AV-GAN: Attention-Based Varifocal Generative Adversarial Network for Uneven Medical Image Translation

Zexin Li
SIGS of Tsinghua University
Shenzhen, China
li-zx21@mails.tsinghua.edu.cn

Yiyang Lin
SIGS of Tsinghua University
Shenzhen, China
lyy20@mails.tsinghua.edu.cn

Zijie Fang
SIGS of Tsinghua University
Shenzhen, China
vison307@gmail.com

Shuyan Li
Department of Automation of Tsinghua University
Beijing, China
li-sy16@tsinghua.org.cn

Xiu Li
SIGS of Tsinghua University
Shenzhen, China
li.xiu@sz.tsinghua.edu.cn

Abstract—Different types of staining highlight different structures in organs, thereby assisting in diagnosis. However, due to the impossibility of repeated staining, we cannot obtain different types of stained slides of the same tissue area. Translating the slide that is easy to obtain (e.g., H&E) to slides of staining types difficult to obtain (e.g., MT, PAS) is a promising way to solve this problem. However, some regions are closely connected to other regions, and to maintain this connection, they often have complex structures and are difficult to translate, which may lead to wrong translations. In this paper, we propose the Attention-Based Varifocal Generative Adversarial Network (AV-GAN), which solves multiple problems in pathologic image translation tasks, such as uneven translation difficulty in different regions, mutual interference of multiple resolution information, and nuclear deformation. Specifically, we develop an Attention-Based Key Region Selection Module, which can attend to regions with higher translation difficulty. We then develop a Varifocal Module to translate these regions at multiple resolutions. Experimental results show that our proposed AV-GAN outperforms existing image translation methods with two virtual kidney tissue staining tasks and improves FID values by 15.9 and 4.16 respectively in the H&E-MT and H&E-PAS tasks.

Index Terms—GAN, Virtual Staining, Image Translation

I. INTRODUCTION

In computational pathology, pathologists usually use pathological slides to obtain an accurate diagnosis and pathological staining is the gold standard [21]. Histochemical staining detects the presence of specific antigens in pathological tissues through the combination of antibodies and antigens and pathologists usually need to use different types of histochemical staining to achieve an accurate pathological diagnosis [23]. For example, different types of pathological stains contain different molecular information, pathologists need several types of stainings to observe each specific structure of the kidney [22], [24]. For example, Masson trichromatic staining (MT) slides are for connective tissue observation, and Periodate Schiff Staining (PAS) slides are for the observation of the basement
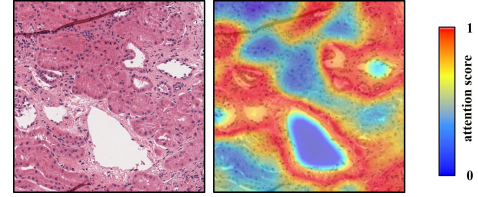

Fig. 1. A H&E stained patch and its attention map. The attention map shows that different regions deserve different extents of attention and structures like the edge of cavities in this patch deserve more attention, which corresponds to the fact that the edge is crucial for the shape of translated images.

membrane. However, the tissue cannot be stained multiple times, as chemical staining destroys the cell structure [25]. To obtain a variety of staining types, existing methods apply virtual staining technology based on image translation [3]. Specifically, they translate the staining slide which is easy to obtain (e.g. H&E) to slides that require more preparation cost (e.g. MT, and PAS).

Although existing virtual staining technologies have achieved promising performance, there are still several challenges. Firstly, as shown in Figure 1, unlike the attention mechanism target in the previous method, which involves finding foreground objects to focus on in the background (such as finding horses in grasslands), in the task of pathological slides virtual staining, we need to find key regions that interact closely with tissues in other areas, as these key regions often require more complex structures to maintain this interaction, leading to increased translation difficulty [27]. For example, in the translation task of renal pathology slides, to efficiently filter the original urine sent from other tissues in a limited space, the glomerulus in the kidney must increase the contact area with the renal tubules, forming a complex cystic structure, making it difficult to translate. Motivated by this point, we design a novel attention mechanism that focuses on finding regions in

pathological slides that are closely related to other regions and treating them as regions with higher translation difficulty.

After that, the existing methods have not been able to effectively translate information of different resolutions separately, making them unable to achieve the desired translation effect for information with high resolution (such as detailed texture, glomerular basement membrane, etc.) and information with low resolution (such as the shape of the tissue and the color appearance) simultaneously and separately [17]. Besides, the relative position of the nucleus may change during the translation of some models, which may lead to the deformation of some regions.

In this paper, we propose an Attention-Based Varifocal Generative Adversarial Network (AV-GAN) for image translation. To provide sufficient attention to the key regions (such as the glomerulus mentioned above) in the image, we adopt the attention mechanism to effectively focus on key regions. Besides, to better translate low-resolution features like the global shape as well as the appearance of the tissue and high-resolution features such as the detailed texture and local edge, we utilize a twin network architecture called varifocal module to achieve the extraction of features at different resolutions and the interaction between these features. In addition, we set the H channel loss to constrain the position of the nucleus to ensure that the shape of the tissue does not change significantly. Our work may do good for several downstream tasks such as object detection [29].

