Bi-directional Feature Fusion Generative Adversarial Network for Ultra-high Resolution Pathological Image Virtual Re-staining

Kexin Sun[1], Zhineng Chen[1, 2*], Gongwei Wang[3], Jun Liu[3], Xiongjun Ye[4], Yu-Gang Jiang[1]
[1]School of Computer Science & Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University
[2]Shanghai Qi Zhi Institute
[3]Peking University People's Hospital
[4]Department of Urology, National Cancer Center & National Clinical Research Center for Cancer, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
sunkx21@m.fudan.edu.cn, {zhinchen, ygj}@fudan.edu.cn,
wgw4300@126.com, hmuliujun@163.com, yexiongjun@cicams.ac.cn

Abstract

The cost of pathological examination makes virtual re-staining of pathological images meaningful. However, due to the ultra-high resolution of pathological images, traditional virtual re-staining methods have to divide a WSI image into patches for model training and inference. Such a limitation leads to the lack of global information, resulting in observable differences in color, brightness and contrast when the re-stained patches are merged to generate an image of larger size. We summarize this issue as the square effect. Some existing methods try to solve this issue through overlapping between patches or simple post-processing. But the former one is not that effective, while the latter one requires carefully tuning. In order to eliminate the square effect, we design a bi-directional feature fusion generative adversarial network (BFF-GAN) with a global branch and a local branch. It learns the inter-patch connections through the fusion of global and local features plus patch-wise attention. We perform experiments on both the private dataset RCC and the public dataset ANHIR. The results show that our model achieves competitive performance and is able to generate extremely real images that are deceptive even for experienced pathologists, which means it is of great clinical significance.

1. Introduction

Pathological examination is the primary method of cancer diagnosis. Different dyes can interact with different components in tissues or cells, making it easier to distinguish different microstructures, abnormal substances and lesions. Among various staining methods, the most common and basic one is the hematoxylin-eosin (HE) staining. However, given the result of HE staining, it is not always enough to make a diagnosis. Therefore, immunohistochemistry (IHC) staining based on specific binding of antigen and antibody is also necessary in diagnosis, even though it is complex, time-consuming and expensive [7, 25].

Due to the cost of IHC, some researchers have tried to generate one type of staining images from another type (usually HE) via computational methods. This can reduce the consumption of materials, money and time during diagnosis. Such a task is usually called virtual re-staining. This task is close to the style transfer of natural images, so it is possible to apply style transfer methods to virtual re-staining. Since pathological images are usually unpaired, virtual re-staining is generally done by unsupervised methods, such as [4, 19, 22]. These approaches are all based on style transfer models for natural images. Researchers made some improvements according to the characteristics of pathological images, and finally achieved better results.

However, on the other hand, pathological images have their own characteristics. For example, the reliability of the results is more critical for this task due to the clinical significance of pathological examination. Meanwhile, the resolution of pathological images is usually higher than that of natural images, reaching $10k \times 10k$ or more. Therefore, virtual re-staining requires additional computational resources as the GPU memory is limited. Most of the existing virtual re-staining models solve this problem by splitting WSI (whole-slide imaging) images into smaller patches for training and inference, and then incorporating these patches into WSI images through post-processing. This results in dif-

*Zhineng Chen is the corresponding author.

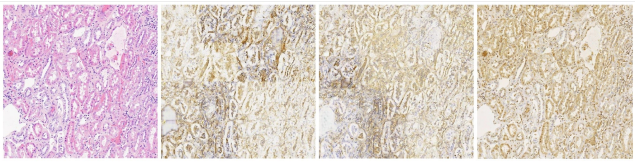

Figure 1. Illustration of the square effect. (a) a real 1600×1600 HE-stained image. (b) a virtually re-stained CK7 image obtained by merging separately re-stained 400×400 patches (using CycleGAN) without overlapping. (c) image obtained by merging 448×448 patches with an overlap of 64 pixels. (d) the result generated by our BFF-GAN.

ferences in color and brightness between adjacent patches, which we call the square effect. As shown in Fig. 1, (a) is a real HE image, (b) and (c) are CK7 images re-stained by CycleGAN without and with an overlap. Generally, overlapping is used to solve this problem, but we can see that the square effect always exists no matter whether there is an overlap. Meanwhile, the result of our BFF-GAN is shown in (d), and it is not easy to find the square effect in it.

Indeed, the square effect exists because patch-based virtual re-staining lacks global information, resulting in mismatches in hue, contrast and brightness between adjacent patches, especially for the regions with different tissue structures and the boundary regions. In addition, since the re-staining of each patch is independent, there may also be color differences between the re-staining results of patches with similar tissue structures. Most existing studies did not consider the global information, leading to serious square effect. To solve this problem, [18] proposed perceptual embedding consistency (PEC) loss, which forces the generator to learn features like color, brightness and contrast. But it is hard to say only using the PEC loss on the patch level can solve this problem to what extent. Subsequently, URUST [11] attempted to correct the mean value and standard deviation of the central patch with those of the surrounding patches. However, the parameters of this method are artificially designed and may not generalize well.

