Kernel-Aware Burst Blind Super-Resolution

Wenyi Lian
Department of Information Technology
Uppsala University, Sweden
wenyi.lian.7322@student.uu.se

Shanglian Peng
School of Computer Science
Chengdu University of Information Technology, China
psl@cuit.edu.cn

Abstract

Burst super-resolution (SR) technique provides a possibility of restoring rich details from low-quality images. However, since real world low-resolution (LR) images in practical applications have multiple complicated and unknown degradations, existing non-blind (e.g., bicubic) designed networks usually suffer severe performance drop in recovering high-resolution (HR) images. In this paper, we address the problem of reconstructing HR images from raw burst sequences acquired from a modern handheld device. The central idea is a kernel-guided strategy which can solve the burst SR problem with two steps: kernel estimation and HR image restoration. The former estimates burst kernels from raw inputs, while the latter predicts the super-resolved image based on the estimated kernels. Furthermore, we introduce a pyramid kernel-aware deformable alignment module which can effectively align the raw images with consideration of the blurry priors. Extensive experiments on synthetic and real-world datasets demonstrate that the proposed method can perform favorable state-of-the-art performance in the burst SR problem.

1. Introduction

With the growing popularity of built-in smartphone cameras, the multi-frame super-resolution (MFSR) has drawn much attention due to its high practical potential to recover rich details from several low-quality images [4, 5, 32]. Compared with single image super-resolution (SISR), MFSR can provide complementary information from subpixel shifts, avoiding aliasing artifacts and losing details [45, 44]. Typically, we explicitly model the multi-frame degradation process as:

$$\boldsymbol{x}_i = (\boldsymbol{k}_i \otimes \mathcal{T}_i \boldsymbol{y})_{\downarrow_s} + \eta_i, \tag{1}$$

where $\boldsymbol{y}$ is the original HR image, $\{\boldsymbol{x}_i\}_1^N$ is the observed low-resolution image bursts. $\boldsymbol{k}_i$ and $\mathcal{T}_i$ denote the blur kernel and scene motion transform, respectively. $\otimes$ represents convolution operation and $\downarrow_s$ is the subsequent downsampling with scale factor s. η_i is an white Gaussian noise that is independent to LR images.

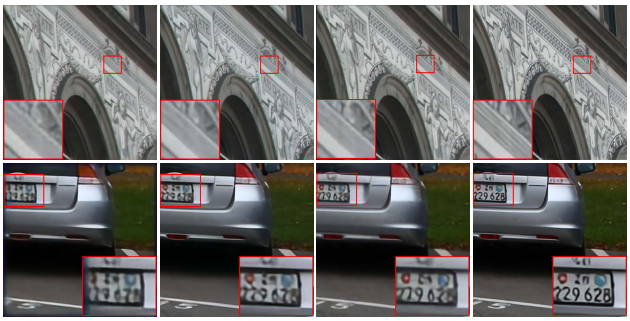


Figure 1: We propose a degradation guided framework to handle the burst super-resolution problem on both real-world dataset (top row) and synthetic dataset (bottom tow). The proposed method outperforms existing state-of-the-art MFSR approaches DBSR [4], EBSR [40] and DeepREP [5]. Our method is effective in restoring edges and details.

Most existing MFSR methods assume that the blur kernels are known (e.g., bicubic) and the same for all frames [4, 5, 32, 40]. Under this assumption, these MFSR methods can achieve dramatic performance to search for the best inverse solution for the bicubicly downsampling degradation. However, they often suffer severe performance drop when applied to real-world applications of that the kernel is actually derived from cameras' intrinsic parameters that are complicated and inconsistent for all burst frames [16, 18, 59]. Moreover, multiple blurry inputs would make the restoration difficult and lose details (See Fig. 1). And transferring the bicubicly designed model to unknown degradation images is also inefficient. To this end, we pay more attention to the model that tackles degradation of multiple unknown blur kernels, *i.e.* burst blind SR.

Single image blind SR has been well studied in recent works [61, 58, 18, 2, 38], which often need to sequentially estimate blur kernel (or its embedding), and then restore

the SR image based on that kernel. However, the overall optimization of blind SR is usually alternating, complicated, and time-consuming [18, 21]. Such a problem could be even more serious when facing burst blind SR, where each frame has a specific blur kernel and an irregular motion displacement. So far little work has focused on the blind property of the burst SR. The common solution is to train a deep model directly on the bicubicly designed synthetic dataset, then finetune it on another real-world dataset [4, 15, 40, 32, 43]. However, it is quite challenging to set up a real-world dataset due to the degradation is always sensor-specific and the images captured by smartphones (LR) and DSLR cameras (HR) usually have different qualities and image signal processors (ISP).

