arXiv:2208.10730v1 [eess.IV] 23 Aug 2022

Ultra-high-resolution unpaired stain transformation via Kernelized Instance Normalization

Ming-Yang Ho*, Min-Sheng Wu, and Che-Ming Wu

aetherAI, Taipei, Taiwan
{kaminyouho,vincentwu,uno}@aetherai.com
https://www.aetherai.com/

Abstract. While hematoxylin and eosin (H&E) is a standard staining procedure, immunohistochemistry (IHC) staining further serves as a diagnostic and prognostic method. However, acquiring special staining results requires substantial costs. Hence, we proposed a strategy for ultra-high-resolution unpaired image-to-image translation: Kernelized Instance Normalization (KIN), which preserves local information and successfully achieves seamless stain transformation with constant GPU memory usage. Given a patch, corresponding position, and a kernel, KIN computes local statistics using convolution operation. In addition, KIN can be easily plugged into most currently developed frameworks without re-training. We demonstrate that KIN achieves state-of-the-art stain transformation by replacing instance normalization (IN) layers with KIN layers in three popular frameworks and testing on two histopathological datasets. Furthermore, we manifest the generalizability of KIN with high-resolution natural images. Finally, human evaluation and several objective metrics are used to compare the performance of different approaches. Overall, this is the first successful study for the ultra-high-resolution unpaired image-to-image translation with constant space complexity. Code is available at: https://github.com/Kaminyou/URUST.

Keywords: Unpaired image-to-image translation, ultra-high-resolution, stain transformation, whole slide image

1 Introduction

Histological staining, highlighting cellular components with dyes, is crucial in clinical diagnosis [1], which enables visualization of cells and extracellular matrix and abnormal identification. Since specific cellular components or biomarkers can be distinguished when particular dyes attach specific molecules in tissues, different staining methods are applied to diagnose various diseases and their subtypes [15,11,4]. The standard stain (or routine stain) is hematoxylin and eosin (H&E). While hematoxylin stains nuclei, eosin can stain cytoplasm. Immunohistochemistry (IHC) protocol is further developed to detect the presence of specific

* Corresponding author

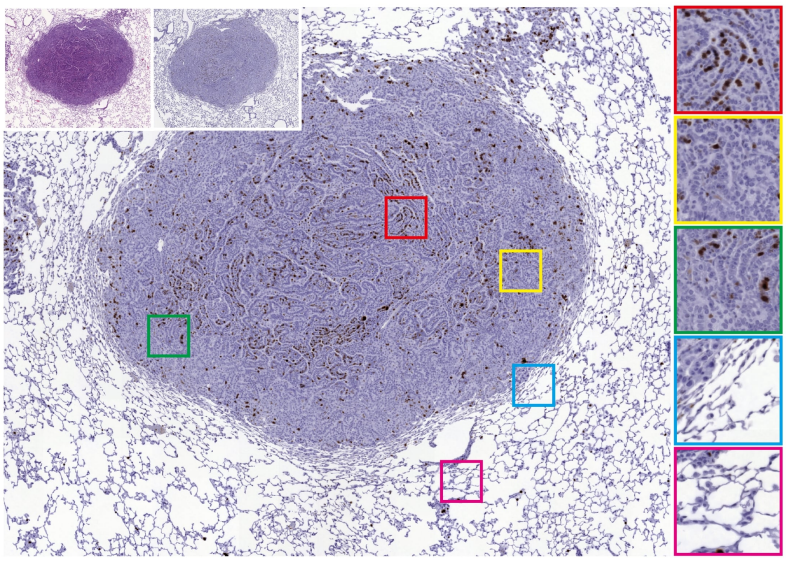


Fig. 1. An ultra-high-resolution translated result ($7,328 \times 8,899$ pixels) from our Kernelized Instance Normalization (KIN). The whole slide image (WSI) was translated from source stain to target stain (on the upper left) with constant space complexity (GPU memory) via KIN, and local appearance was preserved. On the right side, five close-ups demonstrate the detail.

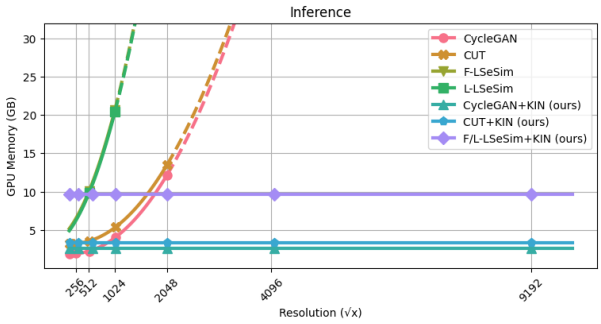


Fig. 2. Comparison of GPU memory usage among different unpaired image-to-image translation approaches. Compared with the models using Instance Normalization (IN), which has limitation (marked by the dashed line) on a 32G GPU (NVIDIA V-100), our Kernelized Instance Normalization (KIN) approach can translate an ultra-high-resolution image with constant GPU memory usage (less than 5GB).

protein markers. For example, Ki-67 and ER staining can quantify the presence of Ki-67 and ER biomarkers, respectively. In clinical practice, high Ki-67 expression is considered a poor prognostic factor [23], while the presence of ER

indicates the suitability of choosing specific target therapies that benefit related disease subtypes [26].

However, compared with H&E staining, the IHC staining process is much more expensive and requires extra biopsies, which are limited materials. With the development of deep learning-based image-to-image translation, virtually translating H&E into different IHC staining can be achieved. For example, de Haan *et al.* . proposed a supervised deep learning approach via CycleGAN [33] to transform stained images from H&E to Masson's Trichrome (MT) staining [13]. While supervised pair-wise training is desirable, this approach requires perfectly paired staining images, necessitating de-staining and re-staining processes, and is not practically efficient in a clinical scenario. Most datasets are composed of unpaired H&E and IHC images from consecutive sections. Several methodologies have been proposed and successfully tackled the unpaired image-to-image translation problem [27]. Regardless of the astonishing performance, the existing methods are limited to low-resolution images and rarely explore the images with ultra-high-resolution. In histopathology, whole slide images (WSIs) are usually larger than $10,000 \times 10,000$ pixels. The main challenge of transforming a WSI is the limitation of GPU memory capacity. Patch-wise inference with an assembly process can tackle ultra-high-resolution image-to-image translation, but tiling artifacts between adjacent patches would be a critical problem. Traditionally, overlapping windows have been leveraged to smooth the transitions but have limited effectiveness. Considering the mean and standard deviation calculated in instance normalization (IN) layers might influence hue and contrast, recently, Chen *et al.* developed Thumbnail Instance Normalization (TIN) for both ultra-high-resolution style transfer as well as image-to-image translation tasks [8]. Unfortunately, while their approach could overcome the resolution limitation, their erroneous assumption that all patches share global mean and standard deviation would lead to dramatically over/under-colorizing according to our comprehensive experiment, which is confirmed in Section 5.3.

To compensate for all the above limitations, we proposed a Kernelized Instance Normalization (KIN) layer that can replace the original IN layer during the inference process without re-training the models. With the help of KIN, images with arbitrary resolution can be translated with constant GPU memory space (as demonstrated in Fig. 1 and 2). Moreover, utilizing the statistics of neighboring patches instead of global ones like TIN, our approach can further preserve the hue and contrast locally, which is especially paramount in stain transformation tasks. Besides the translation of H&E to four IHC staining, we additionally demonstrated the generalizability of KIN with natural images by translating summer to autumn style. Our novel contribution can be summarized as follows:

- To the best of our knowledge, this is the first successful study for the ultra-high-resolution unpaired image-to-image translation with constant space complexity (GPU memory), which manifests state-of-the-art outcomes in stain transformation and can also be generalized to natural images.

- Without re-training the models, our KIN module can be seamlessly inserted into most currently developed frameworks that have IN layers, such as CycleGAN [33], CUT [28], and LSeSim [32].
- With the KIN module, local contrast and hue information in translated images can be well preserved. Besides, different kernels can be further applied to subtly adjust the translated images.

2 Related works

2.1 Unpaired image-to-image translation

Several frameworks have been proposed for unpaired image-to-image translation. CycleGAN [33], DiscoGAN [19], and DualGAN [31] were first presented to overcome the supervised pairing constraint via cycle consistency. However, subtle information was forced to be retained in the translated image to achieve better reconstruction, causing detrimental effects when two domains are substantially different such as dog-to-cat translation. Besides, a reverse mapping function might not always exist, which inevitably leads to artifacts in translated images. Recently, strategies beyond cyclic loss have been developed to reach one-sided unpaired image-to-image translation. While DistanceGAN [3] enforced distance consistency between different parts of the same sample in each domain, CUT [28] leveraged contrastive loss to maximize the patch-wise similarity between domains and achieved remarkable results. LSeSim [32] further utilized spatially correlation to maximize structural similarity and eliminate the domain-specific features.

