REAL-WORLD IMAGE SUPER-RESOLUTION VIA KERNEL AUGMENTATION AND STOCHASTIC VARIATION

Haiyu Zhang1*       Yu Zhu1       Jinqiu Sun2      Yanning Zhang1

1School of Computer Science, Northwestern Polytechnical University, Xi’an 710072, P.R. China
2School of Astronautics, Northwestern Polytechnical University, Xi’an 710072, P.R. China

ABSTRACT

At present, highly satisfactory results have been achieved on the synthetic datasets for most learning-based single image super-resolution (SISR) algorithms in terms of evaluation and visualization. However, in some practical applications, due to the limitation and unity from the most commonly used bicubic degradation kernel, the performance of a model trained by the pre-defined settings has significantly reduced in real-world low-definition photos. Hence, we first propose a novel kernel augmentation (KA) strategy to improve the generalization ability and robustness of the current model. We then aim to restore the widespread stochastic variation (SV) features existing in natural images, which will lead to more authentic and realistic features representation. In the end, extensive experiments demonstrate the feasibility and availability of our methods in dealing with the real-world SISR issue.

Index Terms— Single image super-resolution (SISR), real-world, kernel augmentation (KA), stochastic variation (SV)

1. INTRODUCTION

Super-resolution (SR) plays a significant role in the field of computer vision (CV), as it can be beneficial to many real-world applications. Specifically, given a low-resolution (LR) image, trying to recover or estimate its high-resolution (HR) counterpart refers to single image super-resolution (SISR). As a matter of fact, SISR algorithms based on PSNR-oriented approaches usually can only achieve fairly high peak signal-to-noise ratio (PSNR) scores, while scarcely produce images complied with the human visual system (HVS). Therefore, some perceptual-oriented methods were proposed successively, e.g., SRGAN [1] combined the perceptual loss [2, 3] and GANs [4] to generate images with richer details. And the enhanced edition of it, ESRGAN [5], introduced an architecture consisted of residual-in-residual dense blocks (RRDBs) and became a benchmark for many super-resolution algorithms that focus on perceptual effects.

Albeit their improvements were widely acknowledged, some defects are remained to be optimized. Owing to these perceptual-oriented methods were both training and testing on synthetic datasets, which in most circumstances use the simple and easy bicubic down-sampling kernels to stimulate the sophisticated degradation process of camera-captured photos. However, a bicubic kernel is different from the real camera-blur [6]. And the loss of high-frequency details in camera-captured images is due to several factors like the optical blur, atmospheric blur, camera shake, and lens aberrations [7], which accounted for the performances of previous SISR networks obtained by pre-defined kernels would suffer from a conspicuous declination on these low-definition photos.

Actually, this can be predominantly boiled down to the insufficient generalization ability and robustness when they were applied in real scenarios. Hence, people attempted to extract realistic kernels from these camera-captured photos. Whereas, different camera, lens, aperture, and atmospheric condition combinations may lead to various kernels. Thus, it is a hard thing to produce such a large and diverse enough dataset for training an ideal real-world SISR network.

Given this, we intend to alleviate the dilemma from two sides. For a start, we introduce a kernel augmentation (KA) strategy. Specifically, after extracting kernels from realistic photos, we then implement the augmentation operation and generate additional ones by utilizing the powerful capability from GANs’ variant for fitting and approximating complex

*Corresponding author.
E-mail: zhanghaiyu@mail.nwpu.edu.cn.
distributions. Our objective is to improve the generalization ability and robustness of the real-world SISR model.

On top of that, considering that the wide distribution of features with stochastic variation (SV) properties inside of natural images, whereas the input of generator of original GANs only depends on the input layer and the realization of stochastic variation merely by finding a method to generate pseudo-random numbers on its own. Hence, we set out to add noise into the output of the generator to replace the progress of pseudo-random production. Thus, the generated stochastic variation features could make our results seem more authentic.

In Fig.1, it can be evident to see that we can reconstruct the clear texture and correct trend for this Audi logo, while performances of the rest are sub-optimal and disappointing.

To sum up, our prime contributions are tripartite:
1) We introduced a kernel augmentation (KA) concept to enhance the generalization ability and robustness of the real-world SISR model for unknown camera kernels.
2) We restored the features which are of stochastic variation (SV) properties in the local area to highlight the textures in results and obtained more ideal visual effects.
3) We verified our methods on real-world photos as well as synthetic images and achieved state-of-the-art results both in quantitative and qualitative aspects.

