Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 4, 2026, 609-620.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

HiSGAN: Image-to-Image Translation with Involution based Hybrid-Scale Transformer and Contrastive learning

Farzane Maghsoudi Ghombavani, Mohammad Javad Fadaeieslam and Farzin Yaghmaee*

Electrical and Computer Engineering Department, Semnan University, Semnan, Iran.

Article Info

Article History:
Received 24 October 2025
Revised 01 May 2026
Accepted 01 July 2026

DOI:10.22044/jadm.2026.16833.2815

Keywords:
Image-to-Image Translation, Fast Fourier Convolution, Transformers, Contrastive Learning, Feature Extraction.

**Corresponding author: f_yaghmaee@semnan.ac.ir (F. Yaghmaee).*

Abstract

Image-to-image translation is a highly challenging task, as it requires an accurate understanding of image details and their consistent transformation across domains. Notably, GANs have achieved remarkable success in this field. In essence, convolutional layers are the primary building blocks of these architectures. However, the limited receptive field in shallow layers makes it difficult to capture long-range spatial dependencies and non-local context. In this paper, the HiSGAN architecture is proposed to address this limitation. It combines deep representations with traditional techniques, such as SVD and Fast Fourier Convolution (FFC), to effectively extract style-related information and establish a global receptive field. Furthermore, we introduce the HiS-Transformer block with an involution operator in the bottleneck of the generator. This proposed block utilizes hybrid-scale self-attention to adaptively preserve the global receptive field and fine-grained information in salient regions while maintaining low computational cost. HiSGAN employs a new loss function based on gradient contrastive learning to improve cross-domain feature alignment. Quantitative and qualitative results on four public datasets demonstrate the superiority of the proposed approach over state-of-the-art methods. Importantly, these performance gains are achieved while reducing the parameter count and accelerating training. The code is available at https://github.com/OliverRensu/SG-Former.

1. Introduction

Many problems in computer vision can be naturally formulated as image-to-image translation, which maps a source image to a target image while preserving certain visual features of the original image. Developing a unified framework to solve such problems is its main purpose. Since a huge amount of information received by humans is conveyed through vision, this field of research is important [1]. Generative adversarial network (GAN) [2] is the most successful deep learning method used in image-to-image translation (I2I) [3-5, 6]. It consists of two modules, namely, generator and discriminator. It establishes a zero-sum game between them. Similar to other deep learning approaches, convolution and pooling layers are the basic building blocks of its architecture. In theory, different channels of convolution layers have a high ability to extract several different types of features. However, some inter-channel redundancy may occur inside convolution filters. Moreover, convolution may face difficulty in conveying style. A small receptive field poses challenges in detecting long-range spatial features and non-local contexts [7]. Some papers have tried to overcome this shortcoming by combining features based on linear algebra with those extracted from convolution [3]. Another approach applied for this purpose is the Transformer. However, it uses self-attention, whose computational complexity is quadratic in the input size. Instead of regular

Transformer, Liu et al. employed Swin Transformer with linear computational complexity with respect to input size [8, 9]. Inspired by the above approaches, we propose HiSGAN. The objective of our method is to capture global receptive fields while preserving long-range dependencies and latent information, achieving a favorable balance between training efficiency and predictive accuracy. To enable the model to learn more features, the encoder of its generator consists of two modules: Singular Value Decomposition (SVD) feature module and convolution feature module. The SVD feature map has been utilized to extract style features [3]. Fast Fourier Convolution (FFC) blocks have been embedded in the convolution module to achieve non-local receptive fields [9, 10]. Similar to [6, 11], in order to obtain a more compact network and reduce the computational load, the encoder has been utilized jointly in the generator and in the discriminator. Inspired by SG-Former [7, 12], we introduce a HiS-Transformer block for the bottleneck of the generator. The block plays an important role in preserving the global receptive field. It also adaptively preserves fine-grained information in salient regions while maintaining low computational cost. To establish a better cross-domain feature alignment, instead of cycle consistency loss, some papers have introduced a new loss based on contrastive learning [13, 14]. It retains contents in unsupervised image translation by maximizing the mutual information between the patches of input and output images. The most important contributions of this paper are as follows:

- An end-to-end GAN-based method using contrastive learning has been proposed for an unpaired image-to-image translation called HiSGAN.
- Traditional feature extraction methods, namely SVD and FFC, have been embedded in the encoder module in order to convey style features and better maintain the global receptive field.
- A HiS-Transformer block has been introduced for the bottleneck of the generator. It plays an important role, with low computational cost, in preserving the global receptive field and fine-grained information at salient regions adaptively.

The superiority of the proposed method over existing methods has been demonstrated through quantitative and qualitative experiments.

2. Related Works

2.1. Image-to-Image Translation

The term image-to-image translation (I2I) has gained popularity since the introduction of CGAN [15]. It transfers images from a source domain to a target one while keeping the content properties unchanged. The methods presented in this context deal with two types of datasets: paired [1] and unpaired [3, 4, 16]. From this viewpoint, the methods are divided into two categories: supervised and unsupervised. A pioneering method for paired datasets is Pix2pix [1], which uses CGAN. Preparation of paired datasets is very expensive and time-consuming, and in some cases impossible. Therefore, most of the methods apply unpaired datasets. One of them is CycleGAN [17], which has two generators and two discriminators trained with cycle-consistency loss. The loss is good at training generators and discriminators from unpaired datasets, but the accumulation of errors that may occur during image reconstruction can affect the quality of the generated images. To address this issue, Wang et al. [16] introduced long-short cycle-consistent loss. Due to the limitations of convolutional feature extraction, conventional unpaired image translation performs poorly in generating suitable images. UNTF [3] mixed deep learning with an SVD feature map, a traditional feature extraction algorithm based on linear algebra, to extract style features in a more appropriate way. It transformed 1-D SVD feature maps into 2-D features to learn the potential connections between the various vectors. UNTF can learn the image style well but there are some color deviations in the results. RepairerGAN [18] utilizes CycleGAN framework to segment cracks in images at the pixel-level. ConvNext and an attention mechanism, two blocks in its undamaged-image generator, separately generate an undamaged image and a crack mask, respectively. Using the mask, pixels of the crack area in the original image are filled with synthesized image. Similar to CycleGAN, another generator, namely crack image generator, is trained to produce a crack image from it and cycle-consistency loss is used for training. This method greatly reduces the time complexity of pixel-level labeling compared to fully supervised segmentation methods. NiceGAN [11] redesigned discriminator structure. In order to reduce the computational load, the encoder was jointly used in the generator and the discriminator. Therefore, the discriminator was not discarded after training. U-GAT-IT [19] was proposed by Kim et al., which achieved good results by introducing a new attention module and a new learnable normalization function called AdaLIN.

