Efficient Hybrid Attention Network for Image Super-resolution

[1,*]*Yu Fu* (傅譽) *and* [2,*]*Chiou-Shann Fuh* (傅楸善)

[1]Department of Electrical Engineering,
[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[*]E-mail: b08901039@ntu.edu.tw and fuh@csie.ntu.edu.tw

ABSTRACT

Image super resolution, the task of generating high-resolution images from low-resolution inputs, has been a prominent topic in computer vision research. Recently, transformer-based models have achieved remarkable success in various vision tasks by leveraging self-attention mechanisms. In this paper, we propose a novel approach for image super resolution using transformer-based attention strategies. Our method combines the strengths of transformers and attention mechanisms to effectively capture long-range dependencies and enhance image details. We introduce a transformer-based architecture that incorporates attention modules at different scales, enabling the model to attend to relevant image patches and exploit their contextual information. Additionally, we employ a multi-scale fusion technique to integrate information from multiple resolution levels, allowing the model to generate high-quality high-resolution images. Experimental results on benchmark datasets demonstrate that our proposed method achieves superior performance compared to state-of-the-art approaches. We provide a comprehensive analysis of the model's attention patterns, highlighting the effectiveness of the proposed transformer-based attention strategies in capturing meaningful image features. Furthermore, we conduct ablation studies to investigate the impact of different components and demonstrate their contribution to the overall performance. Our work presents a novel approach for image super resolution, leveraging transformer-based attention strategies to push the boundaries of image quality enhancement.

Keywords: *Image super-resolution, transformer-based attention.*

1. INTRODUCTION

Single image super-resolution (SR) is a fundamental problem in computer vision, aiming to reconstruct high-resolution images from low-resolution inputs. Convolutional neural networks (CNNs) have been widely utilized for SR, demonstrating their efficacy in recovering high-frequency details. However, CNN-based methods often struggle to adaptively capture pixel relations due to their reliance on spatially invariant kernels, and their deep architectures lead to substantial computational resource consumption.

Recently, the success of transformer-based models in natural language processing has inspired their application in computer vision tasks. Transformers excel in modeling self-attention, which allows for effective information integration across long ranges. While transformers have shown remarkable performance in high-level vision tasks, their application to low-level vision tasks, such as SR, has been limited by their quadratic computational complexity, impeding their scalability to large feature sizes.

Efforts have been made to address the computational cost of self-attention in SR. Some approaches divide images into non-overlapping patches, enabling independent local feature extraction and self-attention modeling. However, these methods may suffer from border effects and compromise visual quality. SwinIR, following the design of Swin Transformer, employs cascaded 1x1 convolutions for local feature extraction and employs a shifting mechanism for self-attention within a small-sized window. However, this approach may result in weak feature representations and limited long-range dependency modeling. Restormer adopts channel-based self-attention, but sacrifices spatial information important for high-quality image reconstruction.

To address these limitations, we propose a Hybrid Attention Convolution (HAC) for image super-resolution. HAC combines self-attention and channel attention mechanisms to leverage global information and enhance the representation capacity. Moreover, we introduce an overlapping cross-attention module to promote direct interaction among adjacent window features, overcoming the limitations of the shifting mechanism. Our approach aims to expand the range of utilized information while maintaining efficient computational complexity.

In this paper, we aim to develop an effective and efficient method for exploiting long-range attention in

image super-resolution using a simple network architecture. We gain our thoughts mostly based on ELAN [25] and introduce an efficient hybrid attention block that models image long-range dependencies, and we present an Efficient Hybrid Attention Network (EHAN) that achieves state-of-the-art performance with significantly reduced complexity compared to existing transformer-based SR methods. Our contributions lie in the development of HAC and EHAN, which address the limitations of previous approaches and unlock the potential of transformers for SR tasks.

2. RELATED WORKS

2.1. Deep Networks for Image SR

Since the introduction of SRCNN [2], numerous deep networks have been proposed for SR to further improve reconstruction quality. These networks employ more elaborate convolution module designs, such as residual blocks and dense blocks, to enhance model representation ability. Some methods explore different frameworks like recursive neural networks and graph neural networks. To improve perceptual quality, adversarial learning has been introduced in several approaches. Attention mechanisms have also achieved further improvements in terms of reconstruction fidelity. Recently, Transformer-based networks have been proposed, constantly refreshing the state-of-the-art in SR and demonstrating the powerful representation ability of Transformers.

