Transformers in image super-resolution – a brief review

Serhii Kozlov*[a], Oleh Kolesnytskyi[a], Oleh Korolenko[a], Alexey Zhukov[a],
Dmytro Bondarenko[a], Olena Smetaniuk[a], Aliya Kalizhanova[b], Paweł Komada[c]
[a]Vinnytsia National Technical University, 95 Khmelnytske Shose, Vinnytsia, Ukraine;
[b]Institute of Information and Computational Technologies CS MES RK, Almaty, Kazakhstan;
[c]Lublin University of Technology, Lublin, Poland

ABSTRACT

Since the advent of deep learning a decade ago convolutional neural networks have been the predominant method for approaching computer vision tasks. However, Transformer model, which has shown significant achievements in the field of natural language processing, is increasingly being applied to computer vision tasks and is demonstrating comparable or superior performance. The article discusses the application of Transformer model to the super-resolution task. The direct application of the original Transformer achived performance comparable to the contemporary convolutional neural networks. However, the self-attention mechanism, which underpins Transformer model, involves quadratic computational complexity with respect to the size of the input image, presenting a significant challenge for processing high-resolution images. Further research has significantly improved performance, but these improvements are not exhaustive. An overview and comparative analysis of these studies are presented.

Keywords: super-resolution, transformer, convolution neural network, computer vision

1. INTRODUCTION

The rapid development of digital image processing technologies has led to an increasing demand for high-resolution images in various applications, ranging from medical imaging and surveillance systems to the production of entertainment multimedia content. However, obtaining high-resolution images is often constrained by the capabilities of photosensitive elements or other physical limitations of capture devices, resulting in reduced resolution and, consequently, low detalization and poor quality of images. Super-resolution (SR) – the process of generating a high-resolution (HR) image from a corresponding low-resolution (LR) image – has recently gained significant attention as an effective and cost-efficient way to address this problem.

Traditional SR methods, such as bicubic interpolation, are simple and efficient but are prone to blurring details and producing ringing artifacts. Advanced learning-based methods, such as sparse coding and locally linear regression, have been proposed to address these shortcomings. The rapid growth of computational power and the wide availability of visual data have made it possible to apply deep learning to the SR task. The application of convolution neural network (CNN) and generative adversarial network (GAN) has been extensively investigated for addressing the SR problem over the past decade. These approaches have achieved significant improvements in reconstruction quality and have demonstrated a high degree of adaptability. However, despite the advancements achieved by CNNs, they exhibit inherent limitations due to locality of receptive field, as well as the static nature of convolutional filter weights. GANs focus on generating visually appealing images; however, they are prone to producing artifacts and often suffer from instability during the training process.

Transformer model, recently applied to high-level computer vision tasks, has demonstrated a significant performance boost compared to CNNs. Initially designed for natural language processing (NLP) applications, Transformer leverages a multi-head self-attention mechanism, enabling it to directly model long-range dependencies by examining the relationships between all elements of the input image. However, the self-attention mechanism has quadratic computational complexity with respect to the length of the input sequence, presenting challenges for application of Transformer model to SR task. As an option spiking neural networks[1] (SNN) can potentially be applied with aim to reduce computational complexity.

*e-mail: kolesnytskiy@vntu.edu.ua

Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2024,
edited by Ryszard S. Romaniuk, Andrzej Smolarz, Waldemar Wójcik, Proc. of SPIE
Vol. 13400, 134000J · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3054855

2. SUPER-RESOLUTION

Super-resolution – is a task of restoration of HR digital image from its LR counterpart. Let D be a degradation mapping that represents the relationship between a LR image x and a HR image y:

$$x = D(y, \delta) \tag{1}$$

in which δ – parameters of degradation mapping, for example scaling factor or type and level of the noise. In practice, the exact type and parameters of degradation are typically unknown. Therefore, it is commonly modelled by downscaling the image using bicubic interpolation. So, the SR task can be defined as finding a mapping that reverses the degradation mapping D. Thus, the objective is to find a function M:

$$\hat{y} = M(x, \theta) \tag{2}$$

where $\hat{y}$ is the predicted HR approximation of the LR image x and θ the parameters of M. Since a single LR image can correspond to multiple non-identical reconstructed HR images, the SR is an ill-posed problem.