2. PROPOSED METHOD

Fig. 2 serves as an illustration of our method. By and large, it can be separated into three portions. In the last section of this part, we will describe our exploration of the stochastic variation among real-world images.

\[ L_D^{\text{RGAN}} = \mathbb{E}_{(x_r, x_f) \sim (P, Q)}[f_1(C(x_f) - C(x_r))] + \mathbb{E}_{(x_r, x_f) \sim (P, Q)}[f_2(C(x_f) - C(x_r))] \]  
\[ L_G^{\text{RGAN}} = \mathbb{E}_{(x_r, x_f) \sim (P, Q)}[g_1(C(x_f) - C(x_r))] + \mathbb{E}_{(x_r, x_f) \sim (P, Q)}[g_2(C(x_f) - C(x_r))] \]  

Where \( D \) and \( G \) respectively refer to the discriminative and generative networks. \( x_r \) represents the real data, while \( x_f \) represents the fake ones. For \( P \) and \( Q \) are the distribution of real and fake data, separately. \( f_1, f_2, g_1, \) and \( g_2 \) are scalar-to-scalar functions. As to \( C(x) \) is the original output of the discriminator and it also can be interpreted as how realistic the input data is. If required, see [18] for more details.

Moreover, the standard DCGAN [19] was employed as our kernel augmentation (KA) model architecture, and the generator \( G \) is responsible for producing degraded kernels, while the discriminative network \( D \) takes the sample kernel as input and identifies its real or fake. After acquiring the estimated kernels set \( \Phi_k \), a trained RGANs model will be employed to generate extra kernels which are following the stochastic variation inside of the kernels. Our objective is to improve the generalization capability and robustness of the current model might be a feasible way to mitigate the situation towards real-world SISR problem.

2.2. Kernel Augmentation

Due to the complex and changeable imaging systems, most of the real-world datasets [10-13] are subjected to the given a few camera models, therefore, it is almost impossible to fulfill a coverage of the majority of scenarios. Predictably, the quantity and diversity of kernels collected from the above ways are confined, and train a real-world SISR model only based on them is conspicuously not sufficient. What’s more, some prevalent kernel estimation SISR algorithms [14-16] might be computationally expensive. So, we believe that augment the extracted kernels to a larger kernel pool via GANs’ variant to enhance the generalization capability and robustness of the current model might be a feasible way to mitigate the situation towards real-world SISR problem.

According to [17], Relativistic GANs (RGANs) [18] can simultaneously solve the intractable vanishing gradient and mode collapse problems, furthermore, since what we need here is a simple and convenient GANs. As a result, we choose RGANs to accomplish our kernel augmentation. Its objective functions show as Eq. (1) and Eq. (2).
And we have also found that they could cover a wide range of distribution, including the Gaussian kernels that are a better approximation of the real camera-blur than bicubic-kernel [7]. Last but not the least, since we have already built the augmented kernel pool, then, we selected some of them randomly to degrade the quality of HR images, which ultimately composed LR-HR paired data for the training and testing operations of the real-world SISR network.

2.3. Real-world SISR Model

As a distinguished baseline and fundamental framework for many GAN-based SISR models, ESRGAN [5] has achieved state-of-the-art perceptual effects for the time. Therefore, we build our network based on it. Following [5], we trained a pixel-oriented CNN model in the first place, and then a perceptual-oriented GAN model.

Specifically, the generator adopted RRDB [5] structure, and the magnification times were set as $\times 4$. The model training process was realized by minimizing the weighted combination of pixel loss (L1), perceptual loss (VGG-19), and adversarial loss with weight parameters 0.01, 1, and 0.005, respectively and empirically. To effectively restrain the widespread artifacts phenomenon in GAN-based models, we mimicked the patch discriminator [21, 22] used in [23] and defined a weight coefficient with 0.1. Ultimately, the overall loss function is formulated as Eq. (3).

$$\mathcal{L}_{\text{Overall}} = \lambda_1 \mathcal{L}_{\text{Pixel}} + \lambda_2 \mathcal{L}_{\text{Perceptual}} + \lambda_3 \mathcal{L}_{\text{Adversarial}} + \lambda_4 \mathcal{L}_{\text{PatchGAN}}$$  

By the way, we used Adam optimizer [24] with $\beta_1=0.9$ and $\beta_2=0.999$ for both generator and discriminator training and performed 50,000 training iterations with an initial learning rate of $1\times10^{-5}$ which then halved after 5,000, 10,000, 20,000, and 30,000 iterations.

