PERCEPTUAL QUALITY PRESERVING IMAGE SUPER-RESOLUTION VIA CHANNEL ATTENTION

Wei-Yu Lee[†] *Po-Yu Chuang*[†] *Yu-Chiang Frank Wang*[⋆]

[†]MOXA Lab, MOXA Inc., Taipei, Taiwan
[⋆]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

ABSTRACT

Generative Adversarial Network (**GAN**) has been widely applied on Single Image Super-Resolution (**SISR**) problems. However, there can be quite a variability in the results from the GAN-based methods. In some cases, the GAN-based methods might cause structure distortion, which can be easily distinguished by human beings, especially for artificial structures, because the methods only focus on the perceptual quality of the whole image. On the other hand, PSNR-oriented methods can prevent structure distortion but with overly smoothed context. To overcome these problems, we propose a deep neural net refiner for SISR methods, not only improving perceptual quality but also preserving context structures. In the experiments, our model qualitatively and quantitatively performs favorably against the state-of-the-art SISR methods.

Index Terms— Super-Resolution, Channel Attention, Generative Adversarial Networks

1. INTRODUCTION

Single Image Super-Resolution (**SISR**) is a technology that aims to reconstruct a high-resolution (**HR**) image from a low-resolution (**LR**) image. SISR is an ill-posed problem [1]. The texture details in the reconstructed image is typically absent. Two methods try to optimize reconstructed results in different ways: pixel-wise differences minimization and perceptual quality optimization. Pixel-wise differences minimization, or called PSNR-oriented methods [2–6], are dedicated to minimize the pixel-wise differences between reference and reconstructed images. Although these methods derived outstanding performance on PSNR and Structure Similarity (**SSIM**), the results might have trouble satisfying Human Visual System (**HVS**).

The other approach focused on perceptual quality [7–10]. Ledig et al. [1], and Sajjadi et al. [11] used Generative Adversarial Networks (**GAN**) [12] to reconstruct texture details to make the images look more photo-realistic. Wang et al. [13] introduced Residual-in-Residual Dense Block (**RRDB**) for the generator, which has higher capacity and is easier to train. Although these methods produced better perceptual quality,

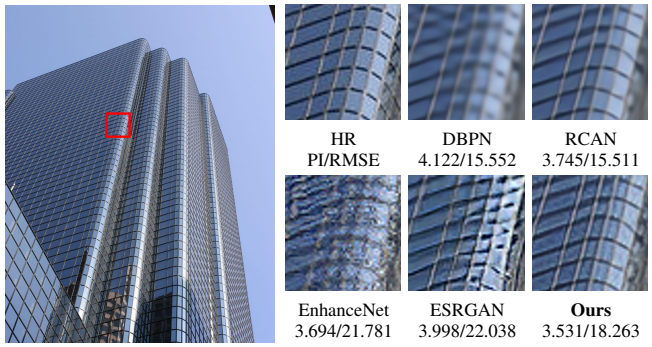

Fig. 1: Qualitative comparisons of our method with DBPN [6], RCAN [5], EnhanceNet [16], and ESRGAN [13] on $\times 4$ super-resolution. Our result is derived by using RCAN results as our input.

these methods easily cause structure distortion, and generate unpleasing artifacts. According to the observations from 2018 PIRM challenge report [14], GAN-based methods generally derived even worse perceptual quality on artificial structures, such as buildings, than PSNR-oriented methods. One possible reason is that GAN-based methods tend to overly enhance edges to make generated images look more ”real” [15]. This approach might work on some natural images, such as animal fur, but can be easily aware when it is applied on artificial structures, which are generally neat and tidy.

