S. L. Kozlov; O. K. Kolesnytskyi, Cand. Sc. (Eng.), Professor

APPLICATION OF TRANSFORMER ARCHITECTURE TO PROBLEM SUPER-RESOLUTION

In the course of the last 15 years convolutional neural networks are basic approach to the solution of the problems, dealing with computer vision and demonstrate high level of productivity. However, transformer architecture, that showed high performance in the field of natural language processing, find wide application in the field, connected with the solution of computer vision problems and demonstrate comparative or even better results. The authors considered the application of the transformer architecture to super-resolution problem, the paper also contains a short review of the previous approaches. Direct application of the original transformer architecture enabled to provide the performance, comparable with the present-day convolutional neural networks. However, efficient application of transformer architecture to the problems of computer vision is connected with the challenges, following from the differences between visual and language domains. The first difference is the scale, as the images contain visual elements of various scales, this complicates their processing, using the transformer architecture, that is analogous to tokens processing in NLP, it operates with the fragments of the same size. The second difference – volume of information, as the computational complexity of self-attention processing is quadratic to the length of the input sequence, that becomes especially critical for processing of the high resolution images.

The given paper analyzes 12 papers devoted to this subject, starting from 2021, they consider various approaches, aimed at elimination of these problems. The following directions of the research may be highlighted in these papers: study of the application of local attention with the windows of different forms, in particular, dispersed attention; study of channel self-attention and its combination with spatial attention; study of the possibilities of transformer architecture enlargement by means of convolutional blocks. The above-mentioned studies enabled to improve considerably the quality of the reconstructed images, but the studies are not exhaustive.

Key words: *super-resolution, transformer architecture, convoluting neural network, computer vision.*

Introduction

Rapid development of the technologies, of digital processing of images led to growing demand for high resolution images in different spheres of human activity, starting from medical visualization and monitoring systems to the production of the entertaining multimedia content. However, obtaining of high resolution images is often ~~is~~ limited by the properties of photosensitive elements or other physical limitations, resulting in the reduction of resolution, detalization and quality of images. Super-resolution (SR) – is the process of reconstruction of the high resolution image from the corresponding low resolution image , this process attracted attention as an efficient and cheap method to approach this problems.

Conventional SR methods, for instance, bicubic interpolation are simple and efficient, but they are prone to detail blurring and ringing artifacts, that negatively affects the quality of the reconstructed image. Improved methods, based on learning, such as methods of the sparse encoding or local linear regression, were suggested for the elimination of these drawbacks. Rapid growth of computational power and wide accessibility of BigData made it possible to apply deep learning to SR problem. Application of convolutional and generative-adversarial networks (CNN and GAN) was actively studied for solving SR problem during last 10 years and allows to reach high level of restoration quality and demonstrated adaptivity. However, in spite of the achievements of CNN, there exist certain limitations, connected with CNN locality property, that does not allow to model for ranging dependences efficiently and with static weights of convoluting filters. GANs are focused on the generation of the images, attractive for eye but inclined to artifacts generation and are non-stable in the process of learning.

Transformer architecture, applied for high level computer vision problems, demonstrated considerable performance improvement as compared with CNN. Transformer, initially developed for the problems of natural language processing (NLP), is based on the mechanism of multi-head self-attention, which enables to model directly far-ranging dependences, analyzing the

interconnections between all the elements of the input image. However, computational complexity of such approach is increased quadratically with the size of the image and this complicates its application for SR problems.

The purpose of this article is to review and analyze existing approaches of transformer architecture application to SR problem.

Super-resolution problem

Super-resolution – is a problem of restoration of HR (high resolution) image from one or several LR (low resolution) images. According to the number of input LR-images it is divided into SISR (single-image super resolution) and MISR (multi-image super resolution). Greater part of studies is concentrated on SISR, due to considerably wider range of potential applications. Besides, techniques, studied for SISR, may be used for MISR. Let D – be the distortion function, representing the connection between LR-images x and HR-image y:

$$x = D(y, \delta), \#(1)$$

where δ – are parameters of the distortion function, for instance, scaling coefficient or the type and level of noise. In practice, type and distortion parameters are usually unknown, that is why it is modeled, for instance, by downscaling the image using bicubic interpolation. The problem of SR can be defined as the search of the function, reverse to the distortion function D, it is necessary to find such function M, that:

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

where $\hat{y}$ – is approximation of the initial HR-image, θ – are the parameters of M function. As for one LR-image there may be several non-identical reconstructed HR-images, then the SR ill-posed problem .

Classification of the available SISR methods is presented in Fig. 1. Early SR methods were based on the application of analytical interpolation namely, linear, bicubic, interpolation by cubic splines or New Edge Directed Interpolation [1]. Main advantage of these methods is simplicity and possibility of their application in real-time, however, simple rule of interpolation leads considerable blurring of the details. Methods of reconstruction [2, 3] use prior knowledge for limitation of the space for possible solutions, this enables to generate more clear details. But, the performance of certain methods of reconstruction rapidly decreases with the increase of scaling coefficient, besides, these methods are resource-consuming. Learning methods for solving SISR problem became widely popular due to their high performance and acceptable computational complexity. Machine learning is used here for searching the statistic dependences between fragments of HR and LR images. Along with the development of machine learning wide variety of models was applied to SR problem: method of nearest neighbors embedding [4], methods of sparse coding [5], methods of local linear regression [6].

With the development of deep learning in 2012 [7] CNNs became a standard in the solution of computer vision problems, in particular, in SR problems. In SRCNN [8] three layers CNN was suggested, it surpassed the results of the available methods of SR learning. Further, its result was improved due to the enhance of the depth of VDSR network [9], or adding residual connections of SRResNet [10]. Architecture of SRResNet was optimized in ESDR [11], this network showed outstanding results and became reference for future research. Subpixel convolutional neural network was proposed in ESPCN [12] as a result it became possible to perform up sampling operation as the last step, this enabled to reduce the requirements, concerning memory and enhance the efficiency. CNN with channel attention is proposed in RCAN [13].