In this paper, we address above issues by proposing a kernel-aware raw burst SR method based on the multi-frame degradation model (Eq. (1)), called *KBNet*, which takes into account the inconsistencies of blur kernels between frames and can learn a practical multi-frame super-resolver using synthetic data. The KBNet consists of two neural networks: a kernel modeling network that estimates blur kernels for each burst frames, and a restoring network that predicts the super-resolved image by fusing the information of all frames and the corresponding estimated kernels. To make full use of the degradation information from kernels, the restoring network employs an Adaptive Kernel-Aware Block (AKAB) to extract deblurred clean features and a Pyramid Kernel-Aware Deformable convolution (Pyramid KAD) module to align and fuse multiple complementary features. Our contributions are summarized as follows:

- We consider the inconsistent degradation of different frames and propose a novel kernel-aware network, named *KBNet*, for raw burst image blind super-resolution, which makes a substantial step towards developing read-world practical MFSR applications.
- We propose a degradation-based restoring network that uses adaptive kernel-aware blocks and a pyramid kernel-aware deformable alignment module to restore SR image based on blur kernels.
- Extensive experiments demonstrate that the proposed method achieves an art performance on various synthetic datasets and real images.

2. Related work

2.1. Single Image Super-Resolution

SISR is a problem of trying to recover a high-resolution image from its degraded low-resolution version. In the past few years, numerous works based on the neural network have achieved tremendous performance gains over the traditional SR methods [13, 14, 22, 29, 33, 55, 64, 35, 34]. Since the pioneering work SRCNN [13], most subsequent works focus on optimizing the network architecture [14, 31, 64, 35, 11] and loss function [27, 33, 55, 36]. These methods are difficult to apply to real-world applications due to their ill-posed nature.

2.2. Multi-Frame Super-Resolution

MFSR is an active area and has been well studied in the last three decades. Tsai and Huang [53] are the first that propose to tackle the MFSR problem. They propose to restore HR image in the frequency domain with known translations between frames. Peleg et al. [46] and Irani and Peleg [25] propose an iterative back-projection approach that can sequentially estimate HR image and synthesizes LR image. Later works [1, 17, 20, 47, 19] improve this method with a maximum a posteriori (MAP) model and a regularization term. Takeda et al. [50, 51] introduce a kernel regression technique for super-resolution and Wronski et al. [57] applies it on fusing aligned input frames.

Recently, several works [12, 28, 42] propose to incorporate deep learning to handle the MFSR problem in remote sensing applications. Bhat et al. [4, 3] introduce a real-world dataset and propose an attention-based based fusion approach for MFSR. And they further improve the model to handle both SR and denoising by transforming the MAP framework to a deep feature space. Luo et al. [40, 39] introduce the deformable convolution to MFSR and show its effectiveness of handling the alignment between frames. Bruno et al. [32] propose an effective hybrid algorithm building on the insight from [57]. Akshay et al. [15] propose to create a set of pseudo-burst features that makes it easier to learn distinctive information of all frames.

2.3. Blind Super-Resolution

Blind SR assumes that the blur kernels of degradation are unavailable. In recent years, the blind SR problem has drawn much research attention since it is close to real-world scenarios [41]. Zhang et al. [61] firstly propose to extract principal components of the Gaussian blur kernels and stretch and concatenate them with LRs to get degradation-aware SR images. Subsequently, Gu et al. [18] modify the strategy in [61] by concatenating kernel embeddings with deep features. Luo et al. [38] and Zhang et al. [60] propose to unfold the blind SR problem as a sequential two-step solution which can be alternately optimized. Hussein et al. [23] propose a closed-form correction filter to transform blurry LR images to bicubicly downsampled LR images. Luo et al. [37] reformulate the kernel to LR space and then they apply a deconvolution to get deblurred SR images. Moreover, ZSSR [48, 2] and MSSR [49] also can be applied to blind SR, where the training is conducted as test time, and it can exploit the internal information of the LR by using an image-specific degradation.