2.2 Image-to-image translation for stain transformation

Transforming one stained tissue into another specific stain will dramatically save laboratory resources and money. Hence, growing research has leveraged unsupervised image-to-image translation to conduct stain transformation in several medical scenarios. Levy *et al.* . translated H&E to trichrome staining via CycleGAN for liver fibrosis staging [22]. Kapil *et al.* translated Cytokeratin to PD-L1 staining to bypass re-staining for segmentation [18]. de Haan *et al.* translated H&E to Masson's Trichrome, periodic acid-Schiff, and Jones silver stain for improving preliminary diagnosis for kidney diseases via CycleGAN with perfectly paired images [13]. Lahiani *et al.* further broke the limitation of 256×256-pixel image patches by applying perceptual embedding consistency for H&E to FAP-CK transformation [21]. However, the lack of detailed description hampers the implementation of their methodology.

2.3 Ultra-high resolution image-to-image translation

Ultra-high-resolution images are ubiquitous in photography, artwork, posters, ultra-high (*e.g.*, 8K) videos, and especially, WSIs in digital pathology (usually,

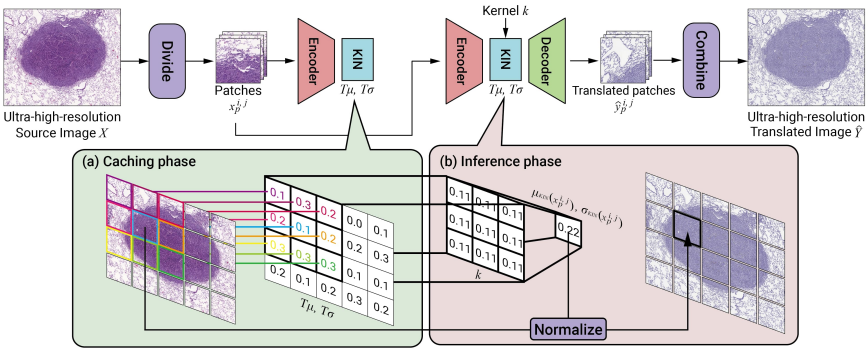


Fig. 3. Overall framework of proposed method. An ultra-high-resolution H&E image is passed through the caching phase and inference phase to be translated into an ultra-high-resolution IHC image. (a) In the caching phase, all mean μ and standard deviation σ values of patches will be cached in caching tables T_μ and T_σ by the Kernelized Instance Normalization (KIN) layer. (b) In the inference phase, a kernel k will convolute the caching table to compute μ_{KIN} and σ_{KIN} for instance normalization. Taking the neighboring statistics into account, our method can preserve local appearance.

larger than $10,000 \times 10,000$). Due to the massive computational costs, transforming these images into different styles will be difficult. Traditionally, strategies have been proposed to address the problem of tiling artifacts created by patch-wise-based methods, including utilizing a larger overlapping window [20] or freezing IN layers during testing time [2]. By providing a patch-wise style transfer network with Thumbnail Instance Normalization (TIN), Chen *et al.* performed ultra-high-resolution style transformation with constant GPU memory usage [8]. Also, they applied their framework to an image-to-image translation task. However, according to our experiment, TIN may result in over/under-colorizing for their fallacious assumption that all patches can be normalized with the same global mean and standard deviation.

3 Proposed method

3.1 Overall Framework

Our framework targets one-sidedly translating an ultra-high-resolution image X in domain $\mathcal{X}$ (e.g., H&E stain domain) into image $\hat{Y}$ in domain $\mathcal{Y}$ (e.g., IHC domain), in which $X, \hat{Y} \in R^{H \times W \times C}$, H and W are the height and width of X, via a mapping function, generator $\mathcal{G}$.

$$\hat{Y} = \mathcal{G}(X), \mathcal{G} : \mathcal{X} \rightarrow \mathcal{Y} \tag{1}$$

Collections of unpaired $\mathbf{X}$ in $\mathcal{X}$ and $\mathbf{Y}$ in $\mathcal{Y}$ would be first cropped into patches with the size of 512×512 pixels to train a generator $\mathcal{G}$.

As our KIN module is only applied during the testing time and can be inserted into any framework with IN layers, we followed the original training process and hyperparameters proposed in the paper of CycleGAN [33], CUT [28], and LSeSim [32] to train the corresponding generators with their specific designed losses without any modification.

During the testing process, all the IN layers in $\mathcal{G}$ are replaced with KIN layers. Given an image X, non-overlapped patches $x_p^{i,j}$ are cropped with the size of 512×512. The coordinates i, j of each patch $x_p^{i,j}$ corresponding to the original X would be recorded simultaneously. For example, an $M \times N$ image would be cropped into $\lfloor M/512 \rfloor \times \lfloor N/512 \rfloor$ patches with coordinates of $\{0, 1, ..., \lfloor M/512 \rfloor\}$ $\times \{0, 1, ..., \lfloor N/512 \rfloor\}$. Two caching tables of size $\lfloor M/512 \rfloor \times \lfloor N/512 \rfloor \times C$, in which C denotes the number of channels, would be initialized in each KIN for caching mean and standard deviation calculated.

As illustrated in Fig 3, we divide the testing process into two phases: caching and inference. During caching phase, each patch $x_p^{i,j}$ with its corresponding coordinates i, j are the input of the generator $\mathcal{G}$, and the calculated mean $\mu(x_p^{i,j})$ and standard deviation $\sigma(x_p^{i,j})$ after passing the KIN will be cached. During the inference phase, $x_p^{i,j}$, its corresponding coordinates i, j and a kernel k are the input of the generator $\mathcal{G}$. The kernel $k \in R^{h \times w}$, where h and w are the height and width of k, is adjustable. When passing through the KIN layer, a region with the same size of kernel k extended from i, j will be extracted from the caching table and convolute with kernel k to compute mean $\mu_{KIN}(x_p^{i,j})$ and standard deviation $\sigma_{KIN}(x_p^{i,j})$ which are used to normalize the feature maps. All the cropped patches will be passed to the $\mathcal{G}$ in the aforementioned manner to yield translated patches $\hat{y}_p^{i,j}$. Eventually, all $\hat{y}_p^{i,j}$ are assembled into an ultra-high-resolution translated image $\hat{Y}$.

3.2 Kernelized Instance Normalization (KIN)

IN [29] has been widely used in GAN-based models for image generation and dramatically improved image quality [27]. Besides, multiple styles can be obtained by conditionally replacing the μ and σ in the IN layer [9]. IN can be formulated by:

$$IN(X) = \gamma(\frac{X - \mu(X)}{\sigma(X)}) + \beta \tag{2}$$

For each instance in a batch, $\mu(X)$ and $\sigma(X)$ are calculated in a channel-wise manner, in which $\mu(X), \sigma(X) \in R^{B \times C}$, B denotes the batch size and γ and β are trainable parameters.

We hypothesize that adjacent patches share similar statistics including the μ and σ computed in IN, and thus proposed KIN that could further alleviate the subtle incongruity that induces the tiling artifacts when adjacent patches are assembled. KIN is the extension of the original IN layer with extra two caching

tables T_μ and T_σ to spatially store $\mu(X)$ and $\sigma(X)$ values and additionally supports convolution operation on the caching tables with a given kernel k. During the caching phase, KIN input a cropped patch $x_p^{i,j}$ with its spatial information, i, j. $\mu(x_p^{i,j})$ and $\sigma(x_p^{i,j})$ are computed as the original IN and cached.

$$T_\mu[i,j] := \mu(x_p^{i,j}), x_p^{i,j} \text{ is cropped from } X_{i,j} \tag{3}$$

$$T_\sigma[i,j] := \sigma(x_p^{i,j}), x_p^{i,j} \text{ is cropped from } X_{i,j} \tag{4}$$

During the inference phase, given a kernel k with the size of $2q+1$, $\mu_{KIN}(x_p^{i,j})$ and $\sigma_{KIN}(x_p^{i,j})$ are computed by convoluting k on cache tables to generate translated images. To address the boundary cases, the cache tables would be padded initially with edge values.

