} - \eta_t\left[U_{\mathbf{p}^{t-1}}^\top(U_{\mathbf{p}^{t-1}}\mathbf{z}^{t-1} - \mathbf{y}) + \mu(\mathbf{z}^{t-1} - \mathbf{x}^{t-1})\right], \quad (5)$$

where $\eta_t > 0$ is some step size, also learned end to end.

Updating the motion parameters p. Let $\mathbf{p}_k$ denote the part of the parameter vector $\mathbf{p}$ responsible for the alignment of $\mathbf{z}^t$ and $\mathbf{y}_k$ in (4). The corresponding optimization problem can be rewritten as

$$\min_{\mathbf{p}_k} \frac{1}{2}\|\mathbf{y}_k - DBW_{\mathbf{p}_k}\mathbf{z}^t\|^2. \quad (6)$$

This is a non linear least-squares problem, which can once again be solved using many different techniques. Here, we pick a Gauss-Newton approach, which corresponds to a variant of the Lucas-Kanade algorithm [1, 29], showing again that a 40-year old technique can still be relevant today. Specifically, we perform one Gauss-Newton step at each iteration t for each $\mathbf{p}_k$ in parallel:

$$\mathbf{p}_k^t \leftarrow \mathbf{p}_k^{t-1} - \left(\mathbf{J}_k^{t\top}\mathbf{J}_k^t\right)^{-1}\mathbf{J}_k^{t\top}\mathbf{r}_k^t, \quad (7)$$

where $\mathbf{r}_k^t = U_{\mathbf{p}_k^{t-1}}\mathbf{z}^t - \mathbf{y}_k$ is the residual of the non-linear least-squares problem (6), and $\mathbf{J}_k^t = (\partial U_{\mathbf{p}_k^{t-1}}/\partial \mathbf{p}_k)\mathbf{z}^t$ is the Jacobian of the $DBW_{\mathbf{p}_k}$ operator. The only difference with a Lucas-Kanade iteration is the presence of a high-resolution frame $\mathbf{z}^t$ and the downsampling operator DB. This is similar to [16], or more recently [2, 17], which align high-resolution images with low-resolution ones.

Estimating the HR image $\mathbf{x}$. The $\mathbf{x}$ update is obtained as

$$\mathbf{x}^t \leftarrow \arg\min_{\mathbf{x}} \frac{\mu_{t-1}}{2}\|\mathbf{z}^t - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x}),$$

which amounts to computing the proximal operator of the prior ϕ_θ. In practice, we follow a "plug-and-play" approach [6, 35, 42], and replace the proximal operator by a parametric function $f_\theta(\mathbf{z}_t)$ (here, a CNN, see implementation details). Using such an implicit prior has proven very effective in our setting. More traditional image priors such as total variation could of course have been used as well.

3.3. Unrolled optimization and backpropagation

The optimization procedure described so far requires choosing hyper-parameters such as the sequence $(\mu_t)_{t\geq 0}$, and its implicit prior also involves model parameters θ. By using a training set of n LR burst/HR image pairs, we propose to learn all these parameters in a supervised fashion. We denote the training set by $(\mathbf{Y}_i, \mathbf{x}_i)_{i=1}^n$, where $\mathbf{Y}_i = \{\mathbf{y}_j^i\}_{j=1}^K$ is the i-th burst of LR images associated to the HR image $\mathbf{x}_i$. We then unroll the optimization procedure for T steps and, denoting by $\hat{\mathbf{x}}_T(\mathbf{Y}_i)$ the HR image estimated from burst $\mathbf{Y}_i$, we consider the objective function

$$\frac{1}{n}\sum_{i=1}^{n} L(\hat{\mathbf{x}}_T(\mathbf{Y}_i), \mathbf{x}_i), \tag{8}$$

where L is the ℓ_2 or ℓ_1 loss (in practive we have observed that the ℓ_1 loss performs slightly better). Because every step of our estimation procedure is differentiable, we minimize (8) by stochastic gradient descent.

Learned data prior. Good image priors are essential for solving ill-posed inverse problems. As noted earlier, instead of using a classical one, such as total variation (TV) or bilateral total variation (BTV) [13], we learn an implicit prior parameterized by a convolutional neural network f_θ in a data-driven manner. We use the ResUNet architecture introduced in [48] in practice. It involves four scales, each of which has an identity skip connection between downscaling and upscaling operations.

3.4. Implementation details and variants

Downsampling and blurring operators $\mathbf{D}, \mathbf{B}$. We have tried different variants of downsampling/blurring strategies such as Gaussian smoothing. In practice, we have observed that simple averaging, which is differentiable and parameter-free, gives good results in all our experiments. As a consequence, we do not assume any knowledge about the blur used to generate data, corresponding to an operator B that only captures blur due to photon integration on the sensor without addressing optical blur. We argue that this limited model is relevant because modern cameras and smartphone are aliased [45], which may explain the generalization to real images, as soon as the scene is static.