To sum up, our main contributions are as follows:

- We design an Attention-based Key Region Selection Module based on the attention mechanism, which can select the structures in the tissue with higher translation difficulty and more connections to other areas.
- We propose a Varifocal Module, utilizing two generators to deal with global (low-resolution) and local (high-resolution) information respectively to ensure that information with different resolutions will not interfere with each other.
- We conduct experiments on the image translation tasks H&E-MT and H&E-PAS. Experimental results show that our method can translate H&E staining slides into high-quality MT staining slides and PAS staining slides.

II. RELATED WORK

A. Virtual Staining in Histopathological Analysis

Researchers have begun to explore virtual staining through image translation models. Lin et al. use StarGAN [5] as the baseline to convert H&E slides into special staining slides to achieve one-to-many image translation [4]. Yang et al. convert the H&E staining slides of the kidney into special staining slides [30]. Rivenson et al. virtually generate PAS, MT slides from H&E slides. However, this model needs to input pixel-level registered data. Due to the large size of pathological WSIs, it is difficult to implement pixel-level registration [7]. For the existing computational pathology image translation model, due to the differences in the translation difficulty of different regions of the slide (for example, the glomerulus is the key structure in kidney slides that should be focused on more), there is still a lack of an effective attention mechanism to enable the model to focus on these regions. Meanwhile, the input of high-resolution images forces the model to pay attention to details (such as tissue edges and cavities), while the input of low-resolution images forces the model to observe the global information and then handle the spatial relevance (such as the law of nuclear distribution). The existing pathology slides translation model can hardly explore the relationship between multiple resolutions.

B. Generative Adversarial Network

The Generative Adversarial network [9] has become an important framework in image translation [26] [28] [5]. Zhu et al. propose Pix2Pix [9] and CycleGAN [10] for supervised and unsupervised image translation. Kim et al. propose UGATIT [11] for style transfer. Huang et al. propose MUNIT [12] to decouple image style information and content information. Mohammed et al. propose an image translation framework called AI-FFPE [8]. Zhang et al. design the Self-Attention Generative Adversarial Network (SAGAN) [13], utilizing prompts from all feature locations to generate details. However, this scheme fails to pay attention to image information at different resolutions. Apart from that, Hu et al. propose the QS-Attn mechanism [14]. Among the features encoded by Encoder, the features that the author needs to constrain are filtered by the "query-select" method. However, this method only filters at the feature level and does not directly notice the regions in the image that should be concerned about. Zhang et al. successfully introduce the attention mechanism into GAN [16], but it is not applied in the image translation task. Kwon et al. propose a new layered adaptive diagonal spatial attention (DAT) layer [15], which hierarchically processes spatial content and style, but this scheme does not process high-resolution and low-resolution information separately. In a word, the above frameworks fail to pay attention to the multi-resolution information, resulting in some key structures (such as the edge of tissue) not being well translated.

III. METHOD

In medical images, to solve the problem some structures are relatively complex (such as glomerulus, renal tubule, etc.) and cannot be translated precisely using a simple network, as shown in Figure 2, we propose an Attention-Based Varifocal Generative Adversarial Network (AV-GAN), which solves the problem of uneven difficulty distribution in image translation.

A. Attention-based Key Region Selection Module

Correct virtual staining of key structures is of great significance for accurate diagnosis. For example, in PAS slides, the glomerular basement membrane is stained evidently, so the glomerulus is the key structure of the slide. Correct staining of the glomerulus helps the diagnosis and treatment of a variety of kidney diseases (such as the classification of nephritis). To capture the key structures in pathological slides, we develop
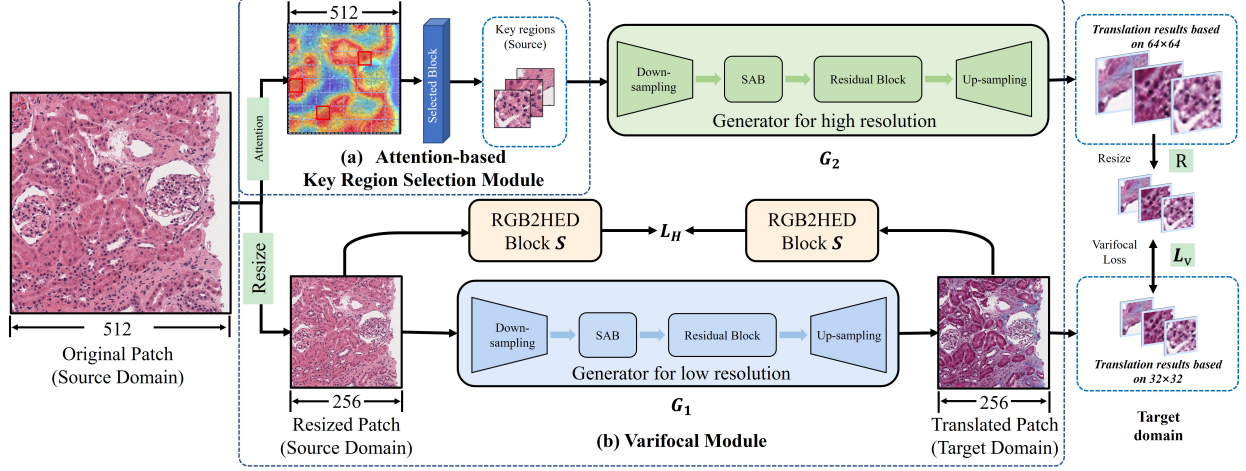
Fig. 2. The structure of AV-GAN. G_1 and G_2 refer to the low-resolution generator and the high-resolution one respectively. D_1 and D_2 are the discriminators of high-resolution and low-resolution, which are not drawn in the figure. The Attention-Based Key Region Selection Module selects the region that is worth attention and the RGB2HED Block converts the RGB image to an HED image, whose H channel (nuclear channel) can be constrained.
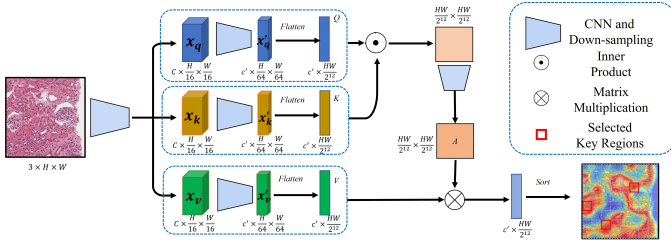

