Use of Attention Mechanism for Decoder in Deep Learning-based Image Super Resolution

Hyeongyu Kim
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
gusrbao33@gmail.com

Byungchan Choi
RF Seeker R&D
LIG Nex1
Yongin, Republic of Korea
byungchan.choi@lignex1.com

Haewoon Nam
Division of Electrical Engineering
Hanyang University
Ansan, Republic of Korea
hnam@hanyang.ac.kr

***Abstract*—One of major issues in training deep neural network for image super resolution is checkerboard artifact. It degrades the quality and resolution of output images. It usually appears at diagonal edges and curved edges of input images. It can also occur when too much information is compressed and lost during encoding process. To address this issues, we propose U-Net based structure with its decoder reinforced with attention modules for image super resolution task. U-Net structure is used to increase the feature reusability and reduce the loss of information during the encoding process. Attention modules are implemented in decoding layers in order to enhance the capability to produce high resolution output at diagonal edges and curved edges. Our proposed method shows improvements in PSNR and SSIM compared to the existing methods.**

***Index Terms*—Image Super Resolution, Deep Neural Networks, Image Attention**

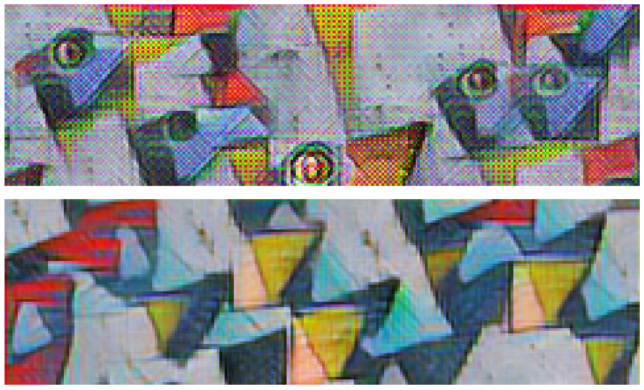

Fig. 1. Checkerboard Artifacts

I. Introduction

Image Super Resolution (SR) is the task of converting a low resolution image into a high resolution image. In recent years, deep learning-based methods have been widely researched for Single Image Super Resolution (SISR). Deep Neural Network (DNN) has shown strong capability to extract and utilize features for enhancing image resolution. DNN for SR task uses encoder-decoder structure. Layers of Convolutional Neural Network (CNN) are used as encoders to extract features from Low Resolution (LR) images and organize them in latent feature vectors. Layers of transposed CNN are used as decoders to produce High Resolution (HR) output images from the compressed information in latent feature vectors.

Checkerboard artifact is one of the greatest obstacles in training DNN for SISR. It is a phenomenon that grid-like noise patterns appear in the output image. It degrades the quality and resolution of SR output. Fig 1 shows the case of checkerboard artifacts [1]. Heavy checkerboard artifacts in upper image of Fig 1 do not only degrade the resolution of the image, but also alter the image by inserting visible noise patterns.

When training the network with CNN and transposed CNN layers for SR task, checkerboard artifacts strongly appear at diagonal edges and curved edges in the image. This is because kernels in transposed CNN move only in straight horizontal and vertical directions. It has structural limitations in handling diagonal edges and curved edges. In addition, deeper the network becomes at the encoder side, stronger the checkerboard artifacts appear. Increase in the number of CNN layers at the encoder compresses the input data into smaller latent feature vectors. Too much compression by encoder can result in the loss of feature information. As a result, decoder will struggle in producing HR images with less information, which can result into checkerboard artifacts. Considering these limitations of CNN-based layers, the network needs to be granted with more flexibility in extracting features and expanding them into HR output. Also, it needs feature reusability to minimize the loss of information during encoding process.

Attention mechanism is one of the recent advancements in deep learning, derived from natural language processing tasks. It is designed to overcome the limitations of encoder-decoder structure by building thorough connections between input and output vectors. In image-based tasks, attention mechanism makes DNN to focus on the important regions of an image and manipulate context information for its intended tasks. Without strict constraints in kernel shape or computation direction, attention mechanism is more flexible in detecting which area of image requires more focus for fine-tuning the output.