Initialization by coarse alignement. To initialize the motion parameters $\mathbf{p}$, we cannot minimize (6) as in the previous section, because no good estimate of the HR image is available. Therefore, we align each LR frame to an arbitrary one from the burst (*e.g.*, the first one) by using the Lucas-Kanade forward additive algorithm [1, 36] which is known to be robust to noise. Note that another difficulty lies in the raw format of images. To overcome this issue, we simply convert raw images into grayscale images by using bilinear interpolation. This is of course sub-optimal, but sufficient for obtaining coarse motion parameters.

Initialization via coarse-to-fine strategy. For extreme upsampling factors ($\times 16$), we found a coarse-to-fine initialization strategy to be useful: We initialize the motion parameters $\mathbf{p}_j^0$ and high-resolution image $\mathbf{z}^0$ by using the output of the algorithm trained at a lower upsampling factor. For instance, $\times 16$ can be obtained by applying twice a $\times 4$ algorithm, or four times $\times 2$ algorithm.

4. Experiments

Experiments were conducted on synthetic and real raw image bursts. We also provide experiments on RGB bursts in the appendix, allowing easier comparison with earlier approaches that cannot handle raw data.

Training procedure and data. For synthetizing realistic *raw bursts* from groundtruth RGB images, we follow the approach described in [4], using the author's publicly available code[4] on the training split of the Zurich raw to RGB dataset [19]. The approach consists of applying the inverse RGB to raw pipeline introduced in [5]. Displacements are randomly generated with Euclidean motions and frames are downscaled with bilinear interpolation in order to simulate LR frames containing aliasing. Synthetic, yet realistic, noise is added to the frames, and color values are discarded according to the Bayer pattern. Then, we train our models for minimizing the loss (8). We perform $100\,000$ iterations of the ADAM optimizer with a batch size of 10, a burst size of 14 and with a learning rate of 3×10^{-5} decaying by a factor 2 after $50\,000$ iterations. Our approach is implemented in Pytorch and takes approximately 1.5 days to train on an Nvidia Titan RTX GPU. We evaluate our models in all our experiments with a burst size of 14 unless specified.

Extreme $\times 16$ upsampling on RGB images. As a proof-of-concept, we also perform experiments for an unusual $\times 16$ super-resolution task, using the coarse-to-fine strategy of Sec. 3.4. A result is presented in Fig. 2, showing impressive reconstruction and additional ones can be found in the appendix. Even though not realistic, we believe the exper-

[4] https://github.com/goutamgmb/NTIRE21_BURSTSR.

iment to be of interest, as it demonstrates the effectiveness of our approach in an idealistic, yet extreme, setting.

Demosaic+SISR ETH [4] Ours

Figure 4: Visual comparison on **synthetic raw image bursts** used in [4]. Demosaic+SISR is our single-image baseline based on the ResUNet architecture [48] (see main text). The two right columns are produced by methods dedicated to raw burst processing, respectively [4] and ours.

Evaluation on synthetic RAW images. The evaluation protocol of [4] allows us to perform quantitative comparison with their state-of-the-art method for processing raw image bursts. An additional comparison with [45] would have been interesting but this method is part of a commercial product that could not be shared with us.

Method	PSNR (db)	Geom (pix)	SSIM
Scores on public validation set			
ETH [4]	39.09	-	-
Ours (refine)	**41.45**	-	0.95
Scores on our own validation set to conduct the ablation study			
Bicubic Single Image	33.45	-	-
Multiframe L2 only	34.21	-	-
Multiframe L2 + TV prior	34.48	-	-
Single Image	36.80	-	-
Ours (no refinements)	40.38	0.55	0.958
Ours (refinements)	**41.30**	0.32	0.963
Ours (known motion)	42.41	0.00	0.971

Table 1: **Results with synthetic raw image bursts** of 14 images generated from the Zurich raw to RGB dataset [19] with synthetic affine motions. Reconstruction error in average PSNR and geometrical registration error in pixels for our models. "known $\mathbf{p}$" is the oracle performance our model could achieve, if motion estimation was perfect.

We provide a quantitative comparison in Table 1 with the model introduced in [4], as well as a single-image upsampling baseline based on the ResUNet architecture [48], which we use as a plug-and-play prior in our model.

To that effect, we first use the validation set of [4] available online (with no overlap with the training set), for which motions are unknown, allowing us to compare with their method, which we outperform by more than 2dBs. In order to perform further comparison and conduct the ablation study, we also build an additional validation set by randomly extracting 266 images from the Zurich raw to RGB dataset, allowing us to generate validation data with known motion. We evaluate variations of our model in the same table, notably comparing the registration accuracy achieved by these variants by using the geometrical error presented in [36]. More precisely, we perform a small ablation study by introducing a simpler baseline that does not perform joint alignement and only exploits the coarse registration module (no refine baseline). Performing motion refinement significantly improves the registration accuracy and subsequently the image reconstruction quality. Last, we also report the oracle performance of our model with known motions.

We provide a visual comparison in Figure 4 with single-image SR baselines and the state-of-the-art method [4] for processing raw image bursts. Only the two approaches processing bursts are able to recover high-frequency details, demonstrating their ability to leverage and remove aliasing artefacts, which are very present in the top image. Significantly better quality results are obtained with our approach.