Fig. 3. The structure of the attention module. The Q, K, and V refer to the Query, Key, and Value network. The matrix A represents the attention map, which is calculated through Q and K. The network will select the region with key structures that need to be translated more precisely and return a cord list with each coordinate group representing a region.

an Attention-based Key Region Selection Module, as shown in Figure 3. Firstly, we feed the input image $x \in \mathbb{R}^{C \times H \times W}$ into three independent convolution neural networks and downsampling blocks, getting the output x_q, x_k, and $x_v \in \mathbb{R}^{C \times \frac{H}{16} \times \frac{W}{16}}$ where x_q, x_k, and x_v are the representations of the image in the Query, Key and Value space, respectively, when C, H and W refer to the number of channels, the height, and width of a patch. To avoid the excessive size of the attention map, we downsample these representations to x'_q, x'_k, and x'_v. Then, to obtain the strength of interaction between different regions, the three representations are flattened and the attention map of the image is calculated by

$$A = softmax({x'_q}^T x'_k), \tag{1}$$

where the softmax function is applied to the second dimension.

Next, a matrix P, the importance of each key region for selection, is calculated by

$$P = A {x'_v}^T. \tag{2}$$

We take the average of C channels for P and continue the following operation. At last, we sort the values in P to find the top n maximum values where n is the hyperparameter that represents the number of key regions. Finally, we obtain the key regions a_1, a_2, ..., a_i, ..., a_n $(1 \leq i \leq n)$ based on P' generated by P, and the generated process is described in the next paragraph.

To transmit gradients to the attention module, we propose a novel attention mechanism, which is a key novelty of our work. After getting the attention map, the selected block calculates $m = sigmoid(1000(P-\theta))$, where P is the average attention value of each region and θ determines how many regions to focus on (i.e., n). The larger the value of the parameter θ, the fewer regions of interest can be selected and the less n is. Then, we calculate the $P' = m \times P$, and finally, the selected block selects the key regions for the varifocal module. The network also adds Spatial Attention Block which has an impact like [8].

B. Varifocal Module

To improve the translation effect of the key regions selected in section III.A, we further propose a varifocal module. The module completes image translation at different resolutions through two independent generators, with each of them interacting with the other through varifocal loss, thus taking advantage of image features at different resolutions.

The low-resolution generator G_1 consists of a downsampling module, a spatial attention module, several bottleneck residual blocks, and an upsampling block. The main function of the low-resolution generator is to focus on the global characteristics of the image such as the shape and color style of the tissue. The low-resolution generator achieves better performance for translating the overall appearance of the image, especially the color styles, into another type of staining.

At the same time, we set up a high-resolution generator G_2, which can better focus on local characteristics such as the local texture and the glomerulus. The high-resolution generator achieves better performance for translating the local structures of the image, especially the local contour of tissues and detailed texture, into another type of staining.

To sum up, the low-resolution generator is better at capturing the global features such as the whole appearance, while the high-resolution generator is better at extracting local features such as some detailed textures. These two generators perform image translation at different resolutions and play a crucial role in the process of stain translation. we impel the generators of different resolutions to interact with each other properly instead of disorderly interference. The module fuses the advantages of two generators using L1 loss. The interaction process can be described as:

$$\sum_{n=1}^{N} \|C_n(G_1(I)) - G_2(a_n)\|_1 \qquad (3)$$

In the formula, I is the resized patch, a_n is the nth key region, $G_1(\cdot)$ is the low-resolution generator, $G_2(\cdot)$ is the high-resolution generator, and N is the total number of key regions. $C_n(\cdot)$ is the operation that crops the region from the resized patch I corresponding to a_n, $\|\cdot\|_1$ is L1 loss.

As a result, the fused network can deal with features in both high and low resolutions, capturing the information both globally and locally. Then we apply Varifocal loss to them as shown in Figure 2. In short, our model can significantly improve the virtual staining performance of key regions.

C. Discriminator

We use two discriminators D_1 and D_2 to distinguish low-resolution images and high-resolution local areas respectively. According to the research results of [1], the structure of the generator and discriminator equipped in networks should match the difficulty of the task. Therefore, we choose the PatchGAN discriminator.

D. RGB2HED

HED is a color space similar to RGB and is commonly used in computational pathology. RGB2HED is the process of transforming an image from RGB space to HED space by a stain deconvolution matrix. In HED space, the H channel highlights the nucleus. If the consistency of the H channel before and after translation can be maintained, the position of the nucleus can be guaranteed, and the structural precision after image translation is guaranteed. If we remove this constraint, we obtain FID values of 80.66 (H&E-MT translation) and 97.85 (H&E-PAS translation), worse than before, demonstrating the effectiveness of the H channel. Please refer to **H channel loss**.