Several works aim to analyze and interpret the mechanisms of SR networks. LAM [3] explores which input pixels contribute the most to the final performance using the integral gradient method. DDR [4] reveals deep semantic representations in SR networks based on deep feature dimensionality reduction and visualization. FAIG [5] aims to find discriminative filters for specific degradations in blind SR. RDSR [6] introduces a channel saliency map to demonstrate the benefits of Dropout in preventing co-adaptation in real-SR networks. SRGA [7] evaluates the generalization ability of SR methods. In this work, we utilize LAM to analyze and understand the behavior of SR networks.

2.2. Vision Transformer

Transformers have attracted attention in the computer vision community due to their success in natural language processing. Transformer-based methods have been developed for various high-level vision tasks, including image classification [8], object detection [9], and segmentation [10, 11]. While vision Transformers excel in modeling long-range dependencies, it has been shown that convolutions can enhance their visual representation. Transformers have also been introduced for low-level vision tasks, including image SR. Existing works have made progress, but there is still untapped potential in Transformers. Our proposed method activates more input pixels for better reconstruction.

2.3. CNN-based SR Methods

CNN-based methods have demonstrated impressive performance in the SR task. SRCNN [2] made the first attempt to employ CNN for image SR by learning a non-linear mapping from the bicubically upsampled LR image to the HR output using only three convolution layers. Kim et al. [12] deepened the network with VGG-19 [13] and residual learning, resulting in much better performance. To address the computational cost associated with the pre-upsampling strategy, FSRCNN [14] adopted a post-upsampling strategy to accelerate the CNN model. Additionally, an enhanced residual module [15] was proposed in to train deep models without batch normalization.

Recent developments in SR methods aim to build more effective models by employing deeper and more complex architectures, as well as attention techniques. Zhang et al. proposed a residual-in-residual structure coupled with channel attention, training a very deep network with over 400 layers. Other works, such as MemNet [16] and RDN [17], utilize dense blocks to leverage intermediate features from all layers. In addition to increasing network depth, methods like SAN [18], NLRN [19], HAN [20], and NLSA [21] exploit feature correlations along the spatial or channel dimension to boost SR performance. Our proposed EHAN model takes advantage of fast local feature extraction while modeling hybrid feature dependencies through efficient group-wise multi-scale self-attention.

2.4. Transformer-based SR Methods

The breakthrough of transformer networks in Natural Language Processing (NLP) has inspired the use of self-attention (SA) mechanisms in computer vision tasks. Transformers effectively model dependencies across data, achieving impressive results in high-level vision tasks such as image classification, detection, and segmentation. Recently, transformers have also been applied to low-level vision tasks, including image SR.

IPT [22] is an extremely large pre-trained model for various low-level vision tasks based on the standard vision transformer. It computes both local features and SA on non-overlapping patches. However, this approach may lose some useful information for reproducing image details. SwinIR [23] addresses this limitation by adapting the Swin Transformer [24] to image restoration, combining the advantages of both CNNs and transformers. Although SwinIR has achieved impressive results for image SR, its network structure, borrowed from the Swin Transformer designed for high-level vision tasks, is redundant for the SR problem. It calculates SA on small fixed-size windows, preventing the exploitation of long-range feature dependencies. Our

