Self-Supervised Burst Super-Resolution

Goutam Bhat[1,*] Michaël Gharbi[2] Jiawen Chen[2] Luc Van Gool[1] Zhihao Xia[2]
[1]CVL, ETH Zurich [2]Adobe

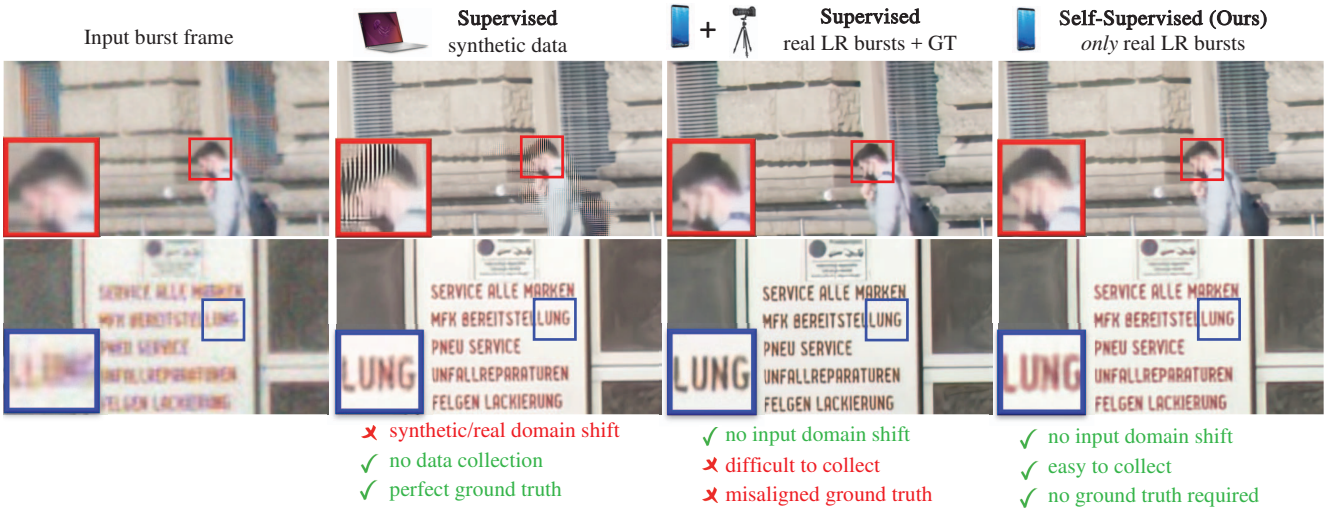


Figure 1: Comparison of our self-supervised approach with existing training strategies on the real-world BurstSR [2] dataset, using the same architecture [3]. Training a model on only synthetically generated data often leads to artifacts on real images, due to the domain shift. The same model trained in a supervised manner using real data provides better results. However, collecting paired ground-truth for real data requires specialized setups and signifiant manual labor, which makes it hard to scale. Furthermore, the smartphone and DSLR used to capture the input and ground-truth often has spatial misalignment and different color responses. Models trained with such "weakly-paired" ground-truth exhibit color shifts in their predictions. Our self-supervised approach alleviate these issues. Despite using *only* easy-to-acquire, unpaired noisy low-resolution bursts during training, our method produces clean, artifact-free high-resolution results,

Abstract

We introduce a self-supervised training strategy for burst super-resolution that only uses noisy low-resolution bursts during training. Our approach eliminates the need to carefully tune synthetic data simulation pipelines, which often do not match real-world image statistics. Compared to weakly-paired training strategies, which require noisy smartphone burst photos of static scenes, paired with a clean reference obtained from a tripod-mounted DSLR camera, our approach is more scalable, and avoids the color mismatch between the smartphone and DSLR. To achieve this, we propose a new self-supervised objective that uses a forward imaging model to recover a high-resolution image from aliased high frequencies in the burst. Our approach does not require any manual tuning of the forward model's parameters; we learn them from data. Furthermore, we show our training strategy is robust to dynamic scene motion in the burst, which enables training burst super-resolution models using in-the-wild *data. Extensive experiments on real and synthetic data show that, despite only using noisy bursts during training, models trained with our self-supervised strategy match, and sometimes surpass, the quality of fully-supervised baselines trained with synthetic data or weakly-paired ground-truth. Finally, we show our training strategy is general using four different burst super-resolution architectures.*

1. Introduction

Recent RAW burst super-resolution pipelines have significantly improved the quality of modern smartphones photos [33, 8]. State-of-the-art algorithms use specialized deep learning models that learn to merge the burst frames into a single high-resolution image [3, 19, 18, 23, 24]. Training them requires paired datasets, in which each noisy burst is matched to a clean reference. Most approaches synthe-

*Work partly done during an internship at Adobe.

size realistic bursts from the reference using carefully tuned degradation models [2, 3, 19, 24]. But because of low-level mismatches between the real and synthetically generated bursts (e.g., noise distribution, blur kernels, camera trajectories, scene motions, etc), models trained synthetically often do not generalize well to real-world inputs (Figure 1). To avoid this, other works collect *weakly-paired* datasets in which the reference is a high-resolution image of the same scene captured using a DSLR and a zoom lens on tripod [2, 39]. However, this capture process is tedious and time-consuming, and the resulting image pairs are often misaligned, exhibit color and detail mismatches because of the different sensors, and permit limited scene motion.