In the natural image domain, some researchers have attempted to get improvement through context aggregation. For example, PSPNet [39] improved the performance of semantic segmentation by increasing the receptive field with pooling kernels of different sizes. HRNet [32] designed parallel branches with different resolutions and integrated feature maps between different branches, achieving impressive performance in a number of visual tasks. GLNet [6] combined feature maps of the entire image with those of patches to improve segmentation performance of high-resolution images. These methods obtained nulti-scale information through feature fusion, and worked well on multiple tasks.

Based on these observations, in this paper, we propose a model that combines global-local features to learn the relationship between patches to solve the square effect, and meanwhile, bypasses the memory constraint for ultra-high resolution images. We design an architecture that consists of a global branch and a local branch, where the former takes the down-sampled whole images as input, and the latter takes the patch-level images coming in batches as input. The two branches perform feature fusion in both directions in the encoder and use patch-wise attention and skip connections to enhance feature expression capability. Finally, we fuse the features of the two branches to output the re-staining results. To verify the effectiveness of the method, extensive experiments were conducted on the private dataset RCC and the public dataset ANHIR [3]. The results show that our model achieves good performance on a variety of metrics, not only significantly eliminating the square effect, but also being generalizable to various datasets. Meanwhile, subjective experiments have also demonstrated the clinical significance of our model. In summary, our main contributions are listed as follows:

- The square effect significantly influences the quality of the virtually re-stained images. Thus, we propose to solve the square effect through the fusion of global and local features. Such an idea can be used in various networks, not only CycleGAN, but also other more advanced style transfer models.

- We propose a model with feature fusion between two branches called BFF-GAN to learn the inter-patch relations. To our knowledge, it is the first network for style transfer of ultra-high resolution images.

- Our proposed BFF-GAN achieves impressive results. It is of great clinical significance and can be generalized to various datasets.

2. Related work

2.1. Style transfer for natural images

Generative adversarial networks (GANs) [9] are commonly used for style transfer. After the original GAN, a conditional GAN (cGAN) [24] is proposed, which adds conditional information to the input so that the generated results can be controlled. Then, Philip Isola et al. proposed supervised style transfer models pix2pix [14] and pix2pixHD [33] based on cGAN. However, due to the lack of paired datasets, more subsequent works chose unsupervised approaches. These methods can be broadly classified into two types, one of which uses only a pair of generator and discriminator, improving performance by proposing various constraints. For example, distanceGAN [2] considered that the distance between the generated images was highly correlated with the distance between the source images. In [27], Alec Radford et al. considered that simple

geometric transformation would not change the semantic information of the image. Therefore, a geometric consistency loss was proposed. CUT [26] improved the performance by combining contrastive learning with style transfer. The other type is to use a dual model to train two symmetric generators with a cycle consistency loss. The earliest dual models included CycleGAN [41], discoGAN [16] and dualGAN [35]. For example, CycleGAN imposes constraints to various unpaired data in the task of style transfer through symmetric generators, thus has greatly inspired many other researchers. Therefore, a lot of work after that was still based on the dual model of CycleGAN. For example, NICE-GAN [5] used a part of the discriminator as the encoder to enhance the ability of the encoder; UGATIT [15] added AdaIN [12] and attention modules to the model. Other works such as ACLGAN [40] and so on were also improvements of the dual model.

2.2. GAN in virtual re-staining

Several methods of virtual re-staining have been proposed to ease the pathological examination. In [1, 28], cGAN was used to achieve virtual staining of unstained and HE-stained images. Rivenson et al. [29] used a supervised UNet-like generator to virtually stain HE images from unstained ones. Haan et al. [8] used paired datasets to transfer HE images to PAS, MT and JMS images. However, in most cases, paired data is not available, therefore most of the researches of virtual re-staining usually use unsupervised models with changes in loss function and model structure suitable for pathological images. For example, Lahiani et al. [17] implemented unsupervised virtual re-staining of Ki67-CD8 to FAP-CK using a Cyclegan-based approach. Li et al. [19] proposed a saliency constraint loss to constrain the model for better results. Liu et al. [22] added a pathological representation network to dig the pathological representation heatmap of the input image and at the same time introduced a cycle structural consistency loss. And Boyd et al. [4] proposed a region of interest discriminator to replace the patch-based discriminator.

In general, pathological images have ultra-high resolution, which makes GPU memory the bottleneck for virtual re-staining tasks. The aforementioned studies split pathological images into small patches, train the model and perform inference at the patch level, then combine the obtained patches to get the final result. Although the square effect can be reduced to some extent through overlapping, this approach still results in obvious boundaries between patches. Several studies have attempted to address these issues. Lahani et al. [18] proposed a perceptual embedding consistency (PEC) loss to force the generator to learn features such as color, brightness and contrast, but the use of PEC loss at patch level alone does not fully resolve the squared effect. After that, URUST [11] solves the problem by averaging the mean value and standard deviation of the surrounding patches. However, the parameters of this method are artificially designed, so it may be less generalizable.