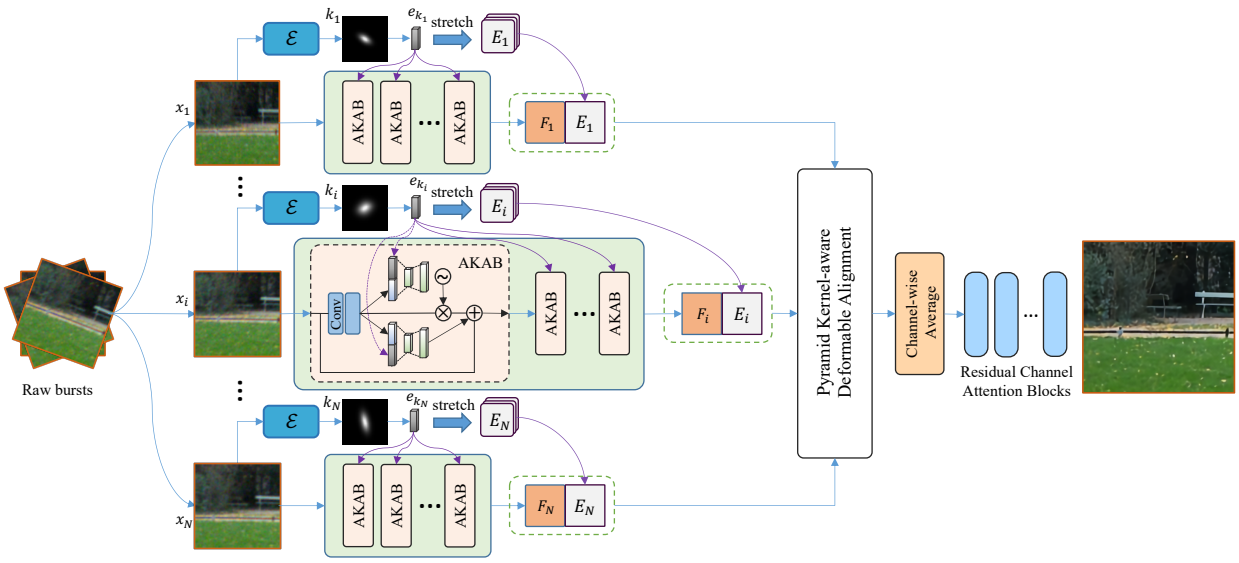


Figure 2: The overview of our method. The inputs are a set of RAW burst images $\{x_i\}_{i=1}^N$. We predict blur kernel for each frame through a simple CNN network, as estimator $\mathcal{E}$. The estimated kernels are reduced to embeddings by PCA and fed into groups of adaptive kernel-aware blocks (AKABs) to extract clean features. And we also stretch the kernel embeddings to degradation maps so as to concatenate them with clean features. These features are then aligned with the reference frame through a pyramid kernel-aware deformable alignment module. We fuse these aligned features with a channel-wise averaging strategy and use residual channel attention blocks to reconstruct the SR result.

3. Method

This section describes the main techniques of the proposed KBNet for raw burst blind super-resolution. As shown in Fig. 2, we first estimate the blur kernel for each frame and obtain its embedding vector through the principal component analysis (PCA). By taking the LR frame and the corresponding degradation kernel embedding as inputs, we can extract clean features with several adaptive kernel-aware blocks (AKAB). Then we stretch the kernel to degradation maps so that the feature and the kernel embedding can be concatenated and sent into the Pyramid kernel-aware deformable (KAD) alignment module. After that, we fuse these aligned clean features by a channel-wise average operation and then restore the HR image like a traditional SISR model through several residual channel attention blocks (RCAB) [63].

3.1. Problem Formulation

Given raw burst images $\{x_i\}_{i=1}^N, x_i \in \mathbb{R}^{h\times w\times 1}$ captured from the same scene and the scale factor s, our goal is to extract and fuse the complementary information between bursts and restore a high-quality image $y \in \mathbb{R}^{sh\times sw\times 3}$ with rich details. In our scenario, each input x_i is a one-channel raw image, while the output is usually a RGB image. The degradation of burst SR is model as Eq. (1). We assume that the degradation adopts anisotropic Gaussian kernels and the noise η_i is independent to the LR image x_i. The burst blind SR problem can be tackled by solving the following maximum a posteriori (MAP) problem:

$$\arg\min_{k,y} \sum_{i=1}^{N} ||\boldsymbol{x}_i - (\boldsymbol{k}_i \otimes \mathcal{T}_i \boldsymbol{y})_{\downarrow_s}||_2^2 + \phi(\boldsymbol{y}) + \psi(\boldsymbol{k}_i), \quad (2)$$

where $\phi(\boldsymbol{y})$ and $\psi(\boldsymbol{k})$ are the parameterized prior regularizers. Since all of the kernels of multiple frames are unknown variables, the overall problem is extremely difficult and challenging. Inspired by recent success of single image blind SR [18, 38], we decompose this problem into two sequential steps:

$$\begin{cases} \boldsymbol{k}_i = \mathcal{E}(\boldsymbol{x}_i; \theta_e) \\ \boldsymbol{y} = \mathcal{R}(\{\boldsymbol{x}_i, \boldsymbol{k}_i\}_{i=1}^N; \theta_r), \end{cases} \quad (3)$$

where $\mathcal{E}(\cdot)$ denotes the kernel estimator that predicts kernels for each frame of the raw bursts and $\mathcal{R}(\cdot)$ denotes the restorer that restores HR image based on LR frames and the estimated kernels. θ_e and θ_r are the parameters of the estimator and restorer, respectively.

3.2. Kernel Estimator

In order to obtain the degradation kernel and help the SR model produce visual pleasant images, we introduce an estimator $\mathcal{E}$ to predict blur kernels for all frames. The network

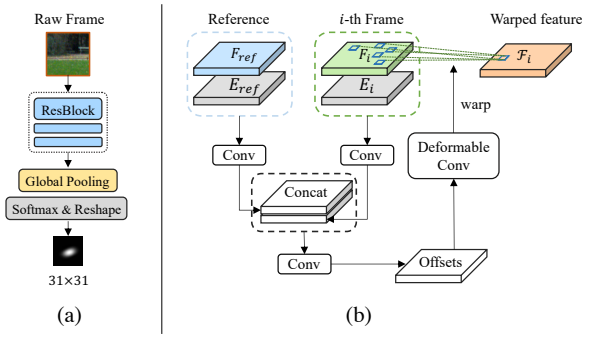


Figure 3: (a) Network architecture of the kernel estimator. (b) Kernel-aware deformable (KAD) alignment module.