The classification of existing SR methods is shown in Figure 1. Early SR methods relied on analytical interpolation techniques such as linear, bicubic, cubic spline interpolation, etc. The main advantage of these methods is their simplicity and real-time applicability; however, the simplistic interpolation rules often lead to significant blurring of details. Reconstruction-based methods utilize prior knowledge to constrain the space of possible solutions, enabling the generation of sharper details, but those methods are resource-intensive. Learning-based methods have gained widespread popularity in solving the SR problem due to their high performance and acceptable computational complexity. These methods employ machine learning to discover statistical relationships between HR and LR image patches. As machine learning has advanced, a wide variety of models have been applied to SR tasks, including neighbor embedding methods, sparse coding methods and locally linear regression methods.

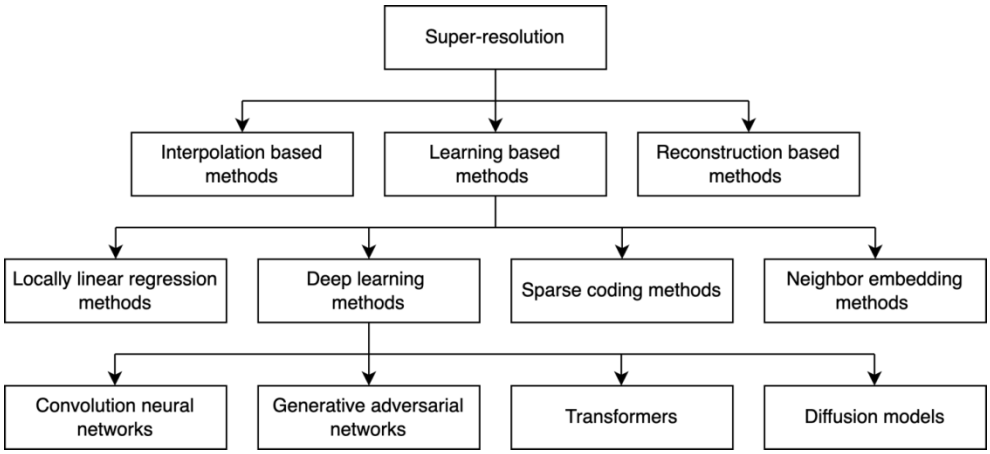

Figure 1. Classification of existing super-resolution methods.

With the advent of deep learning in 2012, CNNs have become the de facto standard for solving computer vision tasks, including SR. Thus, SRCNN[2] introduced a three-layer CNN that surpassed the performance of existing learning-based SR methods. Subsequent enhancements were achieved by increasing the network depth in VDSR[3] and incorporating residual connections in SRResNet[4]. The architecture of SRResNet was further optimized in ESDR[5], which achieved outstanding results and established itself as a benchmark for future research. ESPCN[6] introduced sub-pixel convolution, enabling the upscaling operation to be performed in the final step, thereby reducing memory requirements and improving efficiency. RCAN[7] advanced the field further by introducing channel attention mechanism.

An alternative to CNN-based methods are generative approaches, particularly those based on generative adversarial networks (GANs) and diffusion models. SRGAN[4] introduced a GAN architecture that combines an adversarial loss function with a content loss function, enabling the generation of high-quality images with enhanced perceptual fidelity. ESRGAN[8], an evolution of SRGAN, has set the standard for GAN-based methods. Diffusion models, such as SRDiff[9], represent a relatively new approach that further narrows the gap between the quality of reconstructed images and their subjective perception by humans, albeit at the cost of significant computational resources.

Since 2017, Transformer has achieved significant breakthroughs in the field of NLP, with the self-attention mechanism and novel network structure proving highly effective for processing sequential data. In 2020, the ViT[10] (Vision Transformer) was introduced, adapting the Transformer model for computer vision tasks. ViT has demonstrated superior performance and competitiveness compared to CNNs. Its application in computer vision, particularly in image restoration, is currently a subject of active research.