2.3. Context aggregation

In the CNN, the shallow network contains more geometric information, and the deep network contains more semantic information. Images of different resolutions and receptive fields of different sizes can also obtain different features. Thus, the idea of multi-stage, multi-scale and context aggregation are widely used. For example, [13, 30, 37] etc. achieved better performance with multi-stage training and inference. Many approaches also improved the performance through multi-scale fusion. ICNet [38] introduced cascaded feature fusion units, which merged multi-resolution branches; RefineNet [20] proposed refine blocks, fusing feature maps of different resolution levels; PSPNet [39] used pooling kernels of different sizes to increase the receptive field; FPN [21] fused deep and shallow features; HRNet [32] designed parallel branches of different resolutions and did feature fusion between those branches. Moreover, context aggregation was also commonly used in frontier research. ParseNet [23] used global context to achieve semantic segmentation; BiSeNet [36] preserved spatial information and generated high-resolution features with the spatial path, and a context path was designed to obtain sufficient receptive field. GLNet [6] chose to fuse global features and local features of high-resolution images, whose idea was truly instructive and also used in MedT [31].

3. Method

3.1. Base model

Inspired by CycleGAN, we propose a new bi-directional feature fusion generative adversarial network (BFF-GAN), which includes a global branch (G) and a local branch (L), as shown in Fig.2. Given a $H \times W$ image X, for the global branch, X is down-sampled to X_g as input, and the resolution of X_g is $h \times w$. For the local branch, X is divided into patches $x_0, ..., x_{\frac{H}{h'} \times \frac{W}{w'}}$ whose sizes are $h' \times w'$. The global branch and the local branch have the same basic structure. The encoder contains three convolutional blocks, each of which contains a convolutional layer, an instance norm layer and a ReLU. The kernel of the first convolutional block is 7, while the others are 3. Each convolutional block is followed by a patch-wise attention module (PAM), which will be described in detail in Section 3.3. After the encoder, there are 6 resblocks [10], and then the decoder contains two up-sampling blocks, in which there are a 3×3 transposed convolution layer, an instance norm layer and a ReLU. The output head contains a 7×7 convolutional layer and a tanh function. Besides, skip connections are added between the encoder and the decoder.

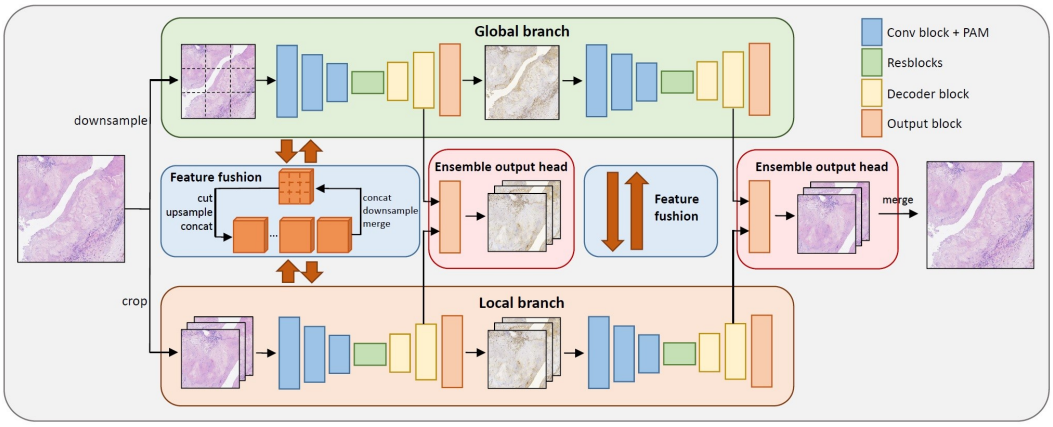

Figure 2. Overview of BFF-GAN, which consists of two branches, i.e. the global branch and the local branch. Through bi-directional feature fusion between these two branches, the model could get inter-patch information and solve the problem of the square effect.

BFF-GAN has two heads for auxiliary and the final output is obtained with an ensemble output head. It has the same structure as the auxiliary heads, but fuses feature maps from skip connections and feature maps of the global and the local branches.

3.2. Feature fushion

In the encoder, the feature fusion of two directions (G→L, L→G) is done after each patch-wise attention module (PAM). The feature maps of the global branch are cut and up-sampled according to the corresponding positions and sizes of each patch in the local branch, and then concatenated with the feature maps of the local branch. In this way, each time the feature fusion in the G→L direction is done, the local branch will obtain more global information. In addition, the feature maps of the local branch are combined together, then down-sampled and concatenated with the feature maps of the global branch. The parameters of the global branch are continuously optimized by feature fusion in the L→G direction, and the local branch is affected in the next fusion step in the G→L direction. The two branches then complete the process of feature fusion, making the local branch learn the information of global features, and the global branch learn the information of local features.

Our model fuses features three times in the encoder stage, covering from shallow to deep features, so the local branch can fully consider the global information. Of course, we could apply this feature fusion to each stage of the model if we would like to, but this would consume more computational resources and just three feature fusion steps at the encoder stage are sufficient to solve the problem indeed.

3.3. Patch-wise attention

To help the model focus on important channel, spatial and patch-wise information, we propose a patch-wise attention module based on CBAM [34]. PAM contains channel, spatial and patch-wise attention parts. The channel and spatial attention adopt a similar approach to CBAM. But due to the parameter redundancy of fully connected layer, we replace it with a convolutional layer.