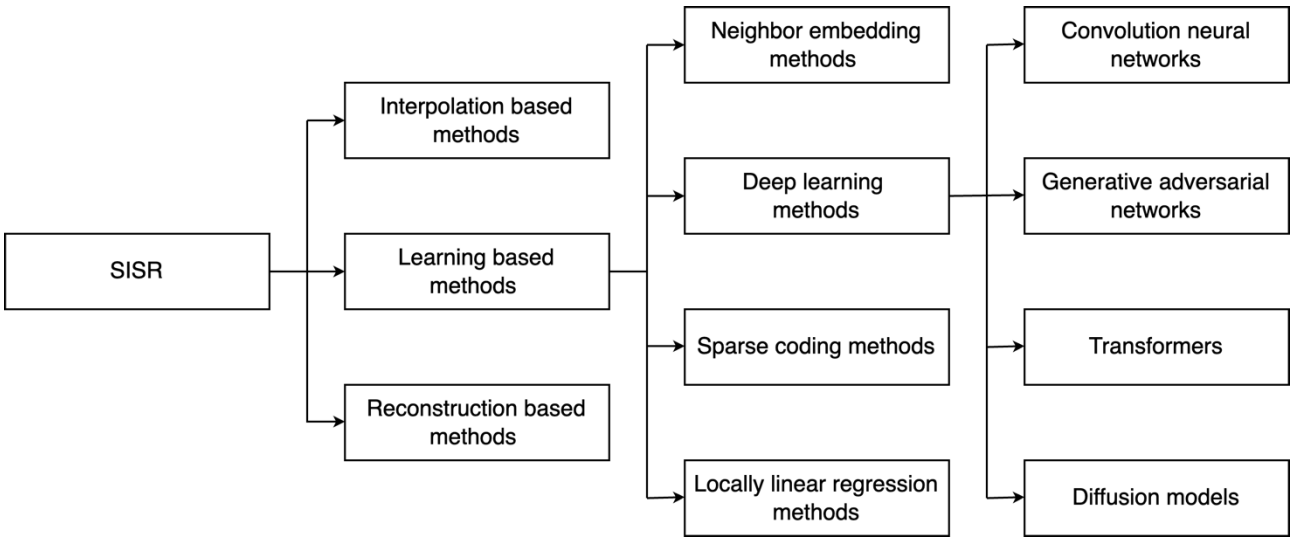

Fig. 1. Classification of the existing SR methods

The alternative to CNN-methods are generative methods, namely, methods, based on GAN and diffusion models. In SRGAN [10] GAN model is suggested, this model enables to obtain the reconstructed images of higher quality from the point of view of human perception due to the combination of the competitive function of losses and function of content losses. ESRGAN [14] is the development of SRGAN and is the reference for GAN-based methods. Diffusion models SRDiff [15] – it is rather new direction, it enables to reduce the gap between the quality of the reconstructed image and human subjective perception of the result, but it requires considerable resources.

Starting from 2017 transformers architecture made a breakthrough in the sphere of NLP. Self-attention mechanism and new architecture of the network proved their efficiency in serial data processing. Later, in 2020, ViT (Vision Transformer) [18] architecture was suggested, it is the adaptation of the transformer architecture to the tasks of computer vision. ViT showed high performance and compatibility as compared with CNN. Its application in the field of computer vision is studied, in particular, in SR problems.

In case of learning methods, SR problem is reduced to the optimization problem, i. e., to the search of such set of function M parameters $\hat{\theta}$, which minimizes the values of losses function L for original HR image y and its approximation $\hat{y}$:

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

MAE (mean absolute error) formula 5, and MSE (mean squared error) formula 6 are the most widely spread losses functions. However, due to high sensitivity of MSE to abnormal values MAE is most often mentioned in literature. Losses function Charbonnier [16] formula 7 is also often used. For the pair of images y, $\hat{y}$ with width w, height h and number of channels c number of points as $N_y = w \cdot h \cdot c$, will be determined and the space of possible positions will be determined as:

$$\Omega_y = \{(i,j,k) \in \mathbb{N}_1^3 | i < h, j < w, k < c\}, \#(4)$$

then, losses functions may be denoted as:

$$L_{MAE}(\hat{y}, y) = \frac{1}{N_y} \sum_{p \in \Omega_y} |y_p - \hat{y}_p|, \#(5)$$

$$L_{MSE}(\hat{y}, y) = \frac{1}{N_y} \sum_{p \in \Omega_y} |y_p - \hat{y}_p|^2, \#(6)$$

$$L_{\text{Charbonnier}}(\hat{y}, y) = \frac{1}{N_y} \sum_{p \in \Omega_y} \sqrt{\left|y_p - \hat{y}_p\right|^2 + \varepsilon^2} \#(7)$$

where $\varepsilon \in (0, 1]$ – is constant that guarantees the distinction of the radical expression from 0.

Assessment of the reconstructed image quality is a complex task, because it is performed by a person who percept it and depends on numerous properties, namely, sharpness, contrast or absence of noise. Obviously, the best result will give methods which are based on the subjective human assessment, for instance, MOS (mean opinion score). However, involvement of human resource is time-consuming and burdening, especially for large data sets. The alternative is application of the reference images and objective assessment. The most widely spread ~~matrix~~ metric of the objective assessment is PSNR (peak signal-to-noise ratio), which is the ratio between maximum signal level *L* (256 for images with 8 bit/channel) and MSE for original and reproduced images:

$$PSNR(\hat{y}, y) = 10 \cdot log_{10} \frac{L^2}{\frac{1}{N_y} \sum_{p \in \Omega_y} \left|y_p - \hat{y}_p\right|}. \#(8)$$

The alternative matrix, that ~~\~~better meets the requirements of images assessment from the point of view of human perception is SSIM (structural similarity index measure). SSIM is based on the comparative assessment of three components: brightness, contrast and structural similarity [17]:

$$SSIM(\hat{y}, y) = \frac{\left(2\mu_{\hat{y}}\mu_y + c_1\right)\left(2\sigma_{\hat{y}y} + c_2\right)}{\left(\mu_{\hat{y}}^2 + \mu_y^2 + c_1\right)\left(\sigma_{\hat{y}}^2 + \sigma_y^2 + c_2\right)}, \#(9)$$

where μ_y and $\mu_{\hat{y}}$ – are average values of pixels brightness of the original and reconstructed images, σ_y and $\sigma_{\hat{y}}$ – the root mean square deviation of pixel luminance between the original and reconstructed images, $\sigma_{\hat{y}y}$– covariation, $c_i = (k_i L)^2$– are variables, preventing division on 0, $k_1 = 0.01$, $k_2 = 0.03$.

