

SSTar-TransGAN: Transformer-based Unsupervised Multi-modality Hyperspectral Digital Staining Network

Yukun Wang^{1, a} and Yanfeng Gu^{1, b *}

¹School of Electronic and Information Engineering, Harbin Institute of Technology, Harbin, China.

^aykwang_1207@hit.edu.cn, ^bguyf@hit.edu.cn

Abstract. Immunohistochemistry (IHC) plays an important role in accurate cancer screening and diagnosis, the clinical application of which has been restricted by complex operation process, high costs and demand of professional skills from pathologists. Digital staining methods based on deep learning provide the possibility for Hematoxylin & Eosin (H&E) stained images to be converted into IHC stained images, but it needs to train multiple staining networks for various modalities of IHC images from different cancers, thus weakening the versatility and convenience of digital staining in the process of practical application. At the same time, microscopic hyperspectral imaging technology can provide abundant spectral information for pathological images, which has been proved effective in digital staining tasks to transform staining modalities from one to many, but microscopic hyperspectral imaging has also been troubled by time-consuming acquisition process and huge data storage. In order to overcome the above challenges, we propose SSTar-TransGAN network for digital staining tasks. With the inhabitation of StarGAN structure, SSTar-TransGAN transfers the training burden of generators into lightweight style encoders between different modalities, while in addition to the introduction of transformer structure into the encoder, the application of spectral super-resolution Swin-Spectral Transformer U-Net (SSTU) network enables the conversion from H&E stained RGB images into hyperspectral images as well, which ensure the spatial structure and color information of IHC images under multiple staining modalities and the further conversion between different staining modalities. Qualitative and quantitative experiments prove the performance of SSTar-TransGAN on digital staining tasks superior to other state-of-the-art digital staining methods.

Keywords: Digital Staining; immunohistochemistry (IHC); microscopic hyperspectral imaging; pathology; transformer.

1. Introduction

As the gold standard for cancer diagnosis, pathological examination provides an accurate basis for cancer diagnosis through microscopic observation of pathological tissue, where immunohistochemistry (IHC) technique has proven significant advantages over traditional Hematoxylin & Eosin (H&E) staining methods. IHC utilizes specific antibodies to identify and locate antigens, thus revealing the molecular expression patterns within pathological tissues, which is crucial for cancer classification, staging and prognostic prediction. However, there are still some shortcomings in clinical application of IHC, such as complex operation, high time and economic costs and high requirements for the professional skills of operators, which limits the widespread promotion and application of IHC.

With the continuous progress of machine learning, digital staining has become a feasible alternative to IHC generation [1]. Digital staining technology can not only generate a variety of histological staining images from unstained microscopic slides, which is usually regarded as virtual staining, but can also convert stained microscopic images into other staining modalities, which can be regarded as staining transformation [2]. Compared with traditional research methods based on linear color mapping [3], the introduction of deep learning technology has simplified the training process of digital staining and improved the accuracy of image generation, resulting in a more rational performance in pathological research [4][5].

However, existing digital staining methods based on deep learning are confronted with some challenges. Firstly, due to the difficulty of performing multiple staining reagents on the same tissue

slice, except for a few studies on virtual staining, there are almost no cross-modality staining image pairs with precise registration at the same field of view for the preparation of the training set. In this case, unsupervised staining transformation without the necessity of registered training dataset has become particularly important, while most current unsupervised staining transformation schemes are essentially various optimizations based on CycleGAN [6]. Secondly, deep learning-based digital staining schemes has proven applicable to staining transformations between different types of microscopic cell images, which requires separate staining network trained for different types of cancers. However, as the increase of staining modalities, the number of trained networks for staining transformation also increases, which not only increases the complexity and computing cost of training process, but also limits the practical application of digital staining methods [7]. Thirdly, studies have proved the superiority of hyperspectral imaging (HSI) over RGB images in digital staining. However, existing digital staining methods based on HSI images generally stack spectral features with original RGB images, failing to fully utilize the spectral information provided by HSI [8]. Moreover, using both RGB and HSI data at the same time will lead to an excessive amount of input data for the network, thus increasing the difficulty of convergence and computational burden of model training.

Based on the above situation, this paper proposes a SSTar-TransGAN digital staining network based on the StarGAN v2 [9] network and vision transformer [10] encoder combined with Swin-Spectral Transformer UNet (SSTU) [11] spectral super-resolution network. With the integration of spectral super-resolution technology, high spectral resolution microscopic images stained by H&E can be obtained as the input of digital staining network without additional training costs. StarGAN structure shares a common generator in the generator, enabling the digital staining process of SSTar-TransGAN to be applied to different numbers of staining modalities without additional training costs, while the transformer encoder structure guarantees better preservation of color and spatial structure, resulting in better results in staining.