In order to overcome the limitations of CNN-based encoder-decoder structure and minimize checkerboard artifacts, we propose U-Net structure with its decoder reinforced with attention modules for SR task. We apply attention at the decoder using Efficient Channel Attention Network (ECA-Net) proposed by Wang et al [2]. ECA captures local cross channel interaction by

considering every channel and neighbors. It can improve pixel reconstruction by collecting global information and improving representation ability for latent vector in the decoder. U-Net structure with channel attention at the decoder can increase the overall feature reusability for producing high resolution output [3]. In addition, we attach Contextual Reasoning Attention Network (CRAN) between encoder and decoder to refine the latent vector [4]. Our network is trained and tested with DIV2K Dataset [5] [6].

Our contributions as follows:

- We overcome the structural limitations of transposed CNN at the decoder by reinforcing the decoder's capability to produce high resolution output through the use of ECA-Net.
- We utilize U-Net structure with channel attention for feature reusability across the entire network. Residual connections of U-Net are integrated at the decoder through channel attention. Its end-to-end structure can provide good use of context information while preveting gradient vanishing and improving convergence during training.
- We apply CRAN between the encoder and decoder. Latent feature vectors from the encoder can be refined by CRAN before being processed by the decoder.

II. Related Works

A. Image Super Resolution

Classical image super resolution methods, such as bicubic interpolation and Lanczos resampling, utilize unique predefined mathematical calculations. Recent methods have focused on deep learning-based approaches.

Dong et al. proposed Super Resolution Convolutional Neural Network (SRCNN) for SR task [7]. Increasing the depth of DNN can enhance the network's feature extraction capability and its performance. However, the network with deep structure often fails to reach convergence, because with more layers, it has to struggle with training more weights. Residual learning allows the neural network with deeper structures to cope with gradient vanishing and degradation through feature reusability. With the help of residual learning, Kim et al. proposed Very Deep Super Resolution (VDSR) [8]. Lim et al proposed the removal of batch normalization layer since it cannot improve the model performance while consuming memory and restricting the network's flexibility with normalization process [9]. Goodfellow et al. introduced the Generative Adversarial Networks (GAN), which has been used in various image-based tasks, such as style transfer and image generation [10]. Inspired by GAN, Ledig et al. proposed Super Resolution Generative Adversarial Network (SRGAN) [11]. While its generator is trained to produce HR output from LR input, its discriminator is trained to determine whether the output is the generator's HR output or its HR groundtruth. Leig et al also proposed perceptual loss in order to cope with ill-posed problem of SR task [11].

B. Image Attention

Image attention is mainly used for feature map selection using multi-modal relationship tasks, such as Visual Question Answering or Image Captioning [12] [13]. Attention can be viewed as a dynamic feature map selection according to the input. Hu et al. proposed a squeeze-and-excitation networks (SENet), which is used to collect global information, capture channel-wise relationships, and improve representation ability [14]. SENet reduces the number of channels to avoid high model complexity. However, it fails to directly model correspondence between weight and inputs, reducing the quality of results. In order to overcome this drawback, Wang et al. proposed ECA-Net [2]. Instead of dimensionality reduction, it uses a 1D convolution to determine the interaction between channels. Zhang el al. proposes Context Reasoning Attention Network (CRAN) for SR task. [4]. They point out that although CNN is effective in extracting local features, it lacks the capability to set up global context for its intended task. By integrating attention mechanism into convolution process, it achieves exceptional quantitative and qualitative SR performance.

III. Proposed Methods

Proposed method focuses on enhancing the decoder of CNN-based SR network with U-Net structure and image attention. Fig 2 shows the network structure of the proposed method. Although CNN in the encoder is effective in extracting local features and organizing them as the latent feature vectors, transposed CNN in the decoder lacks the capability in processing diagonal edges and curved edges for producing SR output. Since the decoder is responsible for generating SR output from latent representation, its performance determines the quality of SR output image. It needs to be more flexible in processing diagonal edges and curved edges in order to prevent checkerboard artifacts.

We use U-Net structure in order to grant the network with feature reusability through residual connections between encoders and decoders of corresponding levels [3]. We integrate residual connections from the encoder into the decoder using channel attention. These connections can minimize loss of information during the encoding process by allowing the decoder to fully utilize latent representation across the entire network path. This can prevent gradient vanishing and improve the network's convergence during the training process.

