HNSR: HIGHWAY NETWORKS BASED DEEP CONVOLUTIONAL NEURAL NETWORKS MODEL FOR SINGLE IMAGE SUPER-RESOLUTION

*Ke Li, Bahetiyaer Bare, Bo Yan**

Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, China

Bailan Feng, Chunfeng Yao

Noah's Ark Laboratory, 2012Labs Huawei Technologies Co., Ltd., Beijing, China

ABSTRACT

Convolutional neural networks (CNNs) have been widely used in computer vision community. Single image super-resolution (SISR) is a classic computer vision problem, which aims to output a high-resolution image from a low-resolution one. In recent years, CNNs-based SISR methods emerged and achieved a performance leap. In this paper, we present a highly accurate deep CNNs model for SISR. Inspired by the ideas in highway networks, we propose a highway unit and cascade highway units to ensemble our model. Furthermore, we employ structural similarity index (SSIM) as a part of loss function to enhance the accuracy of trained deep CNNs model. Experimental results show that our proposed model outperforms other state-of-the-art methods.

Index Terms— Super-resolution, Convolutional neural networks, Deep learning, Highway networks

1. INTRODUCTION

Single image super-resolution (SISR), which aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) one, is a classic computer vision problem. HR images have higher pixel density, more detailed information and delicate picture quality. In order to obtain HR images, the most direct approach is to use HR camera. However, in practical applications, due to the production process and engineering costs, many occasions do not use HR camera for image signal acquisition. Therefore, the adoption of super-resolution (SR) technology to obtain HR images from LR images has a certain application requirements.

Many SR methods have been proposed in computer vision community. Early methods such as bicubic interpolation and Lanczos resampling [1] try to reconstruct a HR image when only using the information of a LR one. These methods can not solve the problem well because SR is an ill-posed problem due to each pixel in LR has multiple solutions when mapping on pixels in HR images. Such a problem should use strong prior to constrain the solution space. Recent state-of-the-art methods mostly adopt the example-based [2] strategy for learning the strong prior knowledge.

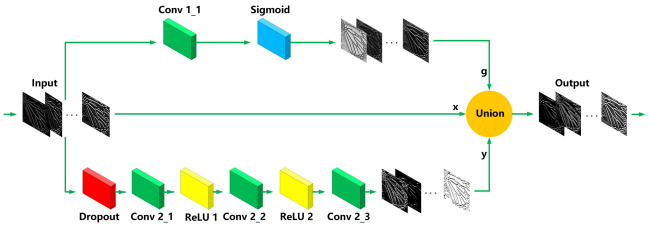

Fig. 1. Highway unit. We cascade highway units to form our deep CNNs model. The highway unit consists of a dropout layer, two ReLU layers, four convolutional layers, a sigmoid layer, and an union layer. The union layer outputs the combination of input signal x, output of lower channel y, and output of upper channel g.

In recent years, with the success of convolutional neural networks (CNNs) [3] in the computer vision community, CNNs-based SR methods [4, 5, 6, 7] have emerged. These methods achieved performance leaps compared to previous methods. Among them, SRCNN [4] and VDSR [5] are typical methods. SRCNN proved that CNNs can be used to learn a mapping between LR and HR images and achieved an end-to-end method, which is trained directly from image patches without any extracted features. In [5], Kim *et al*. proposed a very deep CNNs model for SISR (VDSR) inspired by VGG-net [8] used for ImageNet [9] classification. VDSR demonstrated that deep CNNs model can achieve better performances.

According to the experiments conducted in [10], although SRCNN [4] and VDSR [5] successfully applied deep learning techniques to solve the SR problem and achieved state-of-the-art performance, they suffer from inaccuracy problem of mean squared error (MSE), which is taken as loss function.

*This work was supported in part by NSFC (Grant No.: 61522202; 61772137) and Huawei Technologies Co., Ltd (Contract No.: YBN2017050058).