E. Loss function

Our loss function has five terms and they are Adversarial loss L_{adv}, Identity loss L_{idt}, PatchNCE loss L_{NCE}, H channel loss L_H, and Varifocal loss L_v. Here we only introduce L_H and L_v. You may refer to other terms in [8], [10], [26].

1) H channel loss: To further constrain the spatial position relationship of the nucleus and ensure that the cell position distribution in the patch does not receive too much influence, we utilize the H channel separation operator S to convert the source domain and generated target domain images into the HED color space and constrain the H channel of the HED color space. G refers to the generator and $\mathbf{X}$ refers to the source domain. The H channel loss can be calculated by

$$L_H = \|S(\mathbf{X}) - S(G(\mathbf{X}))\|_1. \qquad (4)$$

2) Varifocal loss: To ensure the consistency of the generated images between different resolutions, we set the Varifocal loss L_v to constrain the difference in the generated results between the low-resolution and high-resolution generators. We find the corresponding regions of the selected regions in the low-resolution generated image and constrained those regions with the generation results of the high-resolution generator. To constrain images translated by the high-resolution generator and the low-resolution generator on the same scale, the larger image of the two is resized to the smaller image with the resizing operator R. The Varifocal loss is calculated by

$$L_v = \sum_{i=1}^{n} \|R(a_i) - b_i\|_1, \qquad (5)$$

where a_i refers to the region mentioned in section III.A, and b_i refers to the corresponding location in the translated image $G_1(x)$.

3) Optimization objective: The optimization objective function can be calculated by

$$\begin{aligned} L = &\lambda_{adv}L_{adv} + \lambda_{idt}L_{idt} + \\ &\lambda_{NCE}(L_{NCE}(\mathbf{X}) + L_{NCE}(\mathbf{Y})) + \lambda_H L_H + \lambda_v L_v, \end{aligned} \qquad (6)$$

where the weights λ_{adv}, λ_{idt}, λ_{NCE}, λ_H, and λ_v are set to 1.0, 0.03, 1.0, 2.5, and 1.0 respectively to keep the loss terms in similar magnitude.

IV. EXPERIMENTS AND RESULTS

A. Dataset and Data Preprocessing

We have evaluated our proposed model over the H&E-MT and H&E-PAS image translation datasets of the ANHIR [18]. In the ANHIR dataset, there are 5 groups of high-resolution tissue slides of human kidneys. Each of which has 3 unpaired Whole Slide Images (WSIs) stained with different kinds of stain, namely Hematoxylin-eosin (H&E), Masson's trichrome(MT), and Periodic Acid-Schiff (PAS). We divide the dataset according to the existing work [4] and segment all slides into 512×512 size patches. We adopt the overlapping measure in the segmentation process, which sets the stride of segmentation to 64. In the end, we obtained 17,964 H&E patches, 20,139 MT patches, and 16,048 PAS patches as the training set 3,013 H&E patches, 3,577 MT patches, and 2,997 PAS patches as the testing set.

B. Experimental Details and Evaluation Criteria

During training, we use Adam [2] optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The initial learning rate is set to 0.0002. The batch size is set to 1. We adopt Fréchet Inception Distance (FID) [19], Kernel Inception Distance (KID) [20], and Collaborative Structural Similarity (CSS) [4] to evaluate our model. The smaller the FID and KID values, the more similar the translated images are to real images. However, CSS only measures the similarity between the translated image and the source image, and higher CSS does not mean higher quality of the generated image.

C. Quantitative Results

1) Comparative Experiments of H&E-MT Task: We test 3013 H&E images in the dataset and compare the results with starGAN [5], MW-GAN [31], UGATIT [11], CycleGAN [10], AI-FFPE [8], SSPRVS [6], and UMDST [4]. The results are shown in Table I.

TABLE I
COMPARISON OF THE EFFECT OF AV-GAN IMAGE TRANSLATION WITH OTHER IMAGE TRANSLATION METHODS (H&E-MT IMAGE TRANSLATION TASK). THE VALUES IN BOLD ARE THE BEST ONES, AND THE RESULTS UNDERLINED ARE THE SECOND-BEST ONES.

Model	FID↓	CSS↑	100KID↓
StarGAN(2017)	284.37	0.62	29.11
MW-GAN(2019)	194.32	0.64	18.18
UGATIT(2019)	142.50	0.33	7.33
CycleGAN(2017)	95.22	0.69	4.34
AI-FFPE(2022)	<u>89.58</u>	**0.81**	<u>3.56</u>
UMDST(2022)	94.43	0.69	4.27
SSPRVS(2022) [6]	140.84	0.68	6.69
AV-GAN (ours)	**73.68**	<u>0.75</u>	**2.41**

In Table 1, the generated images of our model have better quality, indicating that they are more similar to the target domain image. From the effect of the H&E-MT image translation task, we can conclude that our model has further decreased the FID value compared with existing methods. The 100KID also decreases by 1.15 compared with the previous SOTA value. The larger the CSS value, the better the target domain image can maintain the object structure in the source domain. Our CSS value exceeds most image translation methods largely, which shows that our model can effectively maintain the structural information in the source domain. This ensures the authenticity of cell structure in image translation tasks.