Visual Transformer architecture

Architecture ViT (Visual Transformer) suggested in [18] is direct adaptation of transformer architecture, proposed in [19], ~~for~~ to computer vision problems. Fig. 2 presents basic elements of ViT architecture. Visual Transformer consists of N blocks, analogous to the coding blocks of the original transformer. Each block consists of two serial subblocks with residual connections: block of multi-head self-attention and fully-connected network of direct spreading.

Input image is presented in the form of the fragments embeddings similar to the tokens embeddings in case of NLP. For this purpose it is divided into parts, each of them is presented in the form of 1D vector, embedding of the fragments are obtained by means of linear transformation of mentioned above 1D vectors. Alternative approach to the formation of the embeddings is application of one or several convoluting layers [20].

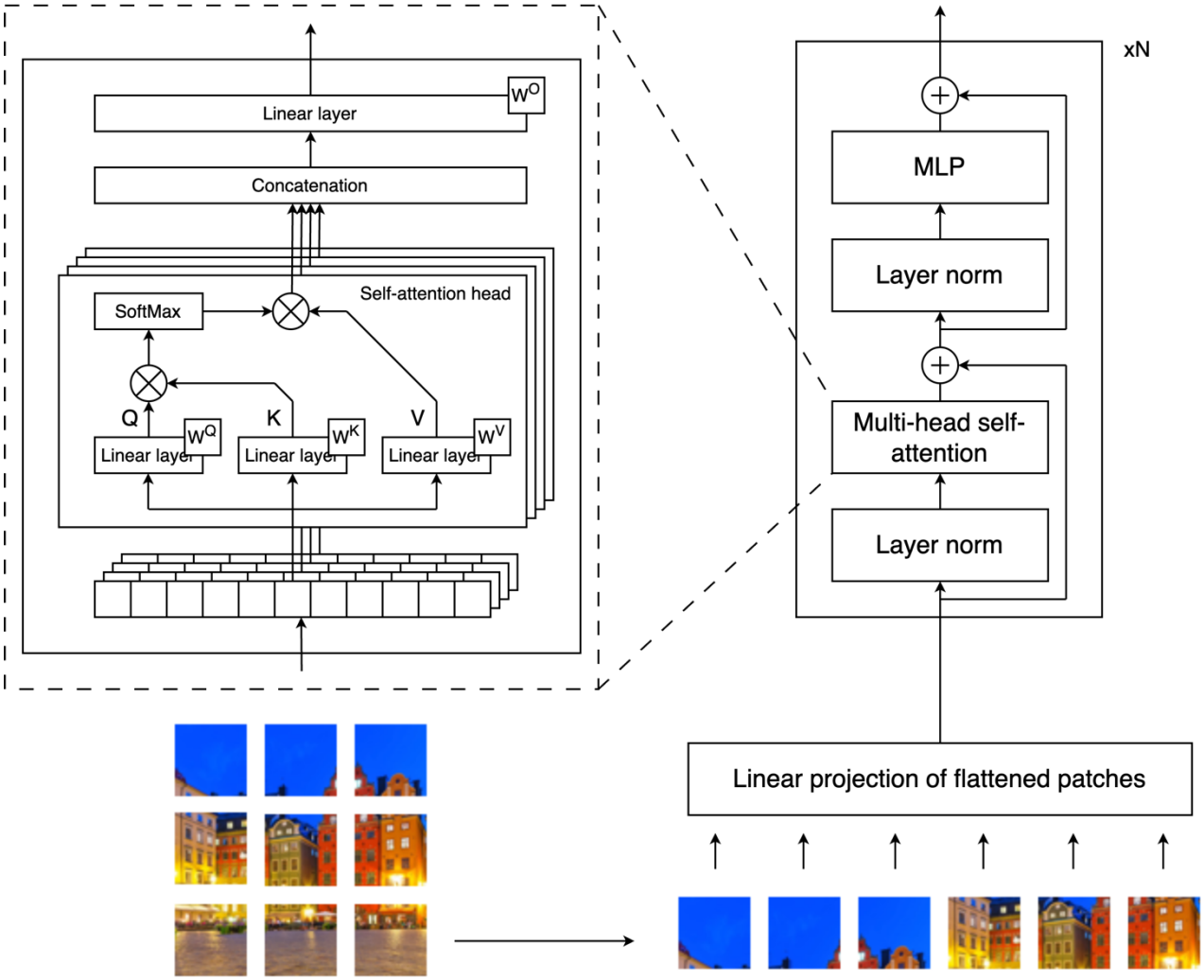

Fig. 2. General diagram of Visual Transformer

Transformer architecture is developed for processing of serial data, but it does not take into account the position of each fragment in the series. To eliminate this limitation the embedding of position that code the position of each fragment on the image is used. Fragment embeddings and corresponding position embeddings are united before the transition at the input to the blocks of the transformer.