Impact of burst length and cropping size. The dataset Zurich rgb-to-raw [19] was very useful for training our models, but it unfortunately features relatively small image crops of size 96×96 without giving access to the original megapixel images. By experimenting with real raw data, it became apparent to us to our method was performing better with larger crops (*e.g.*, more than 200×200 pixels), achieving better registration and visually better results. To study the impact of the crop size and burst length, we have thus synthesized additional raw bursts from the DIV2K dataset, and report our experimental results in Figure 5, confirming our findings. Note that this does not appear to be a strong limitation of our approach, since in real-life scenarios, we can always assume that the original megapixel image is available. As expected, the performance of our approach is also increasing with the burst size, even though our models were trained with bursts of size 14.

Results on real raw image bursts, dataset of [4]. In Figure 6, we show a comparison with [4] using their dataset featuring small crops of size 96×96. As discussed previously, this setup is suboptimal for our approach, but still produces visually pleasant results. Choosing which method performs best here is however very subjective and we found conclusions hard to draw on this dataset. Whereas the im-

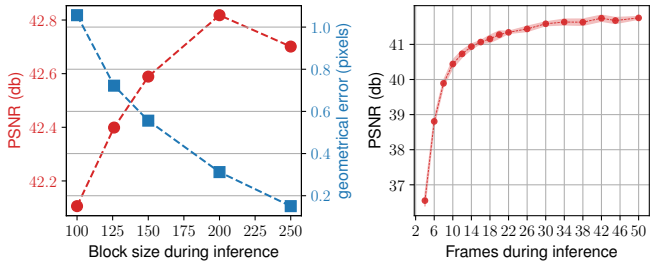

Figure 5: Left: Impact of the crop size on the registration and reconstruction performance. Right: Impact of the burst length, see main text for details.

ages produced by [4] may sometimes look slightly sharper, one may argue that our approach seems to recover more reliable details, *e.g.*, the text is perhaps easier to read. Note that our models were trained on synthetic data only and we leave fine-tuning with real data on this dataset for future work. There is an attempt in [4] to address the open problem of quantitative evaluation with real data using a custom metric, but, like any other attempt so far, it is flawed since (i) it is based on the alignment method of [4], with an unavoidable slight bias in its favor, and (ii) it assumes ground truth from a particular Canon camera. Interestingly, this score improvement does not always correlate with visual quality, as shown by Figure 6. This is by no means a criticism of [3]: we believe instead that quantitative evaluation on real images is an extremely challenging problem, far from being solved. Since the submission of our paper, the results of the NTIRE 2021 burst super-resolution challenge have been published [3]. Our method ranked third quantitatively in the "synthetic data" part of the challenge that we entered.

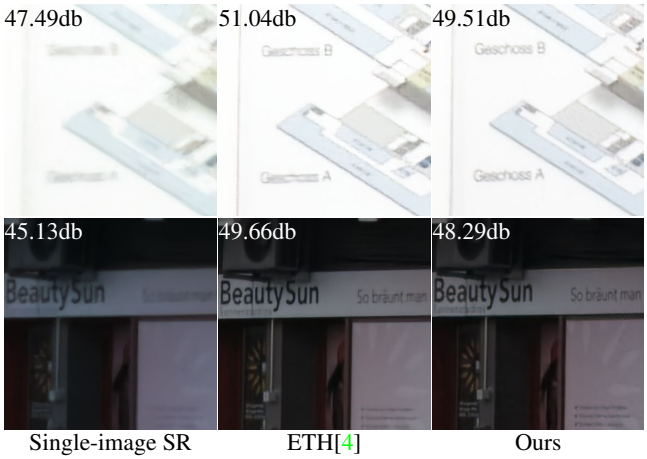

Figure 6: **Results from real raw bursts** from dataset of [4] including Aligned PSNR score (see main text).

Results on real raw image bursts from various devices. Finally, we demonstrate the effectiveness of our approach on real raw bursts acquired by different devices. We consider a Panasonic Lumix GX9 camera, which is interesting for SR as it does not feature an optical anti-aliasing filter, a Canon Powershot G7X camera, a Samsung S7 and a Pixel 4a smartphones. Results obtained in high noise regimes have already been presented in Figure 1, showing that our approach is surprisingly robust to noise. We believe that the result is of interest since it may allow photographers to use high ISO settings in low-light conditions, without sacrificing image quality. Other results are presented in Figure 6 on low-noise outdoor conditions with bursts of 20 to 30 raw images. In all cases, the method succeeds at recovering high-frequency details. Many more examples and comparisons with other multiframe methods are provided in the supplementary material. We also present failure cases, corresponding in large parts to scene motion. Last, we remark that our method is relatively fast at inference time. Processing a burst of 20 raw 300×300 images takes for instance about $1s$ on an Nvidia Titan RTX GPU, producing an upsampled image of size 1200×1200.

