Spectrum-to-Kernel Translation for Accurate Blind Image Super-Resolution

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep-learning based Super-Resolution (SR) methods have exhibited promising performance under non-blind setting where blur kernel is known; however, blur kernels of Low-Resolution (LR) images in different practical applications are usually unknown. It may lead to a significant performance drop when degradation process of training images deviates from that of real images. In this paper, we propose a novel blind SR framework to super-resolve LR images degraded by arbitrary blur kernel with accurate kernel estimation in frequency domain. To our best knowledge, this is the first deep learning method which conducts blur kernel estimation in frequency domain. Specifically, we first demonstrate that feature representation in frequency domain is more conducive for blur kernel reconstruction than in spatial domain. Next, we present a Spectrum-to-Kernel (S2K) network to estimate general blur kernels in diverse forms. We use a conditional GAN (CGAN) combined with SR-oriented optimization target to learn the end-to-end translation from degraded images' spectra to unknown kernels. Extensive experiments on both synthetic and real-world images demonstrate that our proposed method sufficiently reduces blur kernel estimation error, thus enables the off-the-shelf non-blind SR methods to work under blind setting effectively, and achieves superior performance over state-of-the-art blind SR methods, averagely by 1.39dB, 0.48dB (Gaussian kernels) and 6.15dB, 4.57dB (motion kernels) for scales $2\times$ and $4\times$ respectively.

1 Introduction

Single Image Super-Resolution (SISR) is a low-level visual task to recover High-Resolution (HR) image from its degraded Low-Resolution (LR) counterpart. Deep learning methods have significantly promoted SR research and achieved remarkable results on benchmarks [15, 26, 27, 13, 29, 35, 16]. Most of these works are based on the degradation model assuming that the LR image is down-sampled with blur kernel (*e.g.*, bicubic kernel) and additional noise from its HR source. Recent studies [6, 11] show that the SR model is sensitive to the degradation of LR image. The mismatch of Gaussian kernels may bring artifacts either over-sharpening or over-smoothing, and for motion kernels, it may bring unpleasant patch offset and loss of important edges. Blind SR is to super-resolve a single LR image with unknown kernel. It has attracted more and more attention due to its close relationship with the real scene. However, compared with the mature non-blind SR researches, existing blind SR methods are far from meeting the needs of various degradation processes in general scenes.

Some representative works have been proposed on this challenging problem. Bell-Kligler *et al.* estimate the kernel with internal-Generative Adversarial Network (GAN) using patch recurrence on the image prior and a deep linear network, whose performance is limited because the relatively weak prior can not capture the kernel-related features directly [2]. Gu *et al.* propose a kernel prediction network to estimate the kernel from the mismatch between kernel and SR result in an iterative way [6].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Even with well-designed modules and complex training process, the iterative form can not utilize the maximum performance of both estimate and SR module during the testing process, and the barrel effect is inevitable. Ji *et al.* use the correlation between the blur degree of natural image and its up-and-down sampling results in frequency domain to estimate the kernel in an unsupervised way [11]. However, the training requires a certain amount of training data with the same degradation, and its application in general scenes is limited for failing to capture blur effects from different directions.

In this paper, we capture the convolution degradation feature in the local distribution of LR from a novel perspective. Based on convolution theorem and sparsity analysis, we demonstrate the Fourier frequency spectrum of degraded LR image provides powerful shape structure of the degradation kernel in frequency domain, which can be utilised to obtain the unknown kernel more accurately and robustly. Based on the above analysis, we design a Spectrum-to-Kernel (S2K) translation network to directly obtain the kernel interpolation result from degraded LR's frequency spectrum. As far as we know, this is the first work where Fourier transform spectrum is used to predict the up-sampled kernels. The experimental results demonstrate that our method can obtain more accurate estimation of the target kernel and has a higher universality capacity for diverse kernels. Moreover, compared with existing blind-SR methods, our proposed kernel estimation method achieves the best results in aspects of quality evaluation and visual perception metrics on both synthetic and real images.
In general, our contribution is three-fold:

- We theoretically demonstrate that for common sparse kernel, it is more conducive to reconstruct it in frequency domain instead of the commonly used spatial domain.
- We analyze the impacts of different optimization objectives for blind-SR performance and design a novel network S2K to estimate kernel from Fourier amplitude spectrum.
- When plugged into existing non-blind SR models, our proposed pipeline achieves the best performance on both synthetic and real-world images against existing blind SR methods.

2 Related Work

Non-blind SR Under non-blind setting, HR image $\mathbf{I}_{HR}$ is recovered from a given LR image $\mathbf{I}_{LR}$ and the known kernel. Many successful pieces of research have been proposed where bicubic kernel is used to construct paired training datasets, so they belong to non-blind setting in essence. We divide those kernel-aware SR methods into two categories. (1) Multiple Degradation SR (MDSR) based methods take degraded image $\mathbf{I}_D$ and blur kernel $\mathbf{k}$ as input to get target HR image such as explicitly concatenate blurred image or fuse the intermediate feature with stretched k in the SR model [33, 31, 6, 31]. (2) Kernel Modeling SR (KMSR) based methods incorporating the blur kernel in the model's training parameters by constructing the paired kernel modeling dataset for the known kernel to get a unified SR model for test-specific degradation restoration [25, 36, 10]. Inconsistency between degraded image and blur kernel leads to significant performance deterioration in test phase [33]. We propose a spectrum-to-kernel framework S2K which provides accurate kernel estimation that helps both MDSR and KMSR based methods work well under blind settings.

Blind SR Under blind setting, $\mathbf{I}_{HR}$ is recovered from $\mathbf{I}_{LR}$ degraded with unknown kernel. Since it is more closely related to real application scenarios, blind-SR has attracted research attention for a long time [20]. To address this issue, mainstream methods proposed recently can be divided into two categories: (1) Methods transforming the unknown degraded image into target $\mathbf{I}_{HR}$ bicubic down-sampled image; and (2) non-blind SR methods with kernel estimation from spatial input $\mathbf{I}_{LR}$.