$$\mu_{KIN}(x_p^{i,j}) = \sum_{u=-q}^{q} \sum_{v=-q}^{q} T_\mu[i+u, j+v] \cdot K[q+u, q+v], \forall i, j \tag{5}$$

$$\sigma_{KIN}(x_p^{i,j}) = \sum_{u=-q}^{q} \sum_{v=-q}^{q} T_\sigma[i+u, j+v] \cdot K[q+u, q+v], \forall i, j \tag{6}$$

$$KIN(x_p^{i,j}, i, j) = \gamma\left(\frac{x_p^{i,j} - \mu_{KIN}(x_p^{i,j})}{\sigma_{KIN}(x_p^{i,j})}\right) + \beta \tag{7}$$

4 Datasets

4.1 Automatic Non-rigid Histological Image Registration (ANHIR)

Automatic Non-rigid Histological Image Registration (ANHIR) dataset [5,6,7,10,12,24] consists of high-resolution WSIs from different tissue samples (lesions, lung lobes, breast tissue, kidney tissue, gastric tissue, colon adenocarcinoma, and mammary gland). The acquired images are organized in sets of consecutive tissue slices stained by various dyes, including H&E, Ki-67, ER/PR, CD4/CD8/CD68, etc., with sizes vary from $15,000 \times 15,000$ to $50,000 \times 50,000$ pixels. We randomly sampled three types of tissues to conduct our experiments. Each experiment comprises H&E stain and one target IHC stain: breast tissue (from H&E to PR), colon adenocarcinoma (COAD) (from H&E to CD4&CD68), and lung lesion (from H&E to Ki-67).

4.2 Glioma

The private glioma dataset was collected from H&E ($98,304 \times 93,184$ pixels) and epidermal growth factor receptor (EGFR) IHC ($102,400 \times 93,184$ pixels) stained tissue microarrays, and each comprised 105 tissue samples corresponding to 105 different patients. Totally 105 H&E stained tissue images with their consecutive EGFR counterparts were cropped from the microarrays and the image sizes vary from $7,000 \times 7,000$ to $10,000 \times 10,000$ pixels. We randomly selected 55 samples as the training set while the other 50 pairs as the testing set.

4.3 Kyoto summer2autumn

An extra natural image dataset was used to validate the generalizability of our methodology. We collected 17 and 20 high-resolution (3456 × 5184 pixels) unpaired images taken in Tokyo during summer and autumn, respectively, as the training set and additional four summer images were used as a testing set. This Kyoto summer2autumn dataset[1] was released to facilitate solving ultra-high-resolution-related problems that most computer vision studies might encounter.

5 Experiments

5.1 Experimental settings

Three popular unpaired image-to-image translation frameworks: CycleGAN [33], CUT [28], and L-LSeSim [32], were utilized to verify our approach. We followed the hyperparameter settings described in the original papers during the training process except for the model output size, which was changed from 256 × 256 to 512 × 512. We trained the CycleGAN, CUT, and L-LSeSim for 50, 100, and 100 epochs. Models were trained and tested on three datasets: ANHIR, glioma, and Kyoto summer2autumn. Due to the insufficiency of WHIs in ANHIR dataset, we could only inference on the training set (note that training was in an unsupervised manner) while glioma and Kyoto summer2autumn datasets can be further split into training and testing sets. We replaced all IN layers with KIN layers in the generators during the inference process. One ultra-high-resolution image would be cropped into non-overlapped patches and pass through the KIN module. Translated patches were assembled to the final translated output. Constant and Gaussian kernels with sizes of 3, 7, and 11 were used to generate the best results. Translated images generated with KIN were compared with those from IN and TIN. Due to the GPU memory limitation, translated images generated with IN were also in a patch-wise (512×512) manner, which is the same as the patch-wise IN in Chen *et al.* 's work [8].

5.2 Metrics

In addition to the visualization of the translated images, we calculated Fréchet inception distance (FID) [14], histogram correlation, Sobel gradients [16] in YCbCr color domain, perception image quality evaluator (PIQE) [30], and natural image quality evaluator (NIQE) [25] to comprehensively evaluate the quality of translated ultra-high-resolution images.

However, due to the limitations of the available metrics that the tiling artifacts are difficult to be fairly graded and the unavailability of the perfectly

[1] Kyoto summer2autumn dataset is available at: https://github.com/Kaminyou/Kyoto-summer2autumn

matched counterpart, we conducted two human evaluation studies with five specialists: (a) quality challenge: given one source, one reference, and three translated images generated by patch-wise IN, TIN, and our KIN methods, respectively, specialists were asked to select the best among three translated images in 30 seconds. Since the images generated by CycleGAN and L-LSeSim were atrocious, we only chose images generated by CUT; (b) fidelity challenge: given one real image and one translated image, specialists were asked to select the one which is personally considered realistic in 10 seconds. We followed the protocol of the AMT perceptual studies from Isola *et al.* [17] but adjusted the time limitation as our images are extremely large. Since the data in ANHIR breast, COAD, and lung lesion subdatasets are insufficient, we combined these subdatasets as single ANHIR dataset and randomly selected pairs of real and translated WSIs from it.

5.3 Results

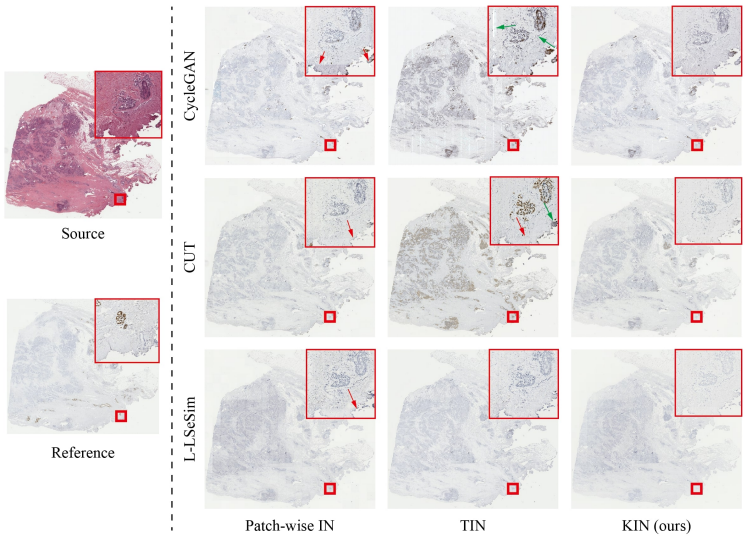


Fig. 4. H&E-to-PR stain transformation results on ANHIR breast dataset ($10,205 \times 10,933$ pixels) generated by different frameworks with IN, TIN, and KIN layers. Red arrows indicate tiling artifacts; green arrows indicate over/under-colorizing. CUT+KIN achieved the best performance. Zoom in for better view.

Stain transformation Figs. 4 to 6 and Fig. S5 and Fig. S6 show the translated images for three ANHIR subdatasets (breast tissue, COAD, and lung lesion) and

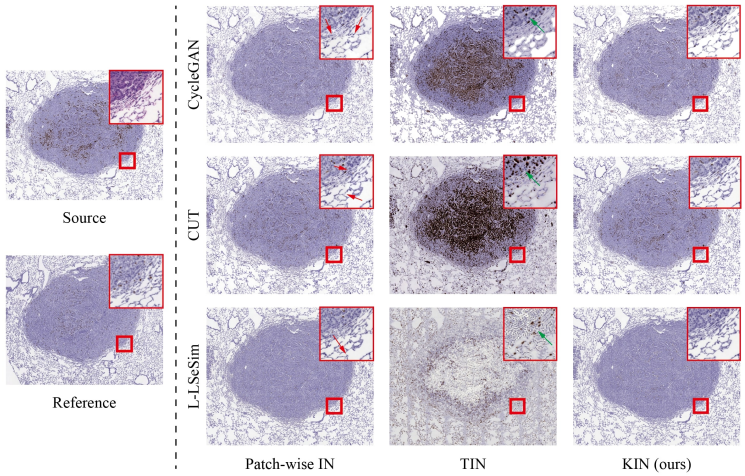


Fig. 5. H&E-to-Ki-67 stain transformation results on ANHIR lung lesion dataset ($7,336 \times 8,915$ pixels) generated by different frameworks with IN, TIN, and KIN layers. Red arrows indicate tiling artifacts; green arrows indicate over/under-colorizing. CUT+KIN achieved the best performance. Zoom in for better view.

Table 1. Quantitative results for ANHIR dataset. For each experiment, the bold shows the best performance; the underline indicates that KIN surpasses IN and TIN.