5. Conclusion

We have presented a simple but effective method for superresolution that combines the interpretability of model-based approaches to inverse problems with the flexibility of data-driven architectures and can be learned from pairs of synthetic LR and real HR images. We plan several extensions, including using multiple cameras to add real LR-burst/HR-image pairs to the training mix, and at test time to take advantage of the multiplicity of imaging devices now available on high-end smartphones. This will open the door to wide-baseline super-resolution applications, such as the construction of high quality panoramas and finely detailed texture maps in multi-view stereo reconstructions. Finally, we plan to explore several other extensions of our approach, including tackling blurry bursts, extending super-resolution to reconstruct HDR images, and pursuing applications in the astronomy and microscopy domains

Acknowledgments

We thank Frédéric Guichard for useful discussions and comments. 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 2020-AD011011252 made by GENCI.

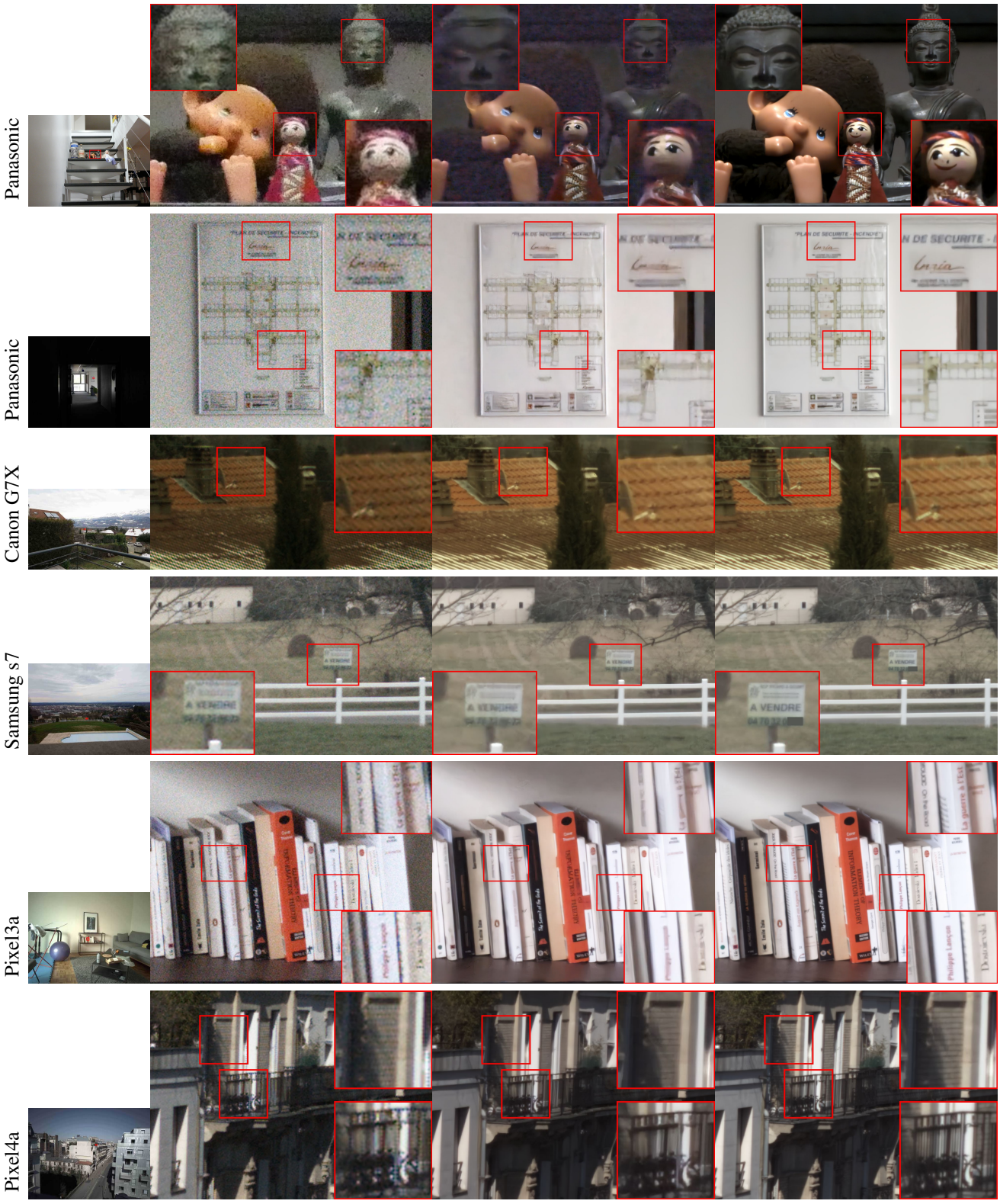

Figure 7: Results from real raw image bursts obtained with various cameras. We provide comparisons with single image and multiframe baselines. Finest restored details can be seen by zooming on a computer screen. The last three digits of the phone number, only legible in our reconstruction, are masked in the figure for privacy concerns.