In the former category, Hussein *et al.* modify the unknown degraded LR image to match the one obtained by bicubic kernel [9]. However, it is much difficult to recover the high frequency part lost in the LR image and get the corresponding target as ideally bicubic down-sampled one for the ill-posedness. Maeda *et al.* train a correction network to get the corresponding clean LR and construct pseudo pair to train the SR model [18]. However, the recovery of high frequency components does not make full use of the SR model, leading to the lack of performance and stability. In the latter category, Bell-Kligler *et al.* utilize the recurrence of patches across scales in the natural image to generate the same degraded down-sampled image as the training input [2]. However, the estimation accuracy is limited even with many handcrafted constrains due to the weak natural prior. Gu *et al.* use a kernel prediction network with a correction network taking the current predict kernel and SR result of the pre-trained SR model as input to correct the kernel prediction iteratively [6]. Huang *et al.* adopt an alternating optimization algorithm with two modules restorer and estimator [8]. However,

Table 1: Comparison with the latest kernel estimaion methods. Our method transfers the degraded image to the blur kernel in *frequency domain*, then obtains the up-sampled blur kernel through cross-domain translation. Compared with recent methods based on image priors or reduced form learning in spatial domain, we get more stable and accurate estimation for *arbitrary kernels* directly.

	Input domain	Estimate target	Task related	Optimization objective	Kernel setting
IKC (CVPR 2019)	Spatial	PCA reduced kernel	Dependent (separated)	SR performance	Gaussian kernel
DAN (NeurIPS 2020)	Spatial	PCA reduced kernel	Dependent (end-to-end)	SR performance	Gaussian kernel
FCA (AAAI 2021)	Spatial	Kernel parameter	Independent	Frequency consistency	Gaussian kernel
FKP (CVPR 2021)	Spatial	Invertible latent distribution	Independent	Normalizing flow based kernel prior	Gaussian kernel
S2K (Ours)	**Fourier frequency**	Up-sampled kernel	Independent	SR-oriented kernel distence	**Arbitrary blur kernel**

the estimated module's performance is affected by the SR module and cannot be optimal at the same time. Ji *et al.* obtain the approximation of the blur kernel by utilizing the frequency domain consistency to estimate kernel in an unsupervised way [11]. However, the handle-able kernel form limits its applicability. Liang *et al.*propose to learn an invertible mapping between the anisotropic Gaussian kernel distribution and a tractable latent distribution as kernel modeling module [14]. We list and compare several recent methods with our proposed methods in Table 1.

Different from these previous spatial input based kernel estimation methods, we demonstrate *Fourier amplitude spectrum* can provide more effective shape structure of kernel, which contributes to accurate kernel estimation in frequency domain. Further, utilizing the shape structure relationship of the sparse kernel between the frequency domain and the spatial domain, our method is capable to estimate diverse blur kernels in an end-to-end manner with implicit cross-domain translation. We obtain more accurate estimation of the unknown blur kernel directly and independently, so as to make the most of existing powerful non-blind SR methods and achieve the best performance on both synthetic and real-world images against existing blind-SR methods.

3 Method

In this section, we first give the description of the problem, then theoretically analyse the advantage of blur kernel estimation from frequency domain and the feasibility of end-to-end translation learning. Then we introduce the details of the proposed S2K network. Finally, the blind SR pipeline combined with the proposed kernel estimation module is presented.

3.1 Problem Formulation

Same as [34, 11], we assume the image degradation process from I_{HR} to I_{LR} is as follows:

$$\mathbf{I}_D = (\mathbf{I}_C{\downarrow_s}) \otimes \mathbf{k} + \mathbf{n}, \tag{1}$$

where $\mathbf{I}_C$ is a sharp and clean image, $\mathbf{k}$ denotes degradation kernel, and $\mathbf{n}$ is additional noise[1]. $\downarrow_s$ and $\otimes$ are down-sampling and convolution operations. Under non-blind setting, both $\mathbf{k}$ and $\mathbf{n}$ are known, denoting the pre-trained SR model as $\mathbf{R}$, the upper bound of SR result $\mathbf{I}^*_{SR}$ can be expressed as:

$$\mathbf{I}^*_{SR} = \mathbf{R}(\mathbf{I}_D \mid \mathbf{k}, \mathbf{n}). \tag{2}$$

Under blind setting, the degradation prior needs to be obtained from $\mathbf{I}_D$. In this paper, we mainly focus on the estimation of degradation kernel from the degraded input $\mathbf{I}_D$, because the extra denoising method can be easily plugged in from the previous noise estimation methods [7, 10]. For the blind-SR performance-oriented kernel estimation, theoretically, the optimal $\mathbf{k}^*$ is defined as follows:

$$\mathbf{k}^* = \arg\min_{\mathbf{k}_{est}} \|\mathbf{R}(\mathbf{I}_D \mid \mathbf{k}_{est}, \mathbf{n}_{est}) - \mathbf{I}^*_{SR}\|. \tag{3}$$

Accurate kernel estimation is the key to give full play to the performance of existing non-blind methods under blind setting. However, it is difficult for its ill-posed property. Meanwhile, the non-blind SR models are very sensitive to the consistency of input $\{\mathbf{I}_D, \mathbf{k}_{est}\}$ pairs and subtle degradation

[1]Above variables are all matrices, and the size of blur kernel is far less than than that of the image.