In order to reinforce the decoder's SR capability with attention, we propose the use of ECA-Net between decoding layers. ECA block has similar function to an SE block. Instead of indirect correspondence, an ECA block only considers direct interaction between each channel and its k-nearest neighbors to control model complexity. This process is described in Equation 1 and 2:

$$s = F_{eca}(x, \theta) = \sigma(Conv1D(GAP(X))) \tag{1}$$

$$Y = sX \tag{2}$$

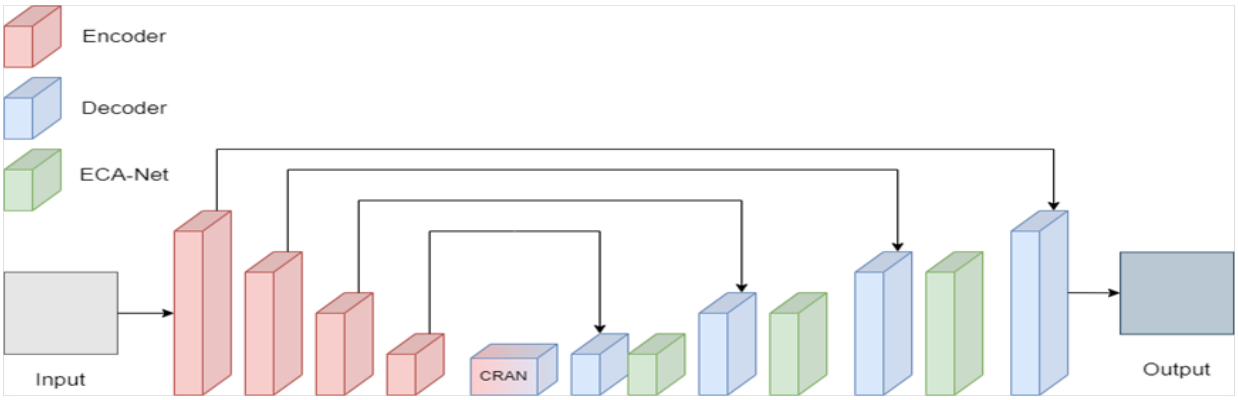

Fig. 2. Structure of Proposed Method with Attention Module

$Conv1D\,(\cdot)$ denotes 1D convolution with a kernel across the channel domain, which models local cross-channel interaction. Compared to SENet, ECA-Net has an improved excitation module.

We apply CRAN for smoother learned information between encoder and decoder. CRAN follows the network design of RCAN [4] [15]. It is composed of Context Reasoning Attention Convolution (CRAC) and Context Residual Attention Block (CRAB) for SR task. The equation of super-resolved output I_{SR} is described as Equation 3. $F_{CRAN}\,(\cdot)$ denotes the function of CRAN.

$$I_{SR} = F_{CRAN}(I_{LR}) \tag{3}$$

We use the combination of two different losses as loss function to train our proposed network. First, the L1 loss is designed to minimize the difference in intensity of pixels and is defined as Equation 4. I_{HR} denotes the high-resolution image. I_{SR} denotes the super-resolution image as output. However, this loss is not enough to handle multi-modal nature of SR task.

$$L_1 = |I_{HR} - I_{SR}| \tag{4}$$

Second, the perceptual loss is used to conserve style and context [16]. It is defined as Equation 5. First four layers in the encoder of Fig 2 are from pre-trained VGG16 network. This loss can compare high level differences and catch the details of images.

$$L_{perc} = \sum_{layer} |VGG(I_{HR}) - VGG(I_{SR})|_{layer} \tag{5}$$

As described in Equation 6, loss function for our propose method is the weighted sum of L1 loss and perceptual loss.

$$L_{final} = \alpha * L_1 + \beta * L_{perc} \tag{6}$$

IV. Experiment Results

A. Training Setup

We trained our proposed network from Section 3 with DIV2k dataset on SR factor x4 [5] [6]. We implemented the network using PyTorch. It is trained on Nvidia Tesla T4 by Adam optimizer with learning learning late =$1e^{-4}$, β_1=0.9, β_2=0.999, and ϵ=$1e^{-8}$. Loss function from Equation 6 is set with α = 0.8, β = 0.2.