Deng et al. [20] in InvolutionGAN used the involution operator and ResBlock to create a lighter and more powerful bottleneck, and to extract local and long-range spatial features of the source images in a more appropriate way. SARCUT [14] has been presented to translate SAR images. A generator with a self-attention mechanism and spectral normalization operation was used in the model. Adversarial discriminant loss has been applied in SARCUT to accelerate model convergence and improve the accuracy of generated images. ARDA-UNIT [6], based on the new adaptive feature combination method and recursive dense self-attention residual block, is able to reduce computational load and time, handle structural changes well, and achieve the desired quality. Despite these advances, it seems that there is a need for a method that overcomes the shortcomings of others and produces better images with more details and higher quality with at a low computational cost.

2.2. Vision Transformer

Transformers [21] have made a major breakthrough in Natural Language Processing (NLP). Inspired by this, Dosovitskiy considered their application in computer vision and overcame the limitation of the local receptive field in convolution by introducing the Vision Transformer (ViT) [22]. Its self-attention module can detect long-range dependencies between image patches. Because self-attention is computed globally on the entire image, its computational cost grows quadratically with respect to the input size. To apply self-attention as a general backbone to high-resolution images, Liu et al. [23] proposed the shifted window Transformer, called Swin Transformer, which builds feature maps hierarchically and computes self-attention only within each local window. To this end, CSWin [10] has also designed cross-shaped attention. However, these methods result in coarse granularity. Ren et al. proposed a model, namely SG-Former [12], that adapts patch size with respect to the saliency map extracted through hybrid-scale self-attention. It tends toward global self-attention and reallocates more patches to salient regions for fine-grained information and fewer patches to background regions. This idea decreases the computational cost while at the same time retaining long-range dependencies between image patches. Sketch2Photo [9] is another image synthesis approach that generates photo-realistic images from weak or partial sketches. It captures long-range global contexts using Swin Transformer block (STB) units and uses a traditional feature extraction method—Fast Fourier Convolution (FFC) residual blocks—to create global receptive fields. Sketch2Photo is not a lightweight model and has a large number of parameters. Torbunov et al. developed UVCGAN [24] to handle long-range spatial dependencies. They incorporated ViT [22] into the CycleGAN generator. In an ablation study, they showed that in addition to architectural changes, gradient penalty and self-supervised pre-training are also needed. In [25], a Transformer-based SAR target image translation network, named SAR-SMT NET, has been proposed to bridge the gap between synthetic and measured SAR target images. In the CSTGAN architecture [8], the generator of CycleGAN has been redesigned with a Swin Transformer block (STB) and convolution layers. It improves the mapping ability from the IR domain to the RGB image domain by combining the advantages of Transformers and convolution.

2.3. Contrastive Learning

Preparing labeled data is a time-consuming task with high costs, while unlabeled data is abundant. Contrastive learning is a powerful unsupervised method that can leverage large amounts of data and learn representations in which similar samples are mapped closer together and dissimilar samples are mapped further apart. Person re-identification, face verification and identification, image retrieval, and zero-shot learning are vision tasks where contrastive learning has yielded good results. It has recently found many applications in image generation [4, 26]. Occasionally, there is a variety of sample images in some source domains. Cao et al. [4] used two conditional contrastive learning loss functions to decompose the domain into multiple classes. Therefore, they transformed image-to-image translation into class-to-class translation. They also applied a conditional contrastive learning stage in their GAN architecture. Mao et al. [26] proposed a visual style metric using the distance between the Gram matrices of deep features of source and reference images. In the second stage, through triplet loss, the real image and its corresponding translated image are forced to be close to each other, and the nearest sample from the source domain is pushed away from them. Zhao et al. [27] applied contrastive learning to deblurring, a pixel-level vision task. They extracted contrastive features in the gradient domain in the generator and introduced an adversarial contrastive loss in the discriminator module. Since their model receives unpaired sharp and blurred images as input, its loss function has been designed similarly to CycleGAN.

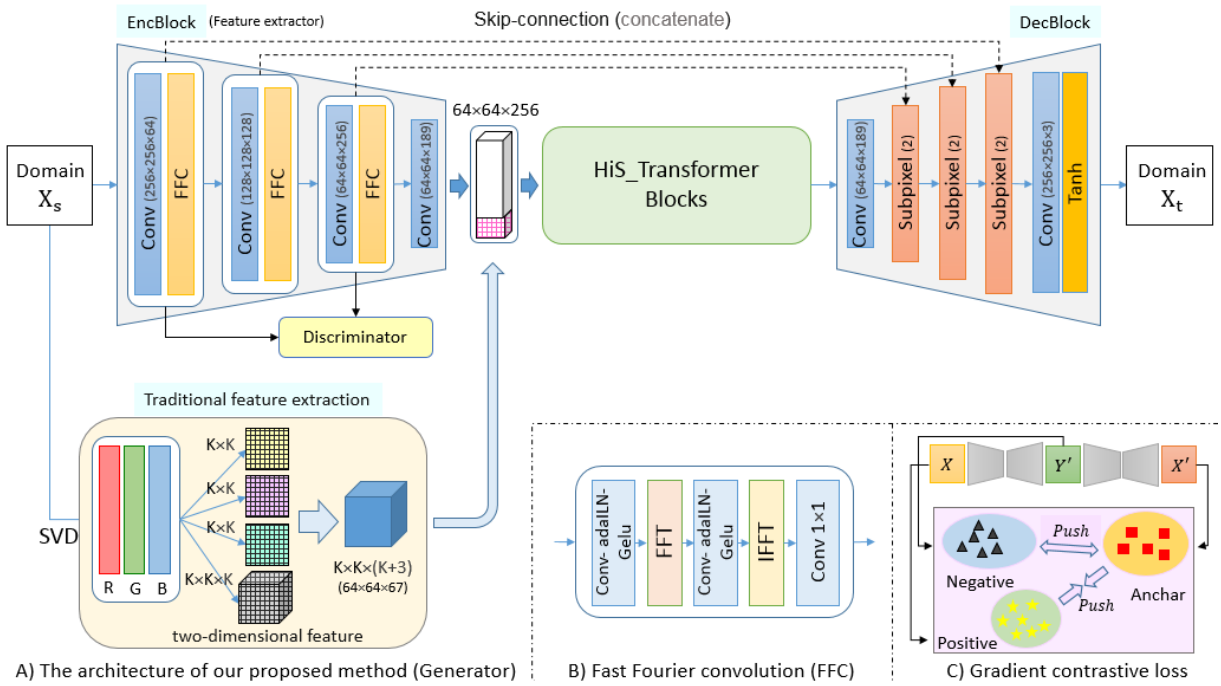


Figure 1. a) The generator structure of HiSGAN architecture. One stream has been shown from domain X to domain Y. b) Fast Fourier convolution c) Gradient contrastive loss.