In this work, we propose a new self-supervised training strategy for burst super-resolution that alleviates the limitations of both synthetic and weakly-supervised datasets. Our approach only requires *real-world noisy bursts* for training, which are easy to collect. It eliminates the data collection complexity of weakly-paired approaches, and, by using real bursts, avoids the domain gap issues that plague synthetically trained models.

We derive a self-supervised reconstruction objective that models the relationship between the noisy burst and the ideal clean reference image we wish to recover. In particular, we exploit the property that burst images are noisy, aliased, and subsampled measurements of a scene, at random spatial offsets due to hand tremor [33], to recover high-frequency image details. Specifically, during training, we randomly split each burst into two sets of images. The first is passed as input to a burst super-resolution network to produce a high-resolution output, from which we derive low-resolution images using a forward image degradation model. We compare these low-resolution frames against the second set of burst images to compute our reconstruction loss. Optimizing this loss on a single burst provides too sparse a signal to train burst super-resolution models. But in a stochastic optimization involving a large dataset of bursts with random camera displacements, our self-supervised objective enables learning a robust image prior, and lets us recover high-resolution merged images.

Our loss function uses an explicit but general parameterized image formation model. Crucially, we do not make any limiting assumption about the *parameters* of this model (e.g., the precise noise distribution, the lens point spread function). Instead, we jointly learn the model's parameters along with the super-resolution network from data. Our training approach is general: it can be used to train any neural network architecture using bursts captured from any sensor. By using *only* noisy low-resolution bursts, which are easy to collect, our approach opens the opportunity to deploy state-of-the-art super-resolution network architectures for various cameras in real-world settings. In short, our contributions are the following:

- To the best of our knowledge, we introduce the first self-supervised training approach for raw burst super-resolution using *only* noisy, low-resolution inputs.
- We develop a robust self-reconstruction loss for training on bursts with dynamic object motion, which are prevalent in real *in-the-wild* bursts.
- Our approach can be used learn a lens blur kernel jointly with a burst super-resolution model, thereby alleviating the need of explicit blur kernel estimation.
- We perform extensive experiments on two real world burst datasets, using four different network architectures. Our approach obtains promising results compared to the model trained using weakly-paired data, despite using only low-resolution bursts.

2. Related Work

Learning real-world single image super-resolution. A number of approaches aim to manually design sophisticated synthetic pipelines which can generate realistic training data [37, 36, 30] for fully-supervised training. However, accurately modeling the real world degradation process is challenging. Thus, a number of approaches aim to learn this degradation process, which can then be used to generate training data [21, 4, 17, 32, 25, 22]. In contrast, a few approaches [39, 5, 7] collect real LR-HR pairs for supervised training using specialized data collection procedures.

Burst Super-Resolution. Classical approaches [16, 1, 14, 12, 13] minimize a reconstruction error computed using the physical image formation model to perform multi-frame super-resolution. Wronski *et al*. [33] propose a kernel regression based approach for burst SR using hand-held cameras. Recent approaches aim to exploit deep learning to perform burst SR [2, 19, 9, 23, 26]. Bhat *et al*. [2] uses a weighted-sum approach to merge the encodings of input images. [24] performs non-local fusion of images in feature space. Bhat *et al*. [3] minimize the classical reconstruction error [16, 13] in a learned feature space. Lecouat *et al*. [19] on the other hand leverage the *plug-and-play* [6, 29] framework to integrate a CNN-based regularizer into the classical model-based objective. Dudhane *et al*. [9] introduce a transformer based architecture for burst restoration.

Learning burst super-resolution. Compared to single image SR, modeling real-world burst SR is even more challenging, as one needs to accurately model the real-world blur kernels, noise distribution, camera motion, and object motion. Bhat *et al*. [2] introduce a synthetic pipeline to generate training data for burst super-resolution. However, this pipeline does not accurately model real world blur kernels, assumes burst frames are related by homographies, and does not consider object motion. Alternatively, Bhat *et al*. [2] also collect a real world dataset containing bursts captured from a handheld smartphone camera, along with

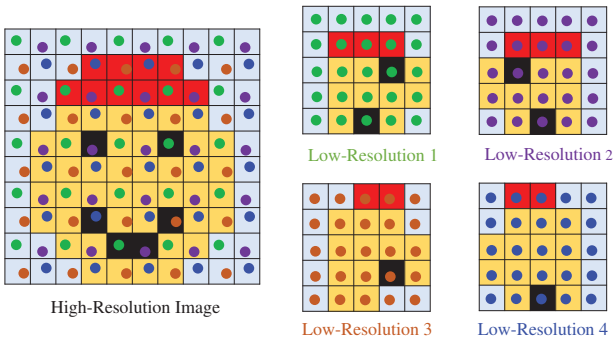


Figure 2: Motivation for our self-supervised training approach. Due to camera motion, the different low-resolution images in a burst provide samples at different spatial locations in the high-resolution image (depicted as colored dots). By comparing downsampled versions of the predicted HR image to the different LR images (after alignment), we can obtain sufficient supervision signal to train a burst SR model. Such supervision is also sufficient for removing any i.i.d. zero-mean noise in the LR images, as each LR image is a noisy observation of the same signal.

corresponding HR reference captured using a DSLR with a zoom lens on a tripod. Sophisticated alignment is done at patch-level to align the reference frame with the smartphone bursts, after which the authors still need to use a flow-based robust loss for handling the remaining misalignments. Furthermore, due to the mismatched camera characteristics between a smartphone and a DSLR (*e.g.*, color), the learned burst SR models tend to alter the output image. By training directly from a dataset of only low-resolution noisy bursts, which we stress are much easier to collect, our approach addresses all the aforementioned issues.

