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] (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 benefit 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-downsampled version of $X_{LR}$ as $X_{bic}^d$ (which is $s$-times smaller than $X_{LR}$ in both width and height) and the bicubically-upsampled version of $X_{bic}$ as $X_{bic}^u$, then $\hat{R}$ is obtained by $X_{LR} - X_{bic}$. 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 $\hat{R}^k$ 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 $\hat{R}^k + X_{bic}^u$.

**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 $L_{CSR}^{bud}$ via the patch discriminator $PD_1$ in the IKE forward pass. Denoting patches sampled from $X_{LR}$ and $X_{LLR}$ as $p$ and $p'$, $L_{CSR}^{bud}$ is defined as

$$L_{CSR}^{bud} = \mathbb{E}_{p,p'}[|PD_1(p) - 1| + |PD_1(p')|].$$

Moreover, we introduce a local energy preservation loss $L_{energy}$ which ensures the downsampling process of IKENet to maintain the local energy between $X_{bic}$ and $X_{LR}$:

$$L_{energy} = |M(X_{bic}^d) - M(X_{LLR})|_1$$

, 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 [13] [19]. First, we obtain $Z'$ (with width and height both $s$-times larger than $Z$) by $U(Z)$, where $U$ is the $Z$-upscale module; then, we concatenate $Z'$ and the bicubic residual map $\hat{R}$ along the channels as the input for $D$ and obtain the upscaled bicubic residual map $\hat{R}^\dagger$ as the output. Finally, the super-resolved $X_{SR}$ is computed by adding up $X_{bic}^u$ (i.e. the bicubically-upscaled version of $X_{LR}$) and $\hat{R}^\dagger$. Note that $U$ is built by a fully convolutional 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 $L_{CSR}^{bwd}$ via the second patch discriminator $PD_2$. By denoting patches sampled from $X_{LR}$ and $X_{SR}$ as $p$ and $p''$, we have

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

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

Furthermore, as inspired by DualSR [17] where 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 $L_{inter}^{bic}$ to encourage $X_{SR}$ being similar to $X_{bic}$ in the homogeneous regions. By having $f_{mask} = 1 - \text{Sobel}(X_{bic})$ where Sobel is the Sobel edge detector, we define

$$L_{inter} = |f_{mask} \times (X_{SR} - X_{bic}^u)|_1. $$

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

$$L_{TV} = \sum_{X_{i,j}} \sqrt{(X_{i,j+1} - X_{i,j})^2 + (X_{i+1,j} - X_{i,j})^2}. $$

**Detailed Architecture of IKENet.** As illustrated in Figure [1], our IKENet is composed of downsampling blocks, where each downsampling 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 $\phi$ and $\sigma$ (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].
**Inverse Function**

**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 $X_{L,L,R}$. Hence, its total objective $L^E_{total}$ is:

$$L^E_{total} = L_{CSR}^{fwd} + \lambda_{energy} L_{energy}$$  \hfill (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 $L_{total}^D$ is defined as:

$$L_{total}^D = \lambda_{CSR}^{fwd} L_{CSR}^{fwd} + \lambda TV L_{TV} + \lambda_{inter} L_{inter}$$  \hfill (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 $\lambda_{energy}$ to 4 and gradually decrease it to 1 over iterations to avoid blurring artifacts on $X_{L,L,R}$. Also, we set $\lambda_{CSR}^{fwd}$, $\lambda TV$, and $\lambda_{inter}$ to 5, 1, and 5 respectively to balance between them. Our source code, datasets, and models would be released upon paper acceptance.

**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 downsampled 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 KernelGAN+ZSSR and DualSR are zero-shot super-resolution baselines. The red and blue colors indicate the best and the second best performances.

<table>
<thead>
<tr>
<th>Method</th>
<th>Upscaling by 2</th>
<th>Upscaling by 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>EDSR [4]</td>
<td>25.008</td>
<td>0.7107</td>
</tr>
<tr>
<td>RCAN [5]</td>
<td>25.007</td>
<td>0.7108</td>
</tr>
<tr>
<td>ESRGAN [23]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KernelGAN+ZSSR [14]</td>
<td>23.599</td>
<td>0.6400</td>
</tr>
<tr>
<td>DualSR [17]</td>
<td>25.295</td>
<td>0.7265</td>
</tr>
<tr>
<td>IKE (Ours)</td>
<td>25.478</td>
<td>0.7298</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th></th>
<th>Bilateral</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upscale by 2</td>
<td>Upscale by 4</td>
</tr>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>EDSR [4]</td>
<td>26.501</td>
<td>0.7761</td>
</tr>
<tr>
<td>KernelGAN+ZSSR [14]</td>
<td>23.784</td>
<td>0.7651</td>
</tr>
<tr>
<td>DualSR [17]</td>
<td>25.764</td>
<td>0.7744</td>
</tr>
<tr>
<td>IKE (Ours)</td>
<td>28.311</td>
<td>0.7906</td>
</tr>
</tbody>
</table>

Anisotropic Diffusion

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upscale by 2</td>
</tr>
<tr>
<td></td>
<td>PSNR</td>
</tr>
<tr>
<td>EDSR [4]</td>
<td>27.227</td>
</tr>
<tr>
<td>KernelGAN+ZSSR [14]</td>
<td>21.120</td>
</tr>
<tr>
<td>DualSR [17]</td>
<td>26.773</td>
</tr>
<tr>
<td>IKE (Ours)</td>
<td>30.055</td>
</tr>
</tbody>
</table>

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.

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.

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