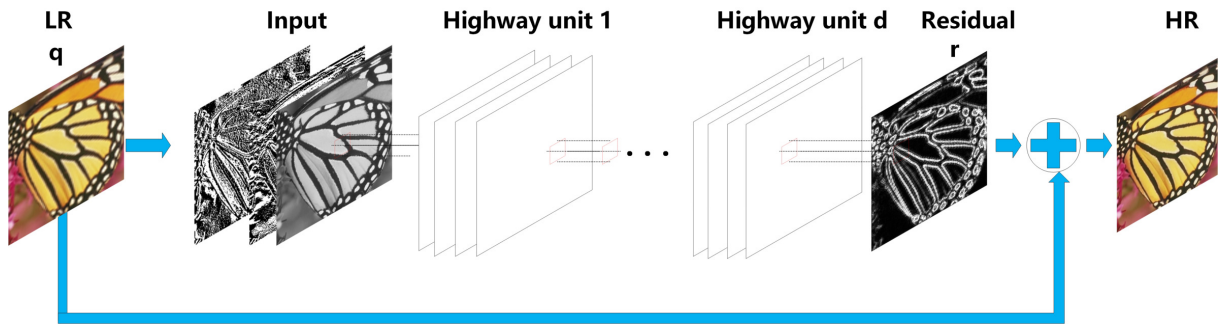

Fig. 2. Our Network Structure. We cascade highway units to obtain our network structure. HR images are predicted from LR images through our deep CNNs model. The output HR image is obtained by combining the input LR image q and the predicted residual image r.

Moreover, the performance of CNNs-based methods can be further improved by designing more appropriate models.

In this paper, in order to improve the performance of CNNs-based SR methods, we propose a deep CNNs model for SISR. Inspired by highway networks [11], we propose a highway unit and cascade highway units to form our deep CNNs model. We name the proposed model as Highway Networks Super Resolution (HNSR). The input of the proposed HNSR consists of vertical gradient map, horizontal gradient map and luminance map. Different from other CNNs-based SR methods, we employ structural similarity index (SSIM) [12] as a part of loss function. The performance of HNSR is verified by benchmark datasets. Experimental results show that HNSR outperforms other CNNs-based SR methods and achieves state-of-the-art performance.

The main contributions of our work are: i) We proposed a highway unit (Shown in Figure 1) based on the ideas in highway networks [11] and cascade highway units to form our deep CNNs model; ii) In order to enhance the accuracy of proposed HNSR, we employed SSIM [12] as a part of loss function. With the aid of SSIM, the proposed HNSR is able to achieve better performance than other CNNs-based SR methods.

The rest of the paper is organized as follows. We describe our proposed model in Section 2. In Section 3, we present experimental results. Finally, we draw conclusions in Section 4.

2. PROPOSED METHOD

2.1. Proposed Network

For the SISR, we proposed a highway unit inspired by highway networks in [11] and cascade highway units to form our network structure. In Figure 1, we demonstrated the structure of proposed highway unit. The input signal x passes through the upper and lower two channels. The lower channel includes three convolutional layers, each of which has 32 kernels size of 5×5, a dropout [14] layer, and two Rectified Linear Unit (ReLU) [15] layers. The upper channel includes a convolutional layer, which also have 32 kernels size of 5×5 , and a sigmoid layer. As shown in Figure 1, the output of the upper channel g, the input signal x, and the output of the lower channel y will pass through a union layer. We define the output of the union layer as:

$$Output = g \times y + (1 - g) \times x, \tag{1}$$

where g denotes the output of upper channel; x denotes the input signal; and y denotes the output of lower channel.

The framework of our proposed network is demonstrated in Figure 2. For an input image, we use bicubic interpolation to upsample it to target resolution and obtain LR image q. In order to faster convergence, our network preprocesses q and obtains three images: horizontal gradient map, vertical gradient map, and luminance map. Then, the proposed network combines these three maps to create a new input signal. The combined new signal passes through several highway units to obtain a residual image r. Finally, the output HR image is obtained by combining the input LR image q and predicted residual image r.

2.2. Loss Function

We designed a new loss function for CNNs-based SISR. Previous CNNs-based methods [4, 5, 6, 7] employed MSE as loss function. Although MSE can well predict the loss between ground truth and predicted image, it has limitations. In some cases [12], MSE fails due to not considering structure similarity of the compared images . SSIM [12] is proposed to address the limitations of previous methods such as MSE or PSNR. SSIM considers three important factors: luminance, contrast, and structure similarity. By taking structure similarity into consideration, SSIM performes better than MSE or PSNR.

The Proposed new loss function is a weighted sum of MSE and SSIM. The new loss function is defined as,

Table 1. Average PSNR\SSIM on Set5, Set14, B100, and Urban100.

