FIND THE WAY BACK: INVERTIBLE KERNEL ESTIMATOR FOR BLIND IMAGE SUPER-RESOLUTION

Ting-Wei Chang *Wei-Chen Chiu* *Ching-Chun Huang*

National Yang Ming Chiao Tung University, Taiwan MediaTek Advanced Research Center, Taiwan

ABSTRACT

We address the task of zero-shot blind image super-resolution, where it aims to recover the high-resolution details from the low-resolution input image under a challenging problem setting of having no external training data, no prior assumption on the downsampling kernel, and no pre-training components used for estimating the downsampling kernel. While existing zero-shot blind super-resolution works follow the strategy of firstly estimating the downsampling kernel via cross-scale recurrence and then learning the non-blind upsampling model, we in turn propose a carefully-designed invertible network for modeling both the downsampling and upsampling operations at once. Specifically, the invertible property enables the use of cross-scale recurrence across more scales and thus further benefits the overall model training. We conduct extensive experiments to demonstrate our proposed method's superior performance over several baselines and its effectiveness in handling the images downsampled by nonlinear kernels.

***Index Terms*—** Blind super-resolution, Flow-based generative model, Zero-shot learning

1. INTRODUCTION

With the recent renaissance of deep learning techniques, we have witnessed the leap in the image super-resolution performance brought by deep-learning-based models [1, 2, 3, 4, 5]. As such improvements typically stem from the supervised learning scenario, it requires plenty of supervised training data composed of the pairs of low-resolution (LR) and high-resolution (HR) images to train the super-resolution models. Moreover, there exists a strong assumption behind these methods: the downsampling kernel to produce the LR images from their corresponding HR ones is known and predefined. However, this assumption limits the generalizability of the learned models when the true downsampling kernel that degrades the LR image differs from the pre-defined kernel used during training.

For tackling the issue caused by the assumption above, the task of **blind image super-resolution** emerges in which the downsampling kernel is not assumed known [6, 7, 8, 9, 10]. Most of the existing blind super-resolution approaches adopt a general strategy: a kernel estimator firstly estimates the downsampling kernel from the LR input image, where the estimated kernel is then utilized by the non-blind super-resolution model [11, 12] to reconstruct the HR output from the LR input. However, the supervised blind super-resolution methods (e.g. [13]), which utilize **supervised datasets** to determine the kernel estimator and construct the non-blind super-resolution models, may again potentially suffer from the issue of generalizability.

To this end, a specific task, **zero-shot blind image super-resolution**, advances to address the setting that none of the modules used in kernel estimation and super-resolution networks is pre-trained or relies on external training datasets for the sake of maximizing its generalizability towards arbitrary downsampling kernels. KernelGAN [14], as a seminal work for such a task, adopts **cross-scale recurrence** [15, 16, 7] (i.e., *an essential assumption that the correct degradation kernel would maximize the patch similarity across different image scales*) and adversarial learning to train a downsampling network which approximates the original degradation procedure to produce the input LR image. Given the estimated degradation kernel, a non-blind SR method (e.g., ZSSR [11]) is then used to produce the super-resolution output. Recently, DualSR [17] follows up to have both kernel estimation and super-resolution models jointly trained in a dual-path framework and achieves state-of-the-art performance.

In this paper, we also focus on the zero-shot blind image super-resolution with novelties: **(1)** Standing on the powerful models of neural flows [18, 19], our framework is novel to approximate the image downsampling and upsampling processes simultaneously via a flow-based generative model, named as **invertible kernel estimator (IKE)**. **(2)** Thanks to the invertibility of IKE; instead of adopting two separate networks for the downsampling and upsampling models (i.e., what DualSR does), the forward and backward flows of our IKE are directly linked to the downsampling and upsampling steps. **(3)** We proposed the objectives of cross-scale recurrence on both the forward and backward paths of IKE. By introducing self-supervision across multiple image scales (SR versus LR in the backward path and LR versus lower-resolution in the forward path), the objectives benefits the model performance. Experimental results under various unknown and nonlinear downsampling kernels show that IKE outperforms the baselines of zero-shot blind super-resolution and supervised super-resolution methods.