2. Methods

2.1 Network Architectures

Inspired by [12], our method is based on StarGAN v2. The difference is that, with the help of SSTU spectral reconstruction network, we convert the input RGB image stained with H&E into a hyperspectral image, which provides rich spectral information for the following staining modality transformation. At the same time, we replace the CNN encoder in the generator network with a vision transformer encoder. Under the participation of the staining modality encoder and modality mapping network, SSTar-TransGAN is trained adversarially by two pairs of generators and discriminators, while the staining transformation is completed under the constraints of the aforementioned loss functions.

Taking the transformation process from H&E to other IHC staining modalities as an example, the modality mapping network F encodes a random latent code \mathbf{z} and a specific IHC staining modality y_{IHC} into a staining code \mathbf{s}_{IHC} . Then, the generator $G_{H\&E}$ generates IHC image \mathbf{x}_{IHC} from the spectrally super-resolved hyperspectral data $\mathbf{x}_{H\&E}$ and the staining code \mathbf{s}_{IHC} . Meanwhile, $\mathbf{x}_{H\&E}$ is input into the staining modality encoder $E_{H\&E}$ and discriminator $D_{H\&E}$, where the staining modality encoder $E_{H\&E}$ evaluates the staining modality of $\mathbf{x}_{H\&E}$ and generates a staining code $\tilde{\mathbf{s}}_{H\&E}$, while generator $G_{IHC}(\mathbf{x}, \tilde{\mathbf{s}})$ generates hyperspectral data $\tilde{\mathbf{x}}_{H\&E}$ from \mathbf{x}_{IHC} and staining code $\tilde{\mathbf{s}}_{H\&E}$. At the same time, \mathbf{x}_{IHC} is input into the staining modality encoder E_{IHC} and the discriminator D_{IHC} , generating a staining code $\tilde{\mathbf{s}}_{IHC}$. This process is shown in Figure 1.

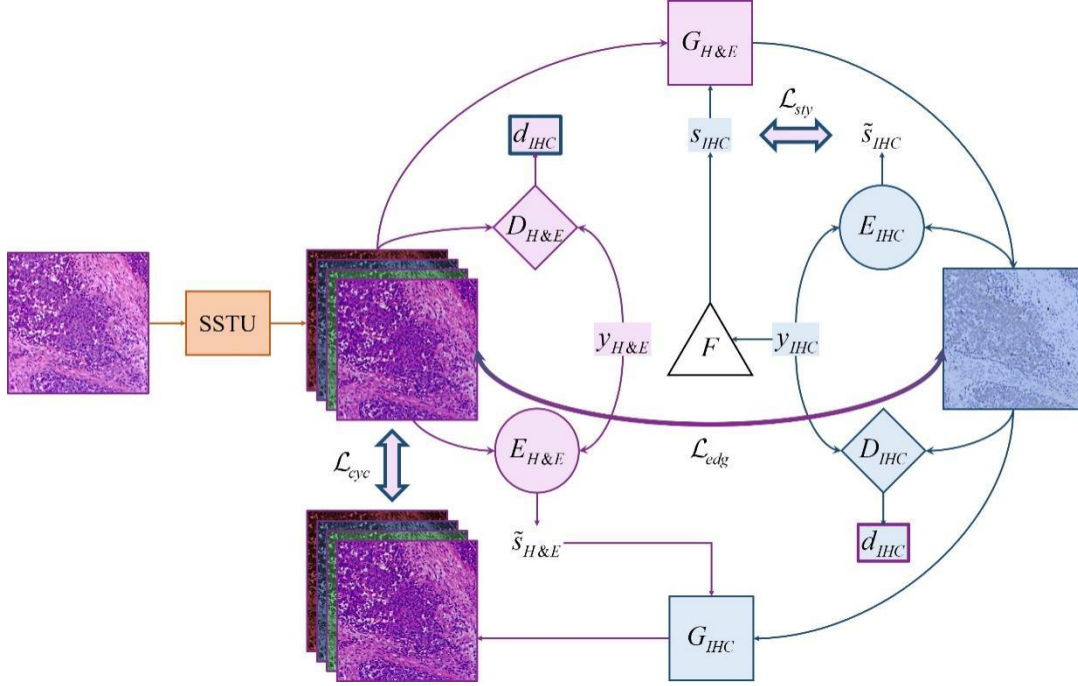


Fig. 1 The workflow of SSTar-TransGAN. Purple arrows indicate the H&E related process, while blue arrows indicate the IHC related process. Cross-colored arrows indicate four loss functions.