3. Method

3.1. HiSGAN Structure

In this paper, a new approach, called HiSGAN, is proposed to enhance the global receptive field, improve the quality of generated images and alleviate computation cost. Figure 1(a) shows the architecture of our proposed generator and its main blocks. The goal is to train function $G_{s\to t}\left(G_{t\to s}\right)$ to map images from the source domain $X_s\left(X_t\right)$ to the target domain $X_t\left(X_s\right)$ using unpaired samples. Two generators (G) and two discriminators (D) are trained, along with a feature extractor to perform the I2I task. Applying FFC in each layer of the feature extractor enables the model to capture global features. The global receptive field is important understanding image structures and non-local features. However, convolutions in low-level layers may cause G to miss global information and ignore global features. Furthermore, traditional algorithms can quickly detect and preserve a large amount of precise information [3]. Therefore, both traditional and deep methods are employed to enhance the feature extraction process. As a way to achieve a more compact network, similar to NiceGAN [11], a joint encoder has been introduced to preserve the required information in the source domain for G, and to extract features for classifying input images in D.

We have introduced a new block based on Transformer, named HiS-Transformer, in the bottleneck of the proposed generator. The block utilizes a hybrid-scale strategy to reduce the computational load and increase the quality of the synthesized images. The hybrid-scale approach tends to adapt the size of the patches to the amount of information in them. In addition, gradient contrastive loss based on contrastive learning has been applied to improve the performance of the model in real-world scenarios. In the following, $G_{s\to t}$ and D_t are explained in more detail.

GELU activation function [28] and AdaLIN normalization function [19] have been employed in all layers of generators (G_s) and discriminators (D_s). GELU has shown a good performance in Natural Language Processing (NLP), especially in Transformer based models and helps alleviate the vanishing gradient problem [8, 29]. AdaLIN enables G to adjust an appropriate ratio between IN and LN during training time to control the amount of change in shape and texture with respect to the dataset.

3.2. Feature Extractor

Conventional image-to-image translation frameworks discard D after training phase. Here, inspired by NICE-GAN [11], a part of the discriminator is distinguished and named as a feature extractor. The part is also used to encode

the images of the target domain. To produce high-quality images, it is essential that G captures long-range dependencies or non-local features. In common generators, stacked convolutional layers are usually responsible for extracting features from the image. However, in the early layers of most convolutional networks, convolution kernels are usually small (e.g. 3×3). Therefore, it is difficult for them to capture long-range contextual information [10]. As a result, G lacks a complete global receptive field. The receptive field of FFC can cover the entire image, so inspired by [9], we embed the FFC blocks [9, 10] in the encoding process to obtain a global receptive field (Figure 1(a (EncBlock))). Figure 1(b) shows the FFC block in detail.

As mentioned in the related work section, Tu et al. have also considered the use of traditional feature extraction methods in their method [3]. They have combined the outputs of the convolution layers with SVD to better preserve boundaries and to make the style transition clearer. SVD largely preserves the full image information, but the obtained features show high integrity and low dimensionality. By adding them to the convolution features, it is possible to prevent the network from learning the relevant hidden information. So it is beneficial to convert these 1D features to 2D to help learn hidden features. Inspired by [3], as shown in Figure 1(a), three color channels of the input image are separated and each channel is decomposed by SVD to obtain singular value features. Then the top k eigenvalues are selected to convert the feature map into 2D features. Similar to [3], the three sets of $k\times k$ feature maps and one set of $k\times k\times k$ feature maps are obtained. Finally the size of the feature map is $k\times k\times(k+3)$. K is a hyper-parameter, if k is small, the network will lose many features. On the other hand, if it is large, the useless noise features will be added to the network operation. Following [3], k is set to 64.

3.3. Generator

DecBlock is a part of the generator that performs the upsampling operation (Figure 1(a)). Common deconvolution methods carry out transposed convolution. It utilizes padding which may cause feature loss. Similar to [3], we replace it with subpixel convolution. It increases the receptive field of the network to provide richer feature information for G and to improve the quality of synthesized images. Also, we have used skip connections between EncBlock and DecBlock in order to allow information to flow more easily. The details of G have been presented in Table 1.

3.4. HiS_Transformer

Convolution blocks in low-level layers often fail to capture global information [12]. Therefore, in addition to embedding traditional features, we rearrange the generator bottleneck to compensate for this limitation. The proposed bottleneck, named HiS-Transformer, is based on SG-former. Inspired by [20], a lightweight involution operator is embedded into the bottleneck [7] which enhances the extraction of local features and long-range dependencies (Figure 2(a)). Transformers introduce a self-attention mechanism. It identifies relationships between different patches of the input images. However, the size of patches in the initial Transformer architecture is fixed which increases computational complexity.

The SG-former Transformer [12] was proposed for global feature awareness with adaptive patch size. Based on a saliency map, it changes the size of the patches. It assigns more patches to salient regions to retain the details and assigns fewer patches to unimportant regions in order to reduce time complexity. Similar to SG-former, HiS-Transformer consists of two blocks: Hybrid-Scale Transformer Block (HSTB) and Self-Guided Transformer Block (SGTB) (The details of SGTB and HSTB can be found in [12]). We embed an involution block between these blocks. The internal structure of these two Transformers is the same a standard Transformer with the difference that SGTB uses the saliency map generated by HSTB. In addition, hybrid-scale attention is applied instead (Figure 2(b)). Table 1 shows the implementation details of HiS-Transformer with three blocks.