architecture of the estimator is illustrated in Fig. 3(a). It consists of three simple steps: feature extraction, global average pooling, and reshape operation. Note that as a widely used kernel prior, we use the softmax function in the last layer so that the kernel could be sum to one. Moreover, we use ground truth kernels as strong supervision to optimize the estimator network. The objective is to minimize the $\mathcal{L}_1$ loss between the estimated kernel and the ground truth as

$$\theta_e = \arg\min_{\theta_e} \sum_{i=1}^{N} ||\boldsymbol{k}_i - \mathcal{E}(\boldsymbol{x}_i; \theta_e)||_1. \tag{4}$$

In practice, the body of the estimator consists 3 residual blocks. The overall network is simple, lightweight yet efficient. Since the estimated kernels are sent to the restorer to help to super-resolve images. We can jointly optimize Eq. (4) with the restorer model together to construct an end-to-end blind burst SR training scheme.

3.3. Adaptive Kernel-Aware Block (AKAB)

Most blind SR methods [61, 18, 38] tend to stretch the kernel embedding to a full-sized degradation map E_i and concatenate it with deep features to extract useful information to predict SR image. However, such a strategy is usually inefficient and computationally costly [21]. In this work, we propose the adaptive kernel-aware block (AKAB) that can utilize low-dimensional embedding and statistical information (such as the mean value of each feature) to extract deep informative features. As illustrated in the center of Fig. 2, the LR feature is firstly sent to two convolution layers, and then squeezed to one-dimensional embedding by a global average pooling layer. After that, the feature embedding $e_{\boldsymbol{x}_i}$ is concatenated with the corresponding kernel embedding $e_{\boldsymbol{k}_i}$ to perform a residual affine attention mechanism which is defined as:

$$\boldsymbol{x}_i^{out} = \gamma(e_{\boldsymbol{x}_i}, e_{\boldsymbol{k}_i}) \odot \boldsymbol{x}_i + \beta(e_{\boldsymbol{x}_i}, e_{\boldsymbol{k}_i}) + \boldsymbol{x}_i, \tag{5}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ denote the scaling and shifting functions, both of which consist of two linear layers as

$$\gamma(e_{\boldsymbol{x}_i}, e_{\boldsymbol{k}_i}) = g(w_2\sigma(w_1\mathcal{C}(e_{\boldsymbol{x}_i}, e_{\boldsymbol{k}_i})), \tag{6}$$

$$\beta(e_{\boldsymbol{x}_i}, e_{\boldsymbol{k}_i}) = v_2\sigma(v_1\mathcal{C}(e_{\boldsymbol{x}_i}, e_{\boldsymbol{k}_i})), \tag{7}$$

where $\mathcal{C}(\cdot)$ represents the concatenating operation across the channel dimension. w_1, w_2 and v_1, v_2 denote the linear layers for γ and β, respectively. σ represents a non-linear activation (e.g., ReLU), and g denotes the sigmoid function. Moreover, the shifting network β focuses on extracting and aggregating channels' information to enrich the deep features to boost the performance of feature alignment.

We compose several AKABs as a powerful feature extractor so as to obtain cleaner features that implicitly embed the degradation information of kernels. Then we could align different frames in the feature level by the following kernel-aware alignment module.

3.4. Pyramid Kernel-Aware Deformable Alignment

Deformable convolution network (DCN) [26, 10, 65] has demonstrated its effectiveness of aligning features from multiple frames [52, 7, 54, 8]. However, training the primitive DCN in the multi-degradation scenario is difficult since the different features of the same scene may have different manifestations. Thus we introduce the degradation information into the alignment process, as a kernel-aware deformable convolution (KAD) module, which could help the DCN to learn accurate offsets without being affected by various degradations.

KAD module. The overview of the KAD alignment module is shown in Fig. 3 (b). Specifically, we first simply stretch the reduced kernel $e_{k_i} \in \mathbb{R}^t$ to degradation map $E_i \in \mathbb{R}^{t\times H\times W}$. Then given the reference feature F_{ref} and the input feature F_i from the i-th LR frame, we concatenate these features with their corresponding kernel embedding maps E_{ref} and E_i to predict the deformable offsets $\Delta\boldsymbol{f}_i$ as

$$\Delta\boldsymbol{f}_i = \mathcal{O}(F_{ref}, E_{ref}, F_i, E_i), \tag{8}$$

where $\mathcal{O}$ is the offset predictor. Then we can get the aligned feature by warping F_i with $\Delta\boldsymbol{f}_i$ [65]:

$$\mathcal{F}_i = \text{warp}(F_i, \Delta\boldsymbol{f}_i). \tag{9}$$

Pyramid KAD. To address large camera motions, we further propose a pyramid alignment structure on the top of the KAD module. Specifically, as shown in Fig. 4, we first downsample all features and kernel maps with convolution layers with strides 1, 2 and 4 to get 3 different levels of pyramid. Then we perform alignment for N burst frames on each pyramid level based on the KAD. And the aligned pyramid features are scaled to be the same sizes and fused as the final aligned features. By doing so, we can effectively