		Breast					COAD					Lung lesion				
		FID↓	Corr.↑	Grad.↓	PIQE↓	NIQE↓	FID↓	Corr.↑	Grad.↓	PIQE↓	NIQE↓	FID↓	Corr.↑	Grad.↓	PIQE↓	NIQE↓
CycleGAN	IN*	98.60	-0.07	13.62	**4.95**	9.39	103.25	75.25	15.08	5.26	**9.08**	**76.15**	-4.49	9.92	**62.69**	13.14
	TIN	179.14	-28.91	14.37	6.16	9.56	**100.78**	**79.53**	16.68	15.79	9.55	239.19	**17.46**	**9.53**	67.79	**11.96**
	KIN	**96.09**	**11.53**	**12.93**	5.29	**7.36**	108.32	43.60	**14.95**	**5.15**	9.69	103.48	-2.16	9.94	63.81	12.16
CUT	IN*	**71.00**	**35.56**	14.55	**3.00**	12.15	95.87	74.64	14.76	4.50	8.96	**54.86**	-5.79	**10.41**	58.32	12.20
	TIN	125.18	39.50	17.04	3.42	10.98	**91.81**	33.29	15.49	11.96	9.02	251.48	**80.14**	11.80	**32.08**	12.20
	KIN	72.59	36.32	**14.05**	3.27	**10.66**	93.68	**76.45**	**14.60**	**4.49**	**8.94**	56.38	-9.63	10.58	60.13	**12.08**
L-LSeSim	IN*	**65.82**	31.57	15.04	**3.03**	13.36	100.42	48.15	13.64	**4.23**	8.45	**56.30**	-3.82	**9.71**	46.40	13.88
	TIN	89.94	22.16	**12.75**	3.29	13.18	100.50	41.37	15.67	9.56	**8.14**	231.13	**59.71**	12.68	**44.29**	**11.13**
	KIN	67.46	**31.58**	14.31	3.19	**12.35**	**100.04**	**51.62**	**13.34**	4.44	8.22	62.74	-4.77	9.91	47.99	13.48

IN*: Patch-wise IN; Corr.: Histogram correlation; Grad.: Sobel gradients; ↓: the lower the better; ↑: the higher the better.

glioma dataset, respectively. With only IN layers, CUT yields the images with best quality with some tiling artifacts, while CycleGAN led to checkerboard artifacts. L-LSeSim powerfully preserves spatial information but compromises color information. With TIN, all the translated images showed dramatically over/under-colorizing. With our KIN, translated images can have minor tiling artifacts and preserve their local features. However, if the original framework generated severe tiling artifacts, our KIN could alleviate but be hard to eliminate. Considering the similarity (FID and histogram correlation) and quality metrics (Sobel gradient, PIQE and NIQE), our KIN is superior to patch-wise IN and TIN in most cases (see Tabs. 1 and 2). Although KIN does not always obtain the best scores, a possible reason is that no appropriate metrics can re-

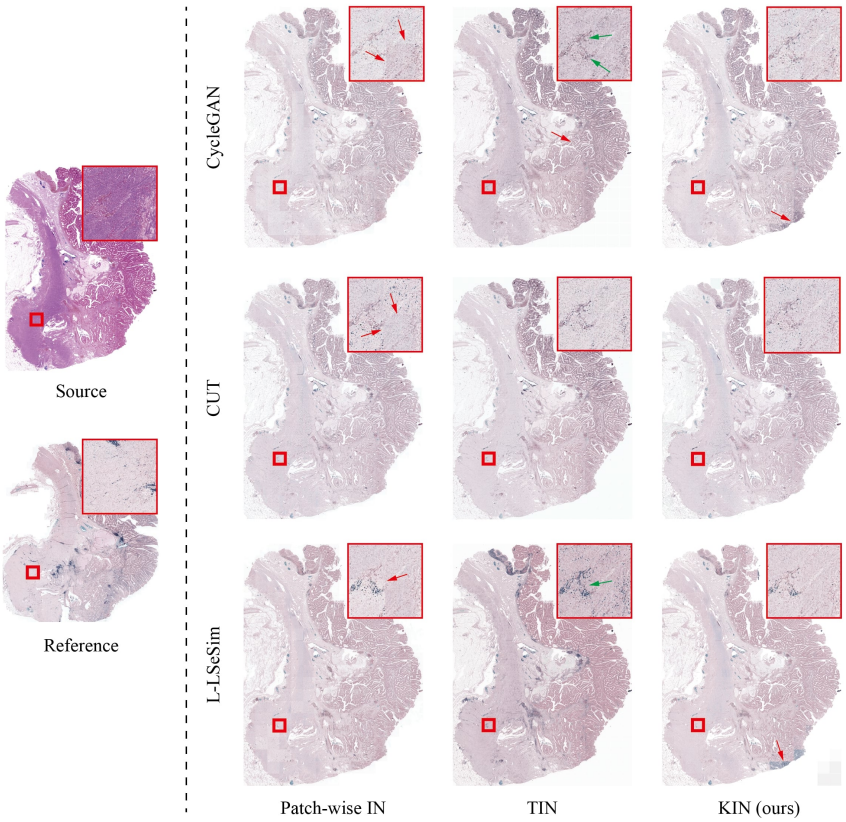


Fig. 6. H&E-to-CD4&CD8 stain transformation results on ANHIR COAD dataset ($9,816 \times 8,433$ pixels) generated by different frameworks with IN, TIN, and KIN layers. Red arrows indicate tiling artifacts; green arrows indicate over/under-colorizing. CUT+KIN achieved the best performance. Zoom in for better view.

flect the performance of such unpaired WSIs stain transformation task. Thus, we established two human evaluation studies to pertinently evaluate the image quality and fidelity. As shown in Fig. 8, our KIN achieved the best performance in both.

Translation for natural images Our KIN module also performed well on natural images (as shown in Fig. 7 and Fig. S7). As described above, KIN can alleviate the tiling artifacts generated by patch-wise IN while TIN would lead to over/under-colorizing. However, when it comes to natural images, over/under-colorizing would not be as obtrusive as in stain transformation cases, since people sometimes prefer over-stylized images. For example, High-Dynamic Range

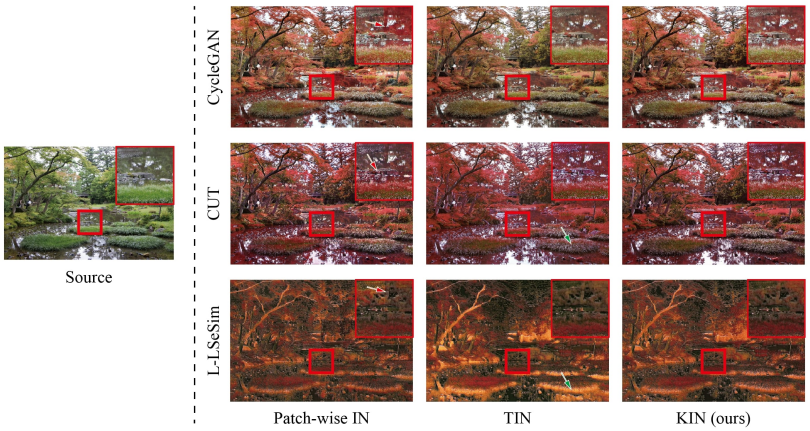


Fig. 7. Image-to-image translation results on Kyoto summer2autumn testing set ($3,456 \times 5,184$ pixels) generated by different frameworks with IN, TIN, and KIN layers. Red arrows indicate tiling artifacts; green arrows indicate over/under-colorizing. CUT+KIN achieved the best performance. Zoom in for better view.

Table 2. Quantitative results for Glioma dataset. For each experiment, the bold shows the best performance; the underline indicates that KIN surpasses IN and TIN.