References

[1] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision (IJCV)*, 56(3):221–255, 2004. 2, 4, 5

[2] Cosmin Bercea, Andreas Maier, and Thomas Köhler. Confidence-aware levenberg-marquardt optimization for joint motion estimation and super-resolution. In *IEEE International Conference on Image Processing (ICIP)*, pages 1136–1140. IEEE, 2016. 3, 4

[3] Goutam Bhat, Martin Danelljan, and Radu Timofte. Ntire 2021 challenge on burst super-resolution: Methods and results. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 613–626, 2021. 7

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. *arXiv preprint arXiv:2101.10997*, 2021. 2, 3, 5, 6, 7

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5

[6] Stanley H Chan, Xiran Wang, and Omar A Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016. 2, 5

[7] Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds. In *Adv. in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[8] Michel Deudon, Alfredo Kalaitzis, Israel Goytom, Md Rifat Arefin, Zhichao Lin, Kris Sankaran, Vincent Michalski, Samira E Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *arXiv preprint arXiv:2002.06460*, 2020. 3

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(2):295–307, 2016. 2, 3

[10] Thibaud Ehret, Axel Davy, Pablo Arias, and Gabriele Facciolo. Joint demosaicking and denoising by fine-tuning of bursts of raw images. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 11

[11] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997. 2

[12] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, 14(2):47–57, 2004. 2

[13] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004. 3, 5

[14] Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE transactions on Image Processing*, 4(7):932–946, 1995. 4

[15] Russell Hardie. A fast image super-resolution algorithm using an adaptive wiener filter. *IEEE Transactions on Image Processing*, 16(12):2953–2964, 2007. 2

[16] Russell C Hardie, Kenneth J Barnard, and Ernest E Armstrong. Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE transactions on Image Processing*, 6(12):1621–1633, 1997. 2, 3, 4

[17] Yu He, Kim-Hui Yap, Li Chen, and Lap-Pui Chau. A nonlinear least square technique for simultaneous image registration and super-resolution. *IEEE Transactions on Image Processing*, 16(11):2830–2841, 2007. 4

[18] Hsieh Hou and H Andrews. Cubic splines for image interpolation and digital filtering. *IEEE Transactions on acoustics, speech, and signal processing*, 26(6):508–517, 1978. 1

[19] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 5, 6

[20] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991. 2

[21] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3224–3232, 2018. 2, 3, 11, 12, 13

[22] Ron Kimmel. Demosaicing: image reconstruction from color ccd samples. *IEEE Transactions on image processing*, 8(9):1221–1228, 1999. 2, 3

[23] Thomas Köhler. Multi-frame super-resolution reconstruction with applications to medical imaging. *arXiv preprint arXiv:1812.09375*, 2018. 2

[24] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. *Adv. in Neural Information Processing Systems (NIPS)*, 2009. 4

[25] Bruno Lecouat, Jean Ponce, and Julien Mairal. Fully trainable and interpretable non-local sparse models for image restoration. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2, 12, 13

[26] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[27] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 2, 3, 12, 13

[29] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In

Proceedings of Imaging Understanding Workshop, 1981. 2, 4

[30] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 3

[31] P. Milanfar. *Super-resolution imaging*. CRC Press, 2011. 2

[32] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3644–3656, 2019. 3

[33] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, 2006. Second edition. 4

[34] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003. 2

[35] Ernest K Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. *Proc. International Conference on Machine Learning (ICML)*, 2019. 5

[36] Javier Sanchez. The inverse compositional algorithm for parametric registration. *Image Processing On Line*, 6:212–232, 2016. 5, 6, 11

[37] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[38] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing*, 16(2):349–366, 2007. 2

[39] R.Y. Tsai and T.S. Huang. Multiframe image restoration and registration. In *Advances in Computer Vision and Image Processing*, page 317–339. JAI Press Inc., 1984. 2

[40] Patrick Vandewalle. Super-resolution from unregistered aliased images. Technical report, EPFL, 2006. 3

[41] Patrick Vandewalle, Luciano Sbaiz, Joos Vandewalle, and Martin Vetterli. Aliasing is good for you: Joint registration and reconstruction for super-resolution. Technical report, 2006. 2

[42] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. 2, 5

[43] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2, 3

[44] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision (ECCV) workshop on Perceptual Image Restoration and Manipulation*, 2018. 1

[45] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019. 1, 2, 5, 6

[46] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 11

[47] J. Yang and T. Huang. Image super-resolution: historical overview and future challenges. In P. Milanfar, editor, *Super-resolution imaging*. CRC Press, 2011. 1, 2

[48] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 5, 6, 11

[49] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 11

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 11

Supplementary material

This supplementary material presents additional qualitative and quantiative results. In Figure A1 we present additional visual comparison with two burst denoising methods on real images. In Table A1 we present additional experiments on RGB images. Figures A2 and A3 are devoted to super-resolution experiments from real raw data from different smartphones (Google Pixel 3a and 4a, Samsung S7 and S10) and cameras (Panasonic Lumix GX9 and Canon Powershot G7X) and comparison with additional baselines. In Figures A4, we present extreme upsampling results by using synthetic RGB image bursts. In Figure A5, we present restoration results obtained from real images with very low SNR to illustrate the efficiency of our method to perform blind denoising. In Figures A6 and A7, we study the effect of the number of frames in the burst on the reconstruction, both in the low SNR and high SNR settings. Finally, we present failure cases in Figure A8, where fast moving objects are present in the scene.