This mechanism enables the model to take into account the relative position of fragments and extract spatial information from the image. The core of the transformer architecture is the mechanism of self-attention which models the interactions and connections between fragments in the input sequence. The result of self-attention function operation may be denoted as the weighted sum of the input values, where the weight given to each value (weight of attention) is determined by the function of the compatibility of the query and corresponding key. Let us consider the sequence of n embeddings $\{X_1, X_2, X_3, \ldots X_n\}$, where $X \in \mathbb{R}^{n \times d}$, and d – is the size of embedding and determined weight matrices, being taught $W^Q \in \mathbb{R}^{n \times d_Q}$, $W^K \in \mathbb{R}^{n \times d_K}$, $W^V \in \mathbb{R}^{n \times d_V}$ for linear projections of the queries, keys and values, correspondingly, then self-attention may be determined in the following way:

$$Q = X \cdot W^Q, \#(10)$$
$$K = X \cdot W^K, \#(11)$$
$$V = X \cdot W^V, \#(12)$$
$$\text{Attention}(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_Q}}\right)V. \#(13)$$

In [19] it is shown that the application of self-attention mechanism several times in parallel the same sequence gives the model the possibility an ability "to concentrate" on the information from different subspaces of representation for different combinations of fragments in the input sequence. In this case self-attention is calculated h times, projecting the input sequence X by means of the separate weight sets W_i^Q, W_i^K, W_i^V. Each mechanism of self-attention, used in such a way is called the head of self-attention, their results are united and projected by means of weight matrix W^O:

$$head = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V), \#(14)$$
$$MultiHead(X) = \text{Concat}(head_1, \dots head_2)W^O, \#(15)$$

where $W_i^Q \in \mathbb{R}^{n \times d_Q}$, $W_i^K \in \mathbb{R}^{n \times d_K}$, $W_i^V \in \mathbb{R}^{n \times d_V}$, $W^O \in \mathbb{R}^{hd_v \times n}$. In order to reduce computational complexity in the process of multi-head self-attention calculation, each head operates only with the part of each ~~of each~~ embedding, i. e.: $d_Q = d_k = d_v = \frac{d}{h}$.

Main advantage of self-attention mechanism, compared with the convolution mechanism is that the weight of attention is calculated dynamically, depending on the input values unlike the weights of filters, which are static for the whole input set of data. In [21] it is shown that self-attention, ~~on~~ under the condition of sufficient number of parameters is rather flexible process and enables to allocate both global and local features. It should be noted that the application of the transformer architecture requires larger sets of training data due to larger capacity of the network.

Application of the transformer architecture in super-resolution problem

For the first time transformer architecture was used for SR problem in the work [22]. The suggested network was called **IPT** (Image processing transformer) and consists of the input component for the allocation of the features from the input image, body and output component for the reconstruction of the image from the set of features. Input and output components differ depending on the type of the problem: noise reduction, SR or rain removal. Body consists of 12 encoder blocks and 12 decoder blocks, each of them is constructed similarly to ViT blocks. Output component consists of the convolution layer and two ResNet layers. Input features are divided into fragments and are presented in the form of embeddings with positional coding similarly to ViT. In case of SR the output component consists of one or two subpixel convoluting layers [12]. The suggested model showed the growth of performance for scaling factors x2, x3, x4 for all datasets in SR problem, compared with actual CNNs, for instance, RCAN. However, it should be noted that the model IPT contains 114M parameters as compared with 16M in RCAN. Besides, it was shown that, in the process of learning on the limited set of data (less than 60% of Image Net data set [7]) IPT shows worse performance than actual CNNs, but the performance increases with the increase of the training dataset size.

Efficient application of the transformer architecture to the computer vision problems is connected with the challenges, following due to the differences between visual and language domains. The first difference – is the scale. Usually, images contain visual elements of different scales, this complicates its processing by transformer architecture which is similar to the processing of the tokens in NLP, operates with the fragments of the same size. The second is the volume of information, because the computational complexity of self-attention calculation is quadratic to the length of the input sequence, this becomes critical for the processing of high resolution images.

In the research [23] Swin Transformer – is presented, it is general purpose visual transformer, which suggests the approaches to the solution of the above-mentioned problems. To improve the efficiency it is suggested to apply the mechanism of local self-attention, in this case, self-attention will be calculated not for the total set of the input embeddings but only for its part $N \times N$ fragments. This enables to obtain the linear computational complexity relatively the image dimensions. To maintain the relations between the visual elements which are in different windows, it is necessary "shift" windows when self-attention is calculated in the deeper layers of the network.

Network **SwinIR** [24] was founded on the ideas with Swin Transformer, it showed the growth of PSNR within the range of 0.08 – 0.28 dB relatively **IPT**, under the condition of far smaller size 11.8 M and learning on the smaller data set, it created foundation for future studies. **SwinIR** is constructed on the architecture, similar to RCAN, as it is shown in Fig. 3 and consists of: module for allocation of the shallow features, module for allocation of the deep features and module for restoration of HR-image. Module for allocation of the shallow features is a convolution layer with a core of 3x3, it provides the allocation of the shallow features and transition of the image in higher dimensionality space for further processing by the deep feature module. Module for allocation of deep features consists of N_{RG}RG (residual groups) and convolution layer. Each RG consists N_{TB} TB (transformer block) and convolution layer. In case of **SwinIR** TB – transformer block of ViT (Fig. 2), with the difference that local self-attention with moving windows is used here. Shallow and deep features are aggregated before the HR-image restoration module , this module in case of SR problem is subpixel layer [12].