2.2 Spectral Reconstruction of HSI Input

As a common tissue staining technique, H&E is often used for clinical observation and diagnosis, but it only provides limited color information, which restricts the potential in-depth analysis of tissue characteristics. Hyperspectral microscopy reconstruction technology provides sufficient spectral information for the input of H&E patches, thus improving the quality of staining transformation. Despite the difficulty in obtaining a large amount of HSI microscopic data, the large amount of data provided by HSI as the input of network will undoubtedly bring difficulties to training and convergence of the network. Considering a great number of success researches of deep learning technology to convert RGB microscopic images into hyperspectral images, we introduced the spectral reconstruction method by using SSTU to convert the input RGB-H&E images into HSI-H&E data. Through a pre-trained SSTU network, the input RGB-IHC patches with the size of $256 \times 256 \times 3$ are converted into patches with the size of $256 \times 256 \times 90$ at the same field of view size, which significantly increases the spectral dimension of the data, thereby providing richer information for further analysis.

2.3 Transformer-based Encoder

Transformer was initially designed for sequential data processing, especially for natural language processing (NLP) tasks. Due to its powerful attention mechanism, transformer-based structures are able to capture long distance dependencies in high dimensional data. When this structure is applied to image processing tasks, it can extract deep-level features from images, which is crucial for generating high-quality image reconstruction. In CycleGAN structure, the generator is responsible for converting input images from one modality to another, such as from low spatial resolution to high, or from one style to another. To achieve this goal, the generator needs to be able to understand the intrinsic structure and content of the input image. Using transformer as the encoder can provide a richer and more detailed feature representation process, which can then be further processed by a CNN based decoder to generate a high-fidelity output image. [13] points out that this combination of transformer and CNN not only improves image quality but also maintains computational efficiency. In this experiment, vision transformer is used as the encoder in the generator, replacing the encoding part of the Adaptive Instance Normalization (AdaIN) module in the original starGAN-v2, resulting in a more effective extraction of the input image.

2.4 Loss Functions

2.4.1 Adversarial Loss

After spectral super-resolution of the SSTU network, the goal of the generator $G_{H\&E}$ is to convert the hyperspectral H&E image $\mathbf{x}_{H\&E}$ into an RGB image \mathbf{x}_{IHC} corresponding to the staining code \mathbf{s}_{IHC} , where \mathbf{s}_{IHC} can be obtained through the mapping network F according to a specified staining modality y_{IHC} within the input of a random latent code \mathbf{z} . To achieve this goal, the generator $G_{H\&E}$ and the discriminator $D_{H\&E}$ collaborate in the adversarial training process, while the task of $D_{H\&E}$ is to distinguish real H&E images from fake images generated by $G_{H\&E}$. The adversarial loss based on binary cross-entropy L_{adv} ensures that the generated image $G_{H\&E}(\mathbf{x}_{H\&E}, \mathbf{s}_{IHC})$ is indistinguishable from other real images under the same staining modality to the discriminator $D_{H\&E}$. L_{adv} can be described as:

$$L_{adv} = \mathbb{E}_{\mathbf{x}_{H\&E}, y_{H\&E}} [\log D_{H\&E}(\mathbf{x}_{H\&E})] + \mathbb{E}_{\mathbf{x}_{H\&E}, y_{IHC}, \mathbf{z}} [\log (1 - D_{IHC}(G_{H\&E}(\mathbf{x}_{H\&E}, \mathbf{s}_{IHC})))] \quad (1)$$

where $D_{H\&E}$, D_{IHC} represents the discriminators on staining modalities of $y_{H\&E}$ and y_{IHC} , respectively.

2.4.2 Cycle Consistency Loss

SSTar-TransGAN adopts the classical cycle consistency mechanism from CycleGAN, where l_1 constraint maintains the spatial structure of input images while undergoing the staining transformation by the generator, making the process of unsupervised staining possible. This process can be described as:

$$L_{cyc} = \mathbb{E}_{\mathbf{x}_{H\&E}, y_{H\&E}, y_{IHC}, \mathbf{z}} [\|\mathbf{x}_{H\&E} - G_{IHC}(G_{H\&E}(\mathbf{x}_{H\&E}, \mathbf{s}_{IHC}), \tilde{\mathbf{s}}_{H\&E})\|] \quad (2)$$

where $\tilde{\mathbf{s}}_{H\&E} = E_{IHC}(G_{H\&E}(\mathbf{x}_{H\&E}, \mathbf{s}_{IHC}))$ denotes the estimated staining code from generated IHC image by $G_{H\&E}$.