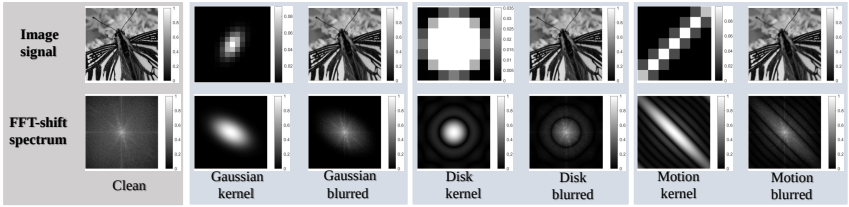


Figure 1: The relationship between the clean image, convolution kernels and convolution results in frequency spectrum. The spectrum of degraded image indicates shape structure of kernel's spectrum, while the degraded images show minor visual differences.

inconsistencies may cause obvious artifacts [6]. We explored the influence of the kernel estimation optimization target on the performance of subsequent non-blind SR, so as to ensure the consistency of $\mathbf{k}_{est}$ and $\mathbf{I}_D$, and get the result close to upper bound $\mathbf{I}^*_{SR}$.

3.2 Spectrum-to-Kernel

Different from previous kernel estimation methods in spatial domain, we are the first to conduct kernel estimation in frequency domain in an end-to-end manner making the best of kernel's shape structure. Shape structure is conducive to the reconstruction task learning: Chen *et al.* obtain better face SR results with shape prior [5], Ma *et al.* maintain the gradients structure, which proves to be beneficial to the SR result [17]; some researchers utilise structure information to improve the performance of image inpainting [24, 21]. We find the kernel representation obtained from Fourier spectrum amplitude map can provide more favorable shape structure information compared with other domains. We below prove that estimation from the Fourier frequency domain is more conducive to the reconstruction of the blur kernel than spatial domain.

Advantages of the Frequency Domain In frequency domain (Fourier frequency domain unless specified), denoting clean images, degraded images and degradation kernels as $\mathbf{S}$, $\mathbf{G}$ and $\mathbf{F}$ respectively. According to the convolution theorem, we have:

$$\mathbf{G} = \mathbf{F} \odot \mathbf{S}. \tag{4}$$

We adopt $\phi(\mathbf{X}, \mathbf{Y})$ to measure the similarity of the shape structure of two matrix variables $\mathbf{X}, \mathbf{Y}$ [12, 32] as follows:

$$\phi(\mathbf{X}, \mathbf{Y}) = \|\delta_\tau(\mathbf{X} - \mathbf{Y})\|_0, \tag{5}$$

where $\delta_\tau(\cdot)$ is a truncated function to ignore the extremely small divergence in each elements and $\|\cdot\|_0$ denotes ℓ_0-norm operator. When the absolute value of element $\mathbf{x}$ in $(\mathbf{X} - \mathbf{Y})$ is less than τ (positive but very small), it returns 0, otherwise returns the absolute value of $\mathbf{x}$. Under the assumption that common kernels are sparse in both the frequency domain and the spatial domain, we compare the shape structure similarity of the input and output for kernel estimation in the frequency domain and the spatial domain respectively. In frequency domain, we have:

$$\begin{aligned}
\phi(\mathbf{G}, \mathbf{F}) &= \sum_{\mathbf{F}_{ij} < \tau} \mathbf{sign}(\delta_\tau(\mathbf{F}_{ij}\mathbf{S}_{ij} - \mathbf{F}_{ij})) + \sum_{\mathbf{F}_{ij} \geq \tau} \mathbf{sign}(\delta_\tau(\mathbf{F}_{ij}\mathbf{S}_{ij} - \mathbf{F}_{ij})) \\
&\leq \sum_{\mathbf{F}_{ij} < \tau} \mathbf{sign}(\delta_\tau(\mathbf{F}_{ij}(\mathbf{S}_{ij} - 1))) + \sum_{\mathbf{F}_{ij} \geq \tau} \mathbf{sign}(\delta_\tau(\mathbf{F}_{ij})) \\
&= 0 + \sum_{\mathbf{F}_{ij} \geq \tau} \mathbf{sign}(\delta_\tau(\mathbf{F}_{ij})) \\
&= \|\delta_\tau(\mathbf{F})\|_0.
\end{aligned} \tag{6}$$

Here, if x is positive, $\mathbf{sign}(\mathbf{x}) > 0$, else if x is negative, $\mathbf{sign}(\mathbf{x}) < 0$, else $\mathbf{sign}(\mathbf{x}) = 0$, and i, j denote position coordinates in spectrum. Eq. 6 gives the upper bound of the distance of the shape structure from the degraded input to the output kernel in the frequency domain. While in spatial domain, denoting the zero-padded degradation kernel with $\mathbf{k}_p$ which has the same size of the

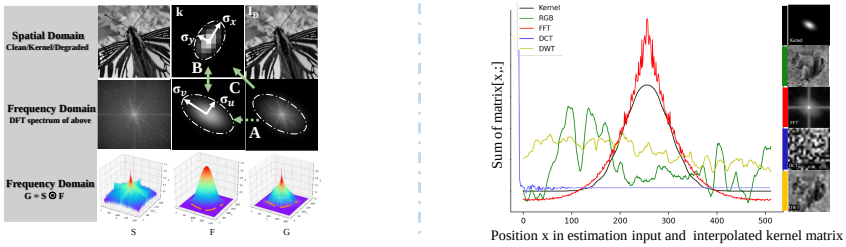


Figure 2: **Lelf**: ***A.*** Due to convolution theorem and the sparsity of the kernel, the shape structure from $\mathbf{G}$ is conducive to reconstruction of kernel spectrum $\mathbf{F}$ in frequency domain. ***B.*** Shape correlation of the Gaussian kernel between Fourier domain and spatial domain: variances are inversely proportional respectively. ***C.*** Based on ***A*** and ***B***, we find it feasible to end-to-end estimate the spatial kernel $\mathbf{k}$ from the degraded image's spectrum $\mathbf{G}$ with implicit cross-domain translation. **Right:** Shape structures of various input domains with interpolated kernel and DFT spectrum is the closest to target kernel.