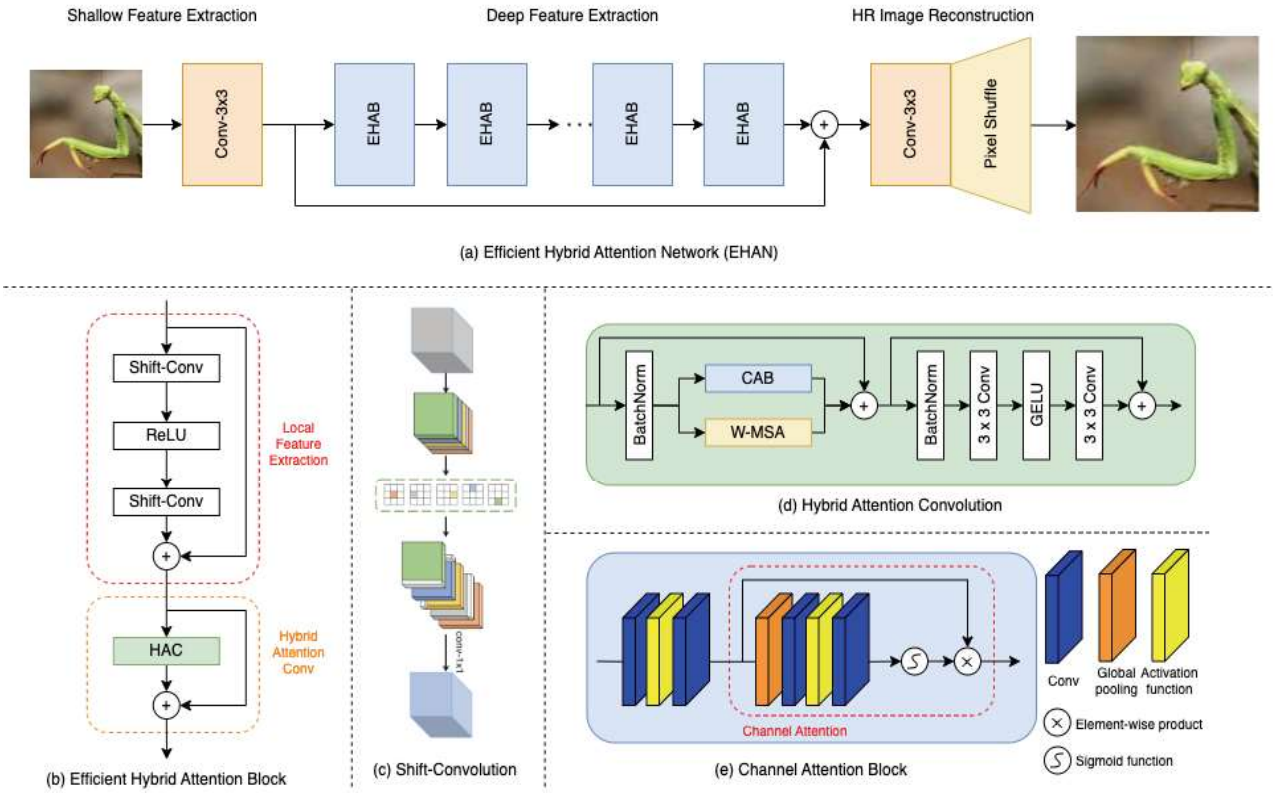

Fig. 1. Illustration of the proposed efficient hybrid attention network (EHAN). (a) The overall pipeline of EHAN, which contains several EHABs, two 3×3 convolutions and one pixel shuffle operator. (b) The architecture of the Efficient Hybrid Attention Block (EHAB). (c) Illustration of shift convolution, which is composed of a shift operation followed by one 1×1 convolution. (d) Illustration of the computation of Hybrid Attention Convolution (HAC). (e) Illustration of channel attention block (CAB).

proposed ELAN model is not only more efficient than SwinIR but also capable of computing SA in larger windows.

3. METHOD

In this section we first present the pipeline of our Efficient Hybrid Attention Network (EHAN) for SR tasks, and then discuss in detail its key component, the Efficient Hybrid Attention Block (EHAB).

3.1. Overall Pipeline of EHAN

The overall pipeline of EHAN is depicted in Figure 1(a) and comprises three modules: shallow feature extraction, EHAB-based feature extraction, and HR image reconstruction. The network follows a simple topology that includes a global shortcut connection, connecting the shallow feature extraction module to the output of the deep feature extraction module before being fed into the HR reconstruction module. Specifically, given a degraded LR image $X_l \in \mathbb{R}^{C \times H \times W}$, where H and W represent the height and width of the LR image, respectively, we initially apply the shallow feature extraction module ($H_{SF}(\cdot)$), which consists of a single 3×3 convolution layer, to extract the local feature $X_s \in \mathbb{R}^{C \times H \times W}$:

$$X_s = H_{SF}(X_l), \tag{1}$$

where C is the channel number of the intermediate feature.

X_s is then passed to the deep feature extraction module, denoted as $H_{DF}(\cdot)$, which consists of M cascaded EHABs. This module produces the output:

$$X_d = H_{DF}(X_s), \tag{2}$$

where $X_d \in \mathbb{R}^{C \times H \times W}$. Utilizing X_d and X_s as inputs, the HR image X_h is reconstructed using the $H_{RC}(\cdot)$ module:

$$X_h = H_{RC}(X_s + X_d). \tag{3}$$

Various options exist for the design of the reconstruction module. To ensure high efficiency, we construct it simply with a single 3×3 convolution and a pixel shuffle operation.

The EHAN can be optimized using commonly employed loss functions for SR, such as L_1, L_2, and perceptual loss. For simplicity, given a number of N ground-truth HR images $\{X_{t,i}\}_{i=1}^{N}$, we optimize the

parameters of EHAN by minimizing the pixel-wise L_1 loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||X_{h,i} - X_{t,i}||_1. \quad (4)$$

We employ the Adam optimizer to optimize our EHAN, leveraging its strong performance in low-level vision tasks.

3.2. Efficient Hybrid Attention Block (EHAB)

Our EHAB (Efficient Hybrid Attention Block) consists of two modules: the local feature extraction module and the Hybrid Attention Convolution (HAC) module, both employing the residual learning strategy.