2.4.3 Style Reconstruction Loss

In order to enhance the influence of the staining code $\tilde{\mathbf{s}}_{H\&E}$ on the generator G_{IHC} , style reconstruction loss L_{sty} has been applied according to the cycling mechanism in CycleGAN so that the other input of G_{IHC} can be iterated. L_{sty} can be described as:

$$L_{sty} = \mathbb{E}_{\mathbf{x}_{H\&E}, y_{IHC}, \mathbf{z}} [\|\mathbf{s}_{IHC} - E_{IHC}(G_{H\&E}(\mathbf{x}_{H\&E}, \mathbf{s}_{IHC}))\|] \quad (3)$$

2.4.4 Edge Loss

Research [12] has confirmed that staining transformation tasks emphasize the invariance of spatial features when converting between different staining modalities, so there is no need for additional loss constraint on style diversity as in other style transformation tasks such as human faces and animals. In SSTar-TransGAN, Canny edge loss L_{edg} has been exploited in order to maintain the spatial edge contours before and after the staining modality transformation:

$$L_{edg} = \mathbb{E}_{\mathbf{x}_{H\&E}, y_{IHC}, \mathbf{z}} [Canny(G_{H\&E}(\mathbf{x}_{H\&E}, \mathbf{s}_{IHC}), \theta_1, \theta_2) - Canny(\mathbf{x}_{H\&E}, \theta_1, \theta_2)] \quad (4)$$

where $Canny(\cdot)$ represents the Canny operator and θ_1, θ_2 represents the hyper-parameters for margin extraction calculation.

The above-mentioned loss functions can be integrated as:

$$\min_{G, F, E} \max_D [L_{adv} + \lambda_{sty} L_{sty} + \lambda_{cyc} L_{cyc} + \lambda_{edg} L_{edg}] \quad (5)$$

where $\lambda_{sty}, \lambda_{cyc}, \lambda_{edg}$ denote the hyperparameters for each individual loss function.

3. RESULTS AND DISCUSSION

3.1 Datasets

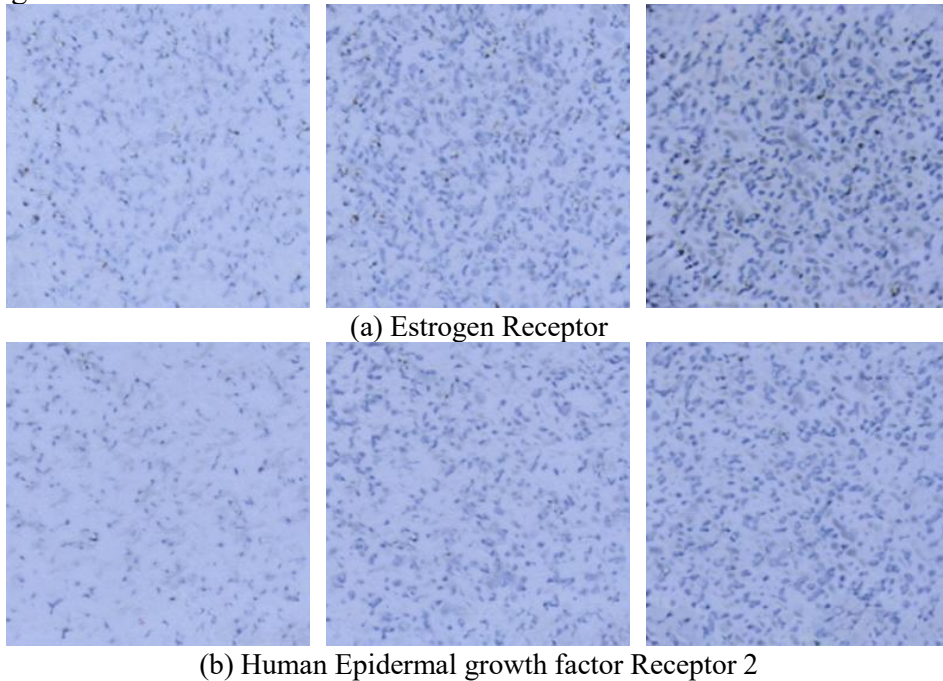
In this experiment, The RGB whole slide imaging data were collected by Leika ICC50 HD metallographic microscope, which provides a magnification of 20 times. The involved biopsy samples were collected from 13 breast cancer patients, the adjacent continuous sections of which were stained with H&E and three types of IHC staining, namely Estrogen Receptor (ER), Human Epidermal growth factor Receptor 2 (HER2), and Progesterone Receptor (PR) staining, thus obtaining image data under four different staining modalities. During the experiment, datasets from the four staining modalities were all segmented into patches with a size of 256×256 pixels. For each staining modality, 520 patches of images have been prepared for the training set. Similarly, the validation and test set were also formed according to the same image size of 256×256 pixels within 40 image patches for each staining modality, to ensure the consistency and effectiveness of the model training and evaluation process.

