Super-Resolution Method for Reconstructing Street Images from Surveillance System based on Real-ESRGAN

Nguyen Quoc Toan
quoctoann3@gmail.com
Department of Electronic and Electrical Engineering,
Hongik University,
94 Wausan-ro, Mapo-gu,
Seoul, Republic of Korea

ABSTRACT

In many cities around the world, large sums of money are invested in surveillance camera systems, but few optimize the benefits and costs of those investments, and thus the overall impact of surveillance cameras on crime rates. In this paper, based on a technique named Real-ESRGAN applied to a practical restoration application that has been enhanced by the efficient ESRGAN. It is a super-resolution method that was developed in blind super-resolution to reinstate low-resolution street images with unknown and complicated degradations. It can be applied for security purposes in surveillance systems. Since video surveillance systems typically capture low-resolution images in many areas, the detection and identification of objects are sometimes required. This task's super-resolution is tough because image appearances vary depending on a variety of factors. The low resolution combined with poor optics is completely insufficient for identifying the subject of interest on the street, from a distance, in bad weather, or under any other limitations. Furthermore, to strengthen discriminator capability and create stable training dynamics, the U-Net discriminator was employed with spectral normalization. Hence, when compared to other experimental techniques, it can be demonstrated that this method delivers the best result. Experiment results show that super-resolution recovery of street images taken from a surveillance system is attainable with the following results: PSNR: 30.36dB and SSIM: 0.86.

KEYWORDS

computer vision, super-resolution, GAN, U-Net, Real-ESRGAN

ACM Reference Format:
Nguyen Quoc Toan. 2022. Super-Resolution Method for Reconstructing Street Images from Surveillance System based on Real-ESRGAN. In *Proceedings of Student Computing Research Symposium (SCORES'22).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

SCORES'22, October 6, 2022, Ljubljana, Slovenia

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Image resolution is a significant factor in calculating image quality. The better the resolution, the more detailed the information in the image, making it more robust for objects on street recognition tasks. Improving image resolution has always been an unstoppable pursuit of industry and academia. Real-ESRGAN [11] has been applied due to its significant improvements compared to other experimental methods. The ESRGAN [12] was extended by Real-ESRGAN, to restore general real-world LR images by combining training sets with a more practical degradation process.

Simply put, Real-ESRGAN extends the classical "first-order" degradation model to "high-order" degradation modeling techniques, i.e., the degradations are modeled with multiple repeated degradation processes, each of which is the classical degradation model. A second-order degradation process is used empirically to obtain a great harmony of simplicity and effectiveness. A recent paper [14] assumes a random shuffling strategy for synthesizing more practical degradations. But even so, it still includes a set amount of degradation processes, and it is unclear whether all the shuffled degradations are useful. High-order degradation modeling, on the other hand, is more adaptable and aims to imitate the real degradation generation process, *sinc* filters in the synthesis procedure are employed to simulate ringing and over-shoot artifacts. Because the degradation space can be much bigger than ESRGAN, training becomes tough. First, the discriminator must be more powerful to distinguish realness from complicated training output. Secondly, the discriminator's gradient feedback should be more precise for local information improvement. In Real-ESRGAN, a VGG-style discriminator was upgraded to a U-Net design [7][8][13]. Thirdly, the U-Net architecture and complex degradations increase training instability. To balance the training dynamics, spectral normalization (SN) regularization [6][8] was used. It is simple to train Real-ESRGAN and obtain stability of local detail advancement and artifact suppression with the dedicated improvements.

2 REAL-ESRGAN

2.1 Classical Degradation Model

Blind SR recovering high-resolution images from low-resolution images that have undergone unknown and intricate degradations. Based on the underlying degradation process, existing techniques can be divided into two categories: explicit modeling and implicit modeling. The classic degradation model [1][4] is widely used in explicit approaches [2][3][15] and includes blur, downsampling, noise, and JPEG compression. To generate the low-resolution input,