Comparison with burst denoising methods. We perform additional qualitative comparison on a real image with two burst denoising methods. We compare our method with [10] which performs joint denoising and demosaicking on a burst of raw images. We use the code and the pretrained model made available online. We also use the code and pretrained model of [46]. However the model is only designed to perform grayscale burst denoising, so we perform denoising independently on each RGB channel and then perform demosaicking to get an RGB image. Despite our best efforts for tuning the parameters of these methods to maximize visual quality, the results obtained are not as good as our method (see Figure A1 below). We believe this is not surprising since each one of these methods only addresses a subset of our problem. Adapting them successfully to our general setting is not trivial.

ISP | BKPN [46]+demosaic | Mosa2mosa [10] | Mosa2mosa [10] + EDSR | Ours

Figure A1: Comparison with joint denoising and demosaicking methods.

Evaluation on RGB images. We compare our approach on the BSD68 dataset against state-of-the-art single-image and video super-resolution algorithms (considering a burst as a video sequence) and report the HR image reconstruction accuracy in terms of average PSNR in Table A1. For the training with RGB data, we perform 80 000 iterations of the ADAM optimizer with a batch size of 10, a burst size of 14 and with a learning rate of 3×10^{-5} decaying by a factor 2 after 40 000 iterations. For evaluating the model VSR-DUF [21], we use the code and the pretrained models made available online by the authors. Other single-image reconstruction results are from [48]. In the present setting, we consistently outperform other baselines, notably demonstrating that burst SR cannot simply be addressed effectively by current video SR approaches. We also note that our models perfom better with less blurring (and more aliasing). Finally, we evaluate variations of our model in the same table, notably comparing the registration accuracy achieved by these variants by using the geometrical error presented in [36]. More precisely, we perform a small ablation study by introducing a simpler baseline that does not perform joint alignement and only exploits the coarse registration module (no refine baseline). Performing joint alignment and image estimation systematically improves motion estimation. Last, we also report the oracle performance of our model with known motions.

Method	Scaling factor / blurring kernel std		
	$\times 2 / \sigma = 0.7$	$\times 3 / \sigma = 1.2$	$\times 4 / \sigma = 1.6$
Single Image SR			
RCAN [50]	29.48	27.30	25.59
IRCNN [49]	29.60	26.89	25.32
USRNet [48]	30.55	27.76	26.18
Video SR			
VSR-DUF[21]	-	31.03	29.24
Ours (no refine)	42.36/0.10	32.63/0.14	**30.00**/0.19
Ours	**43.73**/0.07	**33.10**/0.10	29.87/0.14
Ours (known $\mathbf{p}$)	45.72/0.00	34.47/0.00	31.32/0.00

Table A1: **Results for RGB with synthetic affine motions**, of different methods for different combinations of scale factors and blur kernels. Results are given in term of average PSNR in dBs and geometrical registration error in pixels for our models. "known $\mathbf{p}$" is the oracle performance our model could achieve, if motion estimation was perfect.