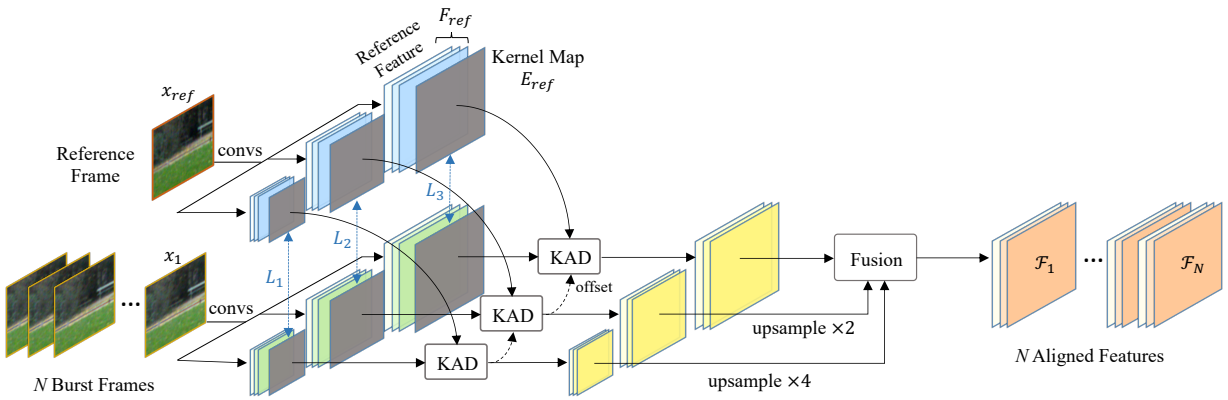


Figure 4: Pyramid kernel-award deformable alignment module. Starting from level L_1, we align all features with reference using KAD, and the predicted offset are sent to the next level to be concatenated with features to improve the offset estimation. Then each level's output are upsampled to have the same sizes so we can fuse them to obtain the aligned output features.

aggregate multi-scale information from multiple images to reconstruct SR with rich details. Empirically, we choose the first frame as the reference.

3.5. Fusion and Reconstruction

Once the features are all aligned, we can combine information across the individual frames to generate a merged feature with rich details. Unlike previous works that use attention-based weights [4] or recursive operation [12] in the fusion, we adopt an effective channel-wise average fusion strategy as shown in Fig. 2. There are two main advantages of averaging features: firstly, the operation is fast and allows us to work with the arbitrary number of frames in both training and inference. Secondly, since the inputs are noisy, averaging all frames can reduce additional noise as a traditional denoiser. Based on the fused feature, we can reconstruct the results with advanced SR networks. Practically, we employ residual channel attention blocks (RCABs) [63] as the reconstruction body. The objective function of the SR reconstruction network is defined via $\mathcal{L}_1$ loss as:

$$\theta_r = \arg\min_{\theta_r} ||\boldsymbol{y} - SR(\{\boldsymbol{x}_i\}_{i=1}^N; \theta_r)||_1. \quad (10)$$

To avoid the phenomenon of kernel mismatching [61, 18], we jointly optimize the estimator and the reconstruction module in an end-to-end manner.

4. Experiments

4.1. Datasets and Implementations

Synthetic dataset. Our method is trained on Zurich RAW to the RGB dataset [24] which consists of 46,839 HR images. For the synthetic setting, we focus on the anisotropic Gaussian kernels. Following [2], we fix the blur kernel size to 31, The kernel width of both axes are uniformly sampled in the range $[0.6, 5]$. And we also rotate the kernel by an angle uniformly distributed in $[-\pi, \pi]$. The RAW burst images are synthesized by randomly translating and rotating a high-quality sRGB image, and blurring and downsampling it with kernels generated from the above procedure as Eq. (1). In the RAW space, we add noises draw from Poisson-Gaussian distribution with sigma 0.26. Then we convert the low-quality images to RAW format using an inverse camera pipeline [6]. The test sets are generated by applying anisotropic Gaussian kernels on 1204 HR images of the validation set in [24] with different kernel width ranges of $[0.0, 1.6]$, $[1.6, 3.2]$ and $[3.2, 4.8]$. We assign different random seed values to ensure that different blur kernels are selected for different images. PSNR, SSIM [56], and the learned perceptual score LPIPS [62] are used as the evaluation metrics on synthetic datasets.

Real-world dataset. For real-world image evaluation, we use the BurstSR dataset [4] which contains pairs of real-world burst images and corresponding ground truth HR images captured by a handheld smartphone camera and DSLR camera, respectively. Each burst in this dataset contains 14 raw images and is cropped to 160×160. We perform super-resolution by a scale factor of 4 in all experiments. Note that the ground truth images are not well aligned with RAW inputs, thus we adopt aligned PSNR, SSIM, and LPIPS as the evaluation metrics as in [4].