3.2 Experimental Setup

In the experiment, the batch size was set to 2, and the model was trained for a total of 150 epochs. The model applied Adam optimizer, while the hyperparameters of the training process, namely learning rate η_{\min} , weight decay ω , coefficients for calculating running averages of gradient β_1 and its square β_2 , were set as $\eta_{\min} = 10^{-4}$, $\omega = 10^{-4}$, $\beta_1 = 0.0$, $\beta_2 = 0.99$. With the aid of 4 NVIDIA Titan XP GPUs, we implemented the proposed model in the PyTorch framework.

3.3 Qualitative and Quantitative Results

To validate the efficiency of our method, we performed qualitative and quantitative analysis between typical and state-of-the-art methods: CycleGAN [6], modified StarGAN [12] and SStar-TransGAN. As Fig. 2 illustrates, the reconstruction results from traditional CycleGAN preserves the least spatial and spectral details among these competing methods. Modified StarGAN is successful in extracting the spatial structures while lack of preserving color information. SStar-TransGAN is capable of maintaining more spatial-spectral details comparing to the other two digital staining methods.



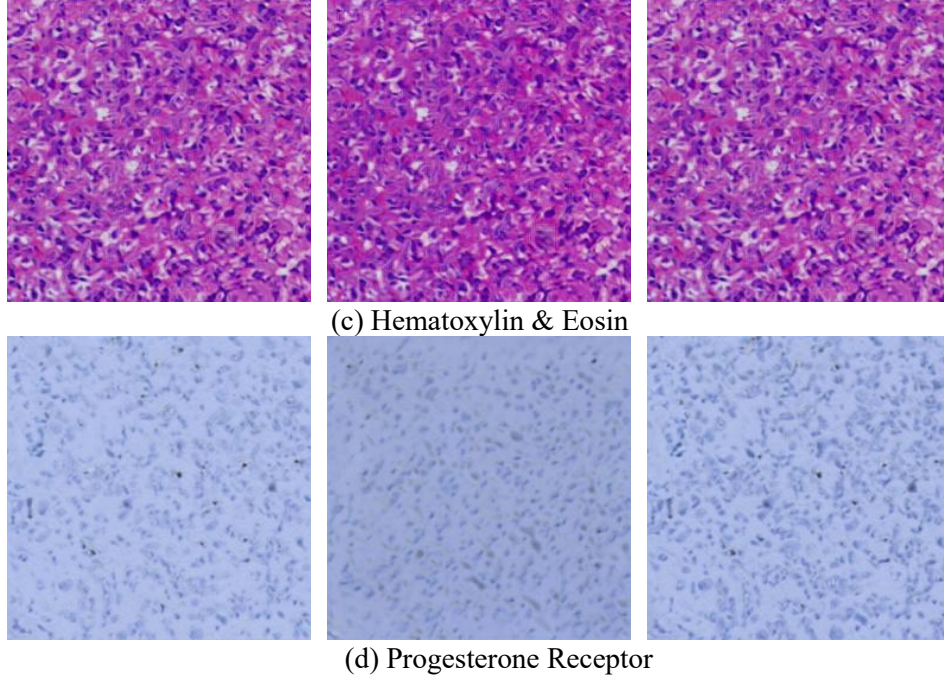


Fig. 2 Digital staining results from CycleGAN (left), StarGAN (middle) and SSTar-TransGAN (right).