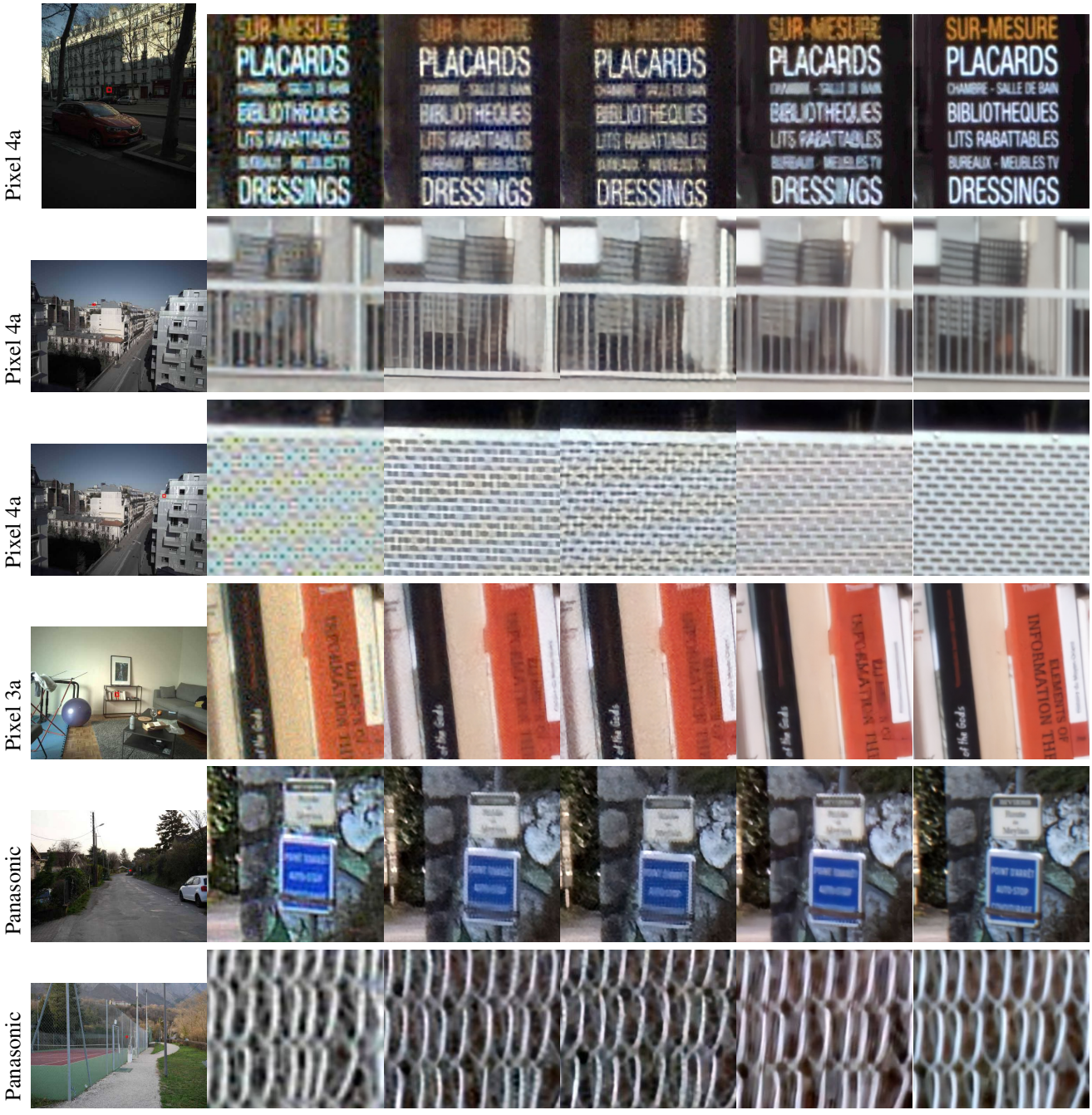

Figure A2: Results from real raw image bursts obtained with various cameras. We provide comparisons with single image and multiframe baselines. Finest restored details can be seen by zooming on computer screen.