2. PROPOSED METHOD

The overview of our proposed framework is shown in Figure 1, which is composed of several subnetworks: invertible kernel estimation network (named as **IKENet**), Z-upsampling module U, and patch discriminators $\{PD_1, PD_2\}$. We now detail our proposed framework in the following.

Our IKENet simultaneously models the downsampling and upsampling processes in a unified flow-based network, in which its forward pass acts as the encoder E to perform downsampling (and kernel estimation) while the backward pass acts as the decoder $D = E^{-1}$ (i.e., the inverse of E) to perform upsampling/super-resolution. Given a test LR image X_{LR}, we assume that it is downsampled from a high-resolution image X_{HR} by a scaling factor s (in both width and height) via a degradation kernel H.

Bicubic Residual. We propose to take the *bicubic residual map* $\hat{R}$, instead of the given test LR image X_{LR}, as the input for E. With denoting the bicubically-downscaled version of X_{LR} as $X^{\downarrow}_{bic}$ (which is s-times smaller than X_{LR} in both width and height) and the bicubically-upscaled version of $X^{\downarrow}_{bic}$ as $X^{\downarrow\uparrow}_{bic}$, then $\hat{R}$ is obtained by $X_{LR} - X^{\downarrow\uparrow}_{bic}$. The motivation of having $\hat{R}$ as the input for E stems from its sparsity (i.e. most pixels in $\hat{R}$ related to the homogeneous regions of X_{LR} will be zero) thus leading to more efficient learning of our IKENet. The IKENet encodes $\hat{R}$ into outputs $\hat{R}^{\downarrow}$ and Z, where the former is the downscaled residual map and the latter are the high-frequency feature maps. We are then able to obtain the H-degraded and downsampled version of X_{LR}, denoted as X_{LLR}, via the computation $\hat{R}^{\downarrow} + X^{\downarrow}_{bic}$.

Multi-Scale Cross-Scale Recurrence. According to the property of *cross-scale recurrence*, the image patches from X_{LR} and X_{LLR} should follow the same distribution. We hence adopt the adversarial learning as KernelGAN [14] to define the adversarial loss $\mathcal{L}^{fwd}_{CSR}$ via the patch discriminator PD_1 in the IKE forward pass. Denoting patches sampled from X_{LR} and X_{LLR} as p and p', $\mathcal{L}^{fwd}_{CSR}$ is defined as

$$\mathcal{L}^{fwd}_{CSR} = \mathbb{E}_{p,p'}[|PD_1(p) - 1| + |PD_1(p')|]. \quad (1)$$

Moreover, we introduce a local energy preservation loss $\mathcal{L}_{energy}$ which ensures the downsampling process of IKENet to maintain the local energy between $X^{\downarrow}_{bic}$ and X_{LLR}:

$$\mathcal{L}_{energy} = |M(X^{\downarrow}_{bic}) - M(X_{LLR})|_1 \quad (2)$$

, where M stands for a 9x9 mean filter.

In order to perform the backward pass D of IKENet for achieving the super-resolution on X_{LR} and recover X_{SR}, the input and output dimensions should be equivalent due to the property of neural flows [18, 19]. First, we obtain $Z^{\uparrow}$(with width and height both s-times large than Z) by $U(Z)$, where U is the $Z-$upscale module; then, we concatenate $Z^{\uparrow}$ and the bicubic residual map $\hat{R}$ along the channels as the input for D and obtain the upscaled bicubic residual map $\hat{R}^{\uparrow}$ as the output. Finally, the super-resolved X_{SR} is computed by adding up $X^{\uparrow}_{bic}$ (i.e. the bicubically-upscaled version of X_{LR}) and $\hat{R}^{\uparrow}$. Note that U is built by a fully convolution network with 8 hidden layers. Each layer has 64 channels followed by a ReLU activation function.

Again, we can apply the cross-scale recurrence on the recovered X_{SR} and X_{LR}, realized by the adversarial loss $\mathcal{L}^{bwd}_{CSR}$ via the second patch discriminator PD_2. By denoting patches sampled from X_{LR} and X_{SR} as p and p'', we have

$$\mathcal{L}^{bwd}_{CSR} = \mathbb{E}_{p,p''}[|PD_2(p) - 1| + |PD_2(p'')|]. \quad (3)$$

Note that $\mathcal{L}^{fwd}_{CSR}$ and $\mathcal{L}^{bwd}_{CSR}$ together impose cross-scale recurrent across multiple image scales, which significantly improves the super-resolution performance in the experiments.