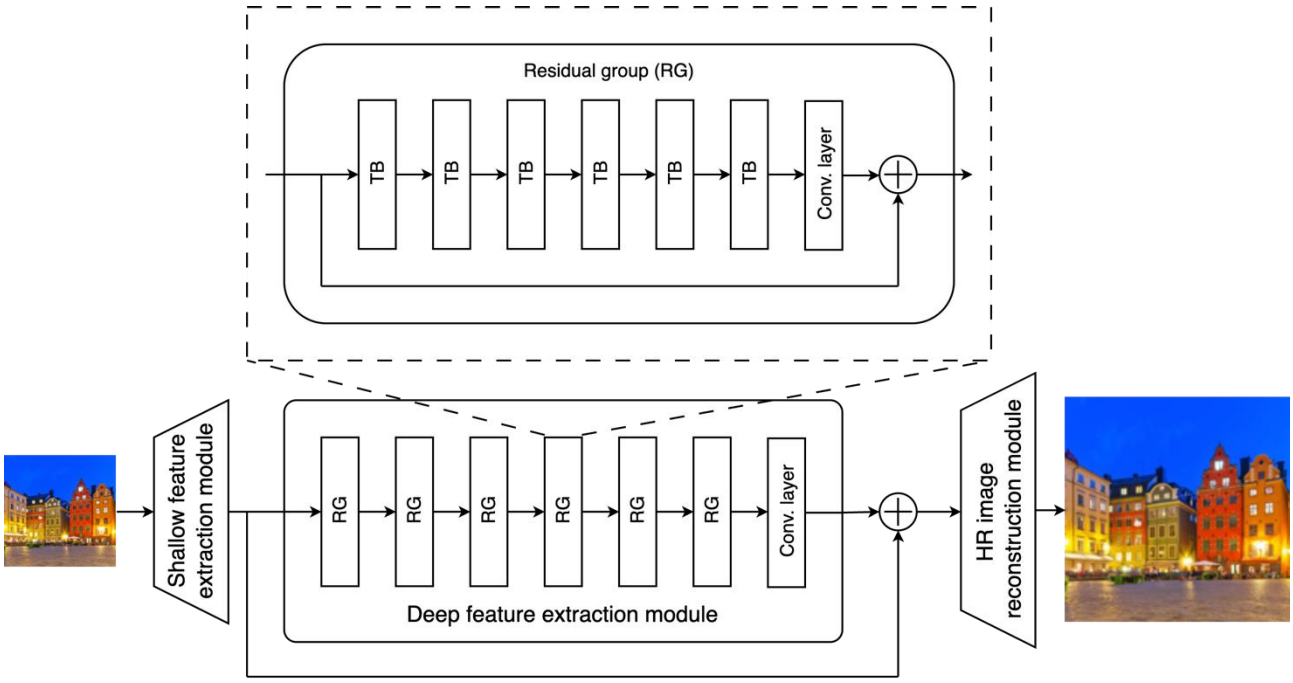

Fig. 3. Architecture of Swin IR network

Networks, suggested in further studies have the architecture, similar to **SwinIR** and are concentrated on searching the efficient method of involving greater part of the global information on condition of maintaining the network size , local self-attention window size and size of the training data set. In the network **EDT** (encode-decoder-based transformer) [25] the approach is suggested where the input map of features is divided into two equal parts according to channel dimension and rectangular windows of self-attention of the vertical or horizontal orientation are applied, forming cross-like receptive field. In [26] **ART** (attention retractable transformer), is suggested, even blocks in RG are replaced by SAB (sparse attention block), where self-attention is applied to the fragments, located through certain interval one from another. Similar approach is suggested in **DWT** (Detailed window transformer) [27], but in this case the interval between the fragments increases with the depth of the network. In the paper [28] RWin-SA (Rectangle-window self-attention) – TB with rectangle self-attention windows which are intersected like **EDT** windows, but windows of different orientation are applied to different heads of self-attention. Rectangular windows of self-attention – network CAT-R, and rectangular windows for which one side equals the height or width of the image – network CAT-A are investigated. Besides, RWin-SA is expanded by means of LCM module (Locality complementary module), which is the convoluting layer by V,

located parallel to self-attention block. Application of the moving windows of self-attention is studied in **Uniwin** [29].

In the paper [30] **SR Former** is suggested, it consists of the blocks PSA (Permuted self-attention), which as a result of decreasing channel dimensionality K and V enable to improve the efficiency of self-attention computation and increase the dimensions of self-attention window maintain the number of the networks parameters and computational complexity.

In the transformer **Swin FIR** [31] the possibility of application of frequency representation of information, replacing the convoluting layer in each RG by SFB (spatial-frequency block), which consists of two branches: frequency and spatial, is investigated. Frequency branch is similar to the branch, suggested in [32] and performs serially direct and reverse Fourier transform to allocate global features. Spatial consists of two serial convolution layers.

Application of channel self-attention is studied in [33]. In the study **HAT** (hybrid attention transformer) is suggested, here CAB (channel attention block) added parallelly to self-attention block, similarly to channel attention block in RCAN-. The last convolution layer in each RG in HAT was replaced by OCAB (over lapping cross-attention block). Unlike self-attention in ordinary TB, where Q, K, V are calculated for the windows of the same size, in OCAB K and V are calculated for the window larger that the window, for which Q is calculated, this may promote the creation of cross-window relations. In the paper [34] **DAT** (dual aggregation transformer) network is suggested, in this networks DSTB (dual spatial transformer block) blocks and DCTB (dual channel transformer block) attention are used in turn of inside RG. Besides, each TB in this network is expanded by the convolution layer parallel to the block of attention, and AIM (adaptive interaction module) module, which allows to unite efficiently the features, obtained from self-attention block and convolution layer. Such approach allows to unite efficiently the features of channel and spatial detentions both on TB level and on the level of the module of deep features allocation, this must have the positive impact on the representative abilities of the network.

In the study [35] the possibility of the preliminary aggregation of global information before the calculation of the local self-attention is studied. For this purpose RGM (recursive generalisation module), is proposed, the given module by means of recursive application of the convolution layer to the input card of features enables to obtain the compressed card of features. Block RA-SA (recursive-generalization self-attention), based on Rwin-SA, contains RGM, and calculate values K and V on the base of the compressed features card but Q on the base of the corresponding window of local self-attention. Network **RGT** (recursive generalization transformer) where RA-SA blocks sequentially alternate with Rwin-SA blocks, is built on the base of RA-SA block.

Table 1 contains the comparison of the characteristics and performance of the considered above networks, based on the Urban100 [36] test dataset with an upscaling factor of x4. For comparison actual CNN with the mechanisms of channel and non-local sparse attention – RCAN and NLSA are presented [37]. As the base for comparison the network **Swin IR** is chosen, in columns ΔPSNR and ΔSSIM the change of the metrics relatively **Swin IR** is shown.

It is quite clear, that the area of the attention window directly influences the performance, this proves the importance of global information for the solution of SR problem. The search of the efficient way of the involvement as much as possible of global information remains actual. Network **DWT** where sparse attention with variable interval was used and **Uniwin,** were sliding attention windows were suggested to use, showed the best results. The approach, used in **RGT** where it is suggested to use the recursive convoluting layer for compression of the input map of features by spatial dimensions prior to computing the self-attention, should be noted.