Self-supervised training for image restoration. A few works have explored unsupervised or self-supervised alternatives for training image restoration models [35, 10, 27]. Yuan *et al*. [35] utilize cycle consistency to train single image SR model using unpaired data. Lehtinen *et al*. [20] show that in case of zero-mean noise, denoising networks can be trained using only independent noisy instances of a clean image. Ehret *et al*. [10] utilize a similar principle to train a video denoising network. This approach, denoted frame-to-frame, is also employed to perform joint denoising and demosaicking using a burst which contains multiple noisy, albeit shifted, versions of the same scene [11]. Nguyen *et al*. [27] extend this idea to train a multi-frame super-resolution model for grayscale satellite images. In contrast to [27], we introduce a general approach for training raw burst SR models for natural images, which is agnostic to the model architecture. Instead of excluding the *base* frame from model input during training, we use a set of held-out burst images to compute a self-supervision loss. [27] assumes a fixed blur kernel when computing the training loss. Instead, our approach learns the blur kernel from data. Furthermore, our approach is robust to the dynamic motions prevalent in real-world bursts.

3. Method

We address the problem of training *self-supervised* models for raw burst super-resolution, and introduce a general strategy for training *any* burst SR model to produce a high-resolution image, given *only* noisy low-resolution data. Our approach is based on the property that the frames in a low-resolution handheld burst provide independent noisy, subsampled versions of an underlying ground truth image because of camera motion due to hand shake (see Fig. 2). These different degraded observations of the same scene can provide information to recover the underlying high-resolution image. Over a large training dataset, this can provide sufficient supervision signal to train a deep learning model to recover the high-resolution images.

Notably, our method is designed to handle challenges with real-world training data. We jointly learn the camera lens blur, along with the SR model parameters, to alleviate the need of lens calibration. We also introduce a robust loss function to handle object motions present in real bursts.

3.1. Self-Supervised Training

Here, we introduce our self-supervised training objective. Given a burst $B = \{b_i\}_{i=1}^{N}$ of N images from the training dataset, as illustrated in Fig. 3, we partition it into two disjoint sets $B_{\text{model}} = \{b_i\}_{i=1}^{K}$ and $B_{\text{unseen}} = \{b_i\}_{i=K+1}^{N}$ containing K and $N-K$ images, respectively. The first set is passed to the burst SR model f, which outputs an HR prediction $\hat{y} = f(B_{\text{model}})$. Next, we use an image formation model $\Pi_{m_i,k}$ to synthesize LR burst frames $\hat{b}_i$:

$$\hat{b}_i = \Pi_{m_i,k}(f(B_{\text{model}})). \tag{1}$$

Our image formation model treats a LR burst image as a shifted and degraded version of HR image y. It is parameterized by the motion m_i from frame i to our prediction, and a spatially-invariant lens blur kernel k, as follows

$$\Pi_{m_i,k}(y) = H D_k \Phi_{m_i}(y)\,, \tag{2}$$

Here, the original image y is first warped by Φ to account for camera motion m_i. The warped image is then blurred by lens blur D_k, then subsampled and mosaicked by the linear operator H to obtain the observation b_i. These synthetic burst frames are then compared to the remaining real frames B_{unseen} using the following reconstruction error:

$$\ell = \frac{1}{N-K} \sum_{i=K+1}^{N} \|b_i - \Pi_{m_i,k}(f(B_{\text{model}}))\|_1. \tag{3}$$