Furthermore, as inspired by DualSR [17] where $X^{\uparrow}_{bic}$ (i.e. the bicubically-upscaled version of X_{LR}) typically gets many artifacts around the edge pixels while having less artifacts in the homogeneous regions, we adopt the interpolation loss $\mathcal{L}_{inter}$ as [17] to encourage X_{SR} being similar to $X^{\uparrow}_{bic}$ in the homogeneous regions. By having $f_{mask} = 1 - \texttt{Sobel}(X^{\uparrow}_{bic})$ where `Sobel` is the Sobel edge detector, we define

$$\mathcal{L}_{inter} = |f_{mask} \times (X_{SR} - X^{\uparrow}_{bic})|_1. \quad (4)$$

Lastly, we have the total variation loss $\mathcal{L}_{TV}$ as regularization to avoid X_{SR} from having checkerboard artifacts. Denoting $X_{i,j}$ as a pixel of X_{SR} at coordinate (i, j), $\mathcal{L}_{TV}$ is

$$\mathcal{L}_{TV} = \sum_{X_{i,j}} \sqrt{(X_{i,j+1} - X_{i,j})^2 + (X_{i+1,j} - X_{i,j})^2}. \quad (5)$$

Detailed Architecture of IKENet. As illustrated in Figure 1, our IKENet is composed of downscaling block(s), where each downscaling block is composed of a pixel-shuffle module and K invertible blocks. The pixel-shuffle module is identical to the sub-pixel convolutional layer proposed by [20], which is popular in many super-resolution networks. In the forward pass, pixel-shuffle module performs the `space-to-depth` permutation to reduce pixels in spatial axes and move them into the channel dimension; while in the backward pass, pixel-shuffle module applies the `depth-to-space` permutation to permute the pixels for reducing the channel size and increasing the spatial size. Regarding the invertible blocks, it is built upon the invertible *tri-channel* coupling layers proposed by us. For most neural flows (e.g. [21, 18]), given the input feature map x, a `split` function firstly splits x into halves along the channel dimension to obtain x_a and x_b, and additive coupling layers are used to mix them via the transformations ϕ and σ (as shown in Table 1). In comparison, we adopt the idea of skip connections for residual learning used by many super-resolution networks. In particular, besides mixing x_a and x_b, our invertible tri-channel coupling layer has an extra channel x_{res} to preserve a copy of x_a for further manipulation. The detailed formulation is provided in Table 1

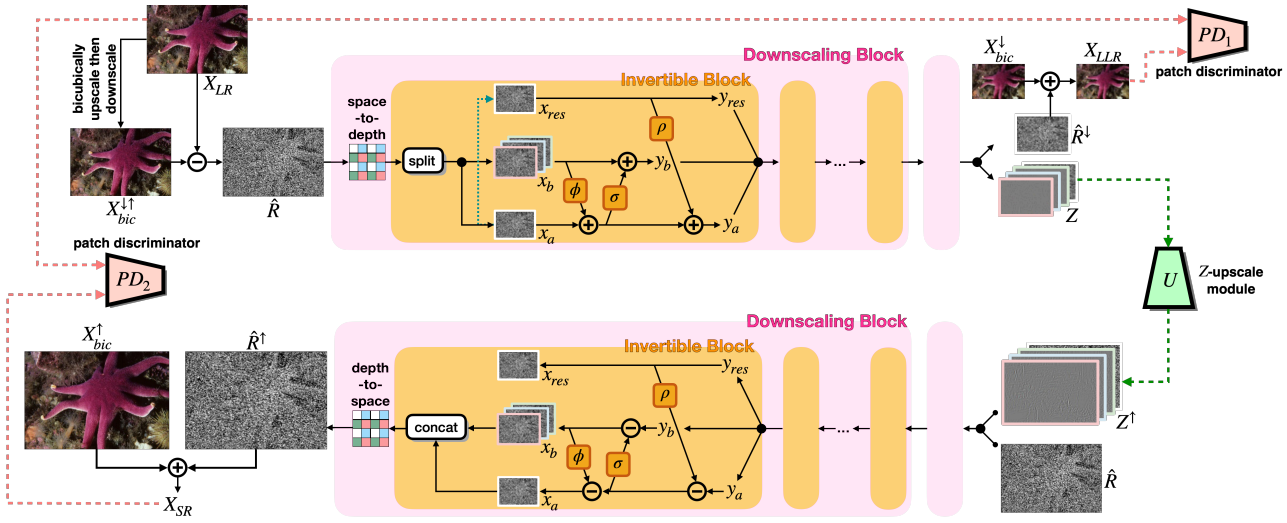

Fig. 1: Illustration of our proposed framework for zero-shot blind image super-resolution.

