RUNet: A Robust UNet Architecture for Image Super-Resolution

Xiaodan Hu[1] Mohamed A. Naiel[1] Alexander Wong[1] Mark Lamm[2] Paul Fieguth[1]
[1]Vision and Image Processing Lab, University of Waterloo, Waterloo, ON, Canada
[2]Christie Digital Systems Canada Inc., Kitchener, ON, Canada
{x226hu, mohamed.naiel, a28wong, pfieguth}@uwaterloo.ca, mark.lamm@christiedigital.com

Abstract

Single image super-resolution (SISR) is a challenging ill-posed problem which aims to restore or infer a high-resolution image from a low-resolution one. Powerful deep learning-based techniques have achieved state-of-the-art performance in SISR; however, they can underperform when handling images with non-stationary degradations, such as for the application of projector resolution enhancement. In this paper, a new UNet architecture that is able to learn the relationship between a set of degraded low-resolution images and their corresponding original high-resolution images is proposed. We propose employing a degradation model on training images in a non-stationary way, allowing the construction of a robust UNet (RUNet) for image super-resolution (SR). Experimental results show that the proposed RUNet improves the visual quality of the obtained super-resolution images while maintaining a low reconstruction error.

1. Introduction

Modern state-of-the-art single image super-resolution (SISR) methods have been deep learning-based methods [1–5], which have demonstrated significant reconstruction quality improvements. For example, generative adversarial network-based SR methods [1, 2] have been able to generate realistic results, but these methods suffer from unstable training. On the other hand, convolutional neural network (CNN) based methods [3–5] have shown effectiveness in learning a nonlinear relationship between low and high resolution images. However, such methods [3–5] underperform when handling images with non-stationary degradations. One of the reasons is that a majority of these methods [3, 4] leverage a Bicubic down-sampling image degradation model for approximating the true degradation [6], which is not true in many practical scenarios such as projector resolution enhancement. Furthermore, such network architectures [3–5] are limited in their ability to learn complex non-stationary degradations.

Motivated by this, we propose a robust UNet (RUNet) architecture for image super-resolution to learn how to treat different image contents in a way that achieves better super-resolution results. More specifically, the proposed RUNet leverages long-range connections to improve learning capabilities, and leverages a degradation model based on spatially varying degradations that force the network to learn handling spatially non-stationary image degradations. Experimental results show that the proposed RUNet offers super-resolution images with improved visual quality while maintains a low reconstruction error.

2. Proposed Method

The proposed resolution enhancement scheme consists of a degradation module and a new UNet architecture as shown in Figure 1. During training, a set of input training images of high resolution are first downsampled by a factor of two in both directions and then blurred, at random, using a Gaussian filter. Next, every blurred image is upsampled by a factor of two in both x and y directions using the Bicubic interpolation for initializing the proposed network. For training the proposed network, every upsampled blurred image and its corresponding image at the original resolution are used. In testing, given a low-resolution input image, an upsampling operator by a factor of two is performed in both the x and y directions, and then the trained network is used to predict the enhanced high-resolution image.

2.1. Network Architecture

The proposed RUNet architecture consists of a number of convolutional layers, batch norms, ReLU activation functions, and tensor operations as shown in Figure 1. Unlike the conventional UNet [7] architecture, the left path shown in Figure 1 consists of a sequence of blocks each followed by a tensor addition operation to feed forward the same block input to the subsequent block, so-called residual block [4]. This allows the network to learn more complex structures. In order to efficiently upscale the low-resolution image, the sub-pixel convolutional layers [8] are

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC-CRD), Christie Digital Systems Inc., and the Ontario Centres of Excellence (OCE-VIPII).

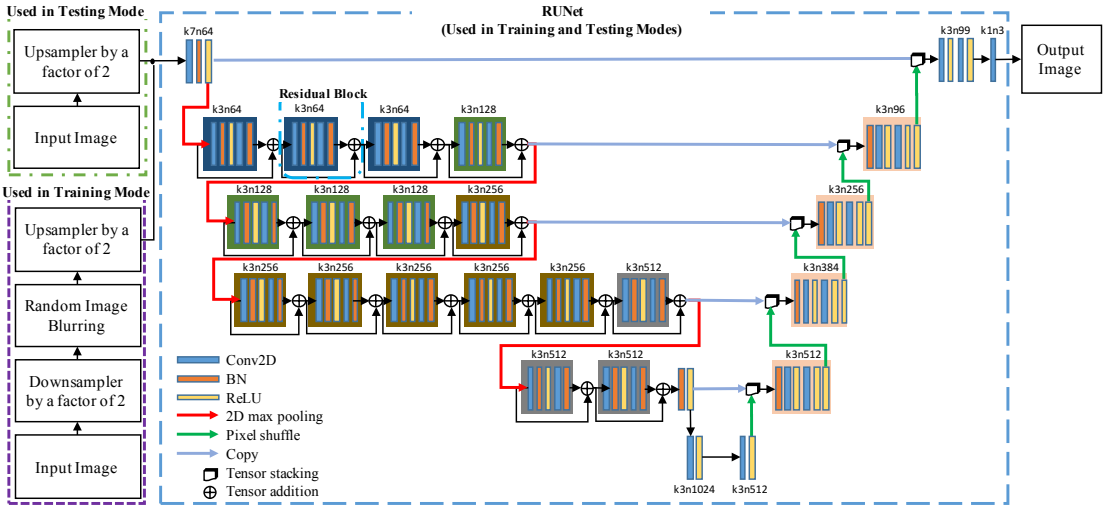

Figure 1. An overview of the proposed robust UNet (RUNet), where kz_1nz_2 represents a convolutional layer with kernel size of $z_1 \times z_1$ and with z_2 feature maps. The test of robustness is based on the presence of degradation in training.