		Glioma (training set)					Glioma (testing set)				
		FID↓	**Corr.↑**	**Grad.↓**	**PIQE↓**	**NIQE↓**	**FID↓**	**Corr.↑**	**Grad.↓**	**PIQE↓**	**NIQE↓**
CycleGAN	IN*	**136.32**	0.26	11.91	**21.83**	13.64	**142.28**	0.28	10.57	**23.73**	13.76
	TIN	220.00	**0.28**	**5.42**	39.05	12.01	207.03	**0.38**	**4.65**	41.16	11.99
	KIN	157.26	0.14	7.22	27.89	**11.27**	150.93	0.19	6.31	29.87	**11.54**
CUT	IN*	**105.22**	**0.85**	14.81	**23.76**	13.99	105.66	**0.85**	13.48	**24.02**	14.05
	TIN	214.22	0.54	**10.02**	34.52	**13.37**	200.56	0.64	**8.64**	35.01	**13.37**
	KIN	108.20	0.81	12.84	31.26	13.70	**100.90**	0.80	11.58	31.94	13.86
L-LSeSim	IN*	**107.74**	0.41	11.83	**21.09**	10.70	**105.59**	**0.48**	10.67	**21.25**	10.62
	TIN	203.70	0.10	**8.22**	24.87	10.94	191.44	0.19	**7.58**	23.34	10.75
	KIN	113.92	**0.41**	8.64	26.00	**10.63**	106.90	0.46	7.69	26.85	**10.40**

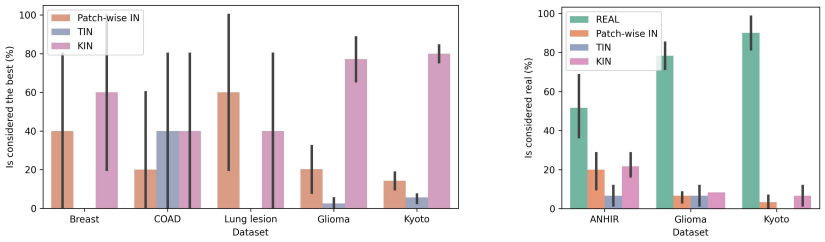
IN*: Patch-wise IN; Corr.: Histogram correlation; Grad.: Sobel gradients; ↓: the lower the better; ↑: the higher the better.

(HDR) or contrast adjustment techniques are popular to beautify photographs and render the photos more attractive. Tab. 3 and Fig. 8 provided the metrics evaluation results, and our KIN obtained the best performance among all methodologies in human evaluation. Considering the FID, CUT with our KIN is superior or competitive to other methods. Although Sobel gradients are higher in some cases, the high contrast level of one image might also contribute to higher gradients. On the other hand, there are only minor differences in PIQE and NIQE between methods. However, none of the metrics can effectively evaluate ultra-high-resolution images with tiling artifacts.

Table 3. Quantitative results for Kyoto summer2autumn dataset. For each experiment, the bold shows the best performance; the underline indicates that KIN surpasses IN and TIN.

		Kyoto (training set)				Kyoto (testing set)			
		FID↓	**Grad.↓**	**PIQE↓**	**NIQE↓**	**FID↓**	**Grad.↓**	**PIQE↓**	**NIQE↓**
CycleGAN	IN*	**79.11**	18.40	**43.62**	12.20	**171.88**	16.52	**37.00**	11.26
	TIN	87.10	**12.39**	54.29	12.24	180.50	**10.49**	52.56	11.92
	KIN	93.60	17.08	44.65	**12.11**	192.25	15.12	39.53	**11.10**
CUT	IN*	77.59	17.87	43.81	13.40	**157.04**	18.29	**37.74**	**11.41**
	TIN	98.37	**17.44**	43.30	**12.11**	181.13	**15.33**	40.97	11.89
	KIN	**75.27**	18.53	**40.21**	12.98	167.15	17.24	38.30	12.31
L-LSeSim	IN*	**178.19**	14.86	19.35	**11.89**	**248.81**	13.05	14.14	11.42
	TIN	178.98	**11.27**	**19.13**	12.07	253.14	**9.74**	**12.21**	11.18
	KIN	192.42	16.41	19.19	12.01	265.07	16.12	13.00	**10.54**

IN*: Patch-wise IN; Grad.: Sobel gradients; ↓: the lower the better



(a) Quality evaluation (b) Fidelity evaluation

Fig. 8. Human evaluation results. In quality evaluation, KIN achieved the best or competitive performance among all datasets while images generated via TIN obtained the worst quality. For the fidelity evaluation, although real consecutive section of tissue is easily to be distinguished from the fake ones, KIN is still the most deceptive among all methods. It can be noticed that translated natural images are hardly to deceive human since their complicated content are difficult to be fabricated.

5.4 Ablation study

Kernel and kernel size To elucidate the effect of different kernels and kernel size on the translated images, we applied constant and Gaussian kernels with the size of 1, 3, 7, 11, and ∞ in the KIN module (see Fig. S8 and S9). It is noteworthy that when kernel size is set to 1, the KIN module will operate in a manner of patch-wise IN, whereas it would be like TIN when kernel size is set to ∞ (bounded by the input image size). KIN is an eclectic approach that combines the advantages of patch-wise IN and TIN and avoids extremes of single and global features calculated in patch-wise IN and TIN. When kernel size increases from one to ∞, the translated results gradually change from patch-wise IN to TIN. On the other hand, the constant kernel can help generate smoother results, while the Gaussian kernel will emphasize local features more.

6 Discussions

Our experiments showed that KIN performed well on multiple datasets, and unseen testing data can even be successfully inferred when sufficient training data are available. The over/under-colorizing problem caused by TIN is also revealed, which might be innocuous when natural images are used but would be detrimental when targeting stain transformation. Pathological features, which are essential for clinical judgment, would be compromised when global mean and standard deviation are applied in the TIN.

Although KIN can be inserted into any IN-based framework, the performance would be compromised if the original framework has amateurish performance, such as CycleGAN, which generates results diversely among adjacent patches. KIN can hardly eliminate all the tiling artifacts undertaking such cases. Interestingly, we found that CUT can yield more consistent results among adjacent patches, especially for the hue. On the other hand, LSeSim meticulously preserves all the structure but ignores the consistency of the hue, which is reasonable as CUT captures domain-specific features, but LSeSim focuses on spatial features according to their loss functions. Despite KIN achieving the best performance surpassing all previous approaches in human evaluation studies, its strength cannot be manifested due to the inadequacy of appropriate metrics for evaluating the quality and fidelity of unpaired ultra-high-resolution WSIs. Finally, ultra-high-resolution images are commonly used in daily life but there is no public dataset available for a fair comparison. To facilitate related researches, we released the Kyoto summer2autumn dataset.

7 Conclusion

This study presents Kernelized Instance Normalization (KIN) for ultra-high-resolution stain transformation with constant space complexity. KIN can be easily inserted into popular unpaired image-to-image translation frameworks without re-training the model. Comprehensive experiments with two WSI datasets were conducted and evaluated by human evaluation studies and appropriate metrics. An extra ultra-high-resolution natural image dataset was also utilized and demonstrated the generalizability of KIN. Overall, KIN surpassed all the previous approaches and generated state-of-the-art outcomes. Henceforth, ultra-high-resolution stain transformation or image-to-image translation, can be easily accomplished and applied in clinical practice.

Acknowledgements We thank Chao-Yuan Yeh, the CEO of aetherAI, for providing computing resources, which enabled this study to be performed, and Cheng-Kun Yang for his revision suggestions.

References

1. Alturkistani, H.A., Tashkandi, F.M., Mohammedsaleh, Z.M.: Histological stains: a literature review and case study. Global journal of health science **8**(3), 72 (2016) 1
2. de Bel, T., Hermsen, M., Kers, J., van der Laak, J., Litjens, G.: Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In: International Conference on Medical Imaging with Deep Learning–Full Paper Track (2018) 5
3. Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. In: NIPS (2017) 4
4. Birkman, E.M., Mansuri, N., Kurki, S., Ålgars, A., Lintunen, M., Ristamäki, R., Sundström, J., Carpén, O.: Gastric cancer: immunohistochemical classification of molecular subtypes and their association with clinicopathological characteristics. Virchows Archiv **472**(3), 369–382 (2018) 1
5. Borovec, J., Kybic, J., Arganda-Carreras, I., Sorokin, D.V., Bueno, G., Khvostikov, A.V., Bakas, S., Eric, I., Chang, C., Heldmann, S., et al.: Anhir: automatic non-rigid histological image registration challenge. IEEE transactions on medical imaging **39**(10), 3042–3052 (2020) 7
6. Borovec, J., Munoz-Barrutia, A., Kybic, J.: Benchmarking of image registration methods for differently stained histological slides. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 3368–3372. IEEE (2018) 7
7. Bueno, G., Deniz, O.: Aidpath: academia and industry collaboration for digital pathology (2019) 7
8. Chen, Z., Wang, W., Xie, E., Lu, T., Luo, P.: Towards ultra-resolution neural style transfer via thumbnail instance normalization. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022) 3, 5, 8
9. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style (2017) 6
10. Fernandez-Gonzalez, R., Jones, A., Garcia-Rodriguez, E., Chen, P.Y., Idica, A., Lockett, S.J., Barcellos-Hoff, M.H., Ortiz-De-Solorzano, C.: System for combined three-dimensional morphological and molecular analysis of thick tissue specimens. Microscopy research and technique **59**(6), 522–530 (2002) 7
11. Fragomeni, S.M., Sciallis, A., Jeruss, J.S.: Molecular subtypes and local-regional control of breast cancer. Surgical Oncology Clinics **27**(1), 95–120 (2018) 1
12. Gupta, L., Klinkhammer, B.M., Boor, P., Merhof, D., Gadermayr, M.: Stain independent segmentation of whole slide images: A case study in renal histology. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1360–1364. IEEE (2018) 7
13. de Haan, K., Zhang, Y., Zuckerman, J.E., Liu, T., Sisk, A.E., Diaz, M.F., Jen, K.Y., Nobori, A., Liou, S., Zhang, S., et al.: Deep learning-based transformation of h&e stained tissues into special stains. Nature communications **12**(1), 1–13 (2021) 3, 4
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 8
15. Inamura, K.: Update on immunohistochemistry for the diagnosis of lung cancer. Cancers **10**(3), 72 (2018) 1
16. Irwin, F., et al.: An isotropic 3x3 image gradient operator. Presentation at Stanford AI Project **2014**(02) (1968) 8