3.2.1. Local feature extraction

The local feature extraction module aims to extract local features from intermediate features X. Previous approaches [1] typically utilize multi-layer perception or two cascaded 1×1 convolutions for local feature extraction. However, these methods only have a small 1×1 receptive field. To overcome this limitation and enable more effective local feature extraction, we introduce the use of two shift-conv layers with a simple ReLU activation in between. Figure 1(c) illustrates the structure of the shift-conv module.

The shift-conv module consists of a set of shift operations and a 1×1 convolution. Specifically, the input feature is divided equally into five groups. The first four groups are shifted along different spatial dimensions (left, right, top, bottom), while the last group remains unchanged. This arrangement allows the subsequent 1×1 convolution to capture information from neighboring pixels, thereby enlarging the receptive field. Importantly, the shift-conv module achieves this with minimal computational cost and without introducing additional learnable parameters. It provides larger receptive fields while maintaining nearly the same computational complexity as a 1×1 convolution.

3.2.2. Hybrid attention convolution

Incorporating channel attention into the network enables the involvement of global information, leading to the activation of more pixels. Additionally, previous research has shown that convolution can enhance the visual representation and optimization capabilities of Transformer-based models. Hence, we introduce a channel attention-based convolution block into the standard Transformer block to improve the network's representation ability.

Figure 1(d) illustrates the integration of a Channel Attention Block (CAB) into the standard Swin Transformer block. The CAB is inserted after the first BatchNorm (BN) layer in parallel with the Window-Based Multi-head Self-Attention (W-MSA) module. Notably, shifted window-based self-attention (SW-MSA) is employed at intervals within consecutive HACBs (Hybrid Attention Convolution Blocks). To mitigate potential conflicts between CAB and MSA in terms of optimization and visual representation, the output of CAB is multiplied by a small constant α.

The computation process of the Hybrid Attention Convolution Block (HACB) is as follows:

$$\begin{aligned} X_N &= BN(X), \\ X_M &= (S)W\text{-}MSA(X_N) + \alpha CAB(X_N) + X, \\ Y &= CNN(BN(X_M)) + X_M. \end{aligned} \quad (5)$$

Here, X_N and X_M represent the intermediate features, and Y represents the output of the HACB. Each pixel is treated as a token for embedding (with a patch size of 1). CNN denotes a multilayer convolution network.

For the self-attention module, the input feature $X_N \in \mathbb{R}^{C\times H\times W}$ is partitioned into HW/M^2 local windows of size $M \times M$. Self-attention is then calculated within each window. Given a local window feature $X_W \in \mathbb{R}^{M^2\times C}$, the query, key, and value matrices (Q, K, and V) are computed through linear mappings. The window-based self-attention is formulated as:

$$Attention(Q,K,V) = SoftMax(QK^T/\sqrt{d} + B)V. \quad (6)$$

Here, d represents the dimension of the query/key, and B denotes the relative position encoding.

To establish connections between neighboring non-overlapping windows, the shifted window partitioning approach is employed, with the shift size set to half of the window size.

The CAB consists of two standard convolution layers with a GELU activation and a Channel Attention (CA) module. Since the Transformer-based structure often requires a large number of channels for token embedding, we apply channel number compression to the two convolution layers using a constant β. The channel number of the output feature after the first convolution layer is reduced to C/β, and then the feature is expanded back to C channels through the second layer. Finally, a standard CA module is utilized to adaptively rescale the channel-wise features.

4. EXPERIMENTAL SETUP

In this section, we present a comprehensive set of experiments to quantitatively and qualitatively evaluate the exceptional performance of our EHAN model in both light-weight and classic single-image super-resolution tasks. We conduct these experiments on three widely-used SR benchmark datasets, aiming to provide a thorough validation of the effectiveness and superiority of our proposed approach.

4.1. Datasets and Evaluation Metrics

For model training, we utilize a subset of the DIV2K dataset [26] that consists of 800 images. To evaluate the performance of our model, we conduct testing on four

well-established super-resolution benchmarks: Set5 [27], Set14 [28] and BSD100 [29] datasets. To assess the quality of the super-resolved images, we employ the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as evaluation metrics. These metrics are computed on the luminance channel (Y) after converting the RGB images to the YCbCr color space.

4.2. Training Details

During the training phase, we employ a network architecture composed of 16 EHABs in our EHAN model. The hidden channel dimension is set to 48, ensuring an appropriate balance between representation capacity and computational efficiency. The training process is carried out for 120 epochs. To optimize the model effectively and achieve faster convergence, we utilize the Adam optimizer. Specifically, we set β_1 to 0.9 and β_2 to 0.999. The initial learning rate is set to 0.0002, and we schedule learning rate reductions at epoch 50, 80, 90, 95, and 100, by halving its value. To ensure diverse and robust training, we adopt a patch size of 128 and a batch size of 32. Additionally, we apply data augmentation techniques, such as random rotations of 90, 180, or 270 degrees and horizontal flips with a probability of 0.5, to enhance the generalization capability of the model.