In the case of learning-based methods, the SR task is framed as an optimization problem, which involves finding a set of parameters $\hat{\theta}$ for the function M that minimizes the loss function L for the original HR image y and its approximation $\hat{y}$:

$$\hat{\theta} = argmin_{\theta} L(\hat{y}, y) \tag{3}$$

The most prevalent loss functions are mean absolute error and mean squared error. Also, Charbonnier loss function is frequently employed.

Evaluating the quality of a reconstructed image is also a challenging task, as it is primarily determined by human perception and depends on various attributes such as sharpness, contrast, and level of noise. Subjective human assessment methods, such as MOS (Mean opinion score), generally provide the most accurate results. However, involving human evaluators is time-consuming and burdensome, particularly for large datasets. An alternative approach involves using reference images as a base of objective assessments. The most commonly used metrics for objective evaluation are PSNR (Peak signal-to-noise ratio) and SSIM (structural Similarity Index Measure).

3. VISUAL TRANSFORMER

The ViT[10] model directly adapts the original Transformer[11] for computer vision tasks. An overview of the model is depicted in Figure 2. The Vision Transformer is composed of N blocks, which are analogous to the encoder blocks of the original Transformer. Each block consists of two sequential subblocks with residual connections: a multi-head self-attention block and a fully connected multilayer perceptron block. The input image is represented as patch embeddings, similar to token embeddings in NLP. This is accomplished by dividing the image into 2d patches, and then applying trainable linear projection for each patch to obtain d-dimensional embeddings.

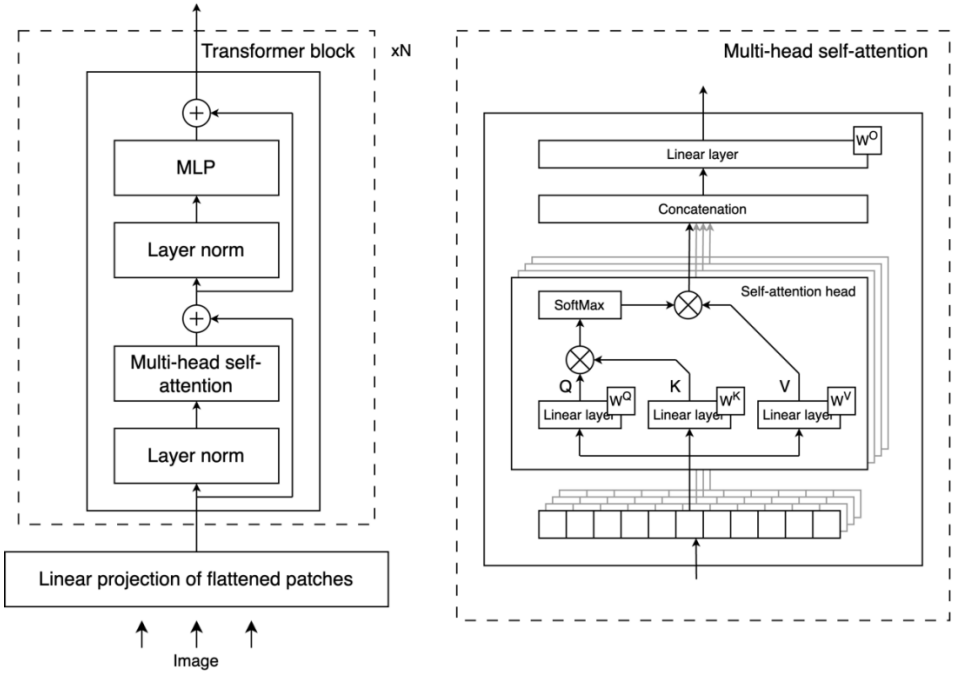

Figure 2. Visual Transformer overview.