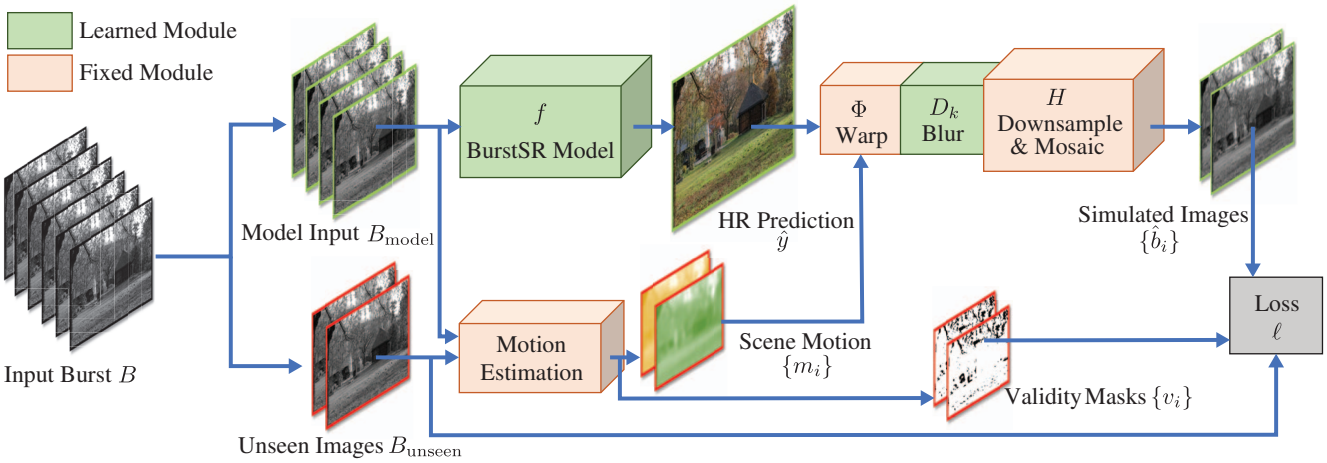


Figure 3: An overview of our self-supervised training framework. Given a noisy low-resolution training burst, we split it into two parts. The first part is passed through a burst SR model f to obtain a putative HR prediction $\hat{y}$. Next, we utilize a simple image formation model to simulate the *unseen* images from the second part of the input burst. This involves estimating the motion m_i for each image b_i. The estimated motion is used to warp the prediction $\hat{y}$ to b_i, followed by blur with kernel k, downsampling, and mosaicking. The simulated images $\hat{b}_i$ are compared with the observed images b_i to compute the training loss ℓ. We additionally compute a validity mask v_i in order to obtain a robust loss in presence of dynamic objects. The burst SR model parameters, as well as the blur kernel k are jointly learned by minimizing the self-supervised objective ℓ.

We use the L_1 loss as it is more robust to clipped noise [10]. Note that in loss (3), we use the observed noisy image as the ground truth for the LR image synthesized using the HR prediction, as in the *Noise2Noise* frameworks [20]. If the camera sensor noise is i. i. d. and zero-mean, as is commonly the case, we can train the burst SR model f to make accurate HR predictions by minimizing the reconstruction error ℓ over a large training dataset [20, 34].

3.2. Motion Estimation

In real-world settings, the motion m_i is unknown and needs to be estimated for each image b_i. We parameterize the motion parameters m_i as pixel-wise optical flow which aligns the prediction $\hat{y}$ to input image b_i. This allows us to better handle object motions and perspective shifts compared to, *e.g.* using a homography. We use an off-the-shelf flow network, PWC-Net [28]. Note that directly computing the optical flow between $\hat{y}$ and b_i is challenging due to differences in spatial resolution and color space (RGB vs. Raw). Furthermore, in the early stages of training, $\hat{y}$ and b_i can be misaligned and have widely different appearances. This can lead to training instabilities and the network may start hallucinating undesired spatial shifts. We avoid this by instead estimating the flow between the first image in the burst b_1 and the i-th image b_i, which we bilinearly upsample to the spatial resolution of $\hat{y}$ to obtain m_i. This strategy constrains the burst SR model f to generate predictions that are aligned w.r.t. the first burst image.

Robust Loss. Since m_i is an optical flow estimate, it inevitably contains errors, especially in the presence of noise. Furthermore, real-world bursts contain dynamic objects and occlusions where accurate alignment is impossible. Consequently, the reconstruction error (Eq. (3)) is invalid in these regions, and naively including them in the loss leads to artifacts in the HR prediction (see Sec. 4.3). We robustly handle errors in motion estimation using a simple binary *validity mask* v_i that indicates which pixels in each image b_i can reliably be used in the reconstruction loss. First, consider the magnitude of the warping residual $|b_i - \Phi_{m_i}(b_1)|$. We expect it to be high at misalignments and occlusions, and low where the flow is accurate. However, it can also be high at well aligned regions due to noise or aliasing. Discarding those regions would discard the very information needed for denoising and super-resolution. To combat gross alignment errors but preserve the useful regions that are challenging to align, we found it sufficient to filter the residual with a low-pass Gaussian F with a standard deviation of 2.7 pixels: $|F(b_i) - F(\Phi_{m_i}(b_1))|$. We threshold this filtered residual and apply morphological dilation to suppress thresholding noise. Our final reconstruction loss is:

$$\ell = \frac{1}{N-K} \sum_{i=K+1}^{N} \| v_i \odot (b_i - \Pi_{m_i,k}(f(B_{\text{model}}))) \|_1, \quad (4)$$

where $\odot$ denotes point-wise multiplication.

3.3. Blur Kernel Estimation

Unlike a synthetic training strategy, we do not assume a known blur kernel or calibrate it for each camera, which is cumbersome and labor-intensive. We instead opt to directly learn the per-camera kernel from data, jointly with

	DBSR [2]			DeepRep [3]			BIPNet [8]			Burstormer [9]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Supervised, Synthetic	36.27	0.918	0.135	36.53	0.920	0.133	36.71	0.924	0.133	36.73	0.923	0.134
Supervised, Weakly-paired	36.01	0.922	0.089	34.62	0.904	0.089	36.77	0.926	0.097	36.84	0.925	0.100
Self-Supervised, Ours	**38.23**	**0.940**	**0.079**	**38.73**	**0.943**	**0.075**	**38.44**	**0.941**	**0.093**	**38.81**	**0.943**	**0.085**

Table 1: Comparison of our self-supervised training with existing training alternatives using synthetic or weakly-paired data, on a synthetic SynBSR benchmark. Results are shown for 4 different architectures, in terms of PSNR, SSIM, and LPIPS.