5. RESULTS

In our evaluation, we compare the performance of our model with two existing super-resolution methods, namely A+ [] and Google RAISR [], on widely used benchmark datasets including Set5 [27], Set14 [28], and BSD100 [29]. By conducting quantitative and qualitative assessments, we aim to comprehensively evaluate the effectiveness and superiority of our approach.

5.1. Quantitative Results

The quantitative results provide objective measurements such as PSNR and SSIM, which quantitatively assess the fidelity and similarity between the reconstructed high-resolution images and the ground truth.

SSIM (Structural Similarity Index) and PSNR (Peak Signal-to-Noise Ratio) are commonly used evaluation metrics in super-resolution (SR) tasks to assess the quality of the reconstructed high-resolution images compared to their ground truth counterparts.

PSNR is a traditional metric that measures the average difference between the pixel values of the reconstructed image and the ground truth image. It calculates the ratio of the maximum possible power of a signal to the power of the noise that affects the signal. Higher PSNR values indicate lower distortion and better image quality, as the reconstructed image closely matches the ground truth. However, PSNR does not always correlate well with perceived image quality, as it does not consider human visual perception factors.

On the other hand, SSIM is a perceptual metric that takes into account the structural information, luminance, contrast, and similarity between the reconstructed image and the ground truth. It compares local image patches to measure the structural similarity between the images. SSIM values range from 0 to 1, with 1 indicating a perfect match between the images. SSIM provides a more comprehensive assessment of image quality, considering both pixel-level fidelity and perceptual similarity.

Table 1. PSNR of upscaling factor 2

	Set5	Set14	BSD100
A+	36.55	32.28	30.78
RAISR	36.15	32.13	-
Ours	38.01	33.63	32.17

Table 2. PSNR of upscaling factor 3

	Set5	Set14	BSD100
A+	32.59	29.13	28.18
RAISR	32.21	28.86	-
Ours	33.96	30.20	28.85

Table 3. PSNR of upscaling factor 4

	Set5	Set14	BSD100
A+	30.29	27.33	26.77
RAISR	29.84	27.00	-
Ours	32.20	28.60	27.57

Table 4. SSIM of upscaling factor 2

	Set5	Set14
RAISR	0.951	0.902
Ours	0.961	0.918

Table 5. SSIM of upscaling factor 3

	Set5	Set14
RAISR	0.901	0.812
Ours	0.924	0.839

Table 6. SSIM of upscaling factor 4

	Set5	Set14
RAISR	0.895	0.781
Ours	0.848	0.738

In SR tasks, both SSIM and PSNR are used to evaluate the performance of super-resolution algorithms. While PSNR is mainly focused on pixel-level fidelity and noise reduction, SSIM provides insights into the perceptual