Transformer was designed for processing sequences; however, it does not explicitly consider the position of each element within the sequence. To mitigate this limitation, positional embeddings are employed to encode the position of each image patch. Patch embeddings are combined with their corresponding positional embeddings before being input to Transformer. This mechanism allows to account for the relative positions of patches and to capture spatial information from the image.

The self-attention, which is central to Transformer, allows to model interactions and dependencies among elements in the input sequence. The output of the self-attention is a weighted sum of the input values, where the weight assigned to each value (the attention weight) is determined by a compatibility function between the query and the corresponding key. Consider a sequence of n embeddings $\{X_1, X_2, X_3, \dots X_n\}$, where $X \in \mathbb{R}^{n \times d}$ and d is the embedding dimension. Let $W^Q \in \mathbb{R}^{n \times d_Q}$, $W^K \in \mathbb{R}^{n \times d_K}$, $W^V \in \mathbb{R}^{n \times d_V}$ be the learnable weight matrices for the linear projections of queries, keys, and values, respectively. Then the self-attention can be defined as:

$$Q = X \cdot W^Q \tag{4}$$

$$K = X \cdot W^K \tag{5}$$

$$V = X \cdot W^V \tag{6}$$

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_Q}}\right)V \tag{7}$$

It was shown[11] that applying the self-attention multiple times in parallel to the same input sequence enables the model to focus on information from different subspaces of the representation for various combinations of input embeddings. Thus, self-attention computed h times, with the input sequence X being projected using distinct sets of weights W_i^Q, W_i^K, W_i^V. Each application of the self-attention mechanism in this manner is termed a self-attention head. The outputs from these heads are then concatenated and projected using a weight matrix W^O:

$$head = Attention\left(XW_i^Q, XW_i^K, XW_i^V\right) \tag{8}$$

$$MultiHead(X) = Concat(head_1, \dots head_2)W^O \tag{9}$$

where $W_i^Q \in \mathbb{R}^{n \times d_Q}$, $W_i^K \in \mathbb{R}^{n \times d_K}$, $W_i^V \in \mathbb{R}^{n \times d_V}$, $W^O \in \mathbb{R}^{hd_v \times n}$. To reduce the computational burden of multi-head self-attention calculation, each head operates on only a part of each embedding, such that: $d_Q = d_k = d_v = \frac{d}{h}$.

4. SOLVING SUPER-RESOLUTION USING TRANSFORMER MODEL

The **IPT**[12] (Image Processing Transformer) is the first application of Transformer to SR. The proposed network, consists of an input component for feature extraction from the input image, a body, and an output component for image reconstruction from the extracted features. The input and output components vary depending on the specific task, such as denoising, SR, or rain removal. The body of the **IPT** is composed of 12 encoding blocks and 12 decoding blocks, each constructed similarly to ViT blocks. The input component includes a convolutional layer and two ResNet layers. For SR tasks, the output component comprises one or two sub-pixel convolutional layers[6]. The **IPT** model demonstrated significant performance improvements for scaling factors of x2, x3, and x4 across all datasets compared to state-of-the-art CNNs, such as RCAN. However, it is noteworthy that the **IPT** model contains 114M parameters, compared to RCAN's 16M. Additionally, it was observed that when trained on a limited dataset (less than 60% of the ImageNet dataset), **IPT** underperforms comparing to CNNs, although its performance increases with larger training datasets.

Effective application of Transformer to computer vision tasks involves challenges that stem from the differences between visual and language domains. The first difference is scale. Images usually contain visual elements of different scales, making it challenging to process them with Transformer, which, similar to token processing in NLP, works with elements of a same size. The second difference is the volume of information, as the computational complexity of self-attention calculating is quadratic with respect to the length of the input sequence, what becomes more critical while processing of high-resolution images.

The Swin Transformer was proposed by Ze Liu et al.[13] – general-purpose visual transformer designed to address these challenges. It improves efficiency by using a local self-attention mechanism, where self-attention is calculated only for window of N×N embeddings instead of whole sequence of embeddings. Such approach allows to reduces

computational complexity to a linear scale. But, to maintain connections between visual elements in different windows, windows should be "shifted" while calculation of self-attention at the deeper layers of the network.