2) Comparative Experiments of H&E-PAS Task: Table II shows the result of the H&E-PAS image translation dataset, compared with other models, our method is significantly ahead. The FID value of the image dataset generated by our method is 88.36, which is better than other methods. One of the key reasons is that the basement membrane in H&E patches is a detailed structure, and the traditional virtual staining image translation model can not recognize the structure well, resulting in a higher FID value. Our model also surpasses most comparison methods in terms of the CSS value, indicating that the structure information in the source domain is also well preserved.

TABLE II
COMPARISON OF THE EFFECT OF AV-GAN IMAGE TRANSLATION WITH OTHER IMAGE TRANSLATION METHODS (H&E-PAS IMAGE TRANSLATION TASK). THE VALUES IN BOLD ARE THE BEST ONES, AND THE RESULTS UNDERLINED ARE THE SECOND-BEST ONES.

Model	FID↓	CSS↑	100KID↓
StarGAN(2017)	219.71	0.60	20.48
MW-GAN(2019)	176.77	0.66	15.35
UGATIT(2019)	156.88	0.31	10.11
CycleGAN(2017)	106.51	0.66	3.95
AI-FFPE(2022)	97.03	**0.77**	<u>3.52</u>
UMDST(2022)	<u>92.52</u>	0.69	4.28
SSPRVS(2022) [6]	131.46	0.68	7.21
AV-GAN (ours)	**88.36**	<u>0.75</u>	**3.16**

D. Ablation Study

1) Sharing the parameters of the two generators G_1 and G_2: We design the high-resolution generator G_2 to let the model focus more on details and avoid artifacts such as cell deformation in the translated images. We utilize the low-resolution generator G_1 to let the model get more global consideration and to avoid errors such as cell mass distribution displacement in the translated images. If the two generators share the parameters, the model can only focus on the common features rather than the features unique to each resolution. The result is shown in Table III. The experiments show that using only one generator has obvious disadvantages over using two.

TABLE III
COMPARE THE RESULTS OF WHETHER THE WEIGHTS OF TWO GENERATORS ARE SHARED.

Measurement	Shared parameters	Unshared parameters
H&E-MT FID↓	128.07	**73.68**
H&E-MT 100KID↓	4.85	**2.41**
H&E-PAS FID↓	132.44	**88.36**
H&E-PAS 100KID↓	5.04	**3.16**

2) Selecting the Proper Number of Key Regions: To verify the importance of the Key Region Selection Module in region selection, we remove the Key Region Selection Module by selecting a fixed region, and then comparing the FID value and KID value of the generated image with the previous results. The results in Table IV show that when the number of regions is the same if the Key Region Selection Module is absent, the FID value and KID value will both increase, the quality of the generated image will decrease, and the distribution of the generated image dataset will be far from the distribution of the real target domain image dataset.

The table also shows the influence of the number of regions on evaluations. When the number of attention regions increases, the CSS value remains almost unchanged, and the FID and KID have decreased to varying degrees. The best FID value is 73.68, and the best KID value is 2.30, which is obtained when adding two regions and three regions, respectively. This shows that in the H&E-MT image translation task, properly increasing the number of the attention regions is conducive to generating higher-quality images.

TABLE IV
ABLATION EXPERIMENT. THE QUALITY OF TRANSLATED PICTURES IN THE CASE OF "USING THE ATTENTION MECHANISM TO SELECT REGION" AND "SELECTING FIXED REGION". (H&E-MT IMAGE TRANSLATION TASK)

Model	FID↓	CSS↑	100KID↓
AV-GAN (1 attention region)	<u>74.25</u>	**0.76**	**2.30**
AV-GAN (2 attention regions)	**73.68**	<u>0.75</u>	<u>2.41</u>
AV-GAN (3 attention regions)	76.01	0.73	2.51
AV-GAN (1 fixed region)	91.16	0.75	4.16
AV-GAN (2 fixed regions)	97.51	0.74	5.31
AV-GAN (3 fixed regions)	83.50	0.74	3.09

TABLE V
ABLATION EXPERIMENT. THE QUALITY OF TRANSLATED PICTURES IN THE CASE OF "USING THE ATTENTION MECHANISM TO SELECT REGION" AND "SELECTING FIXED REGION". (H&E-PAS IMAGE TRANSLATION TASK)

Model	FID↓	CSS↑	100KID↓
AV-GAN (1 attention region)	**88.36**	**0.75**	**3.16**
AV-GAN (2 attention regions)	90.20	0.73	3.83
AV-GAN (3 attention regions)	91.89	0.74	3.50
AV-GAN (1 fixed region)	95.68	0.75	3.41
AV-GAN (2 fixed regions)	96.19	0.73	3.43
AV-GAN (3 fixed regions)	97.96	0.73	3.99

The n needs not to be too large (generally less than 3). In a picture, only key areas such as dense nuclei need to be selected. Once the key areas are selected, other key areas of the same type need not be selected again, because they have similar characteristics due to the homogeneity of pathological images. If an excessively large n is chosen, the effect will not improve significantly except for increasing training time and GPU memory.

Table V shows the results of the H&E-PAS image translation task, which further validates the conclusions drawn previously. Specially, when the number of attention regions is 1, all evaluations are optimal. The result shows that because the overall color style of PAS is close to that of H&E (both of them are soft pink), there is no need for too much attention. As a result, we can translate the H&E image into the PAS image without adding too many attention regions.