2.4. Stochastic Variation

In the natural images, there are abundant stochastic variation textures, such as the cloud shape and tree branch trend. Generally, the traditional way for generators to implement stochastic variation (SV) was by adding random noise into the input vectors. However, considering that the only way of inputting data to the network is through the input layer. As a result, it needs to find a way to generate spatially varying pseudo-random numbers from the previous activations when they are needed. Unfortunately, this process does not merely consume the network capacity, and hiding the periodicity of the generated signal is also difficult, as evidenced by the commonly seen repetitive patterns in generated images. By adding per-pixel noise after each convolution, the StyleGAN [25] avoids these issues skillfully.

As illustrated by Fig. 3, to make our images which are of richer stochastic variations (SV) and realistic details, we introduce Gaussian noise into the output of each dense block of the generator, meanwhile, each of them also couples with a feature scaling factor. It is worth noting that the influence exerted by noise input is very local and does not change the main structure and higher-level information of an image.


Table 1. ×4 amplification factor super-resolved results comparison over the DF2K and DPED datasets. The red color represents the best, the blue color comes second. The arrows mean that whether the high (↑) or low (↓) values are more desirable here.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>PSNR[nm/SSIM/</th>
<th>LPIPS</th>
<th>PLI/NIQE/NRQMI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BICUBIC</td>
<td>25.3780/0.6822/8.5634</td>
<td>8.6322/9.8187/2.5511</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSSR [33]</td>
<td>25.8184/0.6988/7.4603</td>
<td>7.4008/7.7267/2.9874</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>26.2995/0.7193/8.2625</td>
<td>4.1336/4.1156/5.9857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>PSNR[nm/SSIM/</th>
<th>LPIPS</th>
<th>PLI/NIQE/NRQMI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BICUBIC</td>
<td>8.0684/9.1044/2.9143</td>
<td>6.7685/7.1059/3.2223</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RankSRGAN [32]</td>
<td>w/o GT</td>
<td>6.9833/7.4848/3.6044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FSSR [33]</td>
<td>3.4603/4.6742/7.3758</td>
<td>3.4428/4.6513/7.3542</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZSSR [9]</td>
<td>w/o GT</td>
<td>3.4603/4.6742/7.3758</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RealSR [23]</td>
<td>w/o GT</td>
<td>3.4428/4.6513/7.3542</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>w/o GT</td>
<td>3.4428/4.6513/7.3542</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2.2. Qualitative Analyses

We also displayed the visual effects in Fig. 4 and 5, which are the qualitative results over DF2K and DPED datasets of ×4 upscaling factor. As the figures show, both of our results are clear and sharp, besides, we successfully restored some detailed textures while others didn’t make it. For Fig. 6 was utilized for revealing that, compared with other models, we can not only generalize to other unseen camera-captured kernels and but also indeed be able to generate more natural and realistic features which are essential for the real-world super-resolved images.

![Fig. 4. Visual results for ×4 amplification factor super-resolution in comparison with other methods on the DF2K dataset.](image)

![Fig. 5. Visual results for ×4 amplification factor super-resolution in comparison with other methods on the DPED dataset.](image)

![Fig. 6. The super-resolved results ((a) ESRGAN, (b) RankSRGAN, (c) ZSSR, (d) FSSR, (e) RealSR and (f) Ours) from the iPhone, BlackBerry, and Sony smartphones photos respectively in the DPED dataset proved that we could achieve the generalization ability and possess stochastic variation features representation.](image)

3.3. Ablation Study

To further demonstrate the availability of our strategies, we set up the ablation experiments. In Fig. 7, all indicators of referenced and non-referenced fields indicated that we can get the promotion more or less. And the variation trend also reveals that our prime improvements were attached to the perceptual-oriented metrics, confirming the expectation of gaining a better visual effect and perceptual experience.

![Fig. 7. The ablation study for proposed kernel augmentation (KA) and stochastic variation (SV) strategies. Wherein, notations + and – represent whether w/ or w/o this function. We regarded +/- as the baseline SISR model to calculate the differences (absolute values). For better testification, the DF2K dataset was selected to provide results both in pixel-oriented and perceptual-oriented metrics.](image)

4. CONCLUSION

In this paper, we mainly provided two strategies to promote the current real-world super-resolved results. For one thing, we took a kernel augmentation (KA) measure to ameliorate the unsatisfactory performance in generalization ability and robustness for the SISR model on realistic photos, since the limitation and unity from the most commonly used bicubic down-sampling operation kernel. For another, we utilized the stochastic variation (SV) produced by adding Gaussian noise into the generator’s output, leading to more authentic and natural features representation. At last, according to the quantitative and qualitative analyses, we could conclude that our methods conduce to generate images in possession of more natural textures and realistic details.
5. REFERENCES