17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 9
18. Kapil, A., Wiestler, T., Lanzmich, S., Silva, A., Steele, K., Rebelatto, M., Schmidt, G., Brieu, N.: Dasgan–joint domain adaptation and segmentation for the analysis of epithelial regions in histopathology pd-l1 images. arXiv preprint arXiv:1906.11118 (2019) 4
19. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International conference on machine learning. pp. 1857–1865. PMLR (2017) 4
20. Lahiani, A., Gildenblat, J., Klaman, I., Albarqouni, S., Navab, N., Klaiman, E.: Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach. In: European Congress on Digital Pathology. pp. 47–55. Springer (2019) 5
21. Lahiani, A., Klaman, I., Navab, N., Albarqouni, S., Klaiman, E.: Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency. IEEE Journal of Biomedical and Health Informatics **25**(2), 403–411 (2020) 4
22. Levy, J.J., Jackson, C.R., Sriharan, A., Christensen, B.C., Vaickus, L.J.: Preliminary evaluation of the utility of deep generative histopathology image translation at a mid-sized nci cancer center. bioRxiv (2020) 4
23. Luo, Z.W., Zhu, M.G., Zhang, Z.Q., Ye, F.J., Huang, W.H., Luo, X.Z.: Increased expression of ki-67 is a poor prognostic marker for colorectal cancer patients: a meta analysis. BMC cancer **19**(1), 1–13 (2019) 2
24. Mikhailov, I., Danilova, N., Malkov, P.: The immune microenvironment of various histological types of ebv-associated gastric cancer. In: Virchows Archiv. vol. 473, pp. S168–S168. Springer 233 SPRING ST, NEW YORK, NY 10013 USA (2018) 7
25. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012) 8
26. Oshi, M., Tokumaru, Y., Angarita, F.A., Yan, L., Matsuyama, R., Endo, I., Takabe, K.: Degree of early estrogen response predict survival after endocrine therapy in primary and metastatic er-positive breast cancer. Cancers **12**(12), 3557 (2020) 3
27. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications. IEEE Transactions on Multimedia (2021) 3, 6
28. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision. pp. 319–345. Springer (2020) 4, 6, 8
29. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016) 6
30. Venkatanath, N., Praneeth, D., Bh, M.C., Channappayya, S.S., Medasani, S.S.: Blind image quality evaluation using perception based features. In: 2015 Twenty First National Conference on Communications (NCC). pp. 1–6. IEEE (2015) 8
31. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE international conference on computer vision. pp. 2849–2857 (2017) 4
32. Zheng, C., Cham, T.J., Cai, J.: The spatially-correlative loss for various image translation tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021) 4, 6, 8
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017) 3, 4, 6, 8

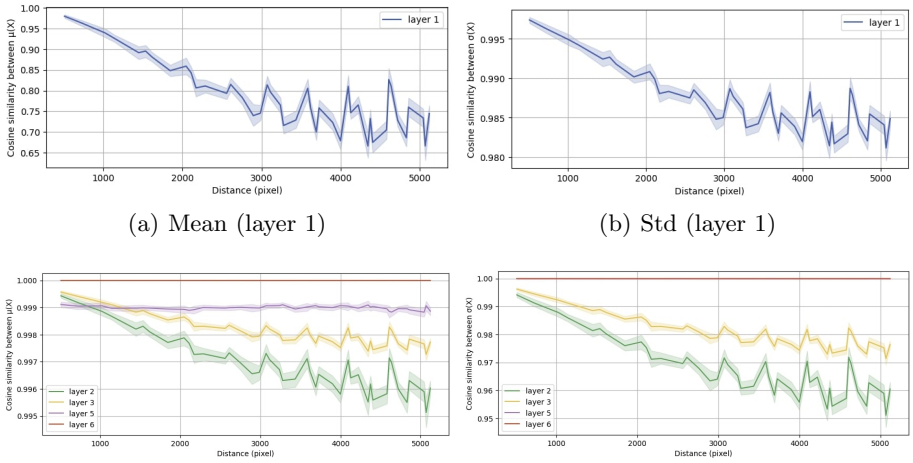
arXiv:2208.10730v1 [eess.IV] 23 Aug 2022

Ultra-high-resolution unpaired stain transformation via Kernelized Instance Normalization (Supplementary Material)

Ming-Yang Ho*, Min-Sheng Wu, and Che-Ming Wu

aetherAI, Taipei, Taiwan
{kaminyouho,vincentwu,uno}@aetherai.com
https://www.aetherai.com/

0.1 Analysis



(a) Mean (layer 1) (b) Std (layer 1)

(c) Mean (layer 2-6) (d) Std (layer 2-6)

Fig. S1. Comparison of mean and std calculated in IN between adjacent patches. Mean, and standard deviation (std) of every two adjacent or nearby patches (up to 5,000 pixels far away) were extracted from the IN in the original CycleGAN model and compared. The CycleGAN model comprises 6 layers and each has one or multiple IN: (1) convolutional layer; (2) down-sampling layer; (3) down-sampling layer; (4) residual backbone; (5) up-sampling layer; (6) up-sampling layer. We analyzed mean and std from IN in all layers except the fourth layer, which is a backbone. It can be noticed that there is a great discrepancy in mean and std between faraway patches in the earlier layers.

* Corresponding author

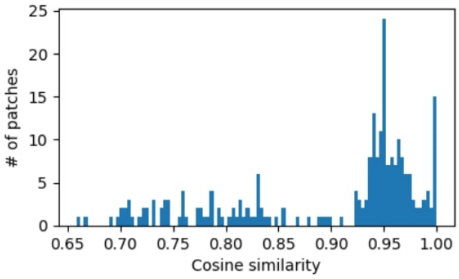


Fig. S2. Distribution of cosine similarity between means of thumbnail and patches calculated in IN. The means of patches and the thumbnail calculated in the IN layer from layer 1 of CycleGAN's generator are extracted and compared. Distribution of the cosine similarity is shown. An obvious discrepancy can be observed, which indicates the inappropriateness of using thumbnail statistics for all cropped patches in the TIN [8].

To verify our hypothesis, we extracted the $\mu(X)$ and $\sigma(X)$ from all the IN layers in $\mathcal{G}$ for patches cropped from one single image, in which $\mu(X), \sigma(X) \in R^{1\times C}$, and C is the number of channels. Then, $\mu(X)$ and $\sigma(X)$ were further flattened into vectors with size C to compute the cosine similarity between every pair. Besides, the Euclidean distances between pairs were recorded.

Fig. S1 demonstrates that cosine similarity of $\mu(X)$ and $\sigma(X)$ between two patches would dramatically decrease when two patches are farther apart, especially in the first few layer blocks.

In addition, we adopted the methodology proposed in TIN [8] and measured the $\mu(X)$ and $\sigma(X)$ between the thumbnail and other cropped patches in Figure S2. It shows that extreme inconsistency occurs in the first few layers, implying local contrast and hue information will diminish if μ and σ of thumbnail are used. On the contrary, the convolution mechanism in our KIN can both alleviate this inconsistency issue and further improve the assembly quality when adjacent patches are combined.