TABLE VI
COMPARATIVE EXPERIMENT. THE QUALITY OF TRANSLATED PICTURES IN THE CASE OF "SELECTING THE REGION OF 128×128" AND "SELECTING THE REGION OF 64×64". (H&E-MT IMAGE TRANSLATION TASK)

The region size	FID↓	CSS↑	100KID↓
64 (1 attention region)	<u>74.25</u>	**0.76**	**2.30**
64 (2 attention regions)	**73.68**	<u>0.75</u>	<u>2.41</u>
64 (3 attention regions)	76.01	0.73	2.51
128 (1 attention region)	80.46	0.75	2.46
128 (2 attention regions)	89.49	0.73	3.52
128 (3 attention regions)	114.16	0.74	4.13

3) Selecting a Proper Size for Key Regions: To further explore the selection strategy of the number of Key Regions and region size, we adjust their values on the H&E-MT image translation task, which contains more structure and color conversion compared with the H&E-PAS task. To be specific,

we experiment with the region size of 64 and 128 and with the number of Key Regions of 1,2, and 3. The results are shown in Table VI. It can be found that when the region size is small, the performance will improve with the increase in the number of Key Regions. On the contrary, when the region size is large, the number of Key Regions should not be too large (Line 6). In summary, when the number of Key Regions and region size reach a balance, better performance can be achieved. Besides, the CSS values are almost all distributed around 0.75, which indicates that the number and size of the regions do not affect the CSS values.

E. Evaluation of Experts

To confirm the image translation effect of AV-GAN, we have designed a Visual Turing test performed by three experts. Since experts often see real stained slides, they have sufficient experience to determine whether a generated patch is realistic. On each task of H&E-MT and H&E-PAS image translation, each patch is translated with 7 models, generating 7 results and each expert is asked to select the most vivid one. We calculate the average of their results. Among the 4,000 best-translated MT patches, 1181.33 patches come from our model. Among the 4,000 best-translated PAS patches, 1149 patches come from our model. The results are shown in the Table VII.

TABLE VII
THE EVALUATION MADE BY THE EXPERTS. THE NUMBERS IN THE TABLE INDICATE HOW MANY OF THE 4000 IMAGES WITH THE BEST STAINING EFFECT COME FROM THE CORRESPONDING METHOD. THE VALUES IN THE TABLE ARE THE AVERAGE VALUES EVALUATED BY THREE EXPERTS.

Method	MT patches	PAS patches
StarGAN (2017)	289.67	236.33
MW-GAN (2019)	310.33	263.67
UGATIT (2019)	476.67	486.33
CycleGAN (2017)	523.33	593.33
UMDST (2022)	493.67	468.67
AI-FFPE (2022)	725.00	802.67
AV-GAN	**1181.33**	**1149**

F. Qualitative Experiment

As shown in the left side of Figure 4 and Figure 6, in the H&E-MT image translation task, the result of UGATIT is relatively poor, which not only has low legibility, but also has the phenomenon of nucleus loss, and the relatively close nuclei cannot be separated from each other (see the yellow circle in Figure 4). The translation of CycleGAN is slightly better, but the nucleus is hidden in the red zone and cannot be distinguished (see yellow circle in Figure 4). Then, the AI-FFPE model is slightly better in terms of structural clarity. Our AV-GAN can also ensure structural clarity and distinguish red muscle fibers better. Furthermore, in the part pointed out by the yellow arrow, there are two glomerular structures. According to the diagnosis of the pathologist, this patient did not suffer from the disease that could lead to the deposition of collagen fibers in the glomerulus and the MT staining slide of the glomerulus should not appear blue, but AI-FFPE translates some areas in the glomerulus into blue (Figure 5).
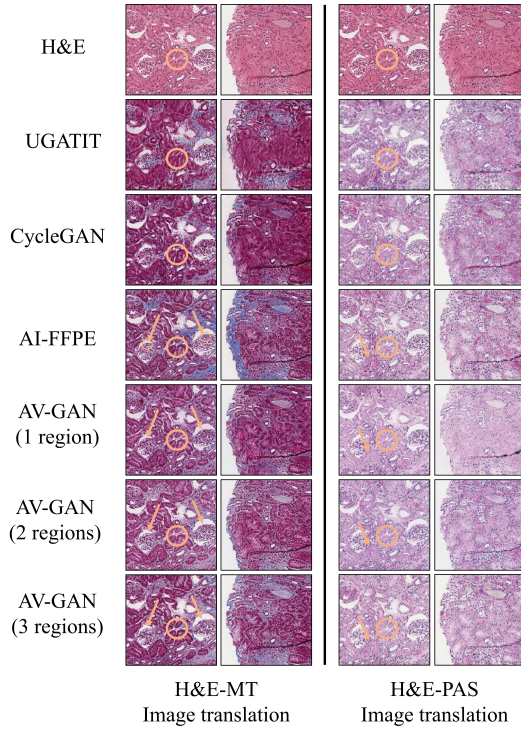
Fig. 4. Image translation results. The first line is the original H&E-stained images, the second line, the third line and the fourth line are the results of image translation using UGATIT, CycleGAN, and AI-FFPE models, and the fifth to seventh lines are the results of image translation using AV-GAN with different numbers of regions. Our task is to translate H&E staining into MT staining and PAS staining.
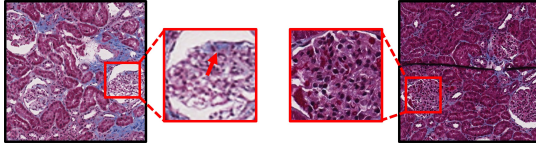

Fig. 5. The leftmost and rightmost are MT patches stained by AI-FFPE and a real patch of the same patient's MT patches (the dataset is an unpaired dataset). A blue zone appears at the location the red arrow points at, while the right patch indicates that there should be no blue in the patient's glomerular.