In BFF-GAN, the information of multiple patches exists at the same time. Thus, to obtain the information of the importance between patches, we add patch-wise attention to the module. In this part, given the input $F_s \in \mathbb{R}^{N\times C\times H\times W}$, the patch-wise attention feature F_{pa}, $F_{pm} \in \mathbb{R}^{N\times 1\times 1\times 1}$ are obtained through average and max pooling, then they are concatenated and passed to a convolutional block and a sigmoid function.

In PAM, spatial and patch-wise attention can be chosen to be used or not depending on the needs of the model. In BFF-GAN, we choose not to use patch-wise attention in the global branch, since its feature map is a singleton.

3.4. Loss function

The loss function includes the global, the local, and the ensemble parts. The global part computes the losses between down-sampled original, re-stained (e.g. CK7) and reconstructed (e.g. recovered HE) images. For local and ensemble parts, they are both designed to contain two parts. The first part computes the losses between real, re-stained and reconstructed patches, while the second one computes the loss between images merged together. Thus, the proposed loss function actually contains five parts.

Thus, the total loss function is shown in Eq.1. L_g is the loss of the global head. L_{l_p} and L_{e_p} are the loss of patches of the local and ensemble head. L_{l_m} and L_{e_m} are the loss of merged images of these two heads.

$$L_{=}L_g + \alpha \cdot (L_{l\_p} + L_{l\_m} + L_{e\_p} + L_{e\_m}) \qquad (1)$$

Here α is a hyperparameter empirically set to 5.

4. Experiment

4.1. Implementation details

In the experiments, we used two datasets: the private dataset RCC and the public dataset ANHIR. RCC contains pathological images of renal cell carcinoma, including two staining types: HE and CK7. ANHIR [3] was proposed for pathological image registration task. It contains pathological images from a variety of tissues, including kidney, breast, lung, stomach, and so on. In fact, some of its sub-datasets are also suitable for virtual re-staining task, so we selected two sub-datasets of ANHIR: breast and lung lesion for experiments. Among them, the staining types of the breast dataset are HE and PR, and the staining types of the lung lesion dataset are HE and Ki67.

Due to the lack of WSI images in the dataset, this paper increased the size of the dataset by splitting the WSI images, and made training and test sets come from different WSI images. For RCC, we used 120 and 50 1600×1600 images as the training and test sets. Our main experiments were performed at $10\times$, but also at $20\times$ and $40\times$. For ANHIR, the resolution of the images used for training and inference was also 1600×1600. The size of the training and test set of the breast set was 72 and 22. As for lung lesion set, because the number of WSI images in it is too few, its training set has to be also used as test set, and the number of images in the training and test set was 68. Of course, it is also feasible to use this model on images larger than 1600×1600.

In our model, the patch size of the local branch and the input size of the global branch was both 448×448, and there was an overlap of 64 pixels between patches. The experiment used a constant learning rate for 70 epochs and a linear-decay learning rate for the remaining 70 epochs.

4.2. RCC

4.2.1 Comparison with other methods

As shown in Fig. 3, we compared our method with some representative style transfer and virtual re-staining methods. We use black boxes to indicate the areas with the square effect. Among these models, most of them could identify the tissue area and the background area except CycleGAN [41], but an obvious square effect could be seen. Besides, some of the generated CK7 images were quite different from real CK7 images. But our method has a significant advantage in visual perception. It eliminate the square effect considerably. Meanwhile, its generated images are quite realistic.

Fig. 4 is an example of a 17580×15798 WSI image generated by our model. It shows that we could obtain a virtual re-staining result with almost no square effect. In fact, we got Fig. 4 through a simple post-processing, merging 1600×1600 images re-stained by BFF-GAN with an overlap of 200 pixels. Due to the insufficiency of data, we had not been able to train on images of larger size, and certainly could not inference directly. However, in general, the cause of the square effect makes the phenomenon more likely to occur in areas with changes in tissue structure and in the boundaries. For smaller patches such as the ones of 448×448 pixels, due to their extremely tiny coverage, the numbers of patches cross the tissue areas or boundaries are usually small. What is more likely is that adjacent patches do not belong to the same type of tissue or the same side of boundaries, resulting in a fairly obvious square effect. Such a problem of the square effect that is easy to appear on adjacent small scale patches can be reduced by BFF-GAN with image size of 1600×1600 pixels. Thus, the square effect would be subtle and inconspicuous in larger scales. So here

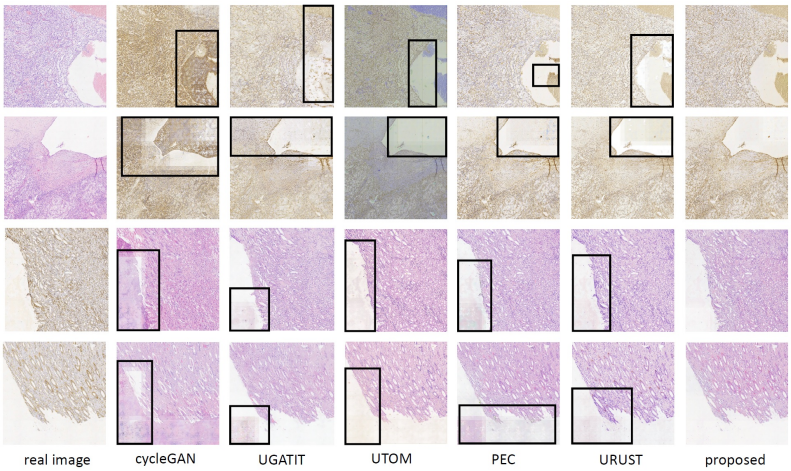

Figure 3. The virtual re-staining result on 10x RCC. The first column is real images in the dataset, following with generated images of different models. The black boxes indicate the areas with square effect. Our model achieves the best result.