Burst image | Supervised, synthetic | Supervised, weakly-paired real | **Self-supervised (Ours)**

Figure 4: Comparison with supervised training strategies using the DeepRep [3] architecture on the real-world BurstSR dataset [2]. The first column shows an input burst image processed using Adobe Camera Raw. The model trained using only synthetic data struggles to handle dynamic scenes. The real data supervised training provides better results, but can introduce small color shifts (second row) due to the color shift in the ground-truth. Furthermore, it requires high-resolution ground truth which is cumbersome to collect. Our self-supervised approach can recover most of the high-frequency details without introducing any artifacts or color shifts, despite using *only* low-resolution noisy bursts for training.

the parameters of the burst SR model, to avoid domain shift. We represent the blur using an *unnormalized* 9×9 kernel, which we pass through a softmax operator to guarantee that the weights are non-negative and sum to 1.

4. Experiments

We perform a thorough evaluation our of approach on synthetic as well as real-world data. First, we compare our self-supervised training method with alternate training strategies. Experiments are performed using 4 different network architectures, highlighting the generality of our approach. Next, we demonstrate our ability to train models for a different sensor, where the high-resolution ground truths are not available. Finally, we perform a detailed analysis of our self-supervised training framework. More details, training parameters, and results are provided in the supplement.

4.1. Comparison With Existing Methods

We propose a self-supervised approach to train burst super-resolution models using only low-resolution, noisy bursts. In this section, we compare our approach with existing training methods using 4 different architectures.

Baselines. As discussed in Section 1, one of the major challenges when training real-world burst super-resolution models is the unavailability of accurately aligned input-ground truth pairs. We compare our approach with the two training alternatives that are commonly employed to tackle this issue. i) **Synthetic**: Training bursts are generated using a synthetic pipeline introduced in [2], using bilinear downsampling kernel and heteroscedastic Gaussian noise model. The model is then trained in a fully supervised manner using the generated bursts [19, 18]. ii) **Weakly-paired real:** The model is trained using real-world training pairs containing unavoidable spatial and color misalignments. An explicit

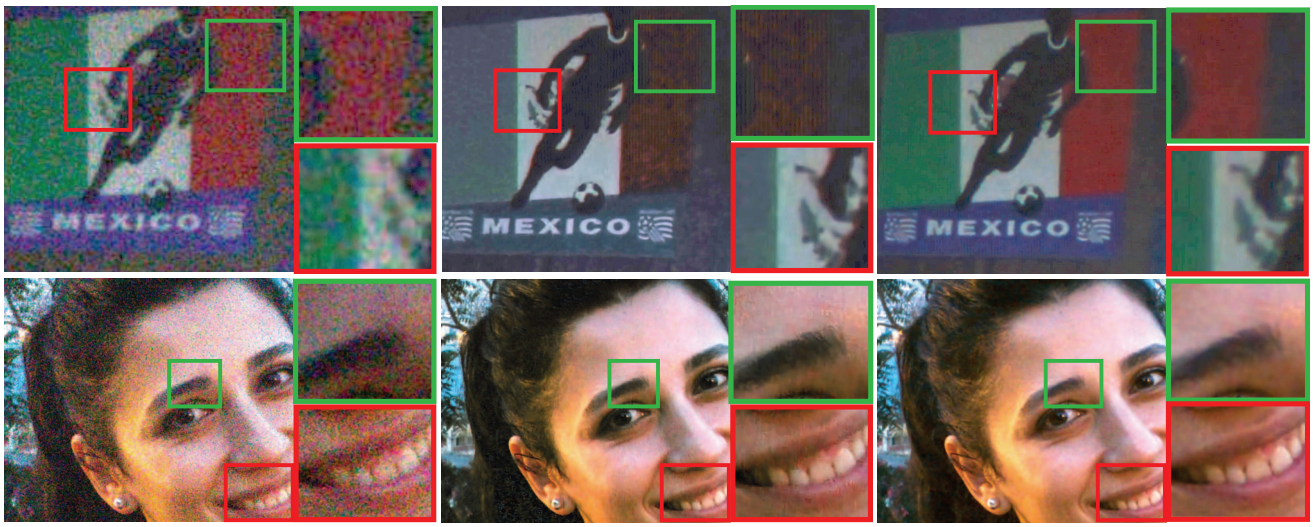
Figure 5: Comparison with alternate training strategies using Burstormer [9], BIPNet [8], and DBSR [2] on BurstSR [2]. The models trained using only synthetic data introduce severe artifacts due to the domain gap. Our approach recovers most of the high-frequency details and provides results comparable to the models trained using weakly-paired HR ground truth.

alignment strategy [2] is employed during training in order to handle these misalignments [2, 3, 8, 9]. Note that this training strategy requires manually collecting ground truth image for each burst, which is cumbersome.