degraded image $\mathbf{I}_D$, we have:

$$\begin{aligned}\phi(\mathbf{k}_p, \mathbf{I}_D) &= \sum_{[\mathbf{k}_p]_{ij}<\tau} \mathbf{sign}(\delta_\tau([\mathbf{k}_p]_{ij} - [\mathbf{I}_D]_{ij})) + \sum_{[\mathbf{k}s_p]_{ij}>\tau} \mathbf{sign}(\delta_\tau([\mathbf{k}_p]_{ij} - [\mathbf{I}_D]_{ij})) \\ &\geq \sum_{[\mathbf{k}_p]_{ij}<\tau} \mathbf{sign}(\delta_\tau([\mathbf{k}_p]_{ij} - [\mathbf{I}_D]_{ij})) \geq \sum_{\substack{[\mathbf{k}_p]_{ij}<\tau, \\ [\mathbf{I}_D]_{ij}>2\tau}} \mathbf{sign}(\delta_\tau([\mathbf{k}_p]_{ij} - [\mathbf{I}_D]_{ij})) \\ &\geq \|\delta_{2\tau}(\mathbf{I}_D)\|_0 - \|\delta_\tau(\mathbf{k}_p)\|_0,\end{aligned} \tag{7}$$

which gives the lower bound of the distance of the shape structure on the spatial domain. Further, since both $\mathbf{F}$ and $\mathbf{k}_p$ are sparse, $\|\delta_\tau(\mathbf{F})\|_0$ and $\|\delta_\tau(\mathbf{k}_p)\|_0$ are significantly smaller than $\|\delta_{2\tau}(\mathbf{I}_D)\|_0$ due to the non-sparsity of $\mathbf{I}_D$. In this case, $\|\delta_\tau(\mathbf{F})\|_0 \ll \|\delta_{2\tau}(\mathbf{I}_D)\|_0 - \|\delta_\tau(\mathbf{k}_p)\|_0$. Thus we have:

$$\phi(\mathbf{G}, \mathbf{F}) \ll \phi(\mathbf{k}_p, \mathbf{I}_D). \tag{8}$$

Eq. 8 clearly reveals that the divergence between degradation kernel and its corresponding degraded image in frequency domain is significantly smaller than that in spatial domain. As illustrated in Figure 1, compared with the spatial domain, $\mathbf{G}$ and $\mathbf{F}$ have a more direct and obvious shape structure correlation, which is conducive for the kernel reconstruction in the frequency domain.

After demonstrating the shape structure advantage for general kernel estimation from frequency domain over spatial domain, we further analyse the cross-domain relationship of kernel's shape structure. Taking the Gaussian kernel as the example, according to Fourier transform property of 2D Gaussian function, kernel in the frequency domain is also in Gaussian function form, and its variances are inversely proportional with that in spatial as shown in Figure 2 left: $\sigma_u = \mathbf{C}_1/\sigma_x, \sigma_v = \mathbf{C}_2/\sigma_y$, $\mathbf{C}_1, \mathbf{C}_2$ refer to two constants[2]. This inspires us that it is feasible to implicitly learn the cross-domain translation, and estimate spatial kernel from degraded image in frequency domain with amplitude only in an end-to-end manner. Experimentally, we find it still works for motion and disk kernels.

3.3 Spectrum based Kernel Estimation

Based on the analysis above, we propose a new kernel estimation method from frequency domain, which makes the most of the shape structure from spectrum amplitude to reconstruct the kernel. We further utilise kernel's shape structure connection to learn the cross-domain translation, thus estimating the unknown spatial kernel from frequency spectrum in an end-to-end manner.

Kernel Generator Different from the previous methods estimating kernel's parameter from spatial input, we obtain the target kernel in spatial from the frequency amplitude. The Generator is an encoder-decoder structure (Figure 3) and takes the 256×256 single-channel amplitude spectrum of the LR image as the input. After passing through an input layer with stride 2, the feature is sent into a U-net with mirror connection between each of the seven down-sampling convolutional layers and up-sampling transposed convolutional layers with the same feature size. Finally, we obtain the

[2]More details provided in supplementary materials.

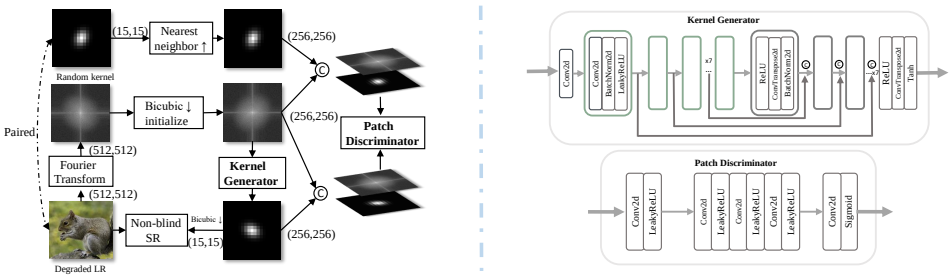


Figure 3: Pipeline of blind SR (left) with proposed S2K network architecture(right).