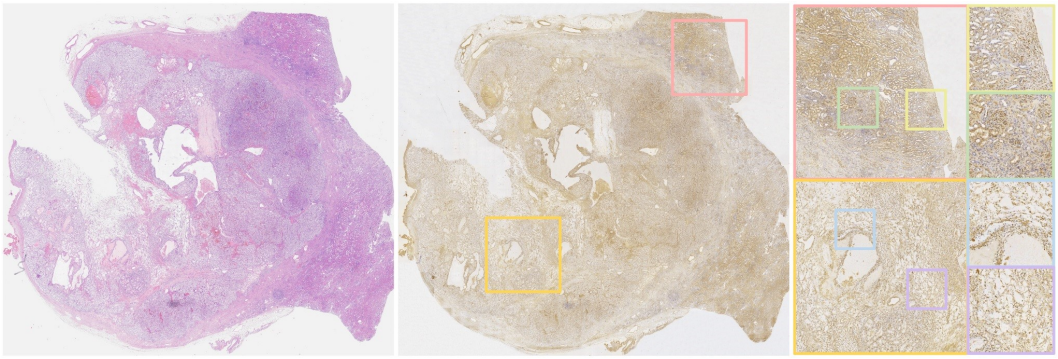

Figure 4. An example of virtually re-stained WSI image and its zoom-in views at different scales. It is generated by merging separately re-stained 1600 × 1600 images with 200 pixels' overlap. The square effect is reduced a lot when images of a larger resolution are considered.

Table 1. Quantitative results on 10×, 20× and 40× RCC, compared with five other models.

Model	10x							20x							40x						
	SSIM↑	PSNR↑	CSS↑	FID↓	$SSIM_p$ ↑	$PSNR_p$ ↑	CSS_p ↑	SSIM↑	PSNR↑	CSS↑	FID↓	$SSIM_p$ ↑	$PSNR_p$ ↑	CSS_p ↑	SSIM↑	PSNR↑	CSS↑	FID↓	$SSIM_p$ ↑	$PSNR_p$ ↑	CSS_p ↑
CycleGAN [41]	93.02	30.46	25.39	148.89	92.95	30.82	25.18	96.01	35.09	11.13	165.34	95.97	35.29	11.02	97.49	36.20	28.34	168.70	97.50	36.50	22.06
UGATIT [15]	91.98	30.26	67.80	**134.63**	91.90	30.46	67.31	96.06	34.50	71.82	148.56	95.98	34.49	71.39	96.38	34.53	76.92	**152.10**	96.30	34.53	76.63
UTOM [19]	**99.65**	**44.31**	66.96	272.09	**99.66**	**44.63**	66.15	**99.84**	**50.13**	42.07	335.70	**99.90**	**54.70**	38.52	**99.80**	**49.39**	50.49	371.99	**99.84**	**52.13**	47.01
PEC [18]	89.58	28.90	79.80	142.47	89.37	29.03	79.70	94.26	32.67	82.82	**144.22**	94.20	32.91	82.81	97.27	34.47	87.17	156.11	97.30	34.90	87.17
URUST [11]	86.97	25.45	79.70	149.79	86.89	25.63	79.69	89.10	25.20	84.69	158.88	89.01	25.36	84.71	91.08	27.28	86.58	179.38	91.02	27.53	86.63
BFF-GAN	95.63	32.79	**85.05**	142.14	95.58	32.81	**84.94**	96.48	35.63	**86.05**	147.39	96.41	35.64	**85.96**	97.17	36.29	**90.30**	155.15	97.12	36.41	**90.30**

we choose to solve the possible color differences between the larger images through such a simple post-processing.

To quantitatively evaluate the performance of the models, we calculated four metrics: SSIM, PSNR, CSS and FID. Among them, SSIM and PSNR are calculated between real and reconstructed images in the same domain (e.g. HE), indicating whether the model has the ability to preserve the structural details. However, as they only consider the reconstructed images, they cannot evaluate the quality of the generated re-stained images indeed. Therefore, we also need other metrics such as FID and CSS to evaluate the performance in other aspects complementarily. FID is calculated between the real and re-stained images in the target domain (e.g. CK7), measuring the similarity of style and indicating whether the model can achieve more realistic virtual re-staining. CSS was proposed in PC-stain GAN [22]. It is inspired by SSIM, but removes the brightness part and only compares the contrast and structure, thus claimed suitable for evaluating the similarity between real images in the source domain and generated re-stained images in the target domain. Besides, we also calculated SSIM, PSNR and CSS at patch level to describe the differences between the two kinds of images more thoroughly.