Table 1

Comparison of the parameters and performance of SR networks, constructed on the base of transformer architecture

Date of publication	Network	Training set	Dimension of self-attention window	Number of parameters $\times 10^6$	Urban100 (x4)			
					PSNR	SSIM	ΔPSNR	ΔSSIM
2018	**RCAN**	DIV2K		16.0	26.82	0.8087	-0.63	-0.0167
12.2020	**IPT**	ImageNet		115.5	27.26		-0.19	
2021	**NLSA**	DIV2K[11]			26.96	0.8109	-0.49	-0.0145
08.2021	**SwinIR**		8x8	11.8	27.45	0.8254	0	0
12.2021	**EDT**	DF2K[11, 38]	6x24	11.7	27.46	0.8246	0.01	-0.0008
05.2022	**HAT**	DF2K	16x16	20.8	27.97	0.8368	**0.52**	**0.0114**
08.2022	**SwinFIR**	DF2K	12x12	14.0	27.87	0.8348	0.42	0.0094
01.2022	**ART**	DF2K	8x8	16.5	27.77	0.8321	0.32	0.0067
11.2022	**CAT-R**	DF2K	4x16	16.6	27.62	0.8292	0.17	0.0038
11.2022	**CAT-A**	DF2K	4xW[H]	16.6	27.89	0.8339	0.44	0.0085
03.2023	**RGT**	DF2K	8x32	13.3	27.98	0.8369	**0.53**	**0.0115**
03.2023	**SRFormer**	DF2K	24x24	10.4	27.68	0.8311	0.23	0.0057
05.2023	**DWT**	DF2K	16x16	12.0	27.81	0.8324	0.36	0.0070
08.2023	**DAT**	DF2K	8x32	14.8	27.87	0.8343	0.42	0.0089
02.2024	**Uniwin**	DF2K	9x9	12.0	27.90	0.8362	0.45	0.0108

Expansion of the transformer block by the convoluting layers parallelly to self-attention blocks, as it is proposed in the networks **CAT**, **HAT** and **DAT** also positively affects the performance in SR problem. It might indicate either about limited possibilities of the transformer in allocation of local features or lack of spatial information and requires further study.

Networks **HAT** and **DAT** also showed high performance that proves positive effect of the usage of channel attention. Thus, features in channel dimensionality also have different weight, this makes the reduction of channel dimensionality an interesting direction of the research, because this will lead to the increase of performance and decrease the time of computations. Similar approach is used in **SRFormer**.

It is worth mentioning the high value of SSIM metric on the condition of the small window of self-attention for **SwinFIR** network, which might indicate the positive effect of frequency representation of information for SR problem and requires further study.

Key aspect of the self-attention mechanism in the transformer architecture is the possibility to concentrate on the important information from the data flow that is an integral property of human biological system [39]. Implementation of the self-attention mechanism based on spiking neural networks (SNN) is very perspective [40, 41]. In the study [42] the combination of the transformer architecture and SNN [43] was suggested for the solution of the problem of images classification. It is expedient to study the similar approach in case of SR problem. Application of SNN [44, 45] enables to provide higher level of energy efficiency and will give the possibility of efficient solution of the problem in real time.

Conclusions

1. Application of the transformer architecture to SR problem allowed to achieve considerable growth of the performance (ΔPSNR: 0.5-1.2 dB, ΔSSIM: 0.0055-0.0234) as compared with actual approaches, based on deep neural networks, such as CNN or GAN.

2. However, application of the transformer architecture to SR problem is connected with a number of challenges, namely: high computational complexity in case of usage of global self-attention, limitation in obtaining spatial information, need of compromise searching between computational complexity and volume of involved global information, high capacity of the

networks, constructed on the transformer architecture and as a consequence, need in large volumes of training data.

3. Analyzed studies are mainly concentrated on searching compromise between the volume of the involved global information and computational complexity. Various forms of local self-attention are implemented and studied to enhance model performance, and, at present it can be stated that sparse self-attention provides the best result. The alternative approach is the method of compression of the input map of features prior the application of self-attention, this approach was suggested in RGT. Combination of the transformer architecture with CNN, application of channel self-attention and usage of frequency representation of information are also promising directions of the research.

4. For efficient solution of SR problem in real-time conditions it is expedient to study the possibility of realization of self-attention mechanism, using SNN.

REFERENCES

1. New edge-directed interpolation [Electronic resource] / Li Xin, M. T. Orchard // IEEE Transactions on Image Processing. – 2001. – Vol. 10, № 10. – P. 1521 – 1527. – Access mode: https://doi.org/10.1109/83.951537 (date of access: 15.02.2024).

2. SoftCuts: A Soft Edge Smoothness Prior for Color Image Super-Resolution [Electronic resource] / Shengyang Dai, Mei Han,Wei Xu[et al.] // IEEE Transactions on Image Processing. – 2009. – Vol. 18, № 5. – P. 969 – 981. – Access mode: https://doi.org/10.1109/tip.2009.2012908 (date of access: 15.02.2024).

3. Image super-resolution using gradient profile prior [Electronic resource] / Jian Sun, Zongben Xu, Heung-Yeung Shum // 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, 23–28 June 2008. – Access mode: https://doi.org/10.1109/cvpr.2008.4587659 (date of access: 15.02.2024).

4. Super-resolution through neighbor embedding [Electronic resource] / Hong Chang, Dit-Yan Yeung, Yimin Xiong // Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Washington, DC, USA. – Access mode: https://doi.org/10.1109/cvpr.2004.1315043 (date of access: 15.02.2024).

5. Image Deblurring and Super-Resolution by Adaptive Sparse Domain Selection and Adaptive Regularization [Electronic resource] / Weisheng Dong [et al.] // IEEE Transactions on Image Processing. – 2011. – Vol. 20, № 7. – P. 1838 – 1857. – Access mode: https://doi.org/10.1109/tip.2011.2108306 (date of access: 15.02.2024).

6. Fast image super resolution via local regression [Electronic resource] / Gu Shuhang, Sang Nong, Ma Fan // Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, 11–15 October 2012. – P. 3128 – 3131. – Access mode: https://ieeexplore.ieee.org/document/6460827 (date of access: 15.02.2024).