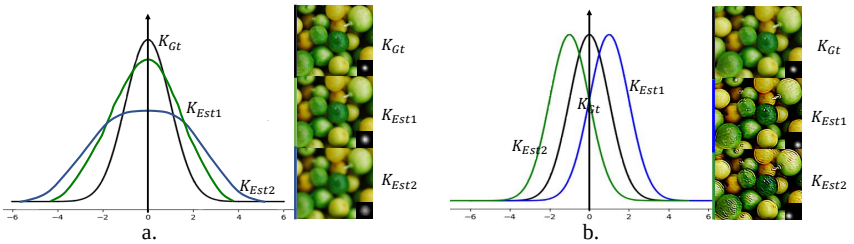


Figure 4: **Illustration of two toy examples to show the limitations of using loss components ℓ_1 (left) and ℓ_{adv} (right) alone respectively, optimization objective of kernel estimation is inconsistent with SR performance.** In case *a*, $\mathbf{K}_{Est1}$ and $\mathbf{K}_{Est2}$ share same ℓ_1 distances to $\mathbf{K}_{Gt}$, but more blurry SR result is obtained with $\mathbf{K}_{Est2}$ than $\mathbf{K}_{Est1}$. In case *b*, $\mathbf{K}_{Est1}$ and $\mathbf{K}_{Est2}$ have the same distribution distance to $\mathbf{K}_{Gt}$, but inconsistent with SR results. We find the combination of the pixel loss and the distribution loss is more beneficial to subsequent SR performance with estimated kernel.

single-channel estimation kernel map through the output layer. As discussed in Figure 4, we use distribution discrimination loss ℓ_{adv} and smoothing loss ℓ_{reg} in addition to the ℓ_1 loss to ensure that the estimated kernel close to ground truth and *favorable for SR results*. For input pair $\{\mathbf{S}_i, \mathbf{k}_i\}$, to make the generator output $G(\mathbf{S}_i)$ closer to target $\mathbf{k}_i$, the generator's loss ℓ_G consists of three parts:

$$\ell_G = \mathbb{E}_{\mathbf{S}_i,\mathbf{k}_i}[\underbrace{\lambda_1\|\mathbf{k}_i - G(\mathbf{S}_i)\|_1}_{\ell_1} + \underbrace{\lambda_2[D(\mathbf{S}_i; G(\mathbf{S}_i)) - 1]^2}_{\ell_{adv}} + \underbrace{\lambda_3\mathcal{R}(G(\mathbf{S}_i))}_{\ell_{reg}}], \tag{9}$$

where (;) denotes concatenate operation and D denotes the discriminator, $\mathcal{R}$ denotes regularization item, λ_i denotes weight of respective loss component. The first reconstruction ℓ_1 loss pushes $G(S_i)$ closer to the target kernel, the second discrimination loss ℓ_{adv} makes $G(S_i)$ satisfy the latent kernel's distribution, and the third total variation loss ℓ_{reg} regularizes the kernel with smoothing regularization.

Patch Discriminator We use a fully-convolutional patch discriminator to learn distribution of different kernel forms. We use the variant of LSGAN [19] to make the training more stable with regression in the last layer of the model. Discriminator's training loss can be written as:

$$\ell_D = \mathbb{E}_{\mathbf{S}_i,\mathbf{k}_i}[(D(\mathbf{S}_i; G(\mathbf{S}_i)))^2 + (D(\mathbf{S}_i; \mathbf{k}_i) - 1)^2]. \tag{10}$$

3.4 Pipeline of Blind-SR

Based on the proposed S2K kernel estimation module, we further propose a basic pipeline for Blind-SR as shown in Figure 3. We crop $\mathbf{I}_{LR}$ with the size of 512. When the size of the input image is not enough, we duplicate to repeat and crop from the result. Then we perform Discrete Fast Fourier Transform (DFFT) on the cropped result and center-shift the spectrum to make the low frequency part in the center. The spectrum is bicubic-down-sampled to 256 as the input of S2K, and then the estimated kernel is obtained by down-sampling S2K's output into the specified kernel size. For anisotropic Gaussian kernel, defined by continuous parameters, both MDSR and KMSR (without kernel expansion [36]) based non-blind models are applicable. For more discrete and irregular motion kernel, we use MDSR based SR model to get target I_{HR} taking I_{LR} and estimated kernel as input.

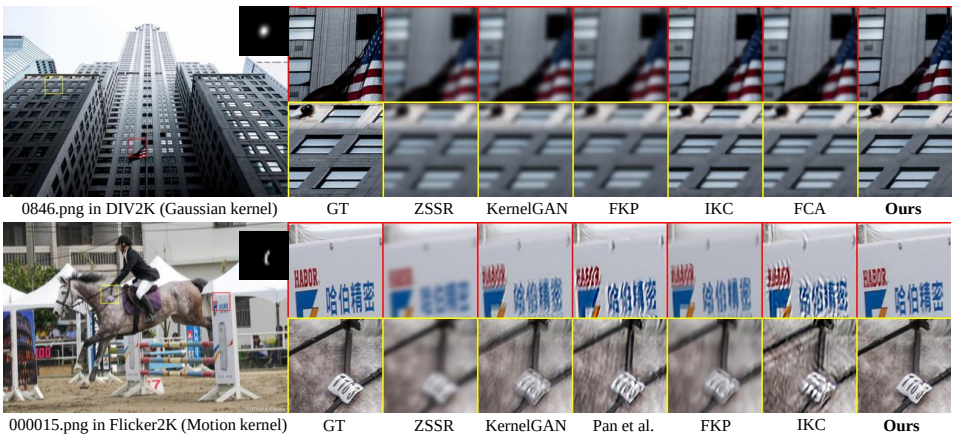


Figure 5: $2\times$ blind SR results on synthetic images with state-of-the-art methods. For *Gaussian* and *motion* blurred images, our method obtains results with *more clarity* and *less shake effects*.