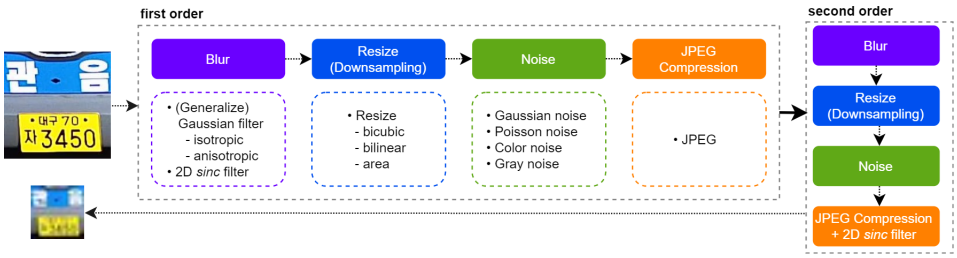


Figure 1: Illustration of High-order data generation Real-ESRGAN

the classical degradation architecture [1][4] is commonly utilized. In general, the ground-truth image y can be convolved with the blur kernel k first. Afterward, with scale factor r, a downsampling operation is applied. Adding noise n yields the low-resolution x. Lastly, JPEG compression is used because it is ubiquitous in real-world images. In general, D represents the process of degradation.

$$x = D(y) = [(y * k) \downarrow_r + n]_{JPEG}, \tag{1}$$

- **Blur.** Gaussian filters, both isotropic and anisotropic, are selected. Although Gaussian blur kernels are typically utilized to model blur degradation, they may still not accurately represent real camera blur. Gaussian blur kernels [5] and a plateau-shaped allocation implement more diverse kernel shapes are generalized as well.
- **Noise.** The additive Gaussian and Poisson noise types are applied. The probability density function of additive Gaussian noise is the same as the Gaussian distribution. The noise intensity is governed by the Gaussian distribution's standard deviation. Color noise occurs when unbiased sampled noise is present for each channel of RGB images. The Poisson distribution is initiated by Poisson noise. It is frequently used to estimate sensor noise caused by statistical quantum fluctuations, or variations in the photon flux sensed at a given exposure level. Poisson noise has a value proportional to image intensity, and noises at the pixel level are self - reliant.
- **Resize (Downsampling).** Downsampling is known as the resize operation. Nearest-neighbor interpolation, area resize, bilinear interpolation, and bicubic interpolation are all resize algorithms. Different resize operations yield different outcomes, some produce blurry images, while others may produce over-sharp ones with overshoot artifacts. To include more diverse and complex resize effects, a random resize operation from the options listed above. Nearest-neighbor interpolation is included because it exposes the misalignment issue and only deems the area, bilinear, and bicubic operations.
- **JPEG compression.** It is a prevalent lossy compression methodology for digital images. It decodes images to the YCbCr color space first, then downsamples the chroma channels. After that, the features are extracted into 8 x 8 blocks, and each block is converted with a 2D discrete cosine transform (DCT). [10] provides more information on JPEG compression algorithms. JPEG compression frequently presents unappealing block artifacts. A quality factor $q \in [0, 100]$ reflects the quality of compressed images, with a lower q indicating a higher compression ratio and lower quality.

2.2 High-order Degradation Model

When we use the above classical degradation model to generate training pairs, the trained model can handle some real-world samples. Nevertheless, it is still unable to handle some complex degradations in the real world, particularly unknown noises, and complex artifacts. This is due to the fact that synthesized low-resolution images still have a significant difference from realistic degraded images. To model more practical degradations, the classical degradation model was extended to a high-order degradation process. The classical degradation model only involves a limited number of fundamental degradations, which can be thought of as first-order modeling. Notwithstanding, real-world degradation is quite various and typically consists of a series of mergers such as imaging systems of cameras, image collection quality from video, and so on. In particular, the original image could be a very limited pixel image that the surveillance camera can get far away from the set up system, which inevitably contains degradations such as camera blur, sensor noise, and low resolution.

The classical first-order model could not model such a complex deterioration process. Therefore, a high-order degradation model is employed. An n-order model consists of n repeated degradation processes (as shown in Eq.2), in which each degradation process precedes the classical degradation model Eq.1 but with different hyperparameters. It should be highlighted that the term "high-order" is deployed differently here than it is in mathematical operations. It primarily refers to the time required to complete the same operation. [14] suggests that the random shuffling strategy encompasses repetitive degradation processes (e.g., double blur or JPEG). The key is the high-order degradation process which indicates that not all of the shuffled degradations are intended. To maintain a reasonable image resolution, the downsampling equation is altered with a random resize execution. Therefore, a second-order degradation process is employed, as it can remedy most real-world problems while remaining simple. The general flow of data generation stream is represented in Fig.1.