We observe PSNR-oriented methods can preserve context structure, and the results from GAN-based methods can better fulfill HSV. Taking the advantages from both methods, we propose a refiner for SISR methods. This refiner aims to overcome the overly smoothed problem of PSNR-oriented methods, and also preserves its advantage on structure restoration. We use the results from general SISR methods as input, which makes proposed refiner focus on the inadequate parts of the input. Inspired by [5], we use Channel Attention (**CA**) on the generator to choose better feature maps to fulfill different input contexts. CA is also applied to the discriminator to make it pay attention to more important inadequate contexts. Since we would like to preserve the structure, which has been restored by the SISR methods, we apply an identity mapping

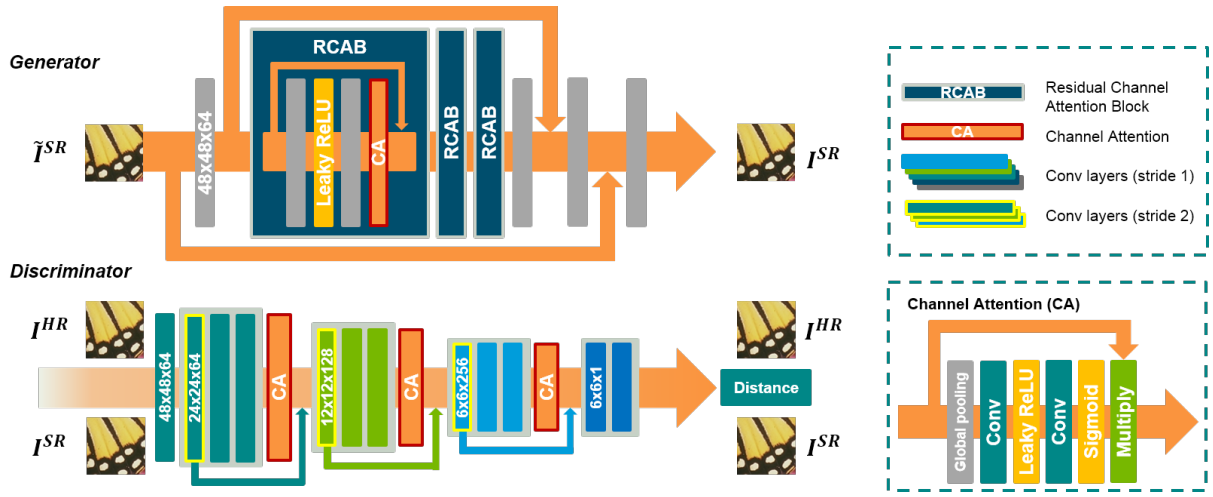

Fig. 2: Framework of our proposed refiner. The input $\tilde{I}^{SR}$ is an image from a general SISR network, and the output I^{SR} is our result. I^{HR} represents the high-resolution image.

shortcut [17] on the generator. By doing so, the refiner can adaptively adjust the reconstruction structure and perceptual quality. In addition, we use the feature maps of our discriminator as part of our perceptual loss. This strategy not only allows our discriminator to focus on the regions that we actually care about, it also encourages our generator to produce patterns that satisfy perceptual quality. We use Perceptual Index (**PI**) and Root Mean Square Error (**RMSE**) as our benchmark [14]. The qualitative and quantitative results show that our proposed method performs favorably against the state-of-the-art methods.

2. PROPOSED METHOD

2.1. Brief Review of Channel Attention

Reconstructing the SISR image is an ill-posed problem. The distribution of natural feature patches might exist in various manifolds. CNN models try to find a universal mapping from low-resolution patches I^{LR} to high-resolution patches I^{SR}. Wang et al. proposed RCAN [5], which used channel attention (**CA**) to adaptively weight the features maps by considering interdependencies among the channels. The main idea of this mechanism is to make the network focus on more informative features according to different input images. However, they only applied CA on LR images before performing SR. Thus, image structures in SR outputs might not be properly preserved.

2.2. Proposed Architecture

An overview of our proposed network is shown in Fig. 2. The input of our framework $\tilde{I}^{SR}$ is an image that has been recovered by a general SISR network, and the output I^{SR} is our results. Based on Wasserstein Generative Adversarial Network (WGAN) [18,19], we apply CA on our generator and discriminator to compensate the missing informative details of input SR images, and to resolve the perceptual quality problem. In addition, in order to preserve the context from inputs, we also use an identity mapping shortcut from the input to the last convolution layer to maintain the main structure of the images and make the network pay more attention to the residual differences between the inputs and the reference images.