Training details. We train the proposed KBNet on aforementioned synthetic datasets for 300 epochs. And as a common practice, we fine-tune the trained model on real-world dataset for 40 epochs. During training, the burst size is fixed to $N = 8$ and the batch size is 16. We use Adam [30] optimizer with β_1=0.9, β_2=0.999 and ϵ=10^{-8}. The learning rate is initialized as 0.0002 and then decreases to half every

Method	AKAB	RCAB	KAD	Pyramid KAD	Synthetic			Real-World		
					PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
KBNet-A	✗	✗	✗	✗	35.24	0.8927	0.1751	47.12	0.9797	0.0320
KBNet-B	✓	✗	✗	✗	36.53	0.9067	0.1422	47.61	0.9818	0.0302
KBNet-C	✓	✓	✗	✗	36.74	0.9142	0.1383	47.68	0.9821	0.0298
KBNet-D	✓	✓	✓	✗	36.87	0.9185	0.1290	47.87	0.9830	0.0278
KBNet-E	✓	✓	✓	✓	37.29	0.9219	0.1203	48.27	0.9856	0.0248

Table 1: Ablation study on our main components. The baseline is a multi-frame SR network that adopts normal deformable convolution to align frames and reconstructs SR results by several residual blocks.

100 epochs. Our models are implemented by the PyTorch framework with 2 Titan Xp GPUs.

4.2. Ablation Study

In this section, We conduct ablation study to analyze the impact of the main components of the proposed framework: adaptive kernel-aware block (**AKAB**), kernel-aware deformable convolution (**KAD**) and **Pyramid KAD**. In addition, we also give an attention to the **RCAB** in reconstruction. To conveniently illustrate the superiority of each module, we implement a baseline model (KBNet-A) which only contains a restorer that adopts normal DCN as the alignment module and uses residual blocks for both feature extraction and HR image reconstruction. All methods are evaluated on both the synthetic data and the real data.

The comparison among the baseline and our methods with different modules (KBNet-B through KBNet-E) are reported in Table 1, from which we have the following observations. First, the AKAB module plays a vital role for extracting useful features by considering multiple degradation information. Compared with the baseline, the model with AKAB improves the results about 1.3+ dB on the synthetic dataset and 0.4 dB on the real dataset. Second, the degradation and multi-scale information is also essential in the alignment. By utilizing the multi-scale features and kernel maps, the Pyramid KAD could achieve an impressive performance even the burst frames are noisy and blurry. Third, although the RCAB performs well on synthetic dataset, the improvement is incremental when finetune it on the real world dataset, which further demonstrates that the AKAB and the Pyramid KAD are key contributions of our work.

We also retrain an EDVR [54] model and our KBNet with the Pyramid, Cascading, and Deformable (PCD) alignment module [54] on the datasets. The results are shown in Table 2. With the proposed pyramid KAD, our model can perform alignment with degradation information in each pyramid level, which leads to important improvement.

Method	Synthetic			Real-world		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS ↓
EDVR [54]	36.34	0.906	0.138	47.48	0.982	0.031
KBNet+PCD	37.02	0.920	0.124	48.14	0.984	0.026
KBNet+Pyramid KAD	**37.29**	**0.922**	**0.120**	**48.27**	**0.986**	**0.025**

Table 2: Comparison of our method with EDVR.

4.3. Comparisons with State-of-the-Art Methods

We compare KBNet with other state-of-the-art learning-based raw bursts SR methods, such as DBSR [4], EBSR [40], and DeepREP [5]. Both DBSR and DeepREP are proposed by Bhat et al. [4, 5]. The former uses a flow-based alignment network with an attention-based fusion to handle the raw burst inputs. The latter employs a deep reparametrization of the MAP to solve image restoration problems. And EBSR is the winner solution of the NTIRE21 Burst Super-Resolution Challenge [3]. All of these methods are implemented from their official code repositories and re-trained with our multiple degradation setting following Sec. 4.1. We also finetune these models and evaluate them on the real-world dataset. The burst sizes of all methods are fixed to 14. In addition, we compare a single image blind SR model DAN [38] which estimates a kernel for the first burst frame and restores the HR image conditioned on that kernel and the first LR frame. Note that DAN adopts an iteratively predicting strategy and chooses to stretch the kernel embedding in reconstruction as the same as IKC [18].

Evaluation on synthetic data. Firstly, we evaluate the proposed KBNet on the synthetic dataset as introduced in Sec. 4.1. Quantitative results are shown in Table 3. Our method achieves the best results and significantly outperforms other burst super-resolution methods. As the table illustrated, all the MFSR methods outperform DAN [38] with great improvements of 3+ dB on kernels width in range $[1.6, 3.2]$ in terms of PSNR. These MFSR mthods do not explicitly utilize degradation information in the restoring and thus are powerless when facing complex degradations. In contrast, the proposed KBNet significantly outperforms other MFSR methods over all kernel width ranges. The

Method	$\sigma = [0, 1.6]$			$\sigma = [1.6, 3.2]$			$\sigma = [3.2, 4.8]$		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DAN* [38]	33.38	0.8543	0.1712	32.69	0.8321	0.2210	31.98	0.8255	0.2688
DBSR [4]	35.52	0.9086	0.1300	35.94	0.9015	0.1625	33.34	0.8633	0.2730
EBSR [40]	35.67	0.9156	0.1149	36.39	0.9095	0.1434	33.57	0.8673	0.2669
DeepREP [5]	36.46	0.9233	0.1104	36.26	0.9082	0.1510	33.49	0.8664	0.2721
KBNet(Ours)	**37.43**	**0.9314**	**0.0967**	**37.27**	**0.9172**	**0.1237**	**35.28**	**0.8924**	**0.1941**

Table 3: Comparison of our method with existing MFSR approaches on the Synthetic test dataset, for scale factor 4. The kernel width σ is split into three ranges. '*' means it is a single image blind super-resolution method.