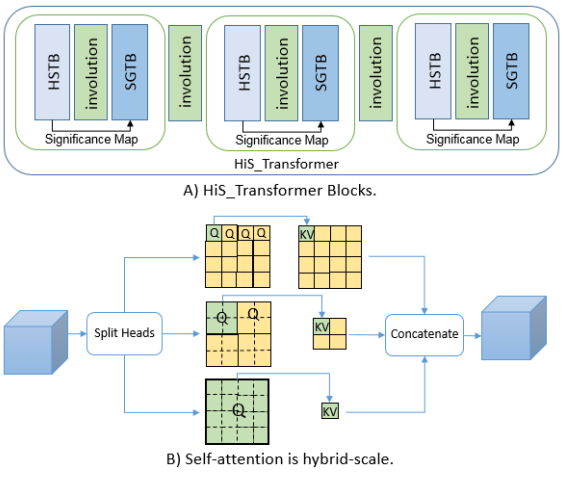


Figure 1. Details of the HiS-Transformer blocks. Query (Q), Key (K) and Value (V).

Table 1. Summary of the Encoder, bottleneck and the Decoder blocks of the generator and discriminator.

Part		Input → Output size	Layer Information
Encoder	Layer1	(3,256,256) →(64,256,256)	Conv2d(3, 64, k=3, s=1, p=1), AdaLIN, GELU, FFC
	Layer2	(64,256,256) →(128,128,128)	Conv2d(64, 128, k=3, s=1, p=1), AdaLIN, GELU, FFC
	Layer3	(128,128,128) →(256,64,64)	Conv2d(128, 256, k=3, s=1, p=1), AdaLIN, GELU, FFC
	Output	(256,64,64) →(256,64,64)	Conv2d(256, 192, k=3, s=1, p=1), concatenate by SVD
Bottleneck	HiS-Trans. Blocks	(256,64,64) →(256,64,64)	HSTB, involution, SGTB (s=8, r=(16, 8, 8, 4)), involution
		(256,64,64) →(256,64,64)	HSTB, involution, SGTB (s=2, r=(8, 4, 4, 2)), involution
		(256,64,64) →(256,64,64)	HSTB, involution, SGTB (s=1, r=(2,1))
Decoder	Up3	(256,64,64) →(256,64,64)	Conv2d(256, 256, k = 3, s = 1, p=1), AdaLIN, GELU,
		(256,64,64) →(512,64,64)	Cat(layer3,up3), Conv2d(512, 512, k = 3, s = 1)
	Up2	(512,64,64) →(128,128,128)	Subpixel(2),), AdaLIN, GELU
		(128,128,128) →(256,128,128)	Cat(layer2,up2), Conv2d(256, 256, k = 3, s = 1)
	Up1	(256,128,128) →(64,256,256)	Subpixel(2),), AdaLIN, GELU
		(64,256,256) →(64,256,256)	Conv2d(64, 64, k = 3, s = 1)
		(64,256,256) →(3,256,256)	Conv2d(64, 3, k = 7, s = 1, p=1), Tanh

Figure 3 shows the discriminator architecture of the proposed method. Similar to NiceGAN, the discriminator utilizes part of the encoder block as a feature extractor. In convolution layers, lower layers capture local features and higher layers extract global features. So, we use the outputs of layers 1 and 3 as inputs to the discriminator to encourage G, via D, to consider both local and global features of the input image. Then, four convolutional layers with different kernel sizes are applied to each feature to further refine these local and global representations and obtain better results. Similar to [18], the learning rate of D is set to one-tenth that of G to prevent D from dominating the training process. The details of the discriminator have been presented in Table 2.

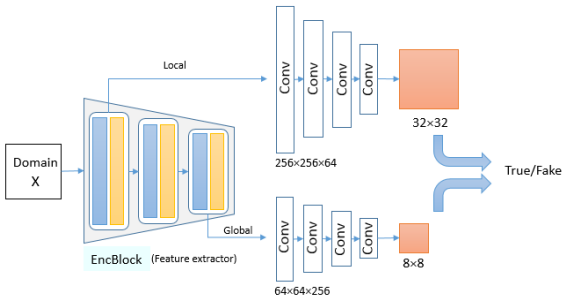


Figure 3. The structure of discriminator in detail.

Table 2. Details of the discriminator module.

Part		Input → Output size	Layer Information
Discriminator	Global	(64,256,256) →(64,128,128)	Conv2d(64, 64, k = 4, s = 2, p=1), SN, GELU
		(64,128,128) →(128,64,64)	Conv2d(64, 128, k = 4, s = 2, p=1), SN, GELU
		(128,64,64) →(256,32,32)	Conv2d(128, 256, k = 4, s = 2, p=1), SN, GELU
		(256,32,32) →(1,32,32)	Conv2d(256, 1, k = 4, s = 1, p = 1), SN
	Local	(256,64,64) →(256,64,64)	Conv2d(128, 128, k = 4, s = 2, p=1), SN, GELU
		(256,64,64) →(512,32,32)	Conv2d(128, 256, k = 4, s = 2, p=1), SN, GELU
		(512,32,32) →(1024,16,16)	Conv2d(256, 512, k = 4, s = 2, p=1), SN, GELU
		(1024,16,16) →(1,16,16)	Conv2d(512, 1, k = 4, s = 1, p = 1), SN

3.6. Loss Function

3.6.1. Gradient Contrastive Loss

The cycle consistency loss works well during the training phase with unpaired datasets, but the accumulation of reconstruction errors can affect the quality of generated images [16]. To improve image quality, we apply gradient contrastive loss (Fig. 1(c)). In contrastive learning, given an anchor sample, the network aims to learn a representation to bring the anchor closer to positive samples in the metric space and move the anchor away from negative samples. The input image is considered as the positive sample (φ^+), the translated image in the target domain as the negative sample (φ^-), and the recovered image reconstructed from the translated image as the anchor point (φ). Contrastive learning consists of two modules: a contrastive learning network and a contrastive loss function. The contrastive learning network includes an encoder. In this work, we use the output of layers 1, 2, and 3 of the encoder in Figure 1(a). Then $\varphi^+ = Enc(X)$, $\varphi^- = Enc(Y')$ and $\varphi = Enc(X')$. Similar to [27], the gradient contrast loss function can be calculated through (1). In this equation, φ represents the feature representation. φ_l (l=1, 2, and 3) is the l-th extracted hidden feature. τ is a temperature parameter to scale the similarity and $S(u,v)$ represents the similarity function between two normalized feature vectors in the gradient domain. It is important to obtain representative features of contrasting samples. Different methods use different separating networks for this purpose [4,

13, 27]. In contrast to these methods, we do not only use the extracted features in the encoder. By transforming the features into the gradient domain, we can capture more image details and structures.