Benchmark		Bicubic PSNR\SSIM	SRCNN [4] PSNR\SSIM	IA [13] PSNR\SSIM	VDSR [5] PSNR\SSIM	Small HNSR PSNR\SSIM	HNSR-SSIM PSNR\SSIM	Large HNSR PSNR\SSIM
Set5	2	33.66\0.9299	36.66\0.9542	37.37\0.9582	37.53\0.9587	37.64\0.9594	37.38**0.9601**	**37.80**\0.9597
	3	30.39\0.8682	32.75\0.9090	33.43\0.9195	33.66\0.9213	33.76\0.9223	33.25\0.9242	**34.03\0.9248**
	4	28.42\0.8104	30.48\0.8628	31.05\0.8781	31.35\0.8838	31.32\0.8824	30.47\0.8864	**31.65\0.8887**
Set14	2	30.24\0.8688	32.42\0.9063	32.83\0.9109	33.03\0.9124	33.12\0.9134	32.84**0.9166**	**33.40**\0.9153
	3	27.55\0.7742	29.28\0.8209	29.63\0.8291	29.77\0.8314	29.77\0.8318	29.29**0.8396**	**30.00**\0.8358
	4	26.00\0.7027	27.49\0.7503	27.85\0.7640	28.01\0.7674	27.88\0.7651	27.36**0.7764**	**28.22**\0.7729
B100	2	29.56\0.8431	31.36\0.8879	31.79\˜	31.90\0.8960	31.88\0.8965	31.57**0.9019**	**32.11**\0.8987
	3	27.21\0.7385	28.41\0.7863	28.76\˜	28.82\0.7976	28.76\0.7969	28.25**0.8086**	**28.95**\0.8014
	4	25.96\0.6675	26.90\0.7101	27.25\˜	27.29\0.7251	27.19\0.7226	26.63**0.7370**	**27.38**\0.7291
Urban100	2	26.88\0.8403	29.50\0.8946	-	30.76\0.9140	30.67\0.9144	30.78\0.9195	**31.43\0.9211**
	3	24.46\0.7349	26.24\0.7989	-	27.14\0.8279	26.93\0.8237	26.88\0.8353	**27.49\0.8362**
	4	23.14\0.6577	24.52\0.7221	-	25.18\0.7524	24.90\0.7430	24.79\0.7604	**25.41\0.7608**

$$loss = MSE + \alpha SSIM, \tag{2}$$

where α is a weight of SSIM. We took different values of α and found that $\alpha = 0.01$ performs well in our experiment.

We present the gradient calculation process of our loss function below. Because it is easy to calculate the gradient of MSE, we only show the gradient calculation process of SSIM. The equation of SSIM can be expressed as,

$$SSIM(X,Y) = \frac{I_1}{I_2} \cdot \frac{S_1}{S_2}, \tag{3}$$

where $\frac{I_1}{I_2}$ denotes luminance measure and $\frac{S_1}{S_2}$ denotes contrast measure multiply by structure similarity [12]. In addition, X and Y denote two compared images. The expressions of I_1, I_2, S_1, and S_2 are given below:

$$I_1 = 2(c * X) \cdot (c * Y) + C_1, \tag{4}$$

$$I_2 = (c * X)^2 + (c * Y)^2 + C_1, \tag{5}$$

$$S_1 = 2(c * (X \cdot Y) - (c * X) \cdot (c * Y)) + C_2, \tag{6}$$

$$S_2 = c*(X \cdot X) - (c * X)^2 + c*(Y \cdot Y) - (c * Y)^2 + C_2, \tag{7}$$

where c denotes a Gaussian kernel with a window size of 11×11 and a variance of one, C_1 and C_2 denote constant values that avoid instability when I_2 and S_2 are very close to zero. C_1 and C_2 are set to 0.01. According to the chain rule, the gradients of SSIM with respect to X can be split to two parts: i) the gradients of SSIM with respect to I_1, I_2, S_1, and S_2, ii) and the gradients of I_1, I_2, S_1, and S_2 with respect to X. We only show the gradient calculation process of SSIM with respect to I_1 and I_1 with respect to X because other parts of SSIM have similar calculation process. The expressions of two gradients are:

$$\frac{\partial SSIM(X,Y)}{\partial I_1} = \frac{S_1}{I_2 \cdot S_2}, \tag{8}$$

$$\frac{\partial I_1}{\partial X} = 2(c * X) \cdot \frac{\partial (c * X)}{\partial X}, \tag{9}$$

where $*$ denotes the convolution operation. The $\frac{\partial (c*X)}{\partial X}$ is easy to calculate and we do not present it in our paper.