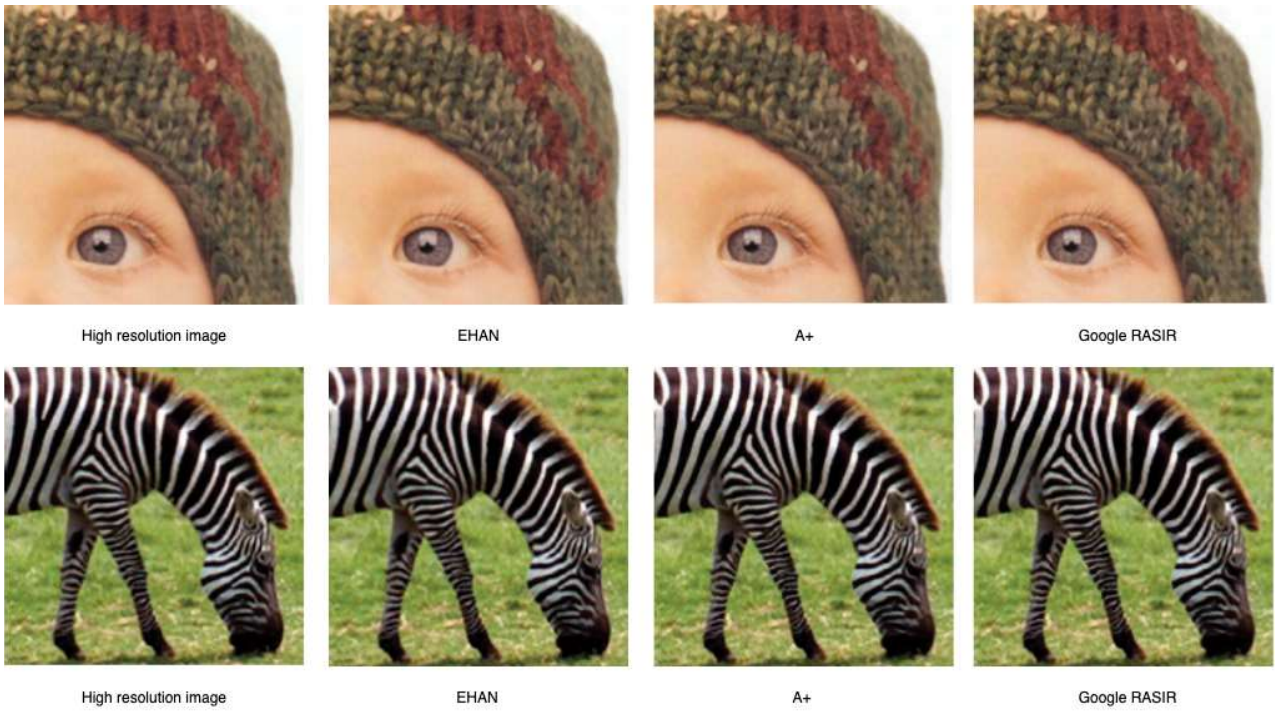

Fig. 2. Qualitative comparison of EHAN, A+, and RASIR with an upscaling factor of 2.

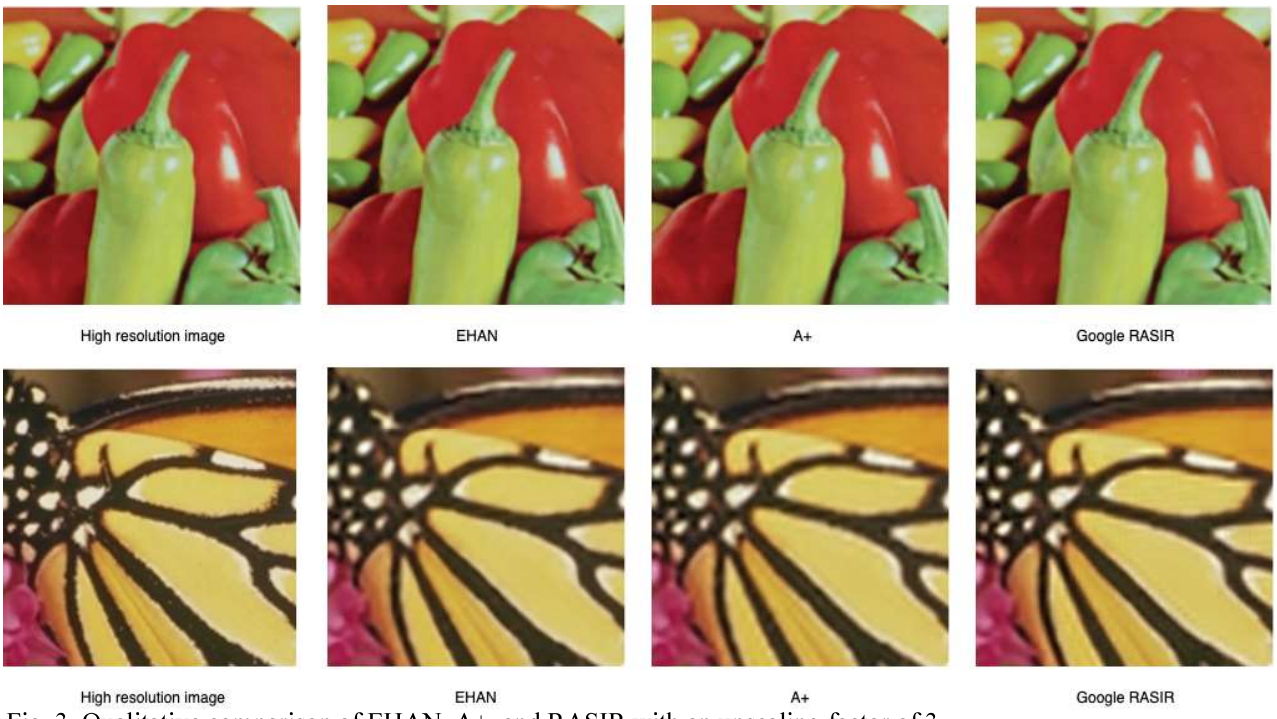

Fig. 3. Qualitative comparison of EHAN, A+, and RASIR with an upscaling factor of 3.

quality and structural similarity of the reconstructed images. By considering both metrics, researchers and practitioners can gain a more comprehensive understanding of the strengths and limitations of different super-resolution techniques and make informed decisions about their suitability for specific applications.