$$l_{gc}\left(\nabla\varphi,\nabla\varphi^{+},\nabla\varphi^{-}\right)= \quad (1)$$
$$-\sum_{l=1}^{L}\log(\exp\left[S\left(\nabla\varphi,\nabla\varphi^{+}\right)/\tau\right]/$$
$$(\exp\left[S\left(\nabla\varphi,\nabla\varphi^{+}\right)/\tau\right]+$$
$$\exp\left[S\left(\nabla\varphi,\nabla\varphi^{-}\right)/\tau\right]))$$

3.6.2. Other Losses

$L_{lsgan}^{s\to t}$ [30]: Adversarial loss originates from the cross entropy between the distribution of real and generated images. Therefore, it penalizes larger errors and in turn leads to large correction rather than non-convergence. In (2), D tries to maximize the whole equation and G tries to minimize $\left(1-D_t\left(G_{s\to t}\left(X\right)\right)\right)^2$.

$$L_{lsgan}^{s\to t}\left(G_{s\to t},D_t\right)=\mathrm{E}_{y\sim X_t}\left[\left(D_t\left(y\right)\right)^2\right]$$
$$+\mathrm{E}_{X\sim Xs}\left[\left(1-D_t\left(G_{s\to t}\left(X\right)\right)\right)^2\right] \quad (2)$$

$L_{identity}^{s\to t}$ [19]: Identity loss is provided to ensure that the color distribution of input images and output images are similar. This function leads to maintaining more features, improving the quality of produced images, stabilizing the training process, and preventing mode-collapse ((3)).

$$L_{identity}^{s\to t}\left(G_{s\to t}\right)=\mathrm{E}_{x\sim X_t}\left[\left|x-G_{s\to t}\left(X\right)\right|_1\right] \quad (3)$$

$L_{GP_A}^{s\to t}$ [24]: Gradient Penalty loss is designed based on WGAN [31] and introduced in [32]. It has little sensitivity to the values of meta-parameters and complements $L_{lsgan}^{s\to t}$ in the discriminator ((4)). The γ-centered GP regularization provides more stable training and is less sensitive to the meta-parameter selection.

$$L_{GP_A}^{s\to t}\left(D_t\right)=\mathrm{E}_{y\sim X_t}\left[\frac{\left(\left\|\nabla_y D_t\left(y\right)\right\|_2-\gamma\right)^2}{\gamma^2}\right] \quad (4)$$

3.6.3. Overall Loss

The final objective functions for G and D are presented in (5). The hyper-parameters λ_1, λ_2, and λ_3 control the contributions of the individual loss terms. Each loss function, except for L_{gc}, is defined as the sum of the forward and backward directions, i.e., $L^{s\to t}+L^{t\to s}$ [33].

$$L_G=\min_{G_{x\to y},G_{y\to x}} L_{lsgan}+\lambda_1 L_{identity}+\lambda_2 L_{gc}$$
$$L_D=\min_{D_{x\to y},D_{y\to x}} L_{lsgan}+\lambda_3 L_{GP} \quad (5)$$

4. Experimental Evaluation

4.1. Implementation Details

To evaluate the effectiveness of HiSGAN, the following datasets were used: horse↔zebra, summer↔winter_yosemite, vangogh↔photo, and selfie↔anime. These datasets are publicly available and widely used in image-to-image translation tasks [11, 17, 34].

In the proposed model, similar to NiceGAN [11] and ARDA-UNIT [6], some parts of the encoder are shared with the discriminator. Therefore, the discriminator and the encoder are first trained together, separately from the generator. In the next step, only the generator parameters are fine-tuned. All models are trained using the Adam optimizer with β1 = 0.5 and β2 = 0.999. During preprocessing, Images are horizontally flipped with a probability of 0.5, resized to 286 × 286, and randomly cropped to 256 × 256. The batch size is set to one in all experiments. The method is implemented in PyTorch and trained in the Google Colab environment. Similar to [6], 10-fold cross-validation is used to tune the loss-function parameters ($\gamma=100$, $\lambda_1=2$, $\lambda_2=5$, and $\lambda_3=0.1$).

4.2. Results and Analysis

Some experiments were arranged to evaluate the proposed model quantitatively and qualitatively. Table 3 shows the number of training parameters, FLOPs, and inference time of different methods. The quantitative evaluation results are reported in Table 4. Three evaluation metrics, namely FID, KID, and IS were applied for this end. The specifications of these three criteria have been given in [6]. As can be seen, HiSGAN scores are almost always better than other models. There are some small limitations in vangogh↔photo and summer↔winter_yosemite when compared to the performance of Sketch2Photo. Considering that Sketch2Photo is a state-of-the-art method in this field, our method has achieved much stronger results with fewer parameters and faster speed.

The first and second rows in Figure 4 show the results of different methods on horse↔zebra dataset. In this collection, the overall style of the image has not changed much. Except for Nice-GAN, zebra stripes are visible in the rest of the methods. In Nice-GAN, the result is associated with distortion. The CycleGAN method produced blurred images in column (b). The comparison of

the results shows that HiSGAN provided sharper details, produced less changes in the background, and made the overall style of the image natural. The third and fourth rows show the results of working on summer↔winter_yosemite, which focuses on the change of color. CycleGAN fails to capture significant semantic transformations. Although NiceGAN and InvolutionGAN are able to show the seasonal transformation well, the mountain and some details remained untouched. Sketch2Photo and the proposed method not only have preserved the background color well, but also applied the necessary changes, resulting in natural images. When the input image is related to summer, the results of other methods are almost similar to each other, except CycleGAN. The vangogh↔photo dataset is also examined in rows 5 and 6. In this set, image generation requires significant changes in the color, texture, and styles of input images. All the mentioned methods have the ability to make the necessary changes, but there are many differences in the resolution of the generated images. CycleGAN and Nice-GAN struggle to capture small features. Although InvolutionGAN and Sketch2Photo achieved better results, HiSGAN can perform the structural transformation in a better and clearer way due to more accurate feature extraction. Rows 7 and 8 deal with the selfie↔anime dataset, which includes female faces. Since humans are very sensitive to facial details, even small distortions are not ignored. Therefore, translating styles and forms presents significant challenges in this dataset. InvolutionGAN, NiceGAN and CycleGAN methods could not apply geometric changes well and could not lead to satisfactory results. The translation results from Sketch2Photo were favorable. But our proposed method can translate the geometric features well and provides very promising results, much closer to real images.

Table 3. Number of model parameters, FLOPs, and inference time.