3. EXPERIMENT

3.1. Datasets

Training dataset In our experiment, we train a small network and two large networks for each scale, respectively. Small networks are trained on 91 images from Yang et al. [17] and large networks are trained on 291 images with additional 200 images from Berkeley Segmentation Dataset [18]. In addition, rotation is used on training images to acquire more training data. We separately crop 36×36 patches with overlaps from LR images and corresponding HR images as inputs and labels. In some experiments, we set small overlapping areas to reduce training data and accelerate training progress.

Testing dataset For benchmark, we use four datasets. Datasets "Set5" [19] and "Set14" [20] are generic datasets used in other works [4, 5, 21, 13]. "Urban100" [16] is a very interesting database, since it contains many challenging images failed by many of the existing methods. "B100" is the testing set of Berkeley Segmentation Dataset [18].

3.2. Training Details

We train three networks for each scale. The first one is a small network with 3 highway units (Small HNSR) and has 14 convolution layers. The second one is a large network with 7 highway units (Large HNSR) and has 30 convolution layers. The last one is a large network which only takes SSIM as the loss function (HNSR-SSIM). In our experiments, the batch size is set to 64, weight decay is set to 0.0001, momentum is set to 0.9, and dropout ratio is set to 0.2.

We train all the networks for 30 epoches. The learning rate is initialized to 0.1 and divided by 10 after every 10 epoches. We use Adjustable Gradient Clipping [5] to limit the parameters' gradient in $[-\theta/\gamma, \theta/\gamma]$, where γ denotes current learning rate and θ is set to 0.1.