The **SwinIR** network[14], inspired by the Swin Transformer, achieved a PSNR improvement of 0.08-0.28 dB over **IPT** while maintaining a significantly smaller model size of 11.8M and being trained on a much smaller dataset, thus establishing a robust foundation for future research. The architecture of **SwinIR**, illustrated in Figure 3, is similar to architecture of RCAN network and consists of three main components: a shallow feature extraction module, a deep feature extraction module, and a high-resolution image reconstruction module. The shallow feature extraction module is a convolutional layer with a 3x3 core, responsible for extracting shallow features and transforming the image into a higher-dimensional space for subsequent processing by the deep feature extraction module. The deep feature extraction module consists of N_{RG} RG (residual group) and a convolutional layer. Each RG consists of N_{TB} TB (transformer block) and a convolutional layer. In **SwinIR**, the TBs are based on the ViT blocks (Figure 2), with the difference that local self-attention with shifted windows applied. Shallow and deep features are fused before passing into the high-resolution image reconstruction module, which implemented as a sub-pixel convolutional layer[6].

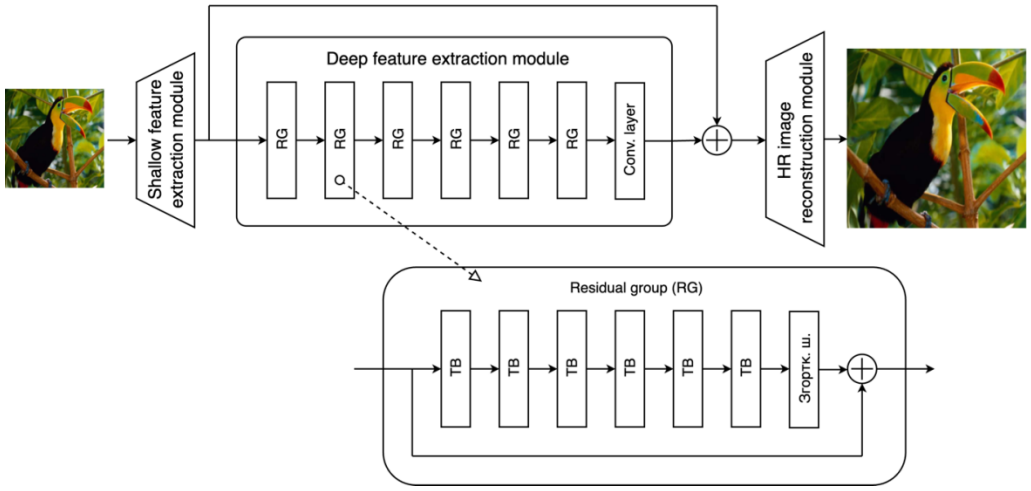

Figure 3. SwinIR architecture.

Subsequent research has offered networks with architectures similar to **SwinIR**, focusing primarily on developing efficient methods for capturing more global information while maintaining network size, local self-attention window size, and training dataset size. In the **EDT**[15] (encoder-decoder-based transformer) network, the approach proposed according to which input feature map divided into two equal parts along the channel dimension and rectangular self-attention windows of different directions applied then to each part creating cross-shaped receptive field. The **ART** (attention retractable transformer) network is proposed by Jiale Zhang et al.[16], here every even TB in RG is replaced by a SAB (sparse attention block), where self-attention is applied to patches spaced at certain intervals. A similar approach is proposed in the **DWT**[17] (Detailed window transformer), but with the difference that interval between patches increases with the increasing of network depth. The RWin-SA (Rectangle-window self-attention) is proposed by Zheng Chen et al.[18], featuring TBs with overlapping rectangular self-attention windows, similar to **EDT**, but with windows of different orientations applied to different self-attention heads. The study includes networks with rectangular self-attention windows – **CAT-R**, and networks where one side of the window is equal to the hight or width of the image – **CAT-A**. Additionally, RWin-SA TB is extended with the LCM (Locality Complementary Module), a convolution operation that it performed on the V in the self-attention block in parallel with the attention part.