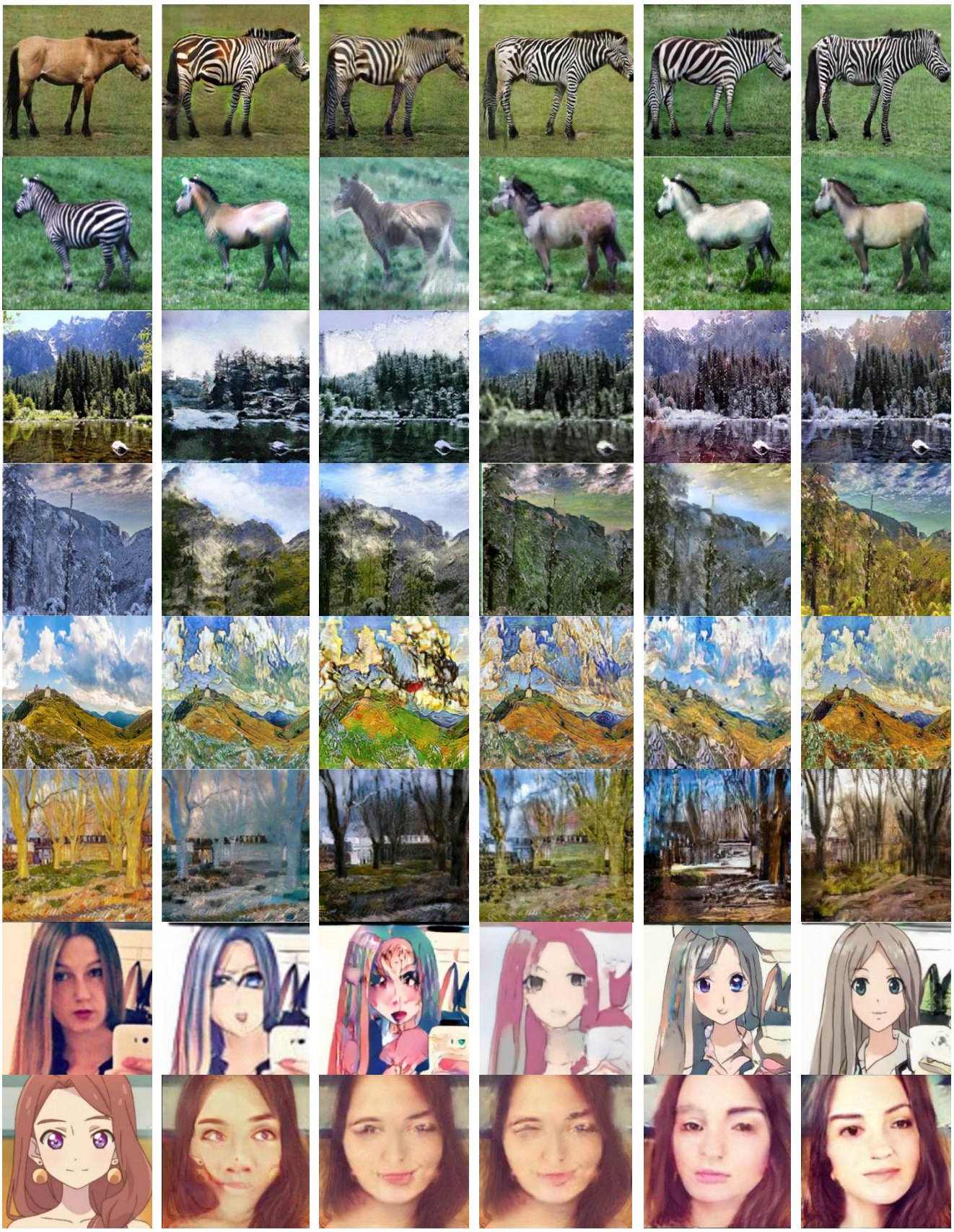
Model	GPara	DPara	AllPara	FLOPs	Inference time
CycleGAN [17]	22.7M	2.5M	28.2M	111G	5.07ms
NICE-GAN [11]	16.2M	93.7M	109.9M	79.6G	22.73ms
Sketch2Photo [9]			158.4M		28.61
InvolutionGAN [20]	-	-	21.6M	-	3.82ms
HiSGAN (ours)	27M	26.27M	53.27M	117.783G	5.72ms

4.4. Ablation Study

Compared to the methods in the literature, the proposed method reduces training time and the number of parameters, and the experimental results demonstrate its ability to translate structural information and to synthesize realistic images. The model utilizes four main components: FFC, SVD, HiS-Transformer, and Involution. To analyze the effectiveness of these blocks, several experiments were conducted. As the FID, KID, and IS metrics, as well as the qualitative results, indicate, the selfie↔anime dataset is more challenging than the others. Therefore, we focus on this dataset to evaluate these components separately. Table 5 reports these results quantitatively. The baseline represents a model in which none of these components is included, similar to U-Net. FFC, SVD, HiS-Transformer with Convolution, and HiS-Transformer with Involution are added one by one from M1 to M4. As can be seen in Figure 5, all of these blocks improve the performance of the model. However, the effect of HiS-Transformer is more significant than that of the other components. Although this study employs three primary feature extraction mechanisms, each approach captures distinct and complementary characteristics of the input image. Accordingly, and as can be seen in Table 5, integrating all three methods can be advantageous, as it facilitates a more comprehensive representation of image features and enhances the overall model performance.

Figure 4 shows the effects of the components on the quality of the generated images. As shown in column (c), the global features have been well transferred. In column (d), the style of the image has been effectively conveyed. In columns (e) and (f), the generated images exhibit very high quality, which shows the power of the HiS-Transformer block in maintaining the global receptive field and long-range dependencies.

We conducted another set of experiments to examine the effects of the loss functions. From column M5 to M8 in Table 6, adversarial loss, identity loss, gradient penalty loss, and gradient contrastive loss were added one by one and, as a result, the quality of the synthesized images gradually improves. Figure 6 confirms these results qualitatively.



a) Input b) CycleGAN c) Nice-GAN d) InvolutionGAN e) Sketch2Photo f) HisGAN

Figure 4. Examples of generated outputs of different methods. From top row to bottom: horse↔zebra, summer↔winter_yosemite, vangogh↔photo, selfie↔anime.

a) Input | b) Baseline | c) M1 | d) M2 | e) M3 | f) M4

Figure 5. Generated images from Ablation study on different blocks.

Table 4. The FID, KID and IS for different models. A↔B reports the mean between A→B and A←B.