Table 1, Table 2, and Table 3 present the PSNR values obtained by the A+, RAISR, and our model, respectively, for various upscaling factors in the super-resolution task. These tables provide a quantitative comparison of the performance of different algorithms. Additionally, Table 4, Table 5, and Table 6 showcase the SSIM scores of RAISR and our model for different upscaling factors. These SSIM tables offer further insights into the perceptual quality and structural similarity of the reconstructed images. Notably, our model consistently outperforms the previously proposed networks, demonstrating its superior performance in terms of both

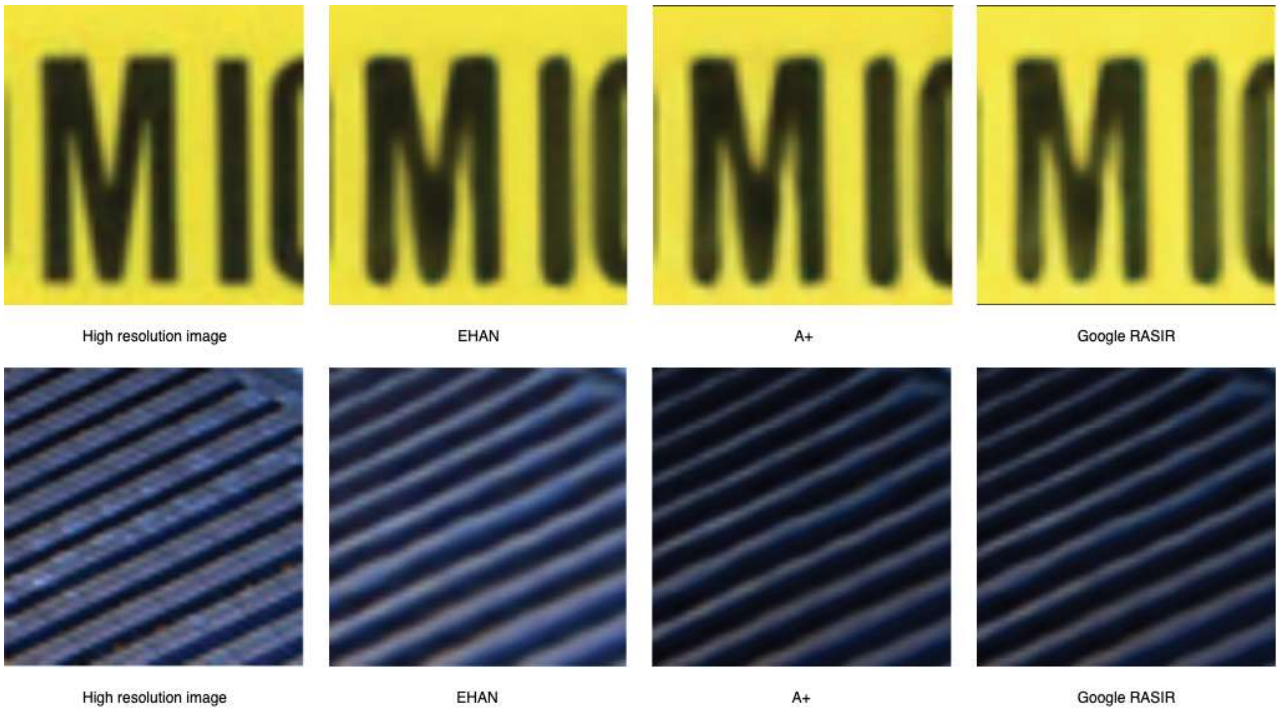

Fig. 4. Qualitative comparison of EHAN, A+, and RASIR with an upscaling factor of 4.

PSNR and SSIM metrics across various upscaling factors. These results highlight the effectiveness and competitiveness of our model in the field of super-resolution.

5.2. Qualitative Results

In addition to the quantitative evaluations, the qualitative results of our EHAN further illustrate its superiority over existing methods such as RAISR and A+. As the upscaling factor increases, both RASIR and A+ tend to generate results where the brightness deviates from that of the ground truth, resulting in images that appear either too bright or too dark. Additionally, these methods often produce distorted lines that are either curved or eroded, which can significantly impact the visual quality and accuracy of the reconstructed images. In contrast, EHAN excels at preserving the brightness and sharpness of lines, ensuring that they maintain their original appearance and remain crisp and well-defined. The enhanced brightness ensures that the output images are visually appealing and well-exposed, avoiding any overexposed or underexposed areas. Moreover, EHAN effectively preserves and enhances image sharpness, resulting in more defined edges and clearer contours. The fine details within the images are also better restored and enhanced by EHAN, revealing intricate textures and subtle features that were previously obscured in the low-resolution inputs. This contributes to the overall visual fidelity and perceptual quality of the enhanced images produced by EHAN. By addressing these limitations and achieving better results in terms of brightness, sharpness, details, and colors, EHAN proves to be an effective solution for super-resolution tasks, outperforming both RASIR and A+ in qualitative assessments. Furthermore, EHAN excels in accurately reproducing vibrant and realistic colors, resulting in visually pleasing and natural-looking images. Overall, the qualitative results demonstrate the superior visual quality achieved by EHAN, surpassing the performance of RAISR and A+ in terms of brightness, sharpness, details, and colors.