The right side of Figure 4 is the result of the image translation experiment of H&E-PAS staining. Similar to the H&E-MT staining image translation experiment, the image translation effect of UGATIT and CycleGAN is poor. The main function of PAS staining is to stain the glomerular basement membrane pink. In the part pointed out by the yellow arrow, AI-FFPE's result appears as a diffuse pink area, which is an error. After applying our model, the error is corrected.

V. CONCLUSION

We propose Attention-Based Varifocal Generative Adversarial Network (AV-GAN), effectively solves the problems of uneven translation difficulty, interference of different resolution information, and nuclear deformation in pathological image translation. In AV-GAN, the Attention-based Key Region Selection Module allows the network to find the area that is most closely related to other regions, which is often the area with higher translation difficulty, and the Varifocal Module decouples the information of regions difficult to translate at multiple resolutions and translate them separately, which improves the translation quality of key regions. This paper also utilizes the H channel loss term to restricts the position distribution of the nucleus and ensures that the tissue morphology does not change after translation. Our model gets the state-of-the-art performance on H&E-MAS and H&E-PAS translation tasks.

REFERENCES

[1] C. Yang, Y. Shen, Y. Xu, D. Zhao, B. Dai, and B. Zhou, "Improving gans with a dynamic discriminator," *arXiv preprint arXiv:2209.09897*, 2022.

[2] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[3] K. de Haan, Y. Zhang, J. E. Zuckerman, T. Liu, A. E. Sisk, M. F. Diaz, K.-Y. Jen, A. Nobori, S. Liou, S. Zhang, *et al.*, "Deep learning-based transformation of h&e stained tissues into special stains," *Nature communications*, vol. 12, no. 1, pp. 1–13, 2021.

[4] Y. Lin, B. Zeng, Y. Wang, Y. Chen, Z. Fang, J. Zhang, X. Ji, H. Wang, and Y. Zhang, "Unpaired multi-domain stain transfer for kidney histopathological images," 2022.

[5] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

[6] B. Zeng, Y. Lin, Y. Wang, Y. Chen, J. Dong, X. Li, and Y. Zhang, "Semi-supervised pr virtual staining for breast histopathological images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 232–241, Springer, 2022.

[7] Y. Rivenson, H. Wang, Z. Wei, K. de Haan, Y. Zhang, Y. Wu, H. Günaydın, J. E. Zuckerman, T. Chong, A. E. Sisk, *et al.*, "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning," *Nature biomedical engineering*, vol. 3, no. 6, pp. 466–477, 2019.

[8] K. Bengisu Ozyoruk, S. Can, G. Irem Gokceler, K. Basak, D. Demir, G. Serin, U. Payam Hacisalihoglu, E. Kurtuluş, B. Darbaz, M. Y. Lu, *et al.*, "Deep learning-based frozen section to ffpe translation," *arXiv e-prints*, pp. arXiv–2107, 2021.

[9] 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*, pp. 1125–1134, 2017.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[11] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *arXiv preprint arXiv:1907.10830*, 2019.

[12] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

[13] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019.

[14] X. Hu, X. Zhou, Q. Huang, Z. Shi, L. Sun, and Q. Li, "Qs-attn: Query-selected attention for contrastive learning in i2i translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18291–18300, 2022.

[15] G. Kwon and J. C. Ye, "Diagonal attention and style-based gan for content-style disentanglement in image generation and translation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13980–13989, October 2021.