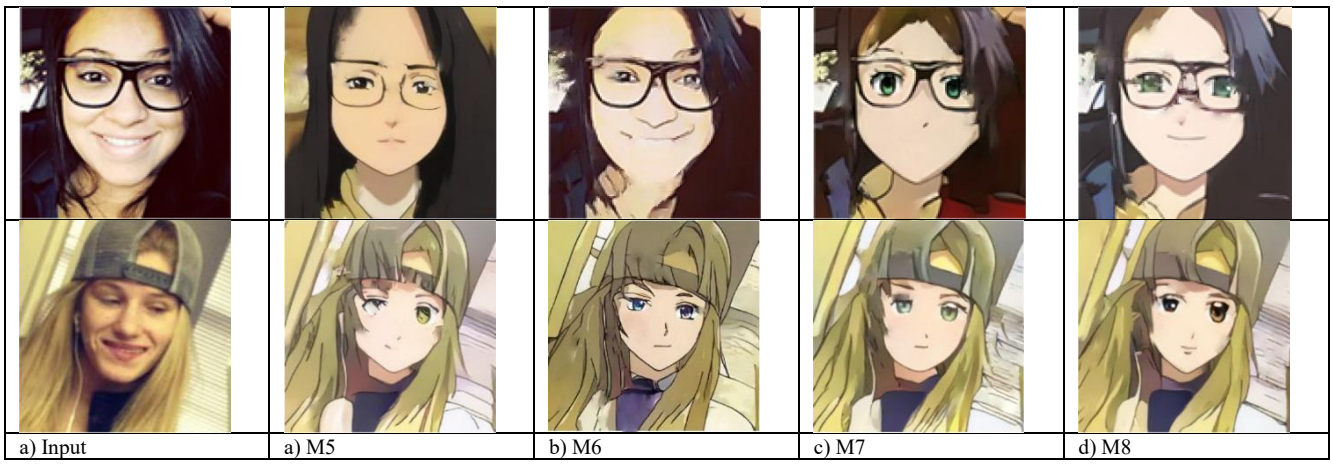
	horse↔zebra			summer↔winter_yosemite		
Model	**FID↓**	**KID*100↓**	**IS↑**	**FID↓**	**KID*100↓**	**IS↑**
CycleGAN [17]	119.57	6.79	1.148	82.72	2.01	1.045
NICE-GAN [11]	89.01	5.42	1.239	79.51	1.04	1.235
Sketch2Photo [9]	70.12	3.67	1.548	67.32	0.87	1.654
InvolutionGAN [20]	88.58	4.98	-	-	-	-
HiSGAN (ours)	69.36	3.38	1.659	61.93	0.71	1.594
	vangogh↔photo			**selfie↔anime**		
Model	**FID↓**	**KID*100↓**	**IS↑**	**FID↓**	**KID*100↓**	**IS↑**
CycleGAN [17]	144.43	7.37	1.021	203.71	12.38	1.019
NICE-GAN [11]	131.91	5.69	1.212	156.48	7.41	1.264
Sketch2Photo [9]	112.08	4.49	1.328	131.49	6.57	1.381
InvolutionGAN [20]	127.54	5.42	-	194.04	7.39	-
HiSGAN (ours)	110.72	5.31	1.349	114.6	5.32	1.628

Table 5. The quantitative result of ablation study on blocks.

Method	**Baseline**	**M1**	**M2**	**M3**	**M4**
FFC	×	✓	✓	✓	✓
SVD	×	×	✓	✓	✓
HiS-Transformer-Conv	×	×	×	✓	✓
HiS-Transformer-Inv	×	×	×	×	✓
FID↓	153.8	141.86	139.25	117.35	114.6
IS↑	1.228	1.352	1.372	1.573	1.628

Table 6. The quantitative result of ablation study on the effects of different losses.

Method	**M5**	**M6**	**M7**	**M8**
Adversarial loss	✓	✓	✓	✓
Identity loss	×	✓	✓	✓
Gradient Penalty loss	×	×	✓	✓
Gradient Contrastive loss	×	×	×	✓
FID↓	253.15	209.37	183.24	114.6
IS↑	0.625	0.781	0.896	1.628



a) Input | a) M5 | b) M6 | c) M7 | d) M8

Figure 6. Generated images from Ablation study on different losses.

5. Discussion and Conclusion

In this paper, a new image-to-image translation method, named HiSGAN, is introduced. The model utilizes an innovative combination of traditional techniques, such as SVD and Fast Fourier Convolution (FFC), to extract style-related features and establish a global receptive field from lower-level layers more effectively. We propose a HiS-Transformer block with an involution operator for the bottleneck of the generator. This block plays an important role in preserving the global receptive field and fine-grained information in salient regions while maintaining low computational cost. It employs a new contrastive-learning-based loss function to improve cross-domain feature alignment. The qualitative and quantitative results demonstrate that the proposed method performs significantly better than existing state-of-the-art methods. However, the method is less effective at preserving the details of small objects and is slightly sensitive to domain gaps. In future work, we will explore how to handle drastic structural transformation alongside minor ones.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks, " in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* Piscataway, 2017, pp. 1125-1134.

[2] I. J. Goodfellow, et al., "Generative adversarial nets, " in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[3] H. Tu, W. Wang, J. Chen, F. Wu, and G. Li, "Unpaired image-to-image translation with improved two-dimensional feature, " *Multimedia Tools and Applications*, vol. 81, no. 30, pp. 43851–43872, Dec. 2022.

[4] Z. Cao, W. Wang, L. Huo, and S. Niu, "Unsupervised class-to-class translation for domain variations, " *Pattern Recognition*, vol. 138, Art. no. 109346, 2023.

[5] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping, " in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, 2006, pp. 1735–1742.

[6] F. M. Ghombavani, M. J. Fadaeieslam, and F. Yaghmaee, "ARDA-UNIT: Recurrent dense self-attention block with adaptive feature fusion for unpaired (unsupervised) image-to-image translation, " *IET Image Processing*, vol. 17, no. 13, pp. 3746–3758, 2023.

[7] D. Li, J. Hu, C. Wang, X. Li, Q. She, L. Zhu, T. Zhang, and Q. Chen, "Involution: Inverting the inherence of convolution for visual recognition, " in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 12316–12325.

[8] M. Zhao, G. Feng, J. Tan, N. Zhang, and X. Lu, "CSTGAN: Cycle Swin Transformer GAN for unpaired infrared image colorization, " in *Proc. 3rd Int. Conf. Control, Robot. Intell. Syst. (CCRIS)*, virtual event, China, Aug. 26–28, 2022, pp. 1–7.