As illustrated in Fig. 2, our CA uses channel-wise global average pooling to compress the feature maps $F : \{f_1, f_2, \ldots, f_n\}$ into n weightings, and a simple gating mechanism with leaky ReLU and Sigmoid activation function is applied to generate the final weightings $W : \{w_1, w_2, \ldots, w_n\}$. The n^{th} feature map is derived by multiplying the weightings respectively:

$$f'_n = W_n \cdot f_n \tag{1}$$

In our discriminator, we use stride to downscale the feature maps and apply CA on multi-scale to preserve the features and the valuable components. In our generator, we only apply single scale CA on the high-resolution feature maps to focus on the informative features. Different from RCAN, our generator dedicates to refine the fine scale images. Because the lower scale context has been reconstructed well by the SISR methods, our refiner only needs to focus on the missing fine details of the SR images. We show the qualitative comparisons of aggregated feature maps before and after being weighted by CA, and show the multi-scale quantitative comparisons in Section 3.

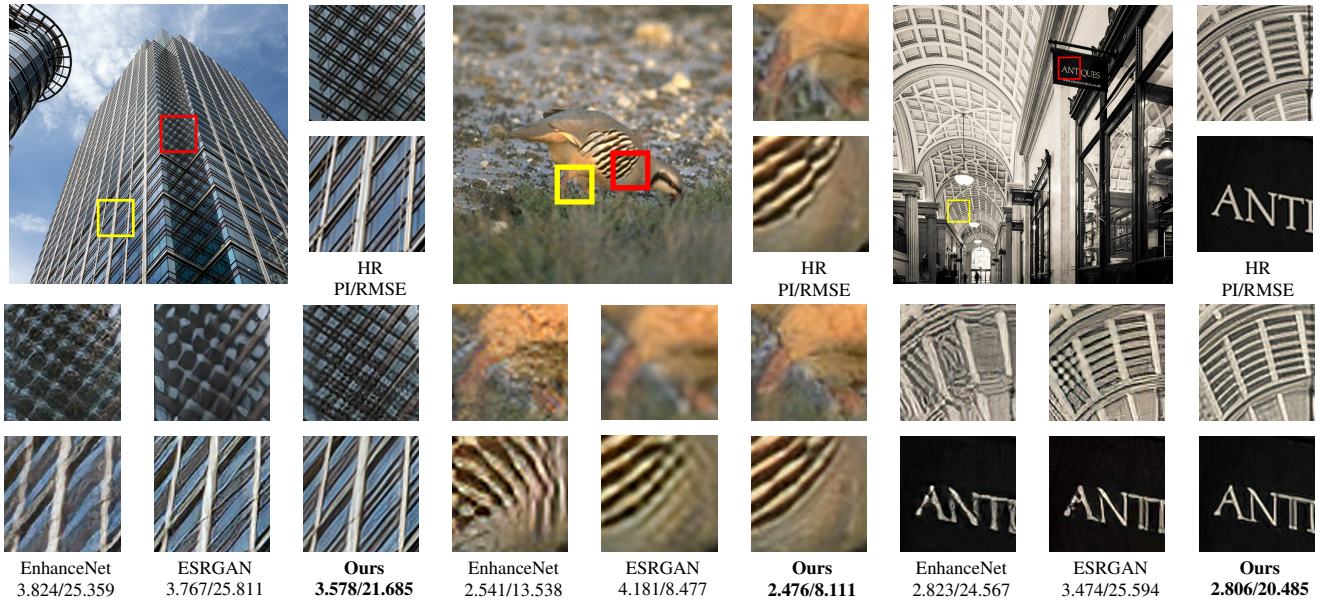

Fig. 3: Qualitative comparisons of EnhanceNet [16], ESRGAN [13], and our method on $\times 4$ super-resolution. Our results are derived by using RCAN results as our input.

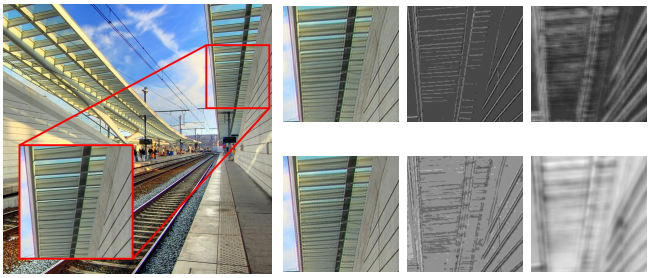

Fig. 4: Qualitative comparisons of aggregated feature maps. (a) High-resolution image, (b) Super-resolution images, (c) Feature maps of the generator, (d) Feature maps of the discriminator. The first and second row show the results with/without channel attention respectively.