7. ImageNet classification with deep convolutional neural networks [Electronic resource] / Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton // Communications of the ACM. – 2017. – Vol. 60, № 6. – P. 84 – 90. – Access mode: https://doi.org/10.1145/3065386 (date of access: 15.02.2024).

8. Learning a Deep Convolutional Network for Image Super-Resolution [Electronic resource] / Dong Chao, Chen Change Loy, Kaiming He[et al.] // Computer Vision – ECCV 2014, 6–12 September 2014. – P. 184 – 199. – Access mode: https://doi.org/10.1007/978-3-319-10593-2_13 (date of access: 15.02.2024).

9. Accurate Image Super-Resolution Using Very Deep Convolutional Networks [Electronic resource] / Jiwon Kim, Jung Kwon Lee, Kyoung Mu Lee // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. – Access mode: https://doi.org/10.1109/cvpr.2016.182 (date of access: 15.02.2024).

10. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [Electronic resource] / Christian Ledig,Lucas Theis; Ferenc Huszár [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July 2017. – Access mode: https://doi.org/10.1109/cvpr.2017.19 (date of access: 15.02.2024).

11. Enhanced Deep Residual Networks for Single Image Super-Resolution [Electronic resource] / Bee Lim, Sanghyun Son, Heewon Kim [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017. – Access mode: https://doi.org/10.1109/cvprw.2017.151 (date of access: 15.02.2024).

12. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network [Electronic resource] / Wenzhe Shi, Jose Caballero, Ferenc Huszár [et al.] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. – Access mode: https://doi.org/10.1109/cvpr.2016.207 (date of access: 15.02.2024).

13. Image Super-Resolution Using Very Deep Residual Channel Attention Networks [Electronic resource] / Zhang Yulun, Kunpeng Li, Kai Li[et al.] // Computer Vision – ECCV 2018, Munich, 8–14 September 2018. – P. 294 – 310. – Access mode: https://doi.org/10.1007/978-3-030-01234-2_18 (date of access: 15.02.2024).

14. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks [Electronic resource] / Wang Xintao, Ke Yu, Shixiang Wu [et al.] // Computer Vision - ECCV 2018 Workshops, Munich, 8–14 September 2018. – P. 63 – 79. – Access mode: https://doi.org/10.1007/978-3-030-11021-5_5 (date of access: 15.02.2024).

15. SRDiff: Single image super-resolution with diffusion probabilistic models [Electronic resource] / Haoying Li, Yifan Yang, Meng Chang [et al.] // Neurocomputing. – 2022. – Vol. 479. – P. 47 – 59. – Access mode: https://doi.org/10.1016/j.neucom.2022.01.029 (date of access: 15.02.2024).

16. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution [Electronic resource] / Wei-Sheng Lai Jia-Bin Huang; Narendra Ahuja[et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July 2017. – 2017. – Access mode: https://doi.org/10.1109/cvpr.2017.618 (date of access: 15.02.2024).

17. Image Quality Assessment: From Error Visibility to Structural Similarity [Electronic resource] / Z. Wang A.C. Bovik,H.R. Sheikh[et al.] // IEEE Transactions on Image Processing. – 2004. – Vol. 13, № 4. – P. 600 – 612. – Access mode: https://doi.org/10.1109/tip.2003.819861 (date of access: 15.02.2024).

18. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [Electronic resource] / A.Dosovitskiy, L. Beyer, A. Kolesnikov[et al.] // International Conference on Learning Representations, 3–7 May 2021.– Access mode: https://openreview.net/pdf?id=YicbFdNTTy (date of access: 15.02.2024).

19. Attention is All you Need [Electronic resource] / Ashish Vaswani, Noam Shazeer, Niki Parmar [et al.] // Advances in Neural Information Processing Systems, 4–9 December 2024. – Access mode: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (date of access: 15.02.2024).

20. Early Convolutions Help Transformers See Better [Electronic resource] / Xiao Tete, Mannat Singh, Eric Mintun[et al.] // Advances in Neural Information Processing Systems: 2021, 6–14 December 2021. – Access mode: https://proceedings.neurips.cc/paper/2021/hash/ff1418e8cc993fe8abcfe3ce2003e5c5-Abstract.html (date of access: 15.02.2024).

21. On the Relationship between Self-Attention and Convolutional Layers [Electronic resource] / Cordonnier Jean-Baptiste, Loukas Andreas, Martin Jaggi // International Conference on Learning Representations , 27–30 April 2020. – Access mode: https://openreview.net/forum?id=HJlnC1rKPB (date of access: 15.02.2024).

22. Pre-Trained Image Processing Transformer [Electronic resource] / Hanting Chen, Yunhe Wang,TianyuGuo [et al.] // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. – Access mode: https://doi.org/10.1109/cvpr46437.2021.01212 (date of access: 15.02.2024).

23. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [Electronic resource] / Ze Liu, Yutong Lin, Yue Cao, Han Hu [et al.] // 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. – Access mode: https://doi.org/10.1109/iccv48922.2021.00986 (date of access: 15.02.2024).

24.SwinIR: Image Restoration Using Swin Transformer [Electronic resource] / Jingyun Liang, Jiezhang Cao, Guolei Sun[et al.] // 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October. 2021. – Access mode: https://doi.org/10.1109/iccvw54120.2021.00210 (date of access: 15.02.2024).

25. On Efficient Transformer-Based Image Pre-training for Low-Level Vision [Electronic resource] / Wenbo Li, Xin Lu, Shengju Qian, [et al.] // International Joint Conference on Artificial Intelligence, Macao, 19–25 August 2024. – Access mode: https://www.ijcai.org/proceedings/2023/0121.pdf (date of access: 15.02.2024).

26. Accurate Image Restoration with Attention Retractable Transformer [Electronic resource] / Jiale Zhang, Yulun Zhang, JinjinGu [et al.] // The Eleventh International Conference on Learning Representations, Kigali, 30 April – 5 May 2023. – Access mode: https://openreview.net/pdf?id=IloMJ5rqfnt (date of access: 15.02.2024).