Referring to the quantitative results for unpaired datasets, we reported the Frechet inception distance (FID) [14] and learned perceptual image patch similarity (LPIPS) [15] of the digital staining results from CycleGAN, StarGAN and SSTar-TransGAN. To be specific:

$$FID = \left\| \mu_g - \mu_r \right\|_2^2 + \text{Tr} \left(\Sigma_g + \Sigma_r - 2 \left(\Sigma_g \Sigma_r \right)^{\frac{1}{2}} \right) \quad (6)$$

$$LPIPS(x_r, x_g) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l) \right\|_2^2 \quad (7)$$

where μ and Σ represents the mean value and covariance corresponding to the generated images x_g and the real images x_r , while H, W, L, w, \hat{y} represents the height, width, layers, scaling weights and unit-normalized features from the reconstruction network. A smaller FID indicates a similar distribution between x_r and x_g , and a lower LPIPS indicates more similarity between x_r and x_g according to universal human cognition. Table 1 displays the reconstruction results of SSTar-TransGAN and other compared methods. The results show that SSTar-TransGAN outperforms the other methods in all evaluation criteria.

Table 1. Three Scheme comparing

	FID	LPIPS
CycleGAN	203.8641	0.1045
StarGAN	234.8697	0.0892
SSTar-TransGAN	228.1115	0.0764

4. Conclusion

This study successfully proposed a digital staining network named SSTar-TransGAN, which solves the limitations of microscopic hyperspectral images to be applied in traditional digital staining works. By introducing SSTU spectral super-resolution network, we were able to convert H&E stained RGB images into hyperspectral images without incurring additional data requirements, thus providing rich spectral information for staining modality transformation process.

The advantage of the SSTar-TransGAN structure lies in the introduction of style encoder, thus enabling a sharing generator between IHC modalities, which can adapt to digital staining tasks with varying numbers of modalities without increasing great computational costs of extra generators. In addition, the introduction of the transformer encoder of generator accelerates the network's convergence speed and provides excellent quantitative and qualitative results while ensuring the preservation of image spatial structure and color information. Experimental results demonstrate that SSTar-TransGAN is capable of generating high-quality images and performs better in tasks of transformation between different staining modalities, which is of significant importance to pathology and life sciences research.

References

- [1] Bai B, Yang X, Li Y, et al. Deep learning-enabled virtual histological staining of biological samples[J]. *Light: Science & Applications*, 2023, 12(1): 57.
- [2] Ortega S, Halicek M, Fabelo H, et al. Hyperspectral and multispectral imaging in digital and computational pathology: a systematic review[J]. *Biomedical Optics Express*, 2020, 11(6): 3195-3233.
- [3] Bautista P A, Yagi Y. Digital simulation of staining in histopathology multispectral images: enhancement and linear transformation of spectral transmittance[J]. *Journal of Biomedical Optics*, 2012, 17(5): 056013-056013.
- [4] Bayramoglu N, Kaakinen M, Eklund L, et al. Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks[C]//*Proceedings of the IEEE international conference on computer vision workshops*. 2017: 64-71.
- [5] Zhang Y, de Haan K, Rivenson Y, et al. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue[J]. *Light: Science & Applications*, 2020, 9(1): 78.
- [6] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//*Proceedings of the IEEE international conference on computer vision*. 2017: 2223-2232.
- [7] Zhang R, Cao Y, Li Y, et al. MVFStain: multiple virtual functional stain histopathology images generation based on specific domain mapping[J]. *Medical Image Analysis*, 2022, 80: 102520.
- [8] Biswas T, Suzuki H, Ishikawa M, et al. Generative adversarial network based digital stain conversion for generating RGB EVG stained image from hyperspectral H&E stained image[J]. *Journal of Biomedical Optics*, 2023, 28(5): 056501-056501.
- [9] Choi Y, Uh Y, Yoo J, et al. Stargan v2: Diverse image synthesis for multiple domains[C]//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020: 8188-8197.
- [10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. *arXiv preprint arXiv:2010.11929*, 2020.
- [11] Wang Y, Gu Y, Nanding A. SSTU: Swin-Spectral Transformer U-Net for hyperspectral whole slide image reconstruction[J]. *Computerized Medical Imaging and Graphics*, 2024, 114: 102367.
- [12] Berijanian M, Schaadt N S, Huang B, et al. Unsupervised many-to-many stain translation for histological image augmentation to improve classification accuracy[J]. *Journal of Pathology Informatics*, 2023, 14: 100195.
- [13] Alimanov A, Islam M B. Retinal image restoration using transformer and cycle-consistent generative adversarial network[C]//*2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, 2022: 1-4.
- [14] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. *Advances in neural information processing systems*, 2017, 30.
- [15] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: 586-595.