Network **SRFormer** proposed by Yupeng Zhou[19] is consists of PSA (Permuted Self-Attention) TBs, here channel dimensions of K and V reduced with aim to enhances efficiency of self-attention computation. This approach allows to obtain larger self-attention window sizes while maintaining the network's parameter count and computational complexity.

In the **SwinFIR**[20] application of frequency domain representation investigated by replacing convolutional layer in each RG with SFB (spatial frequency block). The SFB is composed of two branches: a frequency branch and a spatial branch.

The frequency branch, performs sequential forward and revers Fourier transforms with aim to extract global features. And the spatial branch consists of two consecutive convolutional layers.

The application of channel attention is explored by Xiangyu Chen et al.[21] and the **HAT** (Hybrid Attention Transformer) is proposed. In this model, a Channel Attention Block (CAB), similar to the one used in RCAN, is added parallel to the self-attention block in each TB. In addition, in **HAT**, the final convolutional layer in each RG is replaced with an OCAB (overlapping cross-attention block). Unlike traditional local self-attention in TB, where Q, K, and V are computed for windows of the same size, in OCAB, K and V are computed for a larger window than Q, introducing more inter-window connections. The **DAT** (Dual Aggregation Transformer) network is proposed by Zheng Chen et al.[22], where DSTB (Dual spatial transformer block) and DCTB (Dual Channel Transformer Block) are alternately applied within RGs. Furthermore, each TB in this network is augmented with a convolutional layer on V, in parallel to the self-attention block, and an AIM (Adaptive interaction module), which effectively integrates features obtained from both the self-attention block and the convolutional layer. This approach efficiently combines channel and spatial features at both the TB level and the deep feature extraction module level, enhancing the network's representational capability.

The feasibility of global information pre-aggregating prior to computing local self-attention is studied by Zheng Chen et al.[23] RGM (recursive generalization module) introduced for this purpose. This module utilizes the recursive application of a single convolutional layer to the input feature map with aim to generate a compressed feature map. The RA-SA (Recursive-Generalization Self-Attention) block, built upon Rwin-SA, incorporates RGM and calculates the values of K and V based on the compressed feature map, while Q is derived from the corresponding local self-attention window. The **RGT** (Recursive Generalization Transformer) network consists of on RA-SA and Rwin-SA blocks, which alternating sequentially.

Comparison of the characteristics and performance of the aforementioned networks presented on Table 1. Comparation based on Urban100[24] test dataset with scaling factor x4. State-ot-the-art CNN networks with channel attention and non-local sparse attention mechanisms such as RCAN and NLSA[25] are included for reference. **SwinIR** used as the baseline for comparison, with the columns ΔPSNR and ΔSSIM indicating the changes in the metrics comparative to **SwinIR**.

Table 1. Performance and parameters comparison of SR networks based on transformer architecture.

Training set	Window size	Params. count $\times 10^6$	Publication date	**Network**	Urban100 x4			
					PSNR	SSIM	ΔPSNR	ΔSSIM
DIV2K		16.0	2018	**RCAN**	26.82	0.8087	-0.63	-0.0167
ImageNet		115.5	12.2020	**IPT**	27.26		-0.19	
DIV2K			2021	**NLSA**	26.96	0.8109	-0.49	-0.0145
	8x8	11.8	08.2021	**SwinIR**	27.45	0.8254	0	0
DF2K	6x24	11.7	12.2021	**EDT**	27.46	0.8246	0.01	-0.0008
DF2K	16x16	20.8	05.2022	**HAT**	27.97	0.8368	0.52	0.0114
DF2K	12x12	14.0	08.2022	**SwinFIR**	27.87	0.8348	0.42	0.0094
DF2K	8x8	16.5	01.2022	**ART**	27.77	0.8321	0.32	0.0067
DF2K	4x16	16.6	11.2022	**CAT-R**	27.62	0.8292	0.17	0.0038
DF2K	4xW[H]	16.6	11.2022	**CAT-A**	27.89	0.8339	0.44	0.0085
DF2K	8x32	13.3	03.2023	**RGT**	27.98	0.8369	0.53	0.0115
DF2K	24x24	10.4	03.2023	**SRFormer**	27.68	0.8311	0.23	0.0057
DF2K	16x16	12.0	05.2023	**DWT**	27.81	0.8324	0.36	0.0070
DF2K	8x32	14.8	08.2023	**DAT**	27.87	0.8343	0.42	0.0089
DF2K	9x9	12.0	02.2024	**Uniwin**	27.90	0.8362	0.45	0.0108