Figure 2. Sample frames for the ten video categories in the Visual Projection Assessment Dataset (VPAD).

used for feature expansion in the expansive path, the right path shown in Figure 1. In order to achieve better perceptual performance, we use the perceptual loss function [3] during training as shown in the following section.

2.2. Perceptual Loss Functions

Recently, perceptual loss functions have been used for the tasks of image generation and super-resolution [1, 3, 9]. Rather than considering pixel-wise distance as in [4], the perceptual loss functions [3] map the predicted SR image $\hat{I}$ and the target image I_{HR} into a feature space and then measure the distance between the two mapped images in the feature space. Let $\Phi = \{\phi_j, j = 1, 2, ..., N_p\}$ denote a loss network [10] that extracts features from a given input image and consists of N_p convolutional layers, where $\phi_j(I)$ denotes a feature map of size $C_j \times H_j \times W_j$ obtained at the j^{th} convolutional layer for a given input image I, and $N_p = 5$ is used in this paper. Given a predicted image $\hat{I}$ and a target image I_{HR} fed into the network Φ, the feature distance $\mathcal{L}^j$ at the j^{th} layer can be computed as follows:

$$\mathcal{L}^j = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{I}) - \phi_j(I_{HR})\|_2^2 \qquad (1)$$

3. Experimental Results

3.1. Dataset

A Visual Projection Assessment Dataset (VPAD) is created for image and video resolution enhancement assessment. The VPAD dataset consists of a set of videos for various movies, documentary, sports, and TV news channels with the presence of moving and text-like regions. The video sequences were obtained from a wide range of open websites, such as [11] and [12], and Figure 2 shows a sample frame from each category. The dataset includes a total of 233 video sequences and is publicly released[1] to encourage further research and the assessment of projector resolution enhancement in practical scenarios. More specifically, this dataset includes the following ten categories: Action, Comedy and Romance, Documentary, Fantasy, Graphics and Animation, Horror, News, Sports, TV Shows, and TV Episodes. The videos from the same category share some common features, such as similar background or contents.

3.2. Discussion

The proposed RUNet is evaluated on the VPAD video dataset for 2× super-resolution, and compared with the performance of the Bicubic interpolation and the baseline UNet [7] without using the proposed degradation model. Table 1 summarizes the quantitative results, and example qualitative results are shown in Figure 3. Although it is shown from Table 1 that the proposed RUNet offers the lowest PSNR and SSIM values and the highest MSE value, it can be clearly observed from the example qualitative results shown in Figure 3 that the proposed RUNet can offer significantly improved super-resolution quality with noticeably sharper details than that provided by the other tested methods. This observation suggests the need of developing new evaluation metrics that can assess the SR image enhancement techniques different than the existing metrics. The same conclusion was drawn in past studies when evaluating similar deep-network-based super-resolution methods as in [1, 3, 9], where the use of perceptual loss function led to significant reductions in SSIM and PSNR scores of the reconstructed images similar to what we observed in Table 1.

[1] URL: uwaterloo.ca/vision-image-processing-lab/research-demos/vip-vpad

Figure 3. Example qualitative results for the proposed RUNet, Bicubic Interpolation, and UNet [7] without degradation model (baseline) on four low resolution images sampled from four different VPAD sequences. The proposed RUNet produces super-resolution images with sharper edges and finer details while retaining the quality of un-blurry regions in low-resolution images.

Table 1. Quantitative comparison among the proposed method, the baseline UNet architecture [7] without degradation, and Bicubic interpolation. Consistent with past studies that leveraged perceptual loss [1, 3, 9], it is observed that standard metrics fail to capture perceptual quality of image super-resolution.

Method	SSIM	MSE	PSNR
Bicubic Interpolation	0.800	0.008	25.569
UNet [7] without degradation	0.760	0.016	23.443
RUNet	0.736	0.020	22.753

References

[1] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, 2017.

[2] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. CVPRW*, 2018.

[3] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016.

[4] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. CVPRW*, 2017.

[5] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. CVPR*, 2018.

[6] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. CVPR*, 2018.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015.

[8] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. CVPR*, 2016.

[9] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. CVPR*, 2018.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[11] J. J. Quinlan and C. J. Sreenan, "Multi-profile ultra high definition (UHD) AVC and HEVC 4K dash datasets," in *Proc. ACM-MMSYS*, 2018.

[12] https://archive.org, Last retrieved Mar. 15th, 2019.