As shown in Tab. 1, The performance of our BFF-GAN is relatively balanced and the results on all metrics are good. UTOM [19] has achieved good performance on SSIM and PSNR, but its FID is far higher than other models, which is also consistent with the visual results in Fig. 3. Among other models, BFF-GAN achieves the highest scores in most cases. Its CSS is always the highest, indicating its ability to preserve structural cross-staining details, while SSIM is only slightly lower than CycleGAN at 40×. Meanwhile, though its FID is slightly higher than PEC [18] at 20× and UGATIT [15] at 10× and 40×, but still at a relatively low level. What is more important is that though these methods have a lower FID, our method has a better ability to eliminate the square effect. Compared with the most recent URUST [11], BFF-GAN also has an obvious advantage, which is largely attributed to the more global information included in our model. Its SSIM is about 8.66% higher, PSNR is about 7.34 higher, CSS is about 5.35% higher, and FID is about 7.65 lower than URUST on 10x RCC. The metrics at 20× and 40× also exceed URUST a lot.

Table 2. The result of the validation experiment. The table shows the correct rate of the two pathologists to find out 100 real images apart from 100 fake images.

	P1		P2	
	merged image	patch	merged image	patch
HE	63.5	61.5	64.0	60.5
CK7	68.5	63.5	65.5	62.5

4.2.2 Subjective Experiment

To further demonstrate the clinical value of our model, we invited two pathologists, a senior expert (P1) and a junior expert (P2), to do subjective experiments, including validation and classification experiments as follows.

In the validation experiment, we gave the pathologists

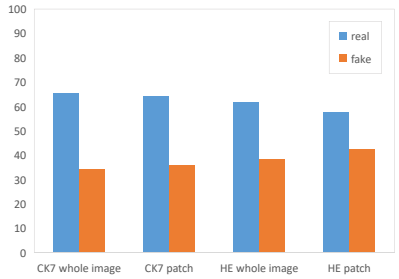

Figure 5. The proportion of real and fake images among the images considered real.

Table 3. The result of the classification experiment. It shows that the reconstructed images are close to the real images.

	real HE	rec HE	real CK7	rec CK7
ResNet50	96.05	90.95	92.01	91.86
P1	95.05	94.06	96.04	96.04
P2	95.05	95.05	94.06	92.08

HE and CK7 images with sizes of both 1600 and 448 pixels. In every single experiment, they would randomly get 100 real and 100 fake images. Pathologists had to choose the images they thought were real. We calculated the correct rate of them, as shown in Tab. 2. No matter in the merged image or patch level, their correct rates are not very high, just at the level of 60-70%. The correct rates of merged images are always higher than that of patches. In fact, this is quite reasonable. In general, the information included in merged images is always more than that included in patches. Therefore, it is easier for pathologists to make an accurate judgement while provided merged images. On the other hand, the correct rate of CK7 images is always higher than HE images. HE staining is the most common staining method, but CK7 staining is an IHC staining method used for specific cancer types which stains specific antibodies. Generally, it is more difficult to capture the specific antibodies than cellular structures. Therefore, obtaining realistic CK7 images through virtual re-staining is usually harder than obtaining HE images. Therefore, it is easier for pathologists to judge whether a CK7 image is true or virtually re-stained, making the result of CK7 images higher. Fig. 5 shows the proportion of real and fake images considered real. This also shows that the images generated by BFF-GAN can deceive experienced pathologists to a certain extent, which means our method has a considerable clinical significance.

In the classification experiment, we asked pathologists to classify the real and reconstructed images into two categories: cancer and non-cancer. We also trained a ResNet50 classifier for comparison. As shown in Tab. 3, the classification accuracy of the reconstructed images of the ResNet50 classifier is lower than that of the real images on the HE domain. The convolutional classifier always focuses more on subtle details which would not be considered by pathologists but are likely to be more different. What is more important is that the accuracy of the two pathologists on the real and the reconstructed images is close, though the accuracy of reconstructed images is a little lower than that of real images. This indicates that the possibility for the pathologists to give different diagnoses according the the real and reconstructed images is low, and the reconstructed images have similar clinical usage with real images.

4.2.3 Ablation study

Table 4. An ablation study of overlap. It indicates that overlapping only affect the result slightly.

	SSIM↑	PSNR↑	CSS↑	FID↓
0 pix overlap	95.02	32.19	84.60	143.07
128 pix overlap	95.24	31.25	84.36	**141.07**
proposed (64 pix overlap)	**95.63**	**32.79**	**85.05**	142.14

Table 5. The results of the ablation study of attention blocks in PAM, indicating the necessity of all the attention blocks.

channel	spatial	patch-wise	SSIM↑	PSNR↑	CSS↑	FID↓
✓	✓		94.31	31.35	78.30	148.49
		✓	94.17	31.60	69.19	147.98
✓	✓	✓	**95.63**	**32.79**	**85.05**	**142.14**

Table 6. The metrics of our model with different structural configurations. All the parts listed are useful for a better result.