It is evident that the size of the self-attention window directly affects performance, highlighting the critical role of global information for SR. Consequently, the search for efficient methods to incorporate extensive global information remains a pertinent research focus. As for now, the most effective results have been demonstrated by **DWT**, where sparse attention with variable intervals was applied, and **RGT** where a recursive convolutional layer is employed to compress the input feature map in spatial dimension before computing self-attention.

The extension of the TBs with convolutional layers in parallel to the self-attention blocks proposed in the **CAT**, **HAT** and **DAT** networks also has a significant positive effect on the performance in the SR. This may indicate either the limited capabilities of local features extraction, or the lack of spatial information and requires further research.

The observed high performance of HAT and DAT networks underscores the efficacy of channel attention mechanisms, highlighting the variability in feature importance across the channel dimension. This variability suggests that reducing channel dimensionality could be a promising research direction. Such an approach has the potential to not only enhance model performance but also decrease computational time. A similar methodology is applied in **SRFormer**.

The **SwinFIR** network demonstrated high value of SSIM metric, achieved despite using a small attention window, which highlights the potential benefits of employing frequency domain representation for SR. So, the further research into this direction is highly warranted.

A key feature of the self-attention mechanism in Transformer model is its capacity to focus on critical information within the data stream, a characteristic that aligns with human biological systems[1]. This makes the implementation of self-attention mechanisms using SNN (spiking neural networks) an intriguing prospect, as both methodologies are biologically inspired. Zhaokun Zhou et al. proposed Spikformer – SNNs based Transformer for image classification. Investigating a similar approach for SR tasks is highly appropriate. The application of SNNs could reduce computational complexity, enhance energy efficiency and facilitate effective real-time processing.

5. CONCLUSIONS

1. The application of transformer architecture to SR task has resulted in substantial performance improvements (ΔPSNR: 0.5–1.2 dB, ΔSSIM: 0.0055-0.0234) compared to state-of-the-art on deep neural networks-based approaches, such as CNNs and GANs.

2. However, the application of Transformer to SR tasks faces next challenges: high computational complexity when global self-attention applied, limitations in capturing spatial information, the need to balance computational complexity with the amount of captured global information, the high capacity of Transformer-based networks, and consequently, need for large volumes of training data.

3. The reviewed works primarily focus on finding a balance between the amount of captured global information and computational complexity. Various forms of local self-attention are being introduced and investigated, with sparse self-attention currently providing the best results. An alternative approach is the method proposed in RGT, which implements compression of the input feature map before applying self-attention. Combining transformer architecture with CNNs, employing channel attention, and utilization of frequency domain representation are also promising research directions.

4. To effectively address SR tasks in real-time scenarios, it is worthwhile to explore the implementation of self-attention mechanisms using SNNs.

REFERENCES

[1] Kolesnytskyj, O. K., Bokotsey, I. V. and Yaremchuk, S. S., "Optoelectronic implementation of pulsed neurons and neural networks using bispin-devices," Opt. Mem. Neural Networks 19(2), 154-165 (2010).