B. Results

1) Quantitative Evaluation: To evaluate the proposed methods, we use two evaluation metrics, PSNR and SSIM. We compared its performance with bilinear interpolation, CNN-only U-Net, and SRCNN. Table I shows that our proposed method achieves higher PSNR and SSIM. CNN-only U-Net achieves higher PSNR ans SSIM with feature reusabiltiy through residual connections. Based on the residual connections of U-Net, our proposed method reaches higher PSNR and SSIM by reinforcing the decoder with ECA-Net and refining latent feature vectors from the encoder with CRAN.

TABLE I
Quantitative Result Comparison

Method	Scale	PSNR	SSIM
Bilinear Interpolation	x4	28.871	0.671
CNN-only U-Net	x4	30.675	0.824
SRCNN	x4	29.540	0.732
Proposed Method	x4	**31.452**	**0.876**

2) Qualitive Results: Fig 3 shows qualitative comparison of SR images produced by our proposed method and existing methods. From Table I, although CNN-only U-Net reaches higher PSNR and SSIM than SRCNN, its output images show signs of jitter noises. Also, compared to HR groundtruth and SRCNN outputs, its outputs show signs of blurriness at the regions of extreme details. This indicates CNN-based layers, especially transposed CNN, have limitations in handling details of HR images. While our proposed method reachers higher PSNR and SSIM than both CNN-only U-Net and SRCNN, it successfully produces SR images with strong details. Reinforcing decoder with attention modules provides more flexibility for the decoder to process various types of edges and details for SR task. This leads to improvements in both evaluation metrics and SR output quality.

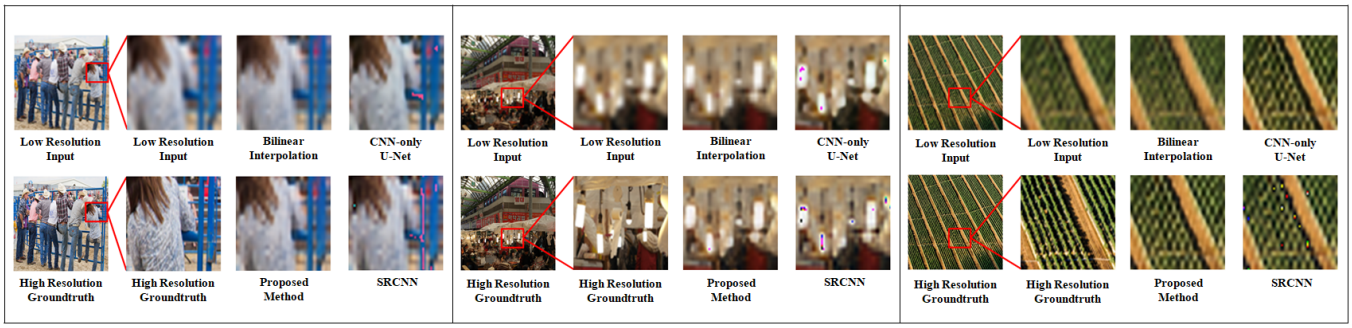

Fig. 3. Qualitative Result Comparison

V. Conclusion

This paper presents the use of attention mechanism in the decoder of SR network to improve SR image generation performance. It points out that structural limitations of transposed CNN at the decoder lead to checkerboard artifacts that degrade the quality and resolution of SR output images. In order to reinforce the decoder's image generation capability, we apply U-Net with channel attention for feature reusability, ECA-Net between decoding layers for more flexible feature processing, and CRAN between encoder and decoder for refining the encoder's latent feature vectors. Our proposed method achieves higher PSNR and SSIM than some of existing methods. It also produces SR image output with strong details. Reinforcing the decoder with attention modules accomplishes improvements in both evaluation metrics and SR output quality by allowing more flexibility in handling various edges and details at the decoder of SR network.

Our proposed method can provide post-processing filter that can enhance details of targets and backgrounds in LR input image. It can be used in the fields that requires resolution enhancements, such as video surveillance and satellite imagery. Our proposed method needs to be further optimized in its computation time and memory usage in order to operate a real-time filter block.

VI. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT of Korea (MSIT) (2022R1A2C1011862).

References

[1] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[2] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] Y. Zhang, D. Wei, C. Qin, H. Wang, H. Pfister, and Y. Fu, "Context reasoning attention network for image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4278–4287.

[5] E. Agustsson and R. 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.

[6] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim *et al.*, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[8] J. Kim, J. K. Lee, and K. M. 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.

[9] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[11] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[12] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[13] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[15] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.