together with its inverse process. Now, in the forward pass of IKENet, we pass $x = \texttt{space-to-depth}(\hat{R})$ through K invertible blocks to get $\hat{R}^{\downarrow} = x_a^K$ and $Z = [x_b^K, x_{res}^K]$, where $\{x_a^K, x_b^K, x_{res}^K\}$ are $\{x_a, x_b, x_{res}\}$ obtained after passing through K invertible blocks.

Table 1: Comparison between additive coupling layers (used in typical neural flows) and our tri-channel coupling layers.

	Forward Function	Inverse Function
Additive	$y_a = x_a + \phi(x_b)$ $y_b = x_b + \sigma(y_a)$	$x_b = y_b - \sigma(y_a)$ $x_a = y_a - \phi(x_b)$
Our Triple-Channel	$y_a = x_a + \phi(x_b)$ $y_b = x_b + \sigma(y_a)$ $y_a = y_a + \rho(x_{res})$ $y_{res} = x_{res}$	$y_a = y_a - \rho(y_{res})$ $x_b = y_b - \sigma(y_a)$ $x_a = y_a - \phi(x_b)$ $x_{res} = y_{res}$

Training Procedure. Our model training has two phases: the *encoding training phase* and the *decoding training phase*, which are iteratively executed until a certain number of iterations (i.e., 3000 in our setting). In the encoding phase, we focus on enhancing the cross-scale recurrence and the energy preservation of X_{LLR}. Hence, its total objective $\mathcal{L}_{total}^{E}$ is:

$$\mathcal{L}_{total}^{E} = \mathcal{L}_{CSR}^{fwd} + \lambda_{energy}\mathcal{L}_{energy}. \quad (6)$$

In the decoding training phase, we adopt the cross-scale recurrence, the interpolation loss, and the total variation loss in the backward pass to jointly train both the IKENet and Z-upscale module U. Its total objective $\mathcal{L}_{total}^{D}$ is defined as:

$$\mathcal{L}_{total}^{D} = \lambda_{CSR}^{bwd}\mathcal{L}_{CSR}^{bwd} + \lambda_{TV}\mathcal{L}_{TV} + \lambda_{inter}\mathcal{L}_{inter}. \quad (7)$$

In our implementation, we use $K = 4$ invertible blocks, and our discriminators in both phases are identical to the ones in KernelGAN [14]. We set λ_{energy} to 4 and gradually decrease it to 1 over iterations to avoid blurring artifacts on X_{LLR}. Also, we set λ_{CSR}^{bwd}, λ_{TV}, and λ_{inter} to 5, 1, and 5 respectively to balance between them. Our source code, datasets, and models would be released upon paper acceptance.

3. EXPERIMENTS AND RESULTS

Datasets. We adopt the DIV2K dataset [22] for our experiments. DIV2K contains 800, 100, and 100 images for training, validation, and testing. We focus on Track 2 of DIV2K, where the LR images are produced by unknown degradation kernels. Moreover, to verify our method's versatility, we also produce our testing datasets where the LR images are downscaled by various nonlinear filters such as bilateral, anisotropic diffusion, median, and random kernels (i.e., randomly assigning values to the elements of a kernel).

Table 2: Comparison based on Track 2 of the DIV2K dataset. The first three baselines are the supervised super-resolution methods, while KeneralGAN+ZSSR and DualSR are zero-shot blind super-resolution baselines. The red and blue colors indicate the best and the second best performances.