[9] H. Liu, Y. Xu, and F. Chen, "Sketch2Photo: Synthesizing photo-realistic images from sketches via global contexts, " *Engineering Applications of Artificial Intelligence*, vol. 117, Art. no. 105608, 2023.

[10] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation, " in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8168–8177.

[12] S. Ren, X. Yang, S. Liu, and X. Wang, "SG-Former: Self-guided transformer with evolving token reallocation, " in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 5980–5991.

[13] Y.-H. Hung, J. Tan, T.-M. Huang, S.-C. Hsu, Y.-L. Chen, and K.-L. Hua, "Unpaired image-to-image translation using negative learning for noisy patches, " *IEEE MultiMedia*, vol. 29, pp. 59–68, 2022.

[14] Y. Zhang, M. Li, W. Cai, et al., "SARCUT: Contrastive learning for optical-SAR image translation with self-attention and relativistic discrimination, " in *Proc. SPIE Int. Workshop Frontiers Graphics Image Process. (FGIP 2022)*, vol. 12644, Art. no. 126440B, May 3, 2023.

[15] M. Mirza and S. Osindero, "Conditional generative adversarial nets, " *arXiv preprint arXiv:1411.1784*, 2014.

[16] G. Wang, H. Shi, Y. Chen, and B. Wu, "Unsupervised image-to-image translation via long-short cycle-consistent adversarial networks, " *Applied Intelligence*, vol. 53, no. 14, pp. 17243–17259, Jul. 2023.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks, " in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2242–2251.

[18] Y. Liu, J. Chen, and J.-a. Hou, "Learning position information from attention: End-to-end weakly supervised crack segmentation with GANs, " *Computers in Industry*, vol. 149, Art. no. 103921, Aug. 2023.

[19] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-

image translation, " in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.

[20] H. Deng, Q. Wu, H. Huang, X. Yang, and Z. Wang, "InvolutionGAN: Lightweight GAN with involution for unsupervised image-to-image translation, " *Neural Computing and Applications*, vol. 35, no. 22, pp. 16593–16605, Aug. 2023.

[21] A. Vaswani, et al., "Attention is all you need, " in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[22] A. Dosovitskiy, et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale, " *arXiv preprint arXiv:2010.11929*, 2020.

[23] Z. Liu *et al*., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 9992-10002.

[24] D. Torbunov, et al., "UVCGAN: UNet vision transformer cycle-consistent GAN for unpaired image-to-image translation, " in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2023, pp. 702–712.

[25] G. Youk and M. Kim, "Transformer-Based Synthetic-to-Measured SAR Image Translation via Learning of Representational Features", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 01 2023.

[26] Q. Mao and S. Ma, "Enhancing Style-Guided Image-to-Image Translation via Self-Supervised Metric Learning", *IEEE Transactions on Multimedia*, vol. 25, pp. 8511–8526, Jan. 2023.

[27] B. Zhao, W. Li, and W. Gong, "Real-aware motion deblurring using multi-attention CycleGAN with contrastive guidance", *Digital Signal Processing*, vol. 135, p. 103953, 2023.

[28] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs) ", *arXiv [cs.LG]*. 2023.

[29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s, " in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 11976–11986.

[30] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks, " in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2794–2802.

[31] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks, " in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, 2017, pp. 214–223.

[32] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation, " in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, 2018.

[33] M. H. Khosravi, "A Siamese Network Based on InceptionV3 with Custom Loss Functions for Document Image Quality Assessment (DIQA) , " *Journal of AI and Data Mining*, vol. 14, no. 3, pp. 291-299, July 2026.

[34] Y. Li, H. Meng, H. Lin, and C. Liu, "AFF-UNIT: Adaptive feature fusion for unsupervised image-to-image translation, " *IET Image Process.*, vol. 15, no. 13, pp. 3172–3188, 2021.

HiSGAN: انتقال محتوای تصویر به تصویر با ترانسفورماتور مقیاس ترکیبی خود هدایت شونده و یادگیری متضاد

فرزانه مقصودی قمبوانی، محمدجواد فدائی اسلام و فرزین یغمایی*

دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

ارسال ۲۰۲۵/۱۰/۲۴؛ بازنگری ۲۰۲۶/۰۵/۰۱؛ پذیرش ۲۰۲۶/۰۷/۰۱

چکیده:

انتقال محتوای تصویر به تصویر یک کار بسیار چالش برانگیز است، زیرا نیاز به درک دقیق جزئیات تصویر و تبدیل مداوم آنها در دامنه‌های مختلف دارد. نکته قابل توجه این است که شبکه‌های مولدتخاصمی (GAN) در این زمینه به موفقیت چشمگیری دست یافته‌اند. در اصل، لایه‌های کانولوشنی بلوک‌های سازنده اصلی این معماری‌ها هستند. با این حال، میدان دریافت محدود در لایه‌های کم‌عمق، ثبت وابستگی‌های مکانی دوربرد و بافت غیرمحلی را دشوار می‌کند. در این مقاله، معماری HiSGAN برای رفع این محدودیت پیشنهاد شده است. این معماری، نمایش‌های عمیق را با تکنیک‌های سنتی، مانند SVD و کانولوشن سریع فوریه (FFC)، ترکیب می‌کند تا اطلاعات مربوط به سبک را به طور موثر استخراج کرده و یک میدان دریافت سراسری ایجاد کند. علاوه بر این، ما بلوک HiS-Transformer را با یک عملگر Involution در گلوگاه مولد معرفی می‌کنیم. این بلوک پیشنهادی از خودتوجهی در مقیاس ترکیبی برای حفظ تطبیقی میدان دریافت سراسری و اطلاعات ریزدانه در مناطق برجسته و در عین حال حفظ هزینه محاسباتی پایین استفاده می‌کند. HiSGAN از یک تابع متضاد جدید مبتنی بر یادگیری یادگیری متضاد برای بهبود هم‌ترازی ویژگی‌های بین دامنه‌ای استفاده می‌کند. نتایج کمی و کیفی روی چهار مجموعه داده عمومی، برتری رویکرد پیشنهادی را نسبت به روش‌های پیشرفته نشان می‌دهد. نکته مهم این است که این افزایش عملکرد در عین کاهش تعداد پارامترها و تسریع آموزش حاصل می‌شود. کد در https://github.com/OliverRensu/SG-Former موجود است.

کلمات کلیدی: انتقال محتوای تصویر به تصویر، کانولوشن فوریه سریع، ترانسفورمر، یادگیری متضاد و استخراج ویژگی.