Datasets and metrics. We evaluate our approach for the $4\times$ super-resolution task on two benchmarks, namely the real-world **BurstSR dataset** [2], and a synthetically generated **SynBSR** benchmark. The real-world BurstSR dataset contains 200 bursts, each with 14 images, captured using a Samsung Galaxy S8 smartphone camera. For each burst, the dataset also contains a high-resolution, weakly-paired ground truth captured using a DSLR. A subset containing 160 bursts can be used to training models, while the remaining are set aside for evaluation. Due to the presence of misalignments between the bursts and the HR ground truth, computing robust evaluation metrics for quantitative comparison is challenging. Hence we only perform a qualitative comparison of different methods on this dataset. In order to perform quantitative comparison of our approach with alternate methods, we introduce a synthetic **SynBSR** benchmark which simulates the real-world training challenges. SynBSR consists of training and test bursts generated using a synthetic pipeline, similar to the one employed in [2]. For each training burst, a slightly misaligned HR ground truth is provided, as in the BurstSR dataset. This models the practical real-world training challenges. We simulate the misalignments in our setup by introducing small random global rotation and translation, and a linear color transformation to the ground truth image. We utilize a test set containing 200 bursts. Since the bursts are synthetically generated, a perfectly aligned groundtruth is available for the test set, which can be used to compute quantitative metrics. We employ PSNR, SSIM [31], and LPIPS [38] metrics for comparisons.

Implementation details. Similar to the strategy employed for weakly-paired training [2], we first pretrain models on synthetic data. These are then further trained on real low-resolution bursts using our self-supervised strategy.

Quantitative Results on SynBSR. We compare our self-supervised training approach to the **synthetic** and **weakly-paired** training baselines on the synthetic SynBSR dataset. The results for 4 different network architectures, DBSR [2], DeepRep [3], BIPNet [8], and Burstormer [9], are reported in Tab. 1. Note that the weakly-paired data training can introduce global spatial shifts in the model predictions since it uses an alignment based loss. In order to not penalize these global shifts, we first align the prediction of the weakly-paired network to the ground truth using a global translation, and then compute the performance metrics. Synthetic data training can lead to sub-optimal results due to domain mismatches between the training and the test data distributions. Predictions from the supervised weakly-paired training can have small spatial deformations and color shifts, due to the presence of misalignments in the training data. This leads to lower performance scores. Our self-supervised training provides the best results for all 4 network architectures. For the Burstormer architecture, the model trained using our self-supervised approach obtains a PSNR of 38.81, outperforming the model trained using weakly-paired data



Burst image | Weakly-paired, BurstSR | **Self-supervised, HDR+**

Figure 6: Our self-supervised training can be used on existing raw burst datasets *with no ground-truth*, such as HDR+ [15] captured using a Google Nexus 6. We achieve superior SNR at higher resolution. In contrast, the weakly-paired model (trained on the BurstSR dataset), produces visible splotches and color shifts due to differences in noise distribution.

by $+1.97$ dB in PSNR. This improvement is obtained despite using less information for training, compared to the weakly-paired method.

Qualitative Results on real-world BurstSR. Here, we evaluate our approach on the real-world BurstSR [2] dataset, using DBSR [2], DeepRep [3], BIPNet [8], and Burstormer [9] architectures. For each architecture, we compare a model trained using our self-supervised strategy with the official models provided by the authors trained using **synthetic** and **weakly-paired** training strategies. A qualitative comparison using the DeepRep architecture is shown in Figure 4. For visualization, we render linear network predictions to sRGB using Adobe Camera Raw. Since the synthetic pipeline models only static scenes, a network trained this way produces artifacts when given a burst with scene motion (third row). The network trained using weakly-paired data provides better results. However, it can introduce small shifts in color (second row). This is because the ground truth high-resolution images are captured using a different camera featuring a sensor with a different color response. Compared to the weakly-paired model, our approach preserves color and recovers most of the high-frequency details, despite not using any high-resolution ground truths for training.

Figure 5 provides further qualitative comparisons using DBSR, BIPNet, and Burstormer architectures. The models trained using only synthetic data introduce severe artifacts, even in static regions in case of Burstormer and BIPNet. The models trained using our approach robustly handle dynamic scenes and exhibits no color shifts, achieving results comparable to the weakly-paired training approach. We reiterate that our approach utilizes only low resolution bursts for training, which are much easier to obtain compared to collecting weakly paired high resolution groundtruths. These results demonstrate the generality of our training framework.

4.2. Deployment On Other Sensor

Since our method only requires low-resolution noisy bursts, it can be adapted to train models for different sensors or domains, obviating current methods' need for extensive data-collection with specialized setups. As a proof of concept, we evaluate our strategy on the HDR+ dataset [15], which consists of 3640 bursts and no HR ground truth. Compared to BurstSR [2], HDR+ bursts cover a greater dynamic range, with bright highlights as well as dark shadows in the same image. To minimize shifts in noise and blur distribution, we use only bursts captured on the Nexus 6 camera. This yields 77 bursts, of which we use 70 for training. This experiment uses the $4\times$ DeepRep architecture.