[16] J. Zhang, L. Yu, D. Chen, W. Pan, C. Shi, Y. Niu, X. Yao, X. Xu, and Y. Cheng, "Dense gan and multi-layer attention based lesion segmentation method for covid-19 ct images," *Biomedical Signal Processing and Control*, vol. 69, p. 102901, 2021.
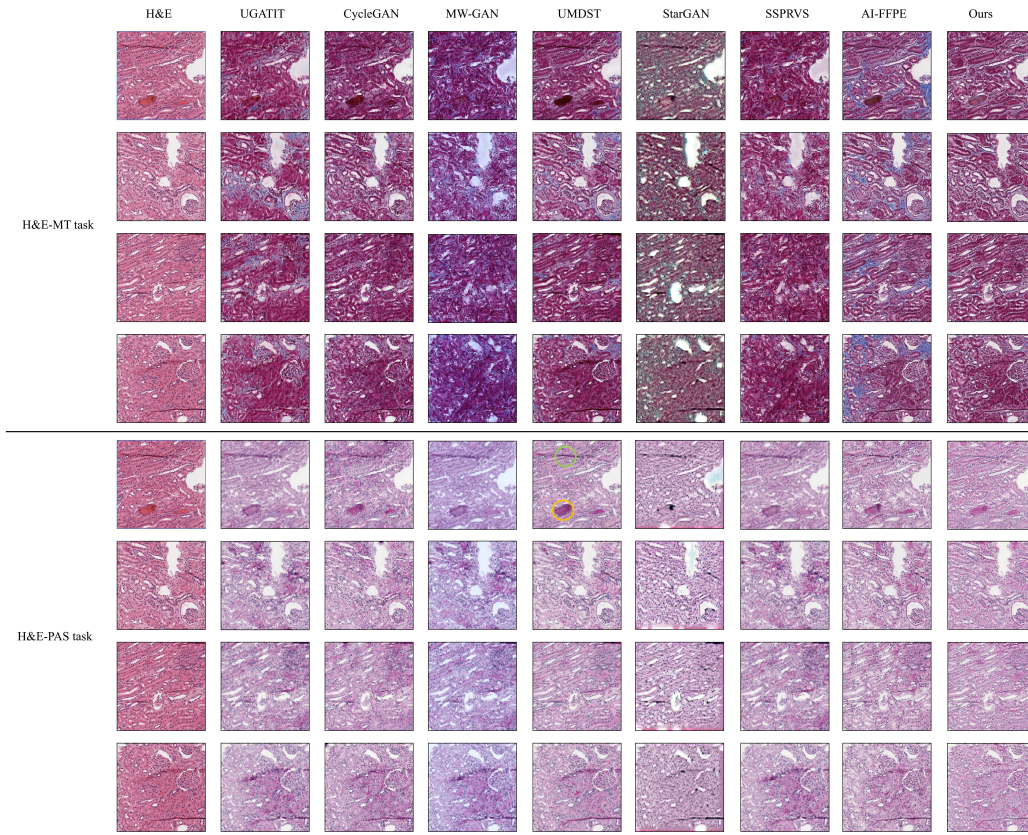
Fig. 6. Results of more qualitative experiments.

[17] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, "Fre-gan: Adversarial frequency-consistent audio synthesis," *arXiv preprint arXiv:2106.02297*, 2021.

[18] J. Borovec, J. Kybic, I. Arganda-Carreras, D. V. Sorokin, G. Bueno, A. V. Khvostikov, S. Bakas, E. I.-C. Chang, S. Heldmann, K. Kartasalo, L. Latonen, J. Lotz, M. Noga, S. Pati, K. Punithakumar, P. Ruusuvuori, A. Skalski, N. Tahmasebi, M. Valkonen, L. Venet, Y. Wang, N. Weiss, M. Wodzinski, Y. Xiang, Y. Xu, Y. Yan, P. Yushkevich, S. Zhao, and A. Muñoz-Barrutia, "Anhir: Automatic non-rigid histological image registration challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3042–3052, 2020.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017.

[20] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *ArXiv*, vol. abs/1801.01401, 2018.

[21] S. Montedonico, P. Cáceres, N. Muñoz, H. Yáñez, R. Ramírez, and B. Fadda, "Histochemical staining for intestinal dysganglionosis: over 30 years experience with more than 1,500 biopsies," *Pediatric surgery international*, vol. 27, no. 5, pp. 479–486, 2011.

[22] W. WANG, Y. LI, T. HUANG, and Y. ZHANG, "Diagnostic value of semi-quantitative analysis of microscopic images of masson's trichrome-stained histological sections in renal interstitial fibrosis," *Chinese General Practice*, vol. 23, no. 33, p. 4202, 2020.

[23] Y. Rivenson, K. de Haan, W. D. Wallace, and A. Ozcan, "Emerging advances to transform histopathology using virtual staining," *BME Frontiers*, vol. 2020, 2020.

[24] Y. Liu, X. Li, A. Zheng, X. Zhu, S. Liu, M. Hu, Q. Luo, H. Liao, M. Liu, Y. He, *et al.*, "Predict ki-67 positive cells in h&e-stained images using deep learning independently from ihc-stained images," *Frontiers in Molecular Biosciences*, vol. 7, p. 183, 2020.

[25] S. Roy-Chowdhuri, S. Dacic, M. Ghofrani, P. B. Illei, L. J. Layfield, C. Lee, C. W. Michael, R. A. Miller, J. W. Mitchell, B. Nikolic, *et al.*, "Collection and handling of thoracic small biopsy and cytology specimens for ancillary studies: guideline from the college of amer- ican pathologists in collaboration with the american college of chest physicians, association for molecular pathology, american society of cytopathology, american thoracic society, pulmonary pathology society, papanicolaou society of cytopathology, society of interventional radiology, and society of thoracic radiology," *Archives of pathology & laboratory medicine*, vol. 144, no. 8, pp. 933–958, 2020.

[26] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European conference on computer vision*, pp. 319–345, Springer, 2020.

[27] C. G. Galizia and R. Menzel, "The role of glomeruli in the neural representation of odours: results from optical recording studies," *Journal of insect physiology*, vol. 47, no. 2, pp. 115–130, 2001.

[28] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10551–10560, 2019.

[29] Q. Chu, S. Li, G. Chen, K. Li, and X. Li, "Adversarial alignment for source free object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 452–460, 2023.

[30] X. Yang, B. Bai, Y. Zhang, Y. Li, K. de Haan, T. Liu, and A. Ozcan, "Virtual stain transfer in histology via cascaded deep neural networks," *ACS Photonics*, vol. 9, no. 9, pp. 3134–3143, 2022.

[31] J. Cao, L. Mo, Y. Zhang, K. Jia, C. Shen, and M. Tan, "Multi-marginal wasserstein gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.