[2] Dong, C., Loy, C. C., He, K. and Tang, X., "Learning a Deep Convolutional Network for Image Super-Resolution," In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T. (eds), [Computer Vision – ECCV 2014], Springer, Cham 184-199 (2014). https://doi.org/10.1007/978-3-319-10593-2_13

[3] Kim, J., Lee, J. K. and Lee, K. M., "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 1646-1654 (2016). https://doi.org/10.1109/CVPR.2016.182

[4] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J. and Wang, Z., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 105-114 (2017). https://doi.org/10.1109/CVPR.2017.19

[5] Lim, B., Son, S., Kim, H., Nah, S. and Lee, K. M., "Enhanced Deep Residual Networks for Single Image Super-Resolution," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 1132-1140 (2017). https://doi.org/10.1109/CVPRW.2017.151
[6] Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D. and Wang, Z., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 1874-1883 (2016). https://doi.org/10.1109/CVPR.2016.207
[7] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B. and Fu, Y., "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," (2018). https://doi.org/10.48550/arXiv.1807.02758
[8] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y. and Loy, C. C., "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," (2019). https://doi.org/10.48550/arXiv.1809.00219
[9] Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q. and Chen, Y., "SRDiff: Single image super-resolution with diffusion probabilistic models," Neurocomputing 479, 47–59 (2022).
[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G. and Gelly, S., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," (2021). https://doi.org/10.48550/arXiv.2010.11929
[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I., "Attention is All you Need," 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA (2017).
[12] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C. and Gao, W., "Pre-Trained Image Processing Transformer," (2021). https://doi.org/10.48550/arXiv.2012.00364
[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," (2021). https://doi.org/10.48550/arXiv.2103.14030
[14] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L. and Timofte, R., "SwinIR: Image Restoration Using Swin Transformer," (2021). https://doi.org/10.48550/arXiv.2108.10257
[15] Li, W., Lu, X., Qian, S. and Lu, J., "On Efficient Transformer-Based Image Pre-training for Low-Level Vision," (2023). https://doi.org/10.48550/arXiv.2112.10175
[16] Zhang, J., Zhang, Y., Gu, J., Zhang, Y., Kong, L. and Yuan, X., "Accurate Image Restoration with Attention Retractable Transformer," (2023). https://doi.org/10.48550/arXiv.2210.01427
[17] Park, S. and Choi, Y. S., "Image Super-Resolution Using Dilated Window Transformer," IEEE Access 11, 60028–60039 (2023). https://doi.org/10.1109/ACCESS.2023.3284539
[18] Chen, Z., Zhang, Y., Gu, J., Zhang, Y., Kong, L. and Yuan, X., "Cross Aggregation Transformer for Image Restoration," (2022). https://doi.org/10.48550/arXiv.2211.13654
[19] Zhou, Y., Li, Z., Guo, C.-L., Bai, S., Cheng, M.-M. and Hou, Q., "SRFormer: Permuted Self-Attention for Single Image Super-Resolution," Proceedings of the IEEE/CVF International Conference on Computer Vision, 12780-12791 (2023).
[20] Zhang, D., Huang, F., Liu, S., Wang, X. and Jin, Z., "SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution," (2023). https://doi.org/10.48550/arXiv.2208.11247
[21] Chen, X., Wang, X., Zhou, J., Qiao, Y. and Dong, C., "Activating More Pixels in Image Super-Resolution Transformer," (2023). https://doi.org/10.48550/arXiv.2205.04437
[22] Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X. and Yu, F., "Dual Aggregation Transformer for Image Super-Resolution," (2023). https://doi.org/10.48550/arXiv.2308.03364
[23] Chen, Z., Zhang, Y., Gu, J., Kong, L. and Yang, X., "Recursive Generalization Transformer for Image Super-Resolution," (2024). https://doi.org/10.48550/arXiv.2303.06373
[24] Huang, J.-B., Singh, A. and Ahuja, N., "Single image super-resolution from transformed self-exemplars," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 5197-5206 (2015). https://doi.org/10.1109/CVPR.2015.7299156
[25] Mei, Y., Fan, Y. and Zhou, Y., "Image Super-Resolution with Non-Local Sparse Attention," EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 3516-3525 (2021). https://doi.org/10.1109/CVPR46437.2021.00352