Method	Upscaling by 2		Upscaling by 4	
	PSNR	SSIM	PSNR	SSIM
EDSR [4]	25.008	0.7107	21.576	0.5474
RCAN [5]	25.007	0.7108	21.571	0.5471
ESRGAN [23]	-	-	21.569	0.5469
KernelGAN+ZSSR [14]	23.599	0.6400	19.623	0.4511
DualSR [17]	25.295	0.7265	20.174	0.5101
IKE (Ours)	25.478	0.7298	21.776	0.5518

Results. We evaluate our method on the Track 2 of DIV2K and the non-linear degradation dataset under the blind zero-shot setting (i.e. only the test set is used) and compare with several supervised super-resolution baselines (i.e. EDSR [4], RCAN [5], and ESRGAN [23]) and two state-of-the-art blind zero-shot super-resolution baselines, KernelGAN+ZSSR [14] and DualSR [17]. The supervised models are trained by using many bicubic LR-HR pairs from the DIV2K training set.

Quantitative results in Table 2 and 3 show that our method achieves superior performance, especially its ability in tackling non-linear degradation kernels (cf. Table 3). It verifies our contributions of having both downsampling and upsam-

Table 3: The performance comparison of our method with KernelGAN and DualSR on the non-linear degradation dataset.

	Bilateral				Median			
	Upscale by 2		Upscale by 4		Upscale by 2		Upscale by 4	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [4]	26.501	0.7761	23.898	0.6821	26.329	0.7884	24.577	0.6897
KernelGAN+ZSSR [14]	23.784	0.7651	20.886	0.6655	25.917	0.7741	20.168	0.6251
DualSR [17]	25.764	0.7744	21.427	0.6557	26.020	0.8023	22.198	0.6843
IKE (Ours)	28.311	0.7906	24.888	0.6806	28.581	0.8132	25.343	0.7096

	Anisotropic Diffusion				Random			
	Upscale by 2		Upscale by 4		Upscale by 2		Upscale by 4	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [4]	27.227	0.8595	22.848	0.6791	25.003	0.7199	24.352	0.6691
KernelGAN+ZSSR [14]	21.120	0.6912	17.212	0.4981	25.210	0.7236	24.326	0.6718
DualSR [17]	26.773	0.8683	20.906	0.6611	23.945	0.7048	22.220	0.6673
IKE (Ours)	30.055	0.8900	24.323	0.6804	26.253	0.7202	25.597	0.6973

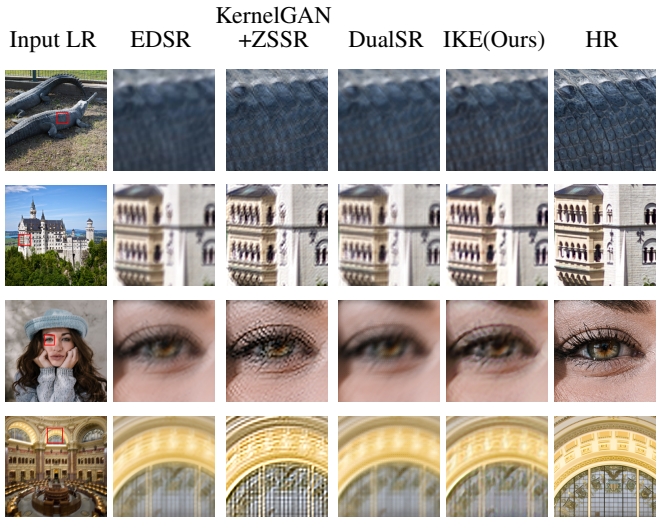

Fig. 2: SR images on DIV2K dataset Track 2. The first two and the last two rows are the results of upsampling by 2 times and 4 times respectively.