2.3. Objective Function

In our generator, we simply choose Mean Square Error (**MSE**) term as the regularizer $L_2(G)$ because the outputs are the SR results with channels properly attended. In addition, inspired by SRGAN [1], we add a perceptual loss term as another regularizer $L_p(G, D)$. Instead of using VGG loss [1], we reuse the output of the discriminator D, which represents a high-level feature of the images, to be our perceptual loss, which itself is a MSE loss of a certain layer output of the discriminator (after activation function). The θ_i indicates the i^{th} layer of the feature map of our discriminator.

$$L_2(G) = E_{I^{HR},\tilde{I}^{SR}}[(I^{HR} - G(\tilde{I}^{SR}))^2] \quad (2)$$

Dataset	Methods (PI/RMSE)		
	EnhanceNet	ESRGAN	Ours
Set5	3.865/9.879	3.871/7.919	3.759/7.750
Set14	3.053/15.495	2.916/15.139	2.987/13.126
BSD100	2.922/16.887	2.489/16.498	2.461/14.824
Urban100	3.654/19.751	3.770/18.939	3.458/16.295

Table 1: Quantitative comparisons of EnhanceNet [16], ESRGAN [13], and our method on $\times 4$ super-resolution. Our results are derived by using RCAN [5] results as our input. Red color indicates the best performance.

$$L_p(G) = E_{I^{HR},\tilde{I}^{SR}}[(\theta_i(I^{HR}) - \theta_i(G(\tilde{I}^{SR})))^2] \quad (3)$$

$$L_{total} = L(G, D) + \lambda_1 L_2(G) + \lambda_2 L_p(G) \quad (4)$$

We optimize the total loss function L_{total} in an alternative manner to solve the adversarial min-max problem. $L(G, D)$ is the WGAN loss, and the coefficients λ_1 and λ_2 in our formula are two fixed values in our proposed models.

3. EXPERIMENTS

3.1. Model Configuration

The input images are from RCAN [5] without any modification, and the patches have size 48×48 with RGB channels. We use three residual channel attention blocks [5] with an identity mapping shortcut as our generator. The discriminator applies three CA blocks with three skip-connections as shown in Fig. 2. The convolution kernels' size are 5×5 with 4 different depth, and we downscale the size with stride 2. The 3^{rd}

CA	IS	PI/RMSE
√	√	3.458/16.285
√	×	3.476/16.292
×	√	3.527/16.582
×	×	3.628/16.715

Table 2: Overall quantitative comparisons for showing the effect of each components on Urban100. **CA**: Channel attention on the discriminator. **IS**: Generator shortcut from the input to the last convolution layer. Red color indicates the best performance.

layer output is used to estimate our proposed perceptual loss.

We use a high quality image datasets DIV2K [20] as our training set, and evaluate our proposed model on various common datasets. The test datasets: Set5, Set14, BSD100 and Urban100, are performed with $\times 4$ scale factor between LR and HR. The minibatch size is 16, and we train our network with ADAM optimizer by setting $\beta_1 = 0.9$, $\beta_1 = 0.999$, $\epsilon = 10^{-8}$. The learning rate is initialized as 10^{-4} to train 10^5 iterations. We update the generator once in every 5 iterations on the discriminator, and the coefficients of the regularizer in generator λ_1 is 10 and λ_2 is 0.25.

3.2. Quantitative and Qualitative Results

In this section, we review the quantitative and qualitative performance to the other methods. Table 1 shows quantitative comparisons on $\times 4$ SR images. Our results are derived by using RCAN restored images as the input, and the results of other methods are released by the authors. Our proposed method has better performance in most of cases, especially in Urban100. There are abundant artificial structures in this dataset, and it is hard to be reconstructed by the previous GAN-based methods. Our method shows superior performance both in RMSE and PI.

In Fig. 3, we show the qualitative comparisons on $\times 4$ SR images. We can observe that the other methods reproduced blur or unpleasing textures on the lines or edges but our method can reconstruct them with correct details. These comparisons show that our refiner can overcome the structure preservation problem, mentioned in [14].