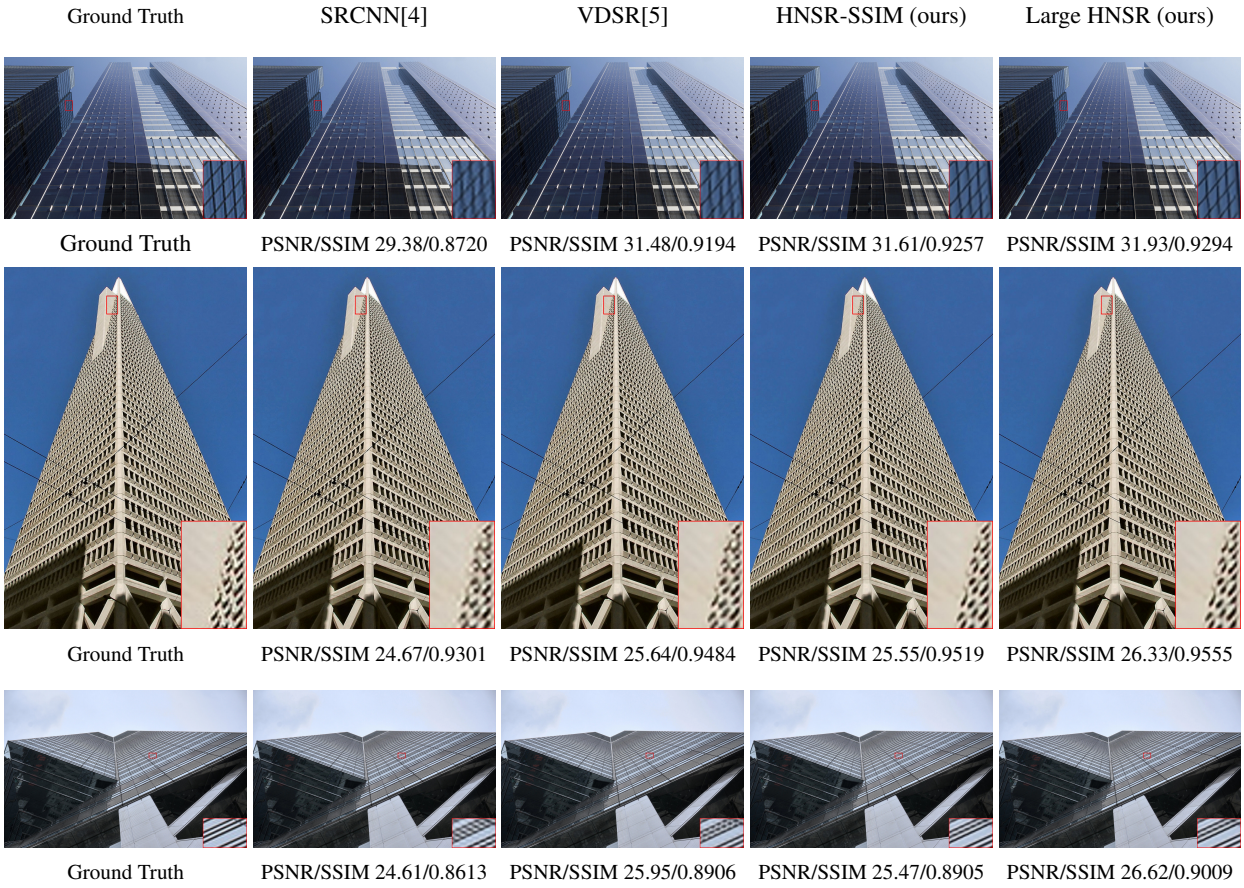

Fig. 3. Visual results for $\times 2$ on dataset "Urban100" [16]. HNSR-SSIM (ours) and Large HNSR (ours) both have 7 highway-units and contain the similar number of parameters as the VDSR [5]. HNSR-SSIM uses only SSIM as the loss function. And Large HNSR employs MSE and SSIM as the loss function and sets the SSIM weight as 0.01.

3.3. Comparison

For comparison with other methods, we follow the framework of Timofte *et al*. [22]. In this framework, we apply our method to luminance component and compute both PSNR and SSIM on luminance component. The color components for HR images are acquired from applying bicubic interpolation on LR images.

Table 1 shows the average PSNR and SSIM performance on Set5, Set14, B100 and Urban100 for magnification factors $\times 2$, $\times 3$ and $\times 4$ of our networks in comparison with SRCNN [4], IA [13], VDSR [5]. As shown in the table, our proposed large HNSR gets higher PSNR and SSIM values than previous methods in these datasets. And HNSR-SSIM almost achieves the best results when SSIM is used as the evaluation standard. Large HNSR and HNSR-SSIM have the same number of parameters as VDSR. In addition, Small HNSR achieves similar results to VDSR and only have 0.3 million parameters which is about half of VDSR's.

In Figure 3, we show the comparison of different methods' results on Urban100 for magnification $\times 2$. As shown in this figure, HNSR and HNSR-SSIM generally restore more sharp details and more clear edges. In addition, as shown in Table 1, HNSR-SSIM has a much lower PSNR and a higher SSIM than VDSR. And obviously HNSR-SSIM has a better visual effect than VDSR as shown in figure 3. Therefore, it can be observed that SSIM is a better measure than PSNR and introducing SSIM into the loss function of deep CNNs-based SR is reasonable.

4. CONCLUSION

In this paper, we proposed a deep CNNs model for single image super-resolution inspired by highway networks. We proposed a highway unit, which is cascaded to form our easy trainable deep convolutional neural networks. In addition, we introduced SSIM into the loss function to make results more consistent with human vision system. Promising high-resolution image reconstruction results are achieved using the trained deep CNNs model. Experimental results confirmed a performance leap relative to compared state-of-the-art methods.

5. REFERENCES

[1] Claude E. Duchon, "Lanczos filtering in one and two dimensions.," *Journal of Applied Meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[2] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang, "Single-image super-resolution: a benchmark," in *European Conference on Computer Vision*. Springer, 2014, pp. 372–386.

[3] Y LeCun, B Boser, J Denker, and D Henderson, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[6] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645.

[7] Gernot Riegler, Samuel Schulter, Matthias Ruther, and Horst Bischof, "Conditioned regression models for non-blind single image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 522–530.

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[10] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.

[11] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[12] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] Radu Timofte, Rasmus Rothe, and Luc Van Gool, "Seven ways to improve example-based single image super resolution," in *Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition*, 2016, pp. 1865–1873.

[14] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[15] Vinod Nair and Geoffrey E. Hinton, "Rectified linear units improve restricted boltzmann machines vinod nair," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[16] J. B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

[17] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2001, vol. 2, pp. 416–423 vol.2.

[19] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[20] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[21] Radu Timofte, Vincent De Smet, and Luc Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.

[22] Radu Timofte, Vincent De Smet, and Luc Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1920–1927.