6. CONCLUSION

In conclusion, our Efficient Hybrid Attention Network (EHAN) stands out as a highly effective and innovative solution for super-resolution tasks. One of the key strengths of EHAN lies in its attention mechanism, which combines hybrid attention and transformer-based architectures. This mechanism allows EHAN to capture and leverage intricate image details effectively, resulting in superior visual quality.

In both qualitative and quantitative comparisons with existing models such as RASIR and A+, EHAN demonstrates clear advantages. In terms of qualitative assessments, EHAN consistently produces images with better brightness, sharpness, details, and colors compared to RASIR and A+. Notably, as the upscaling factor increases, RASIR and A+ tend to generate results with inconsistent brightness in comparison to the ground truth. Additionally, these models may exhibit curved or eroded lines, which can significantly degrade the visual quality. In contrast, EHAN excels at preserving the brightness and sharpness of lines, resulting in visually pleasing and realistic images.

Furthermore, EHAN's superiority is substantiated by quantitative comparisons using metrics such as PSNR

and SSIM. Our model consistently achieves higher PSNR scores on different upscaling factors when compared to RASIR and A+. Similarly, the SSIM values obtained by EHAN surpass those of the competing models. These quantitative assessments further validate the enhanced performance and fidelity of EHAN.

Overall, EHAN's attention mechanism, coupled with its superior qualitative and quantitative results, establishes it as a cutting-edge solution for super-resolution tasks. With its ability to capture fine image details and produce visually compelling results, EHAN holds great potential for a wide range of applications that require high-resolution image generation.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to Professor Chiou-Shann Fuh for his invaluable teaching throughout this lesson. His guidance, advice, and expertise have been instrumental in shaping my understanding and knowledge in the field of computer science and information engineering. I am deeply appreciative of his patience, dedication, and willingness to share his wealth of experience with me.

I would also like to extend my thanks to the Department of Computer Science and Information Engineering at National Taiwan University for providing the necessary resources and support for this lesson. The department's commitment to excellence in education and research has created an enriching learning environment for students like me.

REFERENCES

[1] Chen, X., Wang, X., Zhou, J., Qiao, Y., & Dong, C. (2023). Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22367-22377).

[2] Ward, C. M., Harguess, J., Crabb, B., & Parameswaran, S. (2017, September). Image quality assessment for determining efficacy and limitations of Super-Resolution Convolutional Neural Network (SRCNN). In Applications of Digital Image Processing XL (Vol. 10396, pp. 19-30). SPIE.

[3] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9199–9208, 2021.

[4] Yihao Liu, Anran Liu, Jinjin Gu, Zhipeng Zhang, Wenhao Wu, Yu Qiao, and Chao Dong. Discovering" semantics" in super-resolution networks, 2021.

[5] Liangbin Xie, Xintao Wang, Chao Dong, Zhongang Qi, and Ying Shan. Finding discriminative filters for specific degradations in blind super-resolution. Advances in Neural Information Processing Systems, 34, 2021.

[6] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6002–6012, 2022.

[7] Yihao Liu, Hengyuan Zhao, Jinjin Gu, Yu Qiao, and Chao Dong. Evaluating the generalization ability of superresolution networks. arXiv preprint arXiv:2205.07019, 2022.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[9] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

[10] Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., & Schwing, A. G. (2021). Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764.

[11] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. arXiv preprint arXiv:2304.02643.

[12] Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016).

[13] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[14] Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European conference on computer vision. pp. 391–407. Springer (2016).

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

[16] Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE international conference on computer vision. pp. 4539–4547 (2017).

[17] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2472–2481 (2018).

[18] Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019.

[19] Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. Advances in neural information processing systems 31 (2018).

[20] Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: European conference on computer vision. pp. 191–207. Springer (2020).

[21] Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3517–3526 (2021).

[22] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12299–12310 (2021).

[23] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1844 (2021).

[24] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021).

[25] Zhang, X., Zeng, H., Guo, S., & Zhang, L. (2022, October). Efficient long-range attention network for image super-resolution. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII (pp. 649-667). Cham: Springer Nature Switzerland.