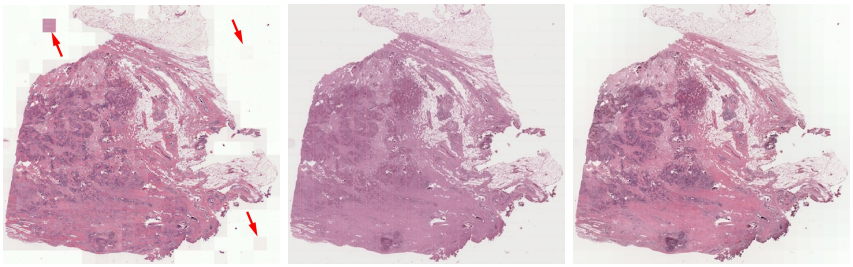
0.2 Performance on the classification downstream task

As there is no well-developed metric that can evaluate unpaired ultra-high-resolution (UHR) images, downstream classification task was experimented to address this issue. We conducted a classification task for the ANHIR dataset (breast, lung lesion, and COAD). A ResNet-50 model was trained on the patches cropped from real WSIs in the IHC domain and tested on the patches cropped from translated WSIs generated by patch-wise IN, TIN, and KIN with the CUT framework. We deliberately cropped patches from the attached boundary to evaluate the influence of tilting artifacts. The accuracies of patch-wise IN, TIN, and **KIN** are 98.8%, 88.4%, and **99.2**%, respectively. The results show that KIN achieves the best performance, which might be due to the reduction of tilting

artifacts that confused the classifier. TIN obtains the worst performance since using global statistics might lead to the loss of local information.

0.3 Evaluated by SSIM and FSIM metrics

To evaluate KIN with SSIM and FSIM metrics, we experimented with pairwise translating gray images of the ANHIR dataset into H&E. However, both SSIM (patch-wise IN: 0.94, TIN: 0.90, KIN: 0.93) and FSIM (patch-wise IN: 0.79, TIN: 0.74, KIN: 0.78) cannot evaluate the presence of tilting artifacts in patch-wise IN (see Fig. S3).

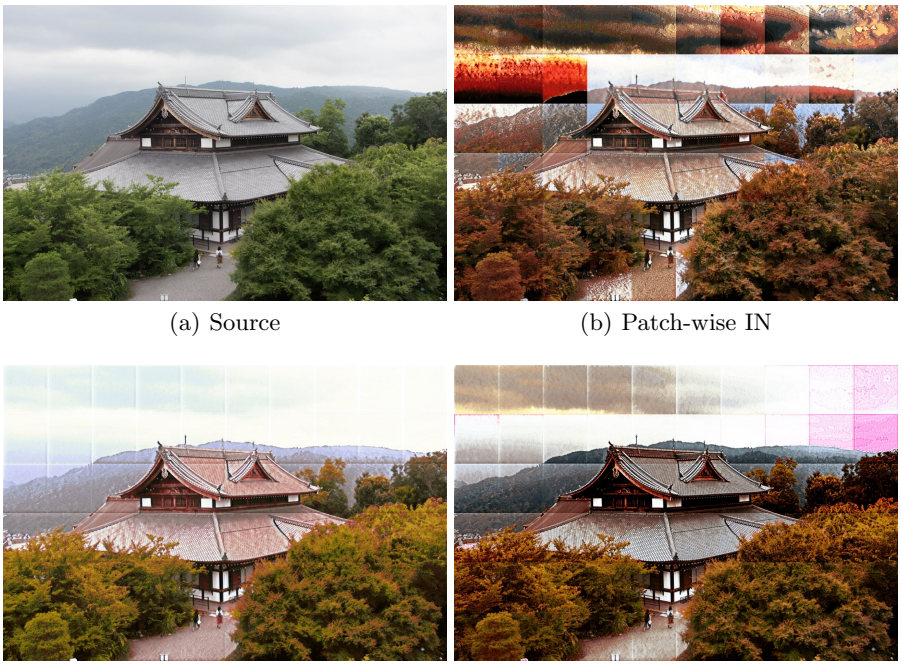


(a) Patch-wise IN (b) TIN (c) KIN

Fig. S3. Generated RGB WSIs by different methods. The presence of tilting artifacts, indicated by red arrows, cannot evaluated by SSIM or FSIM metrics.

0.4 Failure modes of KIN

If the training data lack enough specific scene (e.g., the sky in Kyoto dataset), KIN will be inferior to TIN (see Fig. S4).



(a) Source (b) Patch-wise IN

(c) TIN (d) KIN

Fig. S4. Failure modes. KIN will be inferior to TIN if training data lack enough specific scene.

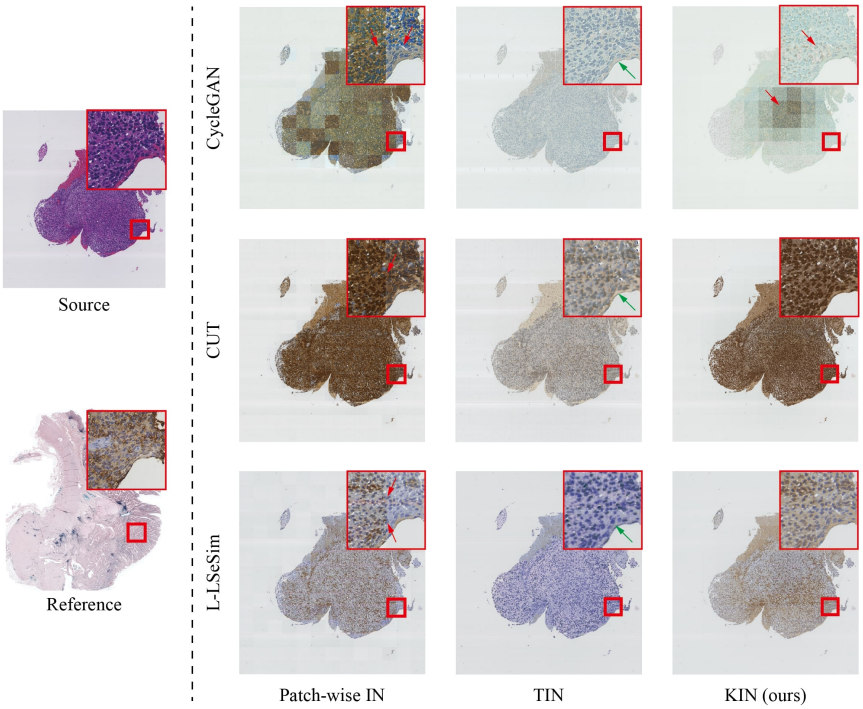


Fig. S5. H&E-to-EGFR stain transformation results on Glioma training set ($7,755 \times 7,109$ pixels) generated by different frameworks with IN, TIN, and KIN layers. Red arrows indicate tilting artifacts; green arrows indicate over/under-colorizing. CUT+KIN achieved the best performance. Zoom in for better view.

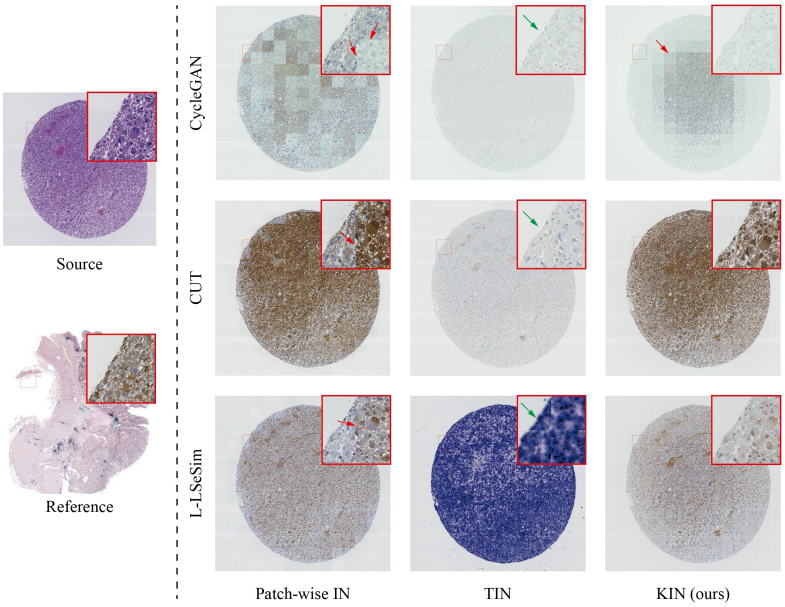


Fig. S6. H&E-to-EGFR stain transformation results on Glioma testing set $(8,078 \times 8,078$ pixels) generated by different frameworks with IN, TIN, and KIN layers. Red arrows indicate tiling artifacts; green arrows indicate over/under-colorizing. CUT+KIN achieved the best performance. Zoom in for better view.