27. Image Super-Resolution Using Dilated Window Transformer [Electronic resource] / Soobin Park, Yong Suk Choi // IEEE Access. – 2023. – P. 1. – Access mode: https://doi.org/10.1109/access.2023.3284539 (date of access: 15.02.2024).

28. Cross Aggregation Transformer for Image Restoration [Electronic resource] / Zheng Chen, Yulun Zhang, JinjinGu [et al.] // Advances in Neural Information Processing Systems, New Orleans, 11–19 December 2022. – Access mode: https://openreview.net/forum?id=wQ2QNNP8GtM (date of access: 15.02.2024).

29. Image Super-Resolution with Unified-Window Attention [Electronic resource] / Gunhee Cho, Yong Suk Choi // IEEE Access. – 2024. – P. 1. – Access mode: https://doi.org/10.1109/access.2024.3368436 (date of access: 15.02.2024).

30. SRFormer: Permuted Self-Attention for Single Image Super-Resolution [Electronic resource] / Yupeng Zhou, Zhen Li, Chun-Le Guo [et al.] // 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023. – Access mode: https://doi.org/10.1109/iccv51070.2023.01174 (date of access: 15.02.2024).

31. SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution [Electronic resource] / Dafeng Zhang, Feiyu Huang, Shizhuo Liu [et al.]. : arxiv.org, 2023. – 14 p. – Access mode: https://arxiv.org/pdf/2208.11247.pdf (date of access: 15.02.2024).

32. Resolution-robust Large Mask Inpainting with Fourier Convolutions [Electronic resource] / Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin [et al.] // 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022. – Access mode: https://doi.org/10.1109/wacv51458.2022.00323 (date of access: 15.02.2024).

33.Activating More Pixels in Image Super-Resolution Transformer [Electronic resource] / Xiangyu Chen, Xintao Wang, Jiantao Zhou [et al.] // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

Vancouver, BC, Canada, 17–24 June 2023. – Access mode: https://doi.org/10.1109/cvpr52729.2023.02142 (date of access: 15.02.2024).

34. Dual Aggregation Transformer for Image Super-Resolution [Electronic resource] / Zheng Chen, Yulun Zhang, JinjinGu[et al.] // 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023. – Access mode: https://doi.org/10.1109/iccv51070.2023.01131 (date of access: 15.02.2024).

35. Recursive Generalization Transformer for Image Super-Resolution [Electronic resource] / Zheng Chen, Yulun Zhang, Jinjin Gu [et al.] // The Twelfth International Conference on Learning Representations, Vienna, 7–11 May 2024. – Access mode: https://openreview.net/forum?id=owziuM1nsR (date of access: 15.02.2024).

36. Single image super-resolution from transformed self-exemplars [Electronic resource] / Jia-Bin Huang, Abhishek Singh, Narendra Ahuja // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. – Access mode: https://doi.org/10.1109/cvpr.2015.7299156 (date of access: 15.02.2024).

37. Image Super-Resolution with Non-Local Sparse Attention [Electronic resource] / Yiqun Mei, Yuchen Fan, Yuqian Zhou // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. – Access mode: https://doi.org/10.1109/cvpr46437.2021.00352 (date of access: 15.02.2024).

38. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results [Electronic resource] / RaduTimofte, EirikurAgustsson, Luc Van Gool [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017. – Access mode: https://doi.org/10.1109/cvprw.2017.149 (date of access: 15.02.2024).

39. Relating transformers to models and neural representations of the hippocampal formation [Electronic resource] / James C. R. Whittington, Joseph Warren, Tim E. J Behrens // International Conference on Learning Representations, 25–29 April 2022. – Access mode : https://openreview.net/forum?id=B8DVo9B1YE0 (date of access: 15.02.2024).

40. Bardachenko V. F. Prospects of pulse neural networks application with timing presentation of information for dynamic images recognition / V. F. Bardachenko, O. K. Kolesnytskyi, S. A. Vasyletskyi // UCM. – 2003. – №6. – P. 73 – 82. (Ukr).

41. Kolesnytskyi O. K. Principles of construction of spiking neurocomputers architecture / O. K. Kolesnytskyi // Bulletin of Vinnytsia Polytechnic Institute. – Vinnytsia: UNIVERSUM-Vinnytsia. – 2014. – №4 (115). – P. 70 – 78. (Ukr).

42. Spikformer: When Spiking Neural Network Meets Transformer [Electronic resource] / Zhaokun Zhou, Yuesheng Zhu, Chao He [et al.] // The Eleventh International Conference on Learning Representations, Kigali, 1–5 May 2023. – Access mode: https://openreview.net/forum?id=frE4fUwz_h (date of access: 15.02.2024).

43. Optoelectronic implementation of pulsed neurons and neural networks using bispin-devices [Electronic resource] / O. K. Kolesnytskyj, I. V. Bokotsey, S. S. Yaremchuk // Optical Memory and Neural Networks. – 2010. – Vol. 19, № 2. – P. 154 – 165. – Access mode : https://doi.org/10.3103/s1060992x10020062 (date of access: 15.02.2024).

44. Optoelectronic spiking neural network [Electronic resource] / V. P. Kozemiako, O. K. Kolesnytskyj, T. S. Lischenko [et al.] // Optical Fibers and Their Applications 2012, Krasnobrod, Poland. – 2013. – Access mode: https://doi.org/10.1117/12.2019340 (date of access: 25.04.2024).

45. Neurocomputer architecture based on spiking neural network and its optoelectronic implementation [Electronic resource] / Oleh K. Kolesnytskyj, Vladislav V. Kutsman, Krzysztof Skorupski [et al.] // Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019, Wilga, Poland, 25 May – 2 June 2019 / ed. by R. S. Romaniuk, M. Linczuk. – [S. l.], 2019. – Access mode: https://doi.org/10.1117/12.2536607 (date of access: 25.04.2024).

Editorial office received the paper 22.02.2024.
The paper was reviewed 25.02.2024.

Kozlov Sergiy – Post Graduate with the Department of computer science, e-mail: serhii.kozlov@gmail.com.

Kolesnytskyi Oleg – Cand. Sc. (Eng.), Professor with the Department of computer science.

Vinnytsia NationalTechnical University.