4 Experiments

Experiment Setup In theory, our method can handle general sparse kernels. We select two most commonly used blur kernels in image processing *i.e.* Gaussian kernels and motion kernels, to conduct blind SR experiments. For Gaussian kernel, we set kernel size as 15 and variance between 1.0 and 3.0, rotation angle between 0 and 2π). For motion blur kernel, we randomly select from 500 kernels using code from [3]. The kernel size is 23 and exposure time is set in $[0.15, 0.3]$, anxiety is 0.005. Both Gaussian kernels and motion kernels are centralized with the sum of 1. All methods compared are trained on the *same* degraded dataset, including the kernel estimation methods (*e.g.*, KernelGAN, IKC and FCA), and those non-kernel-estimation methods (*e.g.*, RCAN and ZSSR). We use DIV2K [28] training set to train S2K and non-blind SR model. Then, we use random blurred DIV2K validation set, 100 images in Flicker2K [1] as test datasets in the synthetic experiment, and conduct $2\times$, $3\times$, $4\times$ SR comparison, and compare the accuracy of kernel estimation as well. Finally, we use our pre-trained estimation model of Gaussian kernels to perform $4\times$ SR test on real image [4] and compare with state-of-the-art methods. We set the weights of loss item in generator $\lambda_1 = 100$, $\lambda_2 = 1$, $\lambda_3 = 1$ respectively. For optimization, we use Adam with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and learning rate is 0.001.

4.1 Blind SR on Synthetic Images

We combine our proposed kernel estimation with MDSR-based SFTMD, KMSR-based RCAN, and provide comparison with existing methods in Table 2. Due to the inconsistency of training and test domain, the performance of kernel-awareless SR models (RCAN, ZSSR) suffer a significant deterioration on the unknown degraded test data. The performance of blind methods trying to correct test LR image to get clear LR image (DeblurGAN w. RCAN) is limited, because the SR degradation process includes the down-sampling process makes the recovery of high frequency components more difficult in LR image. KernelGAN and FKP obtains more accurate prediction than bicubic interpolation kernel, the inadequate estimation accuracy degrades the subsequent SR model's performance. Using the same pre-trained SFTMD model, we can obtain better results against IKC because the training of iterative kernel correction and prediction is based on pre-trained SR model, which largely limits the performance of estimation. We also compare our method with [22] designated to estimate the motion kernel with dark channel prior. Benefit from more effective spectrum input and novel process to obtain kernel directly, our method obtains better results. The requirement for data with the same degradation in training makes FCA unable to deal with the kernel estimation of the specific degraded image. Our method can estimate more general kernels more accurately without these restrictions. Visual comparison results are provided in Figure 5.

Kernel Estimation We center-crop the images with the same size. For visualization, the kernel is divided by the mean of the maximum value of the estimated results of all methods for normalization. For quantitative comparison, we divide each result by the sum (as ground truth) to make the sum for each kernel equals 1. we provide the ℓ_1 distance ($\mathcal{D}_{value}$) and relative shape distance ($\mathcal{D}_{shape}$ as in

Method	Kernel	DIV2K[28]			Flicker2K[1]		
		2×	3×	4×	2×	3×	4×
RCAN finetuned	Gaussian	24.92 / 0.6903	23.39 / 0.6304	22.37 / 0.5977	24.88 / 0.6832	23.46 / 0.6333	22.47 / 0.6013
ZSSR		24.90 / 0.6900	23.36 / 0.6302	22.36 / 0.5974	24.91 / 0.6907	23.43 / 0.6332	22.46 / 0.6013
DeblurGAN w. RCAN		25.38 / 0.7090	23.87 / 0.6463	22.64 / 0.6039	25.39 / 0.7091	23.85 / 0.6465	22.72 / 0.6080
KernelGAN		25.53 / 0.7130	23.57 / 0.6370	22.97 / 0.6130	25.48 / 0.7126	23.66 / 0.6398	23.09 / 0.6174
FCA w. RCAN		28.29 / 0.8160	26.19 / 0.7382	24.93 / 0.6873	28.45 / 0.8160	26.30 / 0.7414	24.81 / 0.6905
FKP		28.13 / 0.8023	26.05 / 0.7219	24.82 / 0.6748	28.40 / 0.8188	26.22 / 0.7412	24.68 / 0.6912
IKC		28.45 / 0.8366	26.84 / 0.7595	25.56 / 0.7100	28.46 / 0.8421	26.72 / 0.7589	25.64 / 0.7146
S2K w. SFTMD		29.31 / 0.8359	27.19 / 0.7590	25.84 / 0.7096	29.38 / 0.8411	26.96 / 0.7582	25.79 / 0.7143
S2K w. RCAN		29.77 / 0.8444	27.47 / 0.7674	26.05 / 0.7139	29.91 / 0.8469	27.39 / 0.7639	26.10 / 0.7145
RCAN finetuned	Motion	23.34 / 0.6522	22.06 / 0.6080	22.04 / 0.5782	23.43 / 0.6535	22.11 / 0.5992	21.49 / 0.5803
DeblurGAN w. RCAN		23.30 / 0.6555	22.24 / 0.6096	21.24 / 0.5778	23.22 / 0.6540	22.10 / 0.6018	21.40 / 0.5801
ZSSR		23.33 / 0.6506	22.25 / 0.6074	21.32 / 0.5785	24.91 / 0.6907	23.43 / 0.6332	22.46 / 0.6013
KernelGAN		22.40 / 0.6080	22.17 / 0.6035	20.45 / 0.5414	22.40 / 0.6115	22.00 / 0.5935	20.56 / 0.5363
IKC		24.96 / 0.7543	22.27 / 0.6357	22.11 / 0.6330	24.70 / 0.7352	21.95 / 0.6244	21.89 / 0.6284
Pan *et al.* [22] w. SFTMD		24.10 / 0.6729	20.04 / 0.5539	19.21 / 0.5318	21.04 / 0.5996	19.90 / 0.5529	18.80 / 0.5168
Pan *et al.* [23] w. SFTMD		21.50 / 0.5673	20.37 / 0.5113	19.63 / 0.4765	21.28 / 0.5238	20.22 / 0.5182	19.83 / 0.4932
FKP		24.38 / 0.6897	22.39 / 0.6273	21.02 / 0.5887	24.43 / 0.6902	22.41 / 0.6268	21.10 / 0.7011
S2K w. SFTMD		30.95 / 0.8870	28.28 / 0.8118	26.95 / 0.7593	31.21 / 0.8971	28.32 / 0.8212	26.70 / 0.7557

Table 2: **Quantitative [PSNR↑ / SSIM↑] comparison results of fidelity-oriented SR model for 2×, 3×, 4× up-sampling.** G, M denote Gaussian kernels and motion kernels respectively. The best performance is shown in red and the second best in blue.