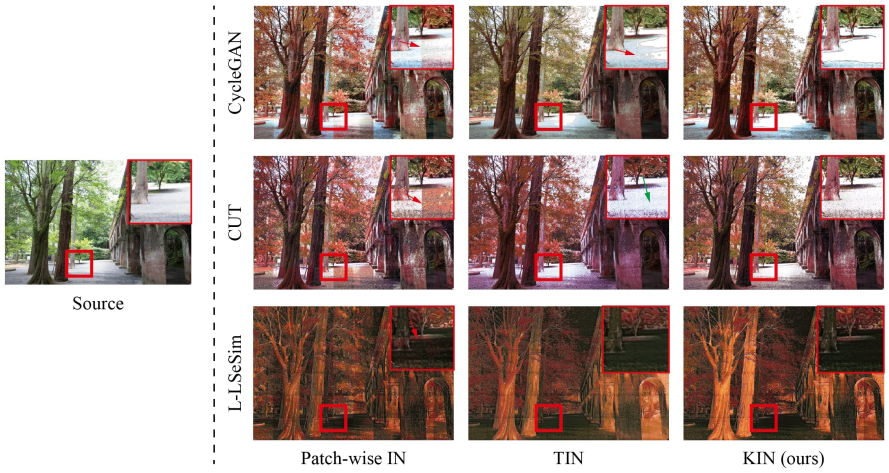


Fig. S7. Image-to-image translation results on Kyoto summer2autumn training set $(3,456 \times 5,184$ pixels) generated by different frameworks with IN, TIN, and KIN layers. Red arrows indicate tilting artifacts; green arrows indicate over/under-colorizing. CUT+KIN achieved the best performance. Zoom in for better view.

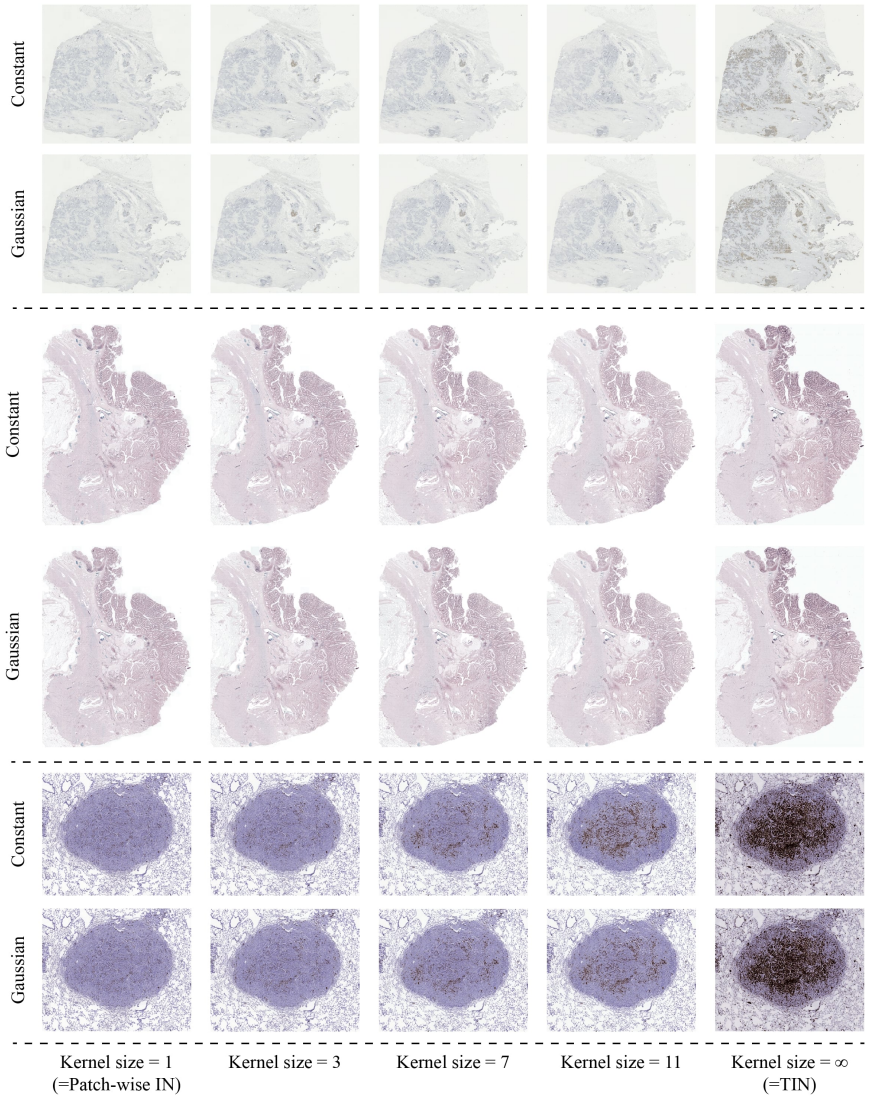


Fig. S8. Ablation study for kernel types on three ANHIR subdatasets. Constant and Gaussian kernels with the size of 1, 3, 7, 11, and ∞ are applied to elucidate the effect of KIN module. When kernel size is set to 1, the KIN module will operate in a manner of patch-wise IN, whereas it would be like TIN when kernel size is set to ∞.

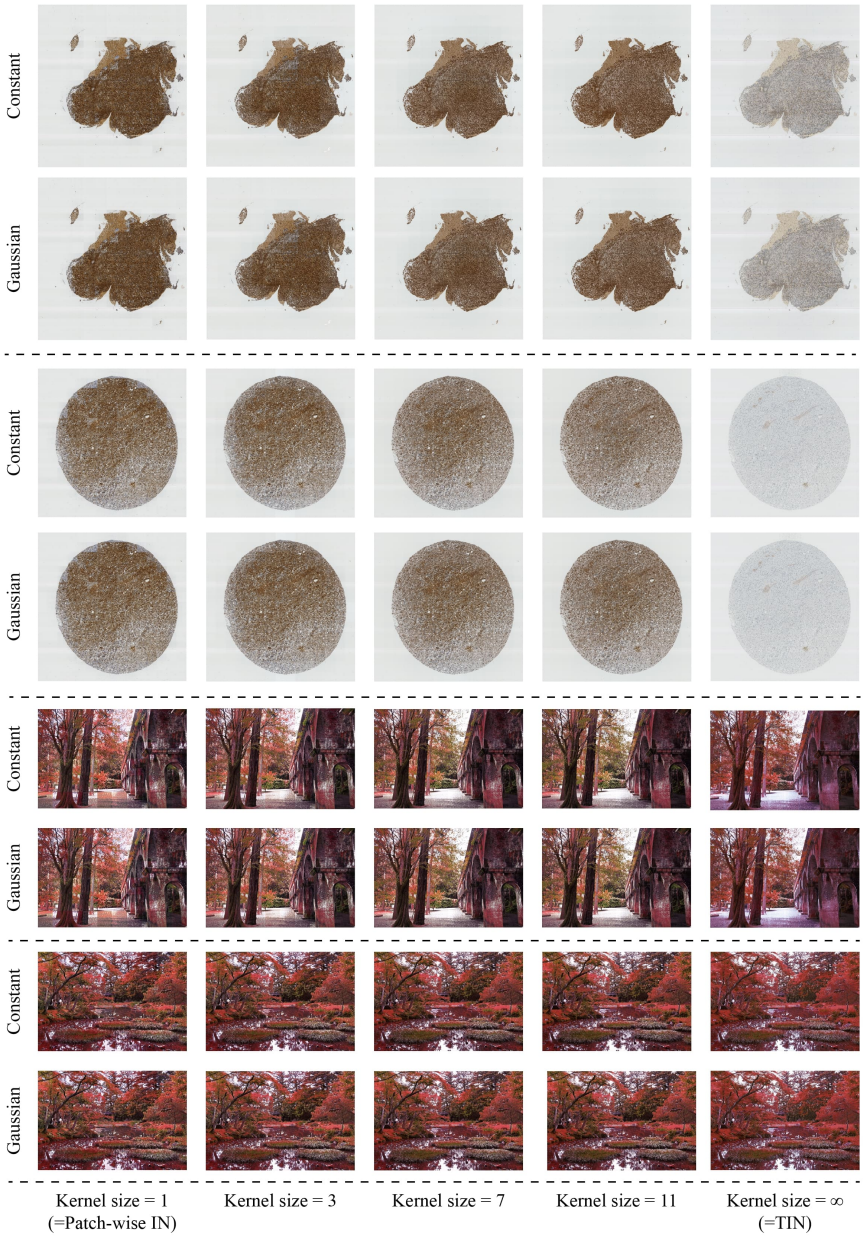


Fig. S9. Ablation study for kernel types on Glioma and Kyoto summer2autumn datasets. Constant and Gaussian kernels with the size of 1, 3, 7, 11, and ∞ are applied to elucidate the effect of KIN module. When kernel size is set to 1, the KIN module will operate in a manner of patch-wise IN, whereas it would be like TIN when kernel size is set to ∞.