G→L	L→G	sc	PAM	SSIM↑	PSNR↑	CSS↑	FID↓
	✓	✓	✓	94.43	31.75	83.40	**140.00**
✓		✓	✓	95.50	32.55	83.64	144.07
✓	✓		✓	91.62	29.96	75.78	147.05
✓	✓	✓		93.95	30.39	79.47	156.16
✓	✓	✓	✓	**95.63**	**32.79**	**85.05**	142.14

As shown in Tab. 4, we did experiments on different overlaps. The gap on the performance with or without overlap is really small, and there is no significant improvement when the overlap is larger. So in other experiments in this paper, we chose a 64 pixels' overlap.

Tab. 5 shows the ablation experiments on the attention blocks in PAM. PAM includes three attention blocks: channel, spatial and patch-wise attention block. When there is only channel and spatial block or only patch-wise block, the metrics are always lower than BFF-GAN with all the blocks available. Such a result prove that all these blocks can improve the performance to some extent.

We also designed ablation experiments on the network structure, and the results are recorded in Tab. 6. Experiments show that single-direction feature fusion and removing the skip connections (sc) plus PAM will all lead to a decline in performance. All of these structures play important roles in the model. The only exception is that when there is

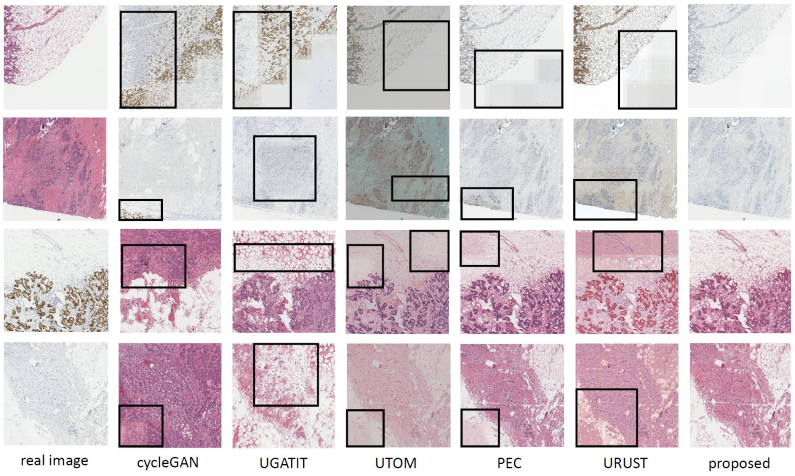

Figure 6. The virtual re-staining results obtained on the breast dataset of ANHIR. The first column is real images, followed by generated images of different models. The black boxes indicate the areas with square effect.

Table 7. Quantitative results on breast and lung lesion subset of ANHIR, which indicates the generalization of our model.

Model	breast							lung lesion						
	SSIM↑	PSNR↑	CSS↑	FID↓	$SSIM_p$ ↑	$PSNR_p$ ↑	CSS_p ↑	SSIM↑	PSNR↑	CSS↑	FID↓	$SSIM_p$ ↑	$PSNR_p$ ↑	CSS_p ↑
CycleGAN [41]	90.20	26.69	42.61	202.35	90.14	27.32	42.62	94.97	34.67	28.20	124.15	94.97	34.94	28.51
UGATIT [15]	78.21	21.33	55.77	204.11	78.19	22.58	55.76	90.14	27.25	76.63	92.63	89.90	28.50	76.31
UTOM [19]	**98.41**	**35.83**	72.28	282.48	**98.37**	**36.49**	70.07	**98.65**	**41.65**	93.04	127.04	**98.67**	**42.12**	92.93
PEC [18]	91.07	26.93	83.84	248.37	91.23	27.71	83.91	95.28	33.26	93.22	**83.16**	95.27	33.77	93.27
URUST [11]	92.07	25.73	86.27	257.22	92.22	26.57	86.30	91.90	27.30	93.17	106.33	91.91	27.59	93.19
BFF-GAN	95.31	28.88	**87.59**	**140.68**	95.20	28.98	**87.56**	96.22	35.41	**93.56**	85.47	96.15	35.42	**93.50**

no feature fusion from the global to the local branch, FID is a little lower compared with our model. However, in this situation, its SSIM, PSNR and CSS have a larger drop.

4.3. ANHIR

To demonstrate the generalization of BFF-GAN, we also did experiments on ANHIR without tuning hyperparameters. Fig. 6 shows the result on the breast dataset. The result of the lung lesion dataset is illustrated in the supplementary materials. Just like on RCC, CycleGAN still cannot identify the foreground and the background area, and the square effect inevitably can be seen in the results of other models, including PEC and URUST which are claimed to be able to eliminate the square effect. However, our model can still reduce the square effect well and generate the most realistic images, far ahead of other models in visual perception.

Tab. 7 records the results of quantitative experiments on ANHIR. As on RCC, though UTOM has a higher SSIM and PSNR, our BFF-GAN achieves a more competitive visual results with relatively good metrics. The CSS of BFF-GAN is also the highest on ANHIR, indicating that BFF-GAN well preserves structural details. Among the rest results, our model has certain advantages in various scores, only lower than PEC in FID on lung lesion dataset, but is still the top two. Compared to URUST, our model is always better than it. Our SSIM, PSNR and CSS are relatively 3.24%, 3.15 and 1.32% higher on breast, and 5.13%, 8.11 and 0.39% higher on lung lesion. Meanwhile, FID is 116.54 and 20.86 lower than URUST. All of these prove that our model is the best one in both visual perception and quantitative experiments, and the impressive results on ANHIR demonstrate that our model has the ability to generalize to different datasets.