Figure 5: Qualitative comparison of our method with other MFSR approaches on **synthetic** dataset.

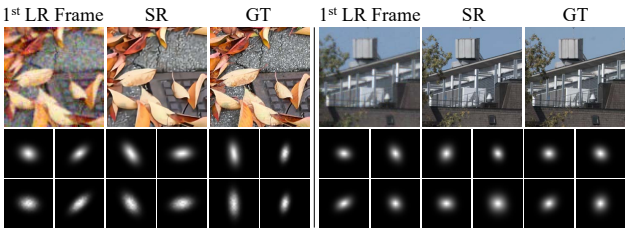
Figure 6: Qualitative comparison of our method with other MFSR approaches on real-world **BurstSR** dataset.

qualitative comparison are shown in Fig. 5. The super-resolved images produced by our KBNet are visually pleasant and have rich details, which demonstrates the superior of our method, and indicates that involving degradation information into restoration can help to obtain informative features and thus improve SR results.

Evaluation on real-world data. Now we conduct the experiment of evaluating models that are pre-trained on synthetic dataset and finetuned on the real-world dataset. Note that the ground-truth kernels of real-world images are not available, which are required by the KBNet in the kernel estimation learning process. Alteratively, we freeze the kernel estimator and only finetune the image restorer. The quantitative results are shown in Table 4. As we can see, the DeepREP [5] significantly outperforms DBSR [4] and EBSR [40], but is still inferior to the proposed KBNet. Visual comparisons on the real-world images are shown in Fig. 6. It is obviously that the results produced by the KBNet have favorable perceptual quality in edges and details, and is robust to real-world noises.

4.4. Analysis about kernels and frame numbers.

We assumes that all frames of a burst sequence are captured by a smartphone under the burst shooting mode in which different degradation kernels can be produced by hand-shaking or different shooting parameters. Visual examples of the estimated kernels are shown in Fig. 7. On the synthetic setting, most kernels can be accurately estimated by our method, which helps us to restore HR images. For real-world images, the degradation kernel should always follow the Gaussian distribution depends on the depth at each pixel and the focal length of the camera [9]. Thus our method prefers to generate some Gaussian-like kernels despite the kernel estimator is not optimized to fit the real



(a) Synthetic images (b) Real-world images

Figure 7: Examples of estimated kernels and super-resolved images. **(a)** The first 6 kernels of the corresponding synthetic image. The top row of the kernel is the estimated kernels and second row of kernel is the ground-truth kernels. **(b)** The first 12 kernels of the corresponding real image.

Method	DAN* [38]	DBSR [4]	EBSR [40]	DeepREP [5]	KBNet(Ours)
PSNR↑	46.18	47.48	47.25	48.15	**48.27**
SSIM↑	0.9777	0.9824	0.9800	0.9842	**0.9856**
LPIPS↓	0.0389	0.0326	0.0356	0.0265	**0.0248**

Table 4: Quantitative comparison of the proposed method with existing MFSR approaches on the real-world BurstSR $\times 4$ dataset. '*' means it is a single image blind SR method.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
KBNet *w/o* using kernel	47.54	0.9824	0.0299
KBNet, fixed Gaussian kernel	48.12	0.9838	0.0264
KBNet, estimated kernel	48.27	0.9856	0.0248

Table 5: Different kernel strategies on real-world dataset.

dataset. The experiment in Table 5 illustrates that the estimated kernels are also useful for restoring HR images.

Moreover, we investigate the impact of multiple frames and compare KBNet with other MFSR methods. Here we don't conduct the results of EBSR because both it's training and testing require fixing the frame number to 14. The results on PSNR are shown in Fig. 8. As the frame number increases, all MFSR methods could achieve higher performances. The proposed KBNet outperforms other methods over all frames and datasets.

4.5. Synthetic Model Transferring

The motivation of this experiment is that making paired SR datasets for real-world photography applications is really challenging since the LR images and GT are usually captured from different devices (e.g., smartphones and DSLR cameras). The mismatch of image qualities and colors would make it extremely difficult to train a SR model across modalities. In such a situation, we would like to train the model only in the multi-degradation environment and apply it to the real scenes directly, which seems like a zero-shot transferring problem.

To illustrate the idea, we provide the comparison of transferring models under bicubic degradation and blind

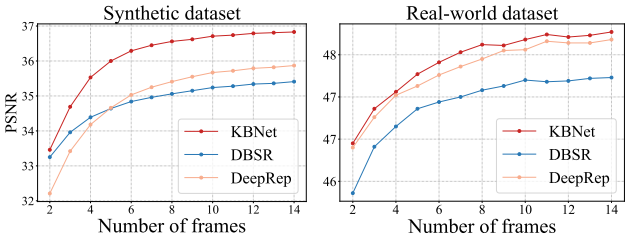


Figure 8: Comparison of the PSNR performances among different MFSR approaches on synthetic and real-world datasets with different number of frames.