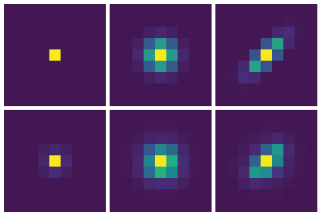
We start with an initial self-supervised model trained on the BurstSR dataset and finetune it on HDR+ data. Figure 6 shows our results as compared to the supervised real data baseline trained exclusively with BurstSR images. Unlike our method, the supervised approach is unable to use any images from HDR+ since high-resolution ground truth is unavailable. Consequently, it struggles with the noise present in the intentionally underexposed images, introducing "splotches" and color shifts.

4.3. Analysis of Self-Supervised training

We analyze different components of our self-supervised training framework in this section. Our analysis is performed on synthetic datasets generated using different blur

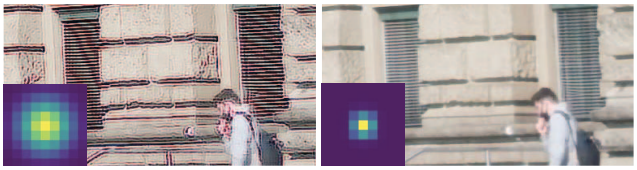
		Known			Unit Impulse	Gaussian Isotropic	Gaussian Isotropic	Gaussian Isotropic	Gaussian Anisotropic
		y	k	m_i	Read & Shot, Mix	Read & Shot, Mix	Read & Shot, Fixed	Only read, Fixed	Read & Shot, Mix
(I)	Fully Sup. SISR	✓			34.681	34.074	35.767	35.557	34.683
(II)	Fully Sup. burst SR	✓			39.810	37.254	39.002	38.761	37.590
(III)	Self Sup. known k, m_i	✗	✓	✓	40.359	36.765	38.469	38.236	37.040
(IV)	Self Sup. known k	✗	✓	✗	38.999	36.347	37.872	37.577	36.544
(V)	Self Sup. fixed k	✗	✗	✗	28.328	28.575	28.970	29.493	34.822
(VI)	Self Sup. learned k	✗	✗	✗	38.451	35.853	37.337	37.201	36.205

Table 2: Analysis of our self-supervised training approach on synthetic datasets generated using different combinations of blur kernels and noise distributions. All numbers are in dB PSNR. We evaluate our approach under different settings where the scene motion m_i and the blur kernel k are assumed to be known or estimated. We also include comparisons with a burst SR model and a single image SR model training using fully-supervised *paired* HR ground truth. Despite using exclusively low-resolution bursts our method is within $\sim$2 dB of the fully-supervised upper bound (second row). This holds even when motion and blur are estimated (bottom row).



Unit Impulse Iso. Gauss. Ani. Gauss.

Figure 7: (Top) Blur kernels used to generate synthetic bursts. (Bottom) The blur kernels learned by our self-supervised method.



Incorrect kernel Learned kernel

Figure 8: Using an incorrect blur kernel (inset) leads to poor results (left). Jointly learning the kernel with the SR model produces much better results (right).

kernels and noise distributions. We consider 3 different blur kernels: the unit impulse, an isotropic Gaussian, and an anisotropic Gaussian (see Figure 7). We also consider 3 noise distributions: i) heteroscedastic Gaussian with varying noise levels, which models camera shot and read noise at different gains (ISO), ii) heteroscedastic Gaussian with a fixed noise level, and iii) homoscedastic Gaussian with a fixed noise level, which only models sensor read noise. For each combination of blur and noise, we synthesize training and test datasets using the pipeline introduced in [2]. We train and evaluate a set of models on exclusively that dataset. Our experiments are conducted on the $2\times$ super-resolution task, using a smaller variant of the DeepRep model [3]. We use a burst size of 14 frames for both training and test.

Analysis of reconstruction loss. We study whether our reconstruction loss (4) provides enough supervision for training a burst SR model. We first consider the ideal scenario where the blur kernel k, as well as the motion m_i are known. For comparison, we also evaluate oracle burst SR and single image SR models, *i.e.* models which are trained in a fully-supervised manner using *paired* HR ground truth. The results of our analysis, in terms of PSNR, are shown in Table 2. Our self-supervised approach obtains promising results for each of the 5 data distributions. When motion and blur kernel are known (third row), the performance of our approach is only 0.55 dB PSNR lower than the fully-supervised upper-bound (second row) in the worst case.

Impact of using estimated motion. We consider the practical scenario where the motion m_i is unknown and must be estimated (Section 3.2). Even when using the estimated scene motion (fourth row), our approach obtains PSNR scores within 1.2 dB of the fully-supervised setting, despite using *only* the noisy LR bursts for training.

Analysis of learning the blur kernel. Next, we consider the scenario where the blur kernel is also unknown, and analyze the impact of jointly learning the blur kernel k with the parameters of the burst SR model. For each data distribution, we initialize the blur kernel to an isotropic Gaussian with large standard deviation (which does not match the true standard deviation). We then jointly learn the kernel with the parameters of the burst SR model by minimizing our self-supervised reconstruction loss (4). As a baseline, we fix the blur kernel to its initial incorrect value and learn only the burst SR model. Table 2 (fifth row) shows that using an incorrect blur kernel drastically degrades performance for nearly all data distributions. In contrast, our approach achieves significantly better performance (last row), while learning accurate blur kernels (Figure 7). A qualitative comparison showing the impact of learning the blur kernel is shown in Figure 8.