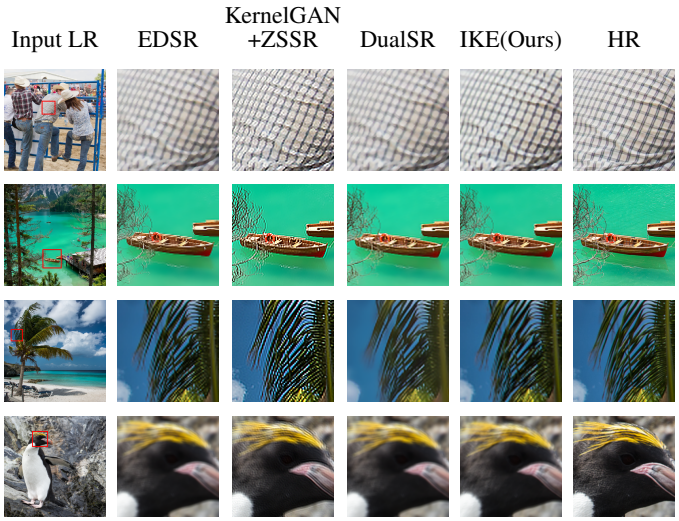

Fig. 3: SR images on the non-linear degradation dataset. The first to the fourth rows are the results against Median, Bilateral, Anisotropic Diffusion and Random degradation kernels.

pling processes modeled in a unified invertible network (i.e., IKENet) as well as several novel model designs. From the qualitative results in Fig. 2 and Fig. 3, we observe that: with using plenty of training data, the supervised models seem to produce super-resolution images with good quality and have good PSNR values. However, the potential mismatch between the bicubic kernel and the true degradation kernel for the testing LR images leads to blurry super-resolution results. For instance, the SR images produced by EDSR shown in Figure 2 have fewer details and lack image contrast. For zero-shot baselines, KernelGAN produces high image contrast but introduces inaccurate high-frequency details and ringing artifacts due to potential error accumulation from the kernel estimation and non-blind super-resolution stages. In contrast, though DualSR seems to well suppress the unnatural artifacts, our IKE is able to recover more HR image details than DualSR. The potential reasons are: (1) the network architecture of DualSR is inherited from KernelGAN, which mainly focuses on tackling linear degradation kernels. Thus, the possible non-linear degradations in the testing images are hard to handle; (2) as DualSR has downsampling and upsampling processes modeled by two sub-networks, the potential inconsistency between two processes leads to worse super-resolution performance. It is also worth noting that even though the zero-shot baselines aim to tackle the unknown degradation kernels, they may still perform worse than supervised methods, while our proposed method clearly alleviates such issue and performs better than supervised baselines. More results can be found here.

4. CONCLUSION

We propose an invertible framework to jointly model the image degradation process and super-resolve LR images under a zero-shot scenario. Neither using any prior knowledge of the degradation kernel nor relying on the external datasets, the proposed IKE is practical and adaptive to help each LR image find its way back to its HR counterpart. The experiments under various unknown and nonlinear downsampling kernels verify the superiority of our method against the state-of-the-art blind zero-shot super-resolution baselines.

5. REFERENCES

[1] 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 (ECCV)*, 2014.

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[6] Isabelle Begin and FR Ferrie, "Blind super-resolution using a learning-based approach," in *International Conference on Pattern Recognition (ICPR)*, 2004.

[7] Tomer Michaeli and Michal Irani, "Nonparametric blind super-resolution," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[8] Qiang Wang, Xiaoou Tang, and Harry Shum, "Patch based blind image super resolution," in *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[9] He He and Wan-Chi Siu, "Single image super-resolution using gaussian process regression," in *CVPR 2011*. IEEE, 2011, pp. 449–456.

[10] Yu He, Kim-Hui Yap, Li Chen, and Lap-Pui Chau, "A soft map framework for blind super-resolution image reconstruction," *Image and Vision Computing*, pp. 364–373, 2009.

[11] Assaf Shocher, Nadav Cohen, and Michal Irani, ""zero-shot" super-resolution using deep internal learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[14] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani, "Blind super-resolution kernel estimation using an internal-gan," *ArXiv:1909.06581*, 2019.

[15] Daniel Glasner, Shai Bagon, and Michal Irani, "Super-resolution from a single image," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[16] Maria Zontak and Michal Irani, "Internal statistics of a single natural image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[17] Mohammad Emad, Maurice Peemen, and Henk Corporaal, "Dualsr: Zero-shot dual learning for real-world super-resolution," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density estimation using real nvp," *ArXiv:1605.08803*, 2016.

[19] Diederik P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *ArXiv:1807.03039*, 2018.

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

[21] Laurent Dinh, David Krueger, and Yoshua Bengio, "Nice: Non-linear independent components estimation," *ArXiv:1410.8516*, 2014.

[22] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[23] 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 *The European Conference on Computer Vision Workshops (ECCVW)*, 2018.