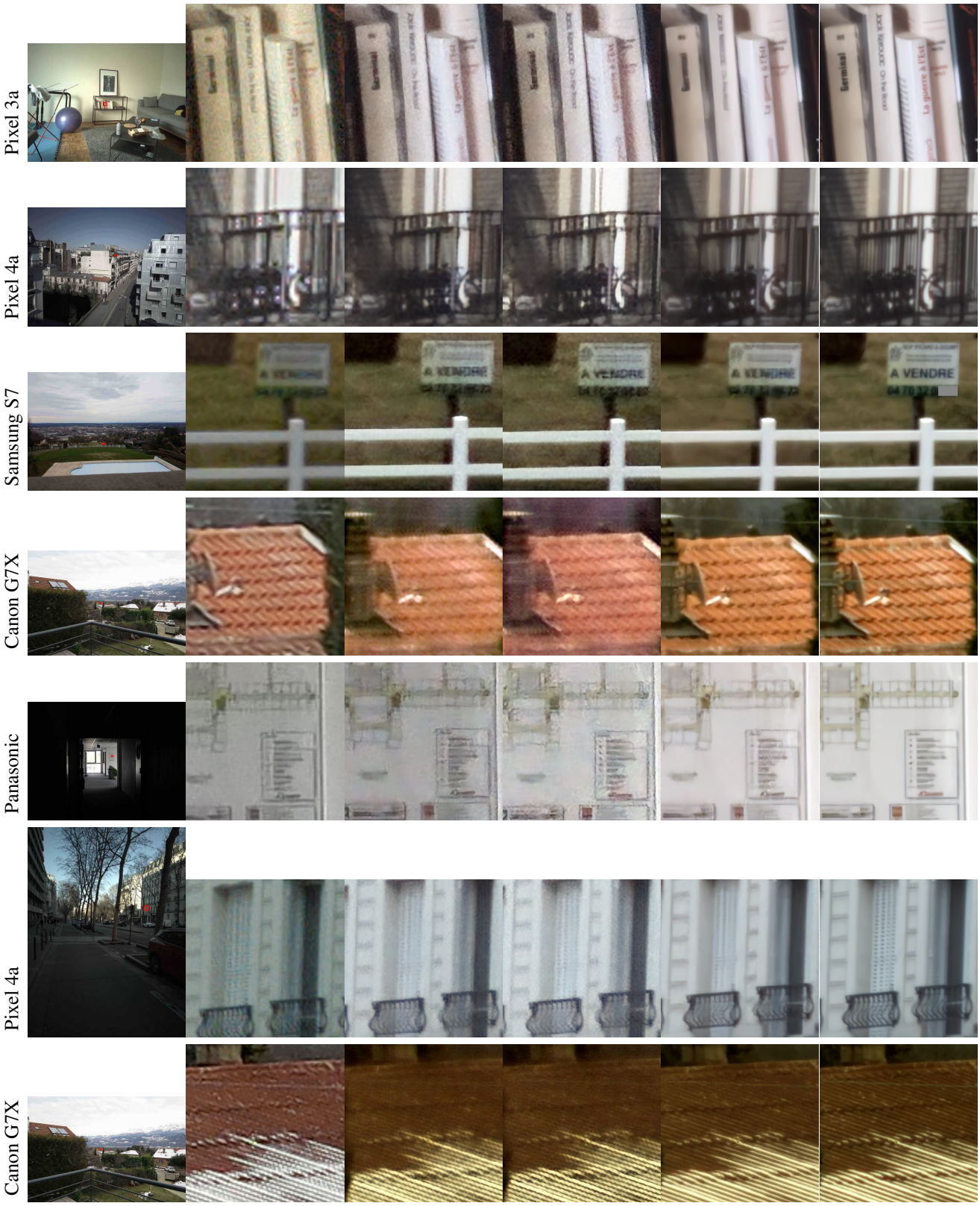

Figure A3: Results from real raw image bursts obtained with various cameras. We provide comparisons with single image and multiframe baselines. Finest restored details can be seen by zooming on computer screen.

Figure A4: Extreme $\times 16$ upsampling experiment. The right image is obtained by processing a burst of 20 LR images presented on the left obtained with synthetic random affine movements and average pooling downsampling

Figure A5: Image restoration of images taken at night with very low signal to noise ratio by using a Panasonic GX9 camera.

Figure A6: Visual differences caused by merging a different number of frames in the case of low SNR scenes. With a larger number of frames we can observe a quality increase and better denoising.

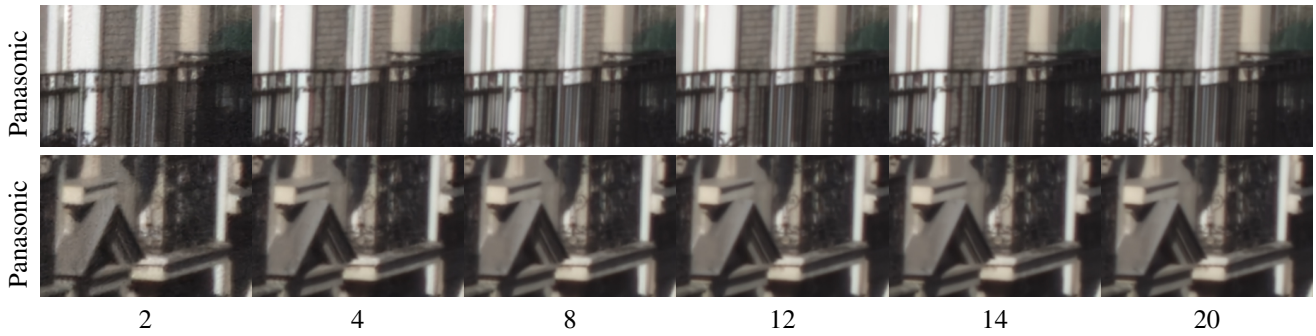

Figure A7: Visual differences caused by merging a different number of frames in the case of high SNR scenes. With a larger number of frames we can observe a quality increase.

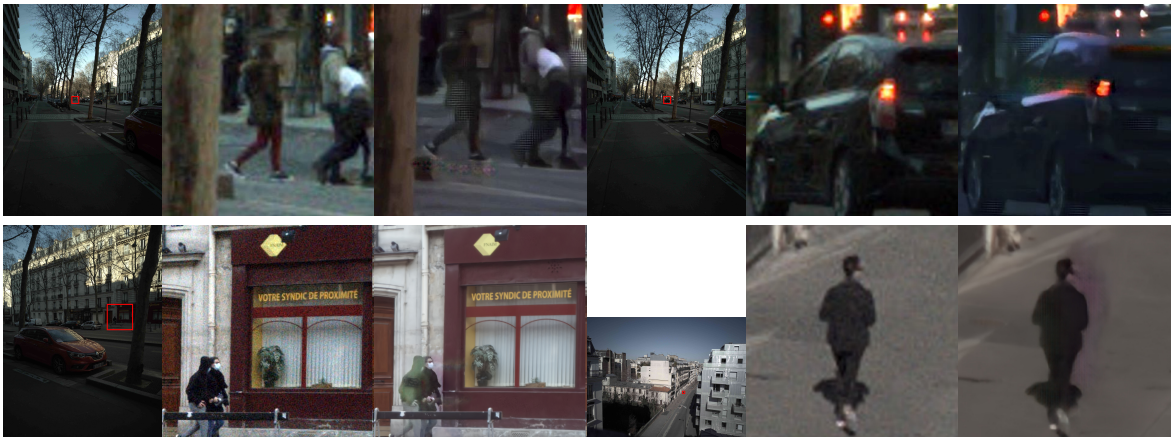

Figure A8: Misalignements artefacts due to moving objects in the scene. Our current implementation does not handle fast moving objects and then generates visual artefacts. Dealing with fast dynamic scenes will be the focus of future work.