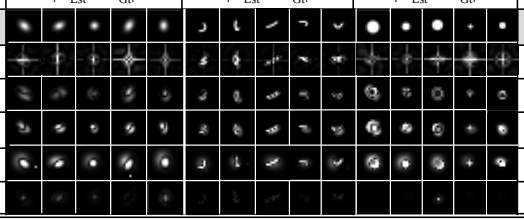
Methods	Gaussian kernel $\lvert K_{Est} - K_{Gt}\rvert$	Motion kernel $\lvert K_{Est} - K_{Gt}\rvert$	Disk kernel $\lvert K_{Est} - K_{Gt}\rvert$	$\sum \mathcal{D}_{value}(K_{Est}, K_{Gt})$ Gaussian/Motion/Disk	$\sum \mathcal{D}_{shape}(K_{Est}, K_{Gt})$ Gaussian/Motion/Disk
K_{Gt}				-	-
Xu et al.				0.2813/0.4555/0.7159	144.2/429.09/335.5
Pan et al.				0.1209/0.3705/0.1696	16.82/446.33/115.4
KernelGAN				0.2146/0.3922/0.3310	27.27/515.76/297.2
FKP				0.3839/0.4851/0.5229	129.0/699.47/577.8
S2K				**0.0274/0.0672/0.0089**	**0.6788/22.27/0.6237**

Figure 6: Kernel estimation comparison on 3× degraded LR of DIV2K dataset with random kernels. Compared with [30, 22, 2, 14], Ours can get more accurate blur kernel estimation results close to GT.

Eq. 11) between predict kernel and the ground truth on 100 test I_{LR} images. Here h, w denote the height and width of the kernel, i, j denote position indices, respectively.

$$\mathcal{D}_{value}(X\ ,\ Y) = \frac{1}{h * w}\sum_{i,j}|X_{i,j} - Y_{i,j}|,\ \mathcal{D}_{shape}(X\ ,\ Y) = \sum_{i,j} X_{i,j}\log(\frac{X_{i,j}}{Y_{i,j}}). \tag{11}$$

The results are shown in Figure 6. Compared with existing single image kernel estimation methods, our proposed method can more accurately estimate results for Gaussian, motion and disk kernels. Further, to verify the generality of estimation in frequency domain, we use different reconstruction model architectures and sizes, respectively, and compare the blind SR performance as shown in Figure 7. Using different model architectures and sizes, estimation in frequency domain is more conducive to accurate results than in the spatial domain. The effective shape structure provided in the frequency domain can still ensure relatively good SR performance for model with less parameters, but spatial estimation is more sensitive to the model size and requires adequate representation capability.

Ablation Study We first conduct ablation study on loss components of proposed generator. The results are shown in Table 3. Then to verify the superiority of obtaining kernel directly from the frequency spectrum, we conduct extra experiments with the same encoder network and training pipeline. Results can be found in Table 4. The ablation experiments are conducted on 2 additional testsets using the setting of 3× blind SR on DIV2K validset degraded with *random* Gaussian kernels.

Table 3: Ablation study on loss components.

Settings			Metrics	
$\mathcal{L}_1$	$\mathcal{L}_{adv}$	$\mathcal{L}_{reg}$	PSNR	SSIM
×	✓	×	8.462	0.1533
✓	×	×	19.46	0.4952
✓	✓	×	26.68	0.7562
✓	✓	✓	**26.94**	**0.7563**

Table 4: Ablation study on input/output forms.

Settings		Metrics	
Input	Output	PSNR	SSIM
Spatial	Para.	20.01	0.5169
Spatial	Kernel↑	22.50	0.6585
DFT	Para.	20.53	0.5417
DFT	Kernel↑	**26.96**	**0.7576**

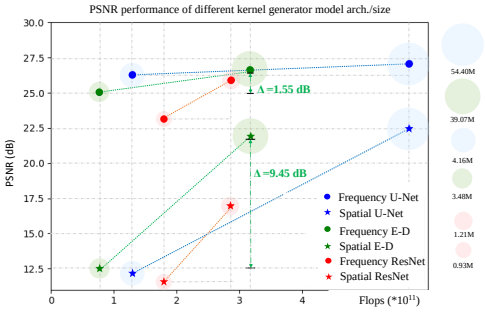
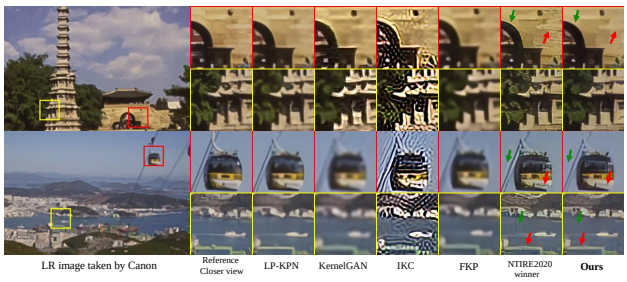


Figure 7: We use different kernel generator architectures and model sizes, take degraded LR images in the spatial and frequency domain as input to estimate the unknown blur kernels respectively, and provide the SR performance (PSNR/SSIM) comparison. Here, E-D denotes U-net removing the mirror skip connection. Model architectures used are Unet-8-64, Unet-5-32; ED-8-64, ED- 5-32; ResNet-9-32, ResNet-5-32 (arch.-downsample/residual blocks-channels).