$$x = D^n(y) = (D_n \circ ... \circ D_2 \circ D_1)(y) \tag{2}$$

2.3 Ringing and overshoot artifact

Ringing artifacts commonly occur as spurious edges near sharp image transitions. They appear as bands near the edges. Overshoot

artifacts are frequently combined with ringing artifacts, which manifest as a higher jump at the edge transition. The primary source of these artifacts is that signal is bandlimited and lacks high frequencies. These artifacts are very common and are typically caused by a sharpening algorithm, JPEG compression, and so on. *sinc* filter assists in cutting off high frequencies and synthesizing ringing and overshoot artifacts for training sets, its filters are deployed twice: during the blurring process and at the end step of the synthesis. The arrangement of the last *sinc* filter and JPEG compression is randomized transferred to cover a larger degradation space because some images may be over-sharpened (with overshoot artifacts) and then JPEG compressed, while others may be JPEG compressed first and then sharpened. The equation of *sinc* is represented in Eq.3, where (i,j) represents the kernel coordinate, ω_c denotes the cutoff frequency, and J_1 means the first order Bessel operation of the first kind:

$$k(i,j) = \frac{\omega_c}{2\pi\sqrt{i^2+j^2}} J_1(\omega_c\sqrt{i^2+j^2}) \tag{3}$$

2.4 Networks and Training

2.4.1 ESRGAN. Firstly, the same generator (SR network) as ESRGAN [12] Fig.2 is used, i.e., a deep network with residual-in-residual dense blocks (RRDB). In conducting super-resolution with scale factors of x2 and x1, the extension with x4 ESRGAN architecture was represented. Since ESRGAN is a large network, the pixel-unshuffle (an inverse function of pixelshuffle [9]) was used before continuing to feed inputs into the main ESRGAN architecture to lower spatial size and increase channel size. As a result, most calculations are conducted in a smaller resolution space, which relieves GPU memory and the computational consumption of resources.

2.4.2 U-Net discriminator with spectral normalization (SN). Since Real-ESRGAN tackles a much bigger degradation space than ESRGAN, the existing discriminator architecture in ESRGAN is no longer appropriate. For complex training outputs, the discriminator in Real-ESRGAN necessitates more discriminative ability. It must ensure an accurate gradient responses for local textures in addition to discriminating global styles. The VGG-style discriminator in ESRGAN was enhanced to a U-Net architecture with skip connections, inspired by [8][13]. The U-Net generates realness values for each pixel that can provide the generator with detailed per-pixel responses. Meanwhile, the U-Net architecture and complicated degradations increase training instability. Regularization of spectral normalization [6] can aid in the stabilization of training dynamics. Furthermore, spectral normalization can help to reduce the oversharpness and annoyance caused by GAN training. Real-ESRGAN training can easily reach a better balance of local detail improvement and artifact suppression with these adjustments.

3 EXPERIMENT

3.1 Dataset

A brand-new dataset collected by HAIL was used for the experiment. It was collected by recording videos on streets in Seoul, South Korea with a Hanwha Techwin PNO-A9081RLP camera, then frames were extracted into images. There are 4500 images in total for use. The resolution is 4K (4096 x 2160). And 3500 crop images in the test set only included license plates, car brands, and text on the car for model evaluation. Fig.3 is samples from the proposed dataset.

Table 1: Results comparison between Real-ESRGAN, BSRGAN and Bicubic

Method	Bicubic	BSRGAN	Real-ESRGAN
PSNR (dB)	25.51	28.09	30.36
SSIM	0.71	0.76	0.86

3.2 Result

The luminance-based evaluation and comparison criteria, SSIM (Structural Similarity) and PNSR (Peak Signal-to-Noise Ratio) are used. To illustrate, two images were used for these evaluations. We refer to them as image 1 and image 2. Image 1 is the original degraded image from the test dataset, while image 2 is a reconstruction of image 1. SSIM is a method for calculating the similarity of two images. The SSIM values range from -1 to 1. PSNR is calculated to compare the quality of image 1 to image 2 which is calculated by using the mean squared error (MSE). MSE is a statistical concept that means that an estimator's mean square error is the mean of the squares of the errors, or the difference between the estimate and what is evaluated, lower value means better performance. On the contrary, the greater the value of SSIM and PSNR, the higher the quality of the reconstructed image followed by Eq.4. Y represents the ground truth (reference) and Y^* describes reconstructed images:

$$PSNR(Y,Y^*) = 10\log_{10}\frac{255^2}{MSE} \tag{4}$$

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(Y^*(i,j) - Y(i,j))^2 \tag{5}$$

The SSIM evaluation between patches P_{Y^*} and P_Y is formulated as:

$$SSIM(P_{Y^*},P_Y) = \frac{(2\mu_{PY^*}\mu_{PY}+c_1)(2\sigma_{Y^*}\sigma_{PY}+c_2)}{(\mu_{PY^*}^2+\mu_{PY}^2+c_1)(\sigma_{PY^*}^2+\sigma_{PY}^2+c_2)} \tag{6}$$

where $\mu_{PY^*}(\mu_{PY})$ and $\sigma_{PY}(\sigma_{PY^*})$ are the knowing and standard deviation of patch $P_{Y^*}(P_Y)$. $c1$ and $c2$ are minor constants. So, the mean score of patch-based SSIM over the image is SSIM (Y^*,Y).