Impact of validity mask. We train models with and without a validity mask v_i on the BurstSR dataset [2]. Figure 9 shows that the validity mask is crucial to avoid ghosting around moving objects.

Without validity mask With validity mask

Figure 9: If trained without a validity mask, our models introduce ghosting artifacts around moving objects.

5. Conclusion

We showed how to train deep learning models for raw burst super-resolution without using any high resolution ground truth. We are able to simultaneously estimate lens defocus and predict a high resolution image while being robust to scene motion. Our strategy compares favorably to the state of the art despite being restricted to data that is significantly easier to collect.

Limitations and Future Work. Our image formation model assumes that the lens blur kernel is fixed over the full dataset. This may not hold in practise. Furthermore, we do not model any motion blur, which may limit the performance of the model. Extending our approach to incorporate spatially varying blur is an interesting future work. Another future direction is to integrate additional supervision in the form of unpaired high-resolution images.

References

[1] B. Bascle, A. Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *ECCV*, 1996. 2

[2] Goutam Bhat, Martin Danelljan, L. Gool, and R. Timofte. Deep burst super-resolution. In *CVPR*, 2021. 1, 2, 5, 6, 7, 8

[3] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021. 1, 2, 5, 6, 7, 8

[4] Adrian Bulat, J. Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a GAN to learn how to do image degradation first. In *ECCV*, 2018. 2

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[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*, 3:84–98, 2017. 2

[7] Chang Wen Chen, Zhiwei Xiong, Xinmei Tian, Zhengjun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, 2019. 2

[8] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022. 1, 5, 6, 7

[9] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burstormer: Burst image restoration and enhancement transformer. In *CVPR*, 2023. 2, 5, 6, 7

[10] Thibaud Ehret, Axel Davy, Pablo Arias, and G. Facciolo. Joint demosaicking and denoising by fine-tuning of bursts of raw images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8867–8876, 2019. 3, 4

[11] Thibaud Ehret, Axel Davy, G. Facciolo, Jean-Michel Morel, and Pablo Arias. Model-blind video denoising via frame-to-frame training. In *CVPR*, 2019. 3

[12] Michael Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 6 12:1646–58, 1997. 2

[13] Sina Farsiu, Michael Elad, and P. Milanfar. Multiframe demosaicing and super-resolution from undersampled color images. In *IS&T/SPIE Electronic Imaging*, 2004. 2

[14] R. Hardie, K. Barnard, John G. Bognar, E. Armstrong, and E. Watson. High-resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. *Optical Engineering*, 37:247–260, 1998. 2

[15] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 35(6), 2016. 7

[16] M. Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP Graph. Model. Image Process.*, 53:231–239, 1991. 2

[17] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[18] Bruno Lecouat, Thomas Eboli, Jean Ponce, and Julien Mairal. High dynamic range and super-resolution from raw image bursts. *ACM Transactions on Graphics (TOG)*, 41(4), July 2022. 1, 5

[19] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2350–2359, 2021. 1, 2, 5

[20] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *ICML*, 2018. 3, 4

[21] Andreas Lugmayr, Martin Danelljan, and R. Timofte. Unsupervised learning for real-world super-resolution. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3408–3416, 2019. 2

[22] Zhengxiong Luo, Yan Huang, , Shang Li, Liang Wang, and Tieniu Tan. Learning the degradation distribution for blind image super-resolution. In *CVPR*, 2022. 2

[23] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu.

Bsrt: Improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 998–1008, 2022. 1, 2

[24] Ziwei Luo, Lei Yu, Xuan Mo, Youwei Li, Lanpeng Jia, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 471–478, June 2021. 1, 2

[25] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, 2020. 2

[26] Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Gated multi-resolution transfer network for burst restoration and enhancement. In *CVPR*, 2023. 2

[27] Ngoc Long Nguyen, Jérémy Anger, Axel Davy, Pablo Arias, and G. Facciolo. Self-supervised multi-image super-resolution for push-frame satellite images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1121–1131, 2021. 3

[28] Deqing Sun, X. Yang, Ming-Yu Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 4

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

[30] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 2

[31] Zhou Wang, A. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. 6

[32] Valentin Wolf, Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. DeFlow: Learning complex image degradations from unpaired data with conditional flows. In *CVPR*, 2021. 2

[33] B. Wronski, Ignacio Garcia-Dorado, M. Ernst, D. Kelly, Michael Krainin, Chia-Kai Liang, M. Levoy, and P. Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38:1 – 18, 2019. 1, 2

[34] Zhihao Xia and Ayan Chakrabarti. Training image estimators without image ground-truth. In *NeurIPS*, 2019. 4

[35] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, 2018. 3

[36] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[37] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018. 2

[38] Richard Zhang, Phillip Isola, Alexei A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[39] X. Zhang, Qi feng Chen, R. Ng, and V. Koltun. Zoom to learn, learn to zoom. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3757–3765, 2019. 2