Estimation error in different positions of the kernel have unequal influences on subsequent SR results, ℓ_1 without position information is inconsistent with SR results when the gap is small. ℓ_{adv} allows the generator to learn the distribution features of the target kernel (*i.e. Gaussian* kernel). The smooth prior ℓ_{reg} is conducive to the initial learning of the generator, and coordinating the learning of pixel features and distribution features. Our method uses the translation from the spectrum to the interpolated kernel, which avoids the learning of local pixel distribution, greatly improves the learning stability, and avoids the constraint of parameters. Compared with other settings, we achieve superior kernel estimation accuracy and better SR results with the same network structure and training pipeline.



Non-reference metrics comparison (Average of 50 test images)

	PIQE ↓	BRISQE ↓
LP-KPN	83.41	43.92
KernelGAN	80.71	50.38
IKC	77.85	42.69
NTIRE20 winner	62.50	45.03
FKP	89.32	52.56
S2K	**52.73**	**36.10**

Figure 8: Visual comparison of $4\times$ blind SR results on real images taken by Canon from RealSR[4].

4.2 Blind SR on Real Images

For real images, we use the paired image taken with a closer view in the dataset [4] as the content-fidelity reference, and take into account the visual perception of the SR results. The comparison methods include KernelGAN [2], LP-KPN [4], IKC [6] and NTIRE20 winner [10]. We adopt Gaussian kernel estimation setting and use the same SR pipeline as [10] but replace the kernel estimation part with our proposed S2K. The result of $4\times$ SR is shown in the Figure 8. From the example results, we obverse that KernelGAN's results are not stable because the accuracy of kernel estimation is not accurate enough, LP-KPN's results can not meet visual perception well, and IKC's results are too sharp. Compared with NTIRE20 winner [10], our results achieve higher fidelity (red indicators) and fewer artifacts (green indicators). There is no ground truth in the strict sense, we provide the non-reference evaluation metrics for the real image SR results as Figure 8 right. Our proposed method achieves the best visual effects and non-reference evaluation metrics performance on real images, proves its effectiveness and presents significant progress towards real-world SR.

5 Conclusion

In this paper, we theoretically demonstrate that feature representation in frequency domain is more conducive for blur kernel reconstruction than in spatial domain. We further propose S2K framework, which makes the most of shape structure from degraded LR image's frequency spectrum and with an feasible implicit cross-domain translation, output the corresponding up-sampled spatial kernel directly. In addition, we apply the kernel estimation optimization objective consistent with the performance of SR. Comprehensive blind SR experiments on synthetic and real-world datasets show that, when plugged into the off-the-shelf non-blind SR model, our proposed method achieves better results both visually and quantitatively against state-of-the-art blind SR methods, by a large margin.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 126–135, 2017.

[2] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 284–293, 2019.

[3] Giacomo Boracchi and Alessandro Foi. Modeling the performance of image restoration from motion blur. *IEEE Transactions on Image Processing*, pages 3502–3517, 2012.

[4] 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 (ICCV)*, pages 3086–3095, 2019.

[5] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1604–1613, 2019.

[7] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33, 2020.

[9] 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 (CVPR)*, pages 1428–1437, 2020.

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

[11] Xiaozhong Ji, Guangpin Tao, Yun Cao, Ying Tai, Tong Lu, Chengjie Wang, Jilin Li, and Feiyue Huang. Frequency consistent adaptation for real world super resolution. *arXiv preprint arXiv:2012.10102*, 2020.

[12] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, 2010.

[13] 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 (CVPR)*, pages 624–632, 2017.

[14] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. Flow-based kernel prior with application to blind super-resolution. *arXiv preprint arXiv:2103.15977*, 2021.

[15] 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 (CVPR) Workshops*, pages 136–144, 2017.

[16] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2359–2368, 2020.

[17] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 291–300, 2020.

[19] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017.

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

[21] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[22] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1628–1636, 2016.

[23] Liyuan Pan, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Phase-only image based kernel estimation for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6043, 2019.

[24] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019.

[25] 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 (CVPR)*, pages 3118–3126, 2018.

[26] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4539–4547, 2017.

[28] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 114–125, 2017.

[29] 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)*, pages 0–0, 2018.

[30] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013.

[31] 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 (CVPR*, pages 12496–12505, 2020.

[32] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):156–171, 2016.

[33] 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 (CVPR)*, pages 3262–3271, 2018.

[34] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Deep plug-and-play super-resolution for arbitrary blur kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1671–1681, 2019.

[35] 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.

[36] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2433–2443, 2019.

Paper Checklist

1. For all authors...
(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
[Yes]
(b) Have you read the ethics review guidelines and ensured that your paper conforms to them?
[Yes]
(c) Did you discuss any potential negative societal impacts of your work?
[No]
(d) Did you describe the limitations of your work?
[Yes]. We describe some limitations in supplementary materials.
2. If you are including theoretical results...
(a) Did you state the full set of assumptions of all theoretical results?
[Yes]. We provide assumptions in Sec.3.2.
(b) Did you include complete proofs of all theoretical results?
[Yes]. Necessary complete proofs are provided in supplementary materials.
3. If you ran experiments...
(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
[Yes]. We will prepare a URL later and our method is totally repeatable.
(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
[Yes]. Please refer to the experimental setup in Subsec. 4.1.
(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
[No].
(d) Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
[Yes].
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
(a) If your work uses existing assets, did you cite the creators?
[Yes].
(b) Did you mention the license of the assets?
[Yes].
(c) Did you include any new assets either in the supplemental material or as a URL?
[Yes]
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
[Yes]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
[No].
5. If you used crowdsourcing or conducted research with human subjects...
N/A.