5. Conclusion

The square effect of the virtual re-staining of ultra-high resolution pathological images has been the pain point so far. To address this problem, we creatively adopted the idea of context aggregation in the natural image field with a simple CycleGAN, proposing BFF-GAN to fuse the global and local features. Meanwhile, we added patch-wise attention to the model to strengthen its ability to learn connections between patches. Extensive experiments were performed on the private dataset RCC and the public dataset ANHIR. We evaluated the performance of the model through quantitative metrics and subjective experiments. The results show that our model surpasses most of the models and gets impressive results. Moreover, the generated images can deceive experienced pathologists to a certain extent, proving the clinical significance of the proposed BFF-GAN.

Acknowledgments The work is supported by the National Natural Science Foundation of China (No. 62172103, 62032006, 62076007), and Beijing Natural Science Foundation (7232132).

References

[1] Neslihan Bayramoglu, Mika Kaakinen, Lauri Eklund, and Janne Heikkila. Towards virtual h&e staining of hyperspectral lung histology images using conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 64–71, 2017. 3

[2] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017. 2

[3] Jiří Borovec, Jan Kybic, Ignacio Arganda-Carreras, Dmitry V Sorokin, Gloria Bueno, Alexander V Khvostikov, Spyridon Bakas, I Eric, Chao Chang, Stefan Heldmann, et al. Anhir: automatic non-rigid histological image registration challenge. *IEEE transactions on medical imaging*, 39(10):3042–3052, 2020. 2, 5

[4] Joseph Boyd, Irène Villa, Marie-Christine Mathieu, Eric Deutsch, Nikos Paragios, Maria Vakalopoulou, and Stergios Christodoulidis. Region-guided cyclegans for stain transfer in whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 356–365. Springer, 2022. 1, 3

[5] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8168–8177, 2020. 3

[6] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8924–8933, 2019. 2, 3

[7] Zhineng Chen, Shuai Zhao, Kai Hu, Jing Han, Yuan Ji, Shaoping Ling, and Xieping Gao. A hierarchical and multi-view registration of serial histopathological images. *Pattern Recognition Letters*, 152:210–217, 2021. 1

[8] Kevin de Haan, Yijie Zhang, Jonathan E Zuckerman, Tairan Liu, Anthony E Sisk, Miguel FP Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, et al. Deep learning-based transformation of h&e stained tissues into special stains. *Nature communications*, 12(1):1–13, 2021. 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[11] Ming-Yang Ho, Min-Sheng Wu, and Che-Ming Wu. Ultra-high-resolution unpaired stain transformation via kernelized instance normalization. In *European Conference on Computer Vision*, pages 490–505. Springer, 2022. 2, 3, 6, 8

[12] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 3

[13] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16755–16764, 2021. 3

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[15] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 3, 6, 8

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 3

[17] Amal Lahiani, Jacob Gildenblat, Irina Klaman, Shadi Albarqouni, Nassir Navab, and Eldad Klaiman. Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach. In *European Congress on Digital Pathology*, pages 47–55. Springer, 2019. 3

[18] Amal Lahiani, Irina Klaman, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency. *IEEE Journal of Biomedical and Health Informatics*, 25(2):403–411, 2020. 2, 3, 6, 8

[19] Xinyang Li, Guoxun Zhang, Hui Qiao, Feng Bao, Yue Deng, Jiamin Wu, Yangfan He, Jingping Yun, Xing Lin, Hao Xie, et al. Unsupervised content-preserving transformation for optical microscopy. *Light: Science & Applications*, 10(1):1–11, 2021. 1, 3, 6, 8

[20] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[22] Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE Transactions on Medical Imaging*, 40(8):1977–1989, 2021. 1, 3, 6

[23] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 3

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[25] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019. 1

[26] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 3

[27] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[28] Aman Rana, Gregory Yauney, Alarice Lowe, and Pratik Shah. Computational histological staining and destaining of prostate core biopsy rgb images with generative adversarial neural networks. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 828–834. IEEE, 2018. 3

[29] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E Zuckerman, Thomas Chong, Anthony E Sisk, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature biomedical engineering*, 3(6):466–477, 2019. 3

[30] Shusuke Takahama, Yusuke Kurose, Yusuke Mukuta, Hiroyuki Abe, Masashi Fukayama, Akihiko Yoshizawa, Masanobu Kitagawa, and Tatsuya Harada. Multi-stage pathological image classification using semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10702–10711, 2019. 3

[31] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2021. 3

[32] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2, 3

[33] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[35] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 3

[36] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 3

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 3

[38] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018. 3

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 3

[40] Yihao Zhao, Ruihai Wu, and Hao Dong. Unpaired image-to-image translation using adversarial consistency loss. In *European Conference on Computer Vision*, pages 800–815. Springer, 2020. 3

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3, 5, 6, 8