4 CONCLUSION

In this paper, by applying Real-ESRGAN, a method for reconstructing low-resolution street images into recognizable images acceptable for recognition information tasks. The model achieved outstanding results (SSIM:0.86, PSNR:30.36dB). It proved that generating degraded real-life scenarios input play an extremely vital role in super-resolution models for street image recognition tasks. Real-ESRGAN performs greatly in the removal of artifacts and the restoration of texture details.

ACKNOWLEDGMENTS

I would like to express my heartfelt appreciation to HAIL (Hongik University - Artificial Intelligence Laboratory, Seoul, Republic of Korea), which is advised by Professor Seongwon Cho, for facilitating me in carrying out this research.

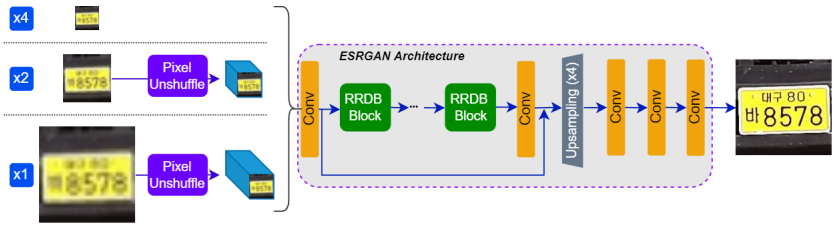
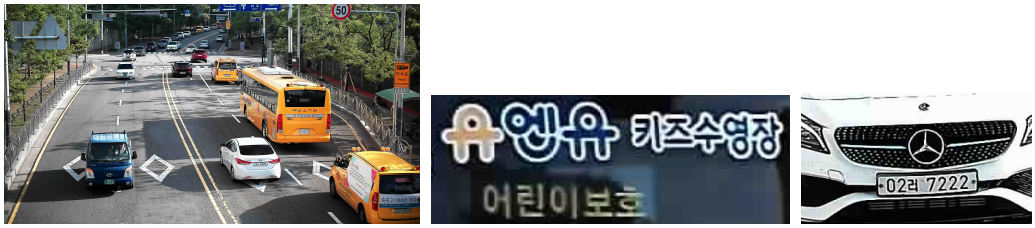


Figure 2: ESRGAN generator network. It first uses a pixel-unshuffle operation to decrease the spatial size and re-arrange information to the channel dimension for scale factors of x1 and x2.



(a) (b) (c)

Figure 3: a is sample from the train set, b and c are samples from the test set

Figure 4: Reconstructed image result comparing between Real-ESRGAN vs BSRGAN and Bicubic

REFERENCES

[1] Michael Elad and Arie Feuer. 1997. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing* 6, 12 (1997), 1646–1658.

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

[3] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. 2020. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems* 33 (2020), 5632–5643.

[4] Ce Liu and Deqing Sun. 2013. On Bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence* 36, 2 (2013), 346–360.

[5] Yu-Qi Liu, Xin Du, Hui-Liang Shen, and Shu-Jie Chen. 2020. Estimating generalized Gaussian blur kernels for out-of-focus image deblurring. *IEEE Transactions on circuits and systems for video technology* 31, 3 (2020), 829–843.

[6] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* (2018).

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[8] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. 2020. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8207–8216.

[9] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1874–1883.

[10] Richard Shin and Dawn Song. 2017. Jpeg-resistant adversarial images. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, Vol. 1. 8.

[11] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1905–1914.

[12] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*.

[13] Yitong Yan, Chuangchuang Liu, Changyou Chen, Xianfang Sun, Longcun Jin, Xinyi Peng, and Xiang Zhou. 2021. Fine-grained attention and feature-sharing generative adversarial networks for single image super-resolution. *IEEE Transactions on Multimedia* 24 (2021), 1473–1487.

[14] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4791–4800.

[15] Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2018. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3262–3271.