The explicit the advantages of CA are shown in Fig. 4. After applying the weightings, we can find that our discriminator puts more attention on the structured high-frequency regions. Otherwise, without CA, the discriminator might focus on the flat surface, which is not the region that we actually care about. As same as the discriminator, Fig. 4 also shows that our generator with CA has the same behavior. They both highlight the same regions and reconstruct better details on the structured parts of the image.

Dataset	PI Before/After Refinement		
	EDSR	DBPN	RCAN
BSD100	5.402/2.585	5.622/2.566	5.136/2.461
Urban100	4.991/3.540	5.264/3.497	4.976/3.458

Table 3: Overall quantitative comparisons for showing the effect of our refiner with different inputs. Red color indicates the best performance.

CA Placement	1^{st} *Scale*	3^{rd} *Scale*	*All Scales*
PI/RMSE	3.547/16.388	3.569/16.384	3.458/16.285

Table 4: Overall quantitative comparisons for applying CA on different scales on Urban100. 1^{st} *Scale* and 3^{rd} *Scale* mean CA is only applied on 24×24 and 6×6 scale of the discriminator. Red color indicates the best performance.

3.3. Ablation Study

In order to clarify the effect of the components in proposed refiner, we further discuss the performance with and without these components in Table 2. All configurations use RCAN output as our input. According to the results, we observe that **CA** component on the discriminator plays an important role in improving PI, and **IS** component is able to significantly reduce RMSE. The configure with **CA** and **IS** achieves the best performance on both PI and RMSE.

Since our proposed refiner is designed for general SISR methods, we compare the results based on different input methods in Table 3. The results show that our refiner did significantly improve PI for all methods.

The discriminator of our refiner aims to focus on inadequate parts of SISR results in multi-scale. In Table 4, we would like to compare the effects of CA placements on the discriminator in different scales. CA in the finest scale (24×24 scale) has the most abundant features from large scale input, and in the coarsest scale (6×6 scale) has the highest level features, extracted by the previous layers. According to the Table 4, the refiner achieves the best performance when CA is applied on all scales. We can conclude that both low-level and high-level features are important to the discriminator, and CA assists our discriminator to extract informative features for all scales.

4. CONCLUSIONS

We propose a refiner with channel attention to improve perceptual quality of SISR methods and overcome the structure preservation problem of GAN-based methods. The weighted feature maps of our refiner demonstrate that CA can help our refiner to focus on the informative context and generate more pleasing details. The ablation studies clarify the effectiveness of the components, and the qualitative and quantitative results show the feasibility of our proposed framework.

5. REFERENCES

[1] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*. 2017, pp. 105–114, IEEE Computer Society.

[2] Ying Tai, Jian Yang, and Xiaoming Liu, "Image super-resolution via deep recursive residual network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 2790–2798.

[3] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *CVPR*. 2017, pp. 5835–5843, IEEE Computer Society.

[4] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPR Workshops*. 2017, pp. 1132–1140, IEEE Computer Society.

[5] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, 2018, pp. 294–310.

[6] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita, "Deep back-projection networks for super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens, "Pixel recursive super resolution," in *ICCV*. 2017, pp. 5449–5458, IEEE Computer Society.

[8] Alexey Dosovitskiy and Thomas Brox, "Generating images with perceptual similarity metrics based on deep networks," in *NIPS*, 2016, pp. 658–666.

[9] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *ICML*. 2014, vol. 32 of *JMLR Workshop and Conference Proceedings*, pp. 1278–1286, JMLR.org.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. 2016, vol. 9906 of *Lecture Notes in Computer Science*, pp. 694–711, Springer.

[11] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 4501–4510.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[13] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.

[14] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor, "2018 PIRM challenge on perceptual image super-resolution," *CoRR*, vol. abs/1809.07517, 2018.

[15] Pablo Navarrete Michelini, Dan Zhu, and Hanwen Liu, "Multiscale recursive and perceptiondistortion controllable image superresolution," in *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.

[16] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 4501–4510.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[18] Martín Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *ICML*. 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 214–223, PMLR.

[19] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville, "Improved training of wasserstein gans," in *NIPS*, 2017, pp. 5769–5779.

[20] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.