Method	Bicubic Degradation			Blind Degradation		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS ↓
DBSR [4]	44.64	0.967	0.079	45.18(+0.53)	0.974(+0.007)	0.048(-0.031)
EBSR [40]	44.32	0.963	0.082	44.96(+0.64)	0.970(+0.007)	0.048(-0.034)
DeepREP [5]	44.80	0.968	0.080	45.39(+0.59)	0.974(+0.006)	0.045(-0.035)
KBNet(Ours)	-	-	-	**45.68**	**0.979**	**0.042**

Table 6: Quantitative results of transferring models from different synthetic environments to real-world images. The improvements on each metric are marked in colors.

degradation in Table 6. Under the multi degradation environment, all MFSR methods can achieve higher performances compared with their bicubicly trained models. And the proposed KBNet remarkablely outperforms other methods and can produce visually pleasant results on real images, which indicate that even if the KBNet is trained on synthesized image pairs, it still has the ability to generalize to images in real applications.

5. Conclusion

In this paper, we present a new framework, named KBNet, to handle the multi-frame super-resolution problem with considering multiple complicated degradations. The proposed burst blind super-resolution task is highly related to real-world applications. To address it, we introduce a kernel-based multi-frame restoration network that includes an adaptive kernel-aware block (AKAB) and a pyramid kernel-aware deformable (Pyramid KAD) alignment module. The blur kernels are first estimated by an estimator and then fed to the LR feature extraction module as well as the feature alignment module to generate a super-resolved clear image. The proposed method can be end-to-end trained on the synthetic dataset and evaluated on both synthetic and real-world images. Experiment results demonstrate that our method can achieve a good performance on various degradations and is beneficial to real-world device applications.

Acknowledgments Thanks to Yanqing Lian, Youqin Zheng, Wenjing Lian, Jingyuan Lian, and Ziwei Luo for their selfless support and help in this work. We also thank the anonymous reviewers for helping improve our work.

References

[1] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*, pages 571–582. Springer, 1996.

[2] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021.

[5] 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.

[6] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019.

[7] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 4:3, 2020.

[8] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021.

[9] Subhasis Chaudhuri and Ambasamudram N Rajagopalan. *Depth from defocus: a real aperture imaging approach*. Springer Science & Business Media, 1999.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019.

[12] 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.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[15] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. *arXiv preprint arXiv:2110.03680*, 2021.

[16] Netalee Efrat, Daniel Glasner, Alexander Apartsin, Boaz Nadler, and Anat Levin. Accurate blur models vs. image priors in single image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2832–2839, 2013.

[17] 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.

[18] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Russell C Hardie, Kenneth J Barnard, John G Bognar, Ernest E Armstrong, and Edward A Watson. High-resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. *Optical Engineering*, 37(1):247–260, 1998.

[20] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019.

[21] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao. Learning the non-differentiable optimization for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2093–2102, 2021.

[22] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018.

[23] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1428–1437, 2020.

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

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

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[28] Michal Kawulok, Pawel Benecki, Krzysztof Hrynczenko, Daniel Kostrzewa, Szymon Piechaczek, Jakub Nalepa, and Bogdan Smolka. Deep learning for fast super-resolution reconstruction from multiple images. In *Real-Time Image Processing and Deep Learning 2019*, volume 10996, page 109960B. International Society for Optics and Photonics, 2019.

[29] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[32] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2370–2379, 2021.

[33] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[34] Wenyi Lian and Wenjing Lian. Sliding window recurrent network for efficient video super-resolution. *arXiv preprint arXiv:2208.11608*, 2022.

[35] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[36] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020.

[37] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17642–17652, 2022.

[38] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[39] 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.

[40] 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*, pages 471–478, 2021.

[41] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952, 2013.

[42] 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.

[43] Rao Muhammad Umer and Christian Micheloni. Rbsricnn: Raw burst super-resolution through iterative convolutional neural network. In *Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS)*, December 2021.

[44] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014.

[45] 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.

[46] Shmuel Peleg, Danny Keren, and Limor Schweitzer. Improving image resolution using subpixel motion. *Pattern recognition letters*, 5(3):223–226, 1987.

[47] Richard R Schultz and Robert L Stevenson. Extraction of high-resolution frames from video sequences. *IEEE transactions on image processing*, 5(6):996–1011, 1996.

[48] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018.

[49] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020.

[50] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Robust kernel regression for restoration and reconstruction of images from sparse noisy data. In *2006 International Conference on Image Processing*, pages 1257–1260. IEEE, 2006.

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

[52] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020.

[53] R Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984.

[54] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[55] 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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[57] 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.

[58] Yu-Syuan Xu, Shou-Yao Roy Tseng, Yu Tseng, Hsien-Kai Kuo, and Yi-Min Tsai. Unified dynamic convolutional network for super-resolution with variational degradations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12496–12505, 2020.

[59] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European conference on computer vision*, pages 372–386. Springer, 2014.

[60] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3217–3226, 2020.

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

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[63] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[64] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[65] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.