# PATHOLOGICAL SECTION STAINING TRANSFERRING WITH TAILORED METRIC-BASED MODEL SELECTION

*Yiming Ji, Suyang Zhu, Dong Zhang, Shoushan Li*

Soochow University

## ABSTRACT

As the important pathological section staining, Immunohistochemistry (IHC) staining uses labeled antibodies to highlight specific antigens, providing clearer results for malignancy identification compared with Hematoxylin and Eosin (H&E) staining. However, obtaining IHC manually is labor-intensive and costly. In this study, we attempt to automatically generate IHC staining image by style transferring from H&E, which is easy to get. Due to the unalignment between single evaluation metric and real generation quality, we propose a tailored metric-based model selection (TMMS) method for pathological section staining transferring (PSST). Our method can fuse multiple traditional metrics and multiple relevant losses to measure the quality of generated IHC and select a best-performed model. Extensive experiments and analysis demonstrate the effectiveness of our method TMMS.

***Index Terms***— H&E-to-IHC Stain Transferring, Pathological Sections, Adversarial Learning, Model Selection And Evaluation

## 1. INTRODUCTION

H&E staining is a common histochemical technique in pathology, used to visualize tissue components in different colors for cancer diagnosis. Its cost-effectiveness makes it widely used, but it lacks the precision needed for differentiating cancer subtypes [1].

Therefore, IHC staining was developed, which uses labeled antibodies to highlight specific antigens, providing clearer results for malignancy identification [2]. However, IHC is labor-intensive and costly, limiting its use in low-resource settings [3]. Researchers aim to convert H&E images into IHC using data-driven methods, but aligning H&E and IHC images is challenging since they can't be pixel-matched. Alternatives include staining serial tissue sections, though this too introduces misalignment. Registration methods exist but face limitations in expertise and efficiency [4, 5]. The other option is to train H&E-to-IHC translation models using unpaired images as shown in Figure 1. The most common methods are CycleGAN-based [6, 7], which maintain structural consistency between generated and source images through inverse generation. However, these approaches normally select the best model by FID [8] or DINO [9] metric, which normally can not align with the best generation quality.

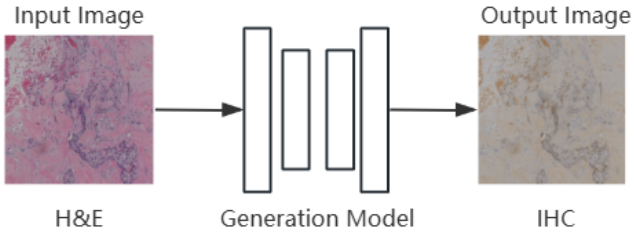



**Fig. 1**. An example of staining image translation.

As well known, for the best model, we want to achieve that: **First**, the distribution of the generated results must be sufficiently close to the distribution of the target domain images, which is typically measured by FID. **Second**, the spatial structure of the generated images must be sufficiently similar to the input images to ensure that the characteristic information of the input images is preserved, which is typically measured by DINO. While, traditional single metric cannot achieve the above both goals simultaneously. For example, if we select the best model by FID, the performance on FID value is usually the best, but the DINO is not and even very worse, resulting poor attributes preservation of input image.

To address the above limitation, we propose to design a new metric by fusing multiple measures and losses to select the best model, which can capture both distribution of the target domain (IHC) and the characteristics of the input images (H&E), namely tailored metric-based model selection (TMMS). Besides the traditional metrics, we also present several losses for metric creation. In summary, our contributions are:

1. We propose a tailored metric-based model selection (TMMS) method for pathological section staining transferring (PSST). Note that this is the first attempt to introduce the fusion of multiple metrics to select the best model.

2. We propose multiple new metrics to conduct model selection by incorporating the adversarial loss in training and testing.

3. Systematic experiments and analysis indicates the effectiveness of our new metrics and TMMS method.

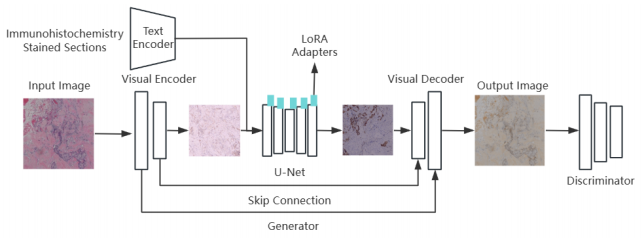



**Fig. 2**. Model architecture.

## 2. RELATED WORKS

**Medical Style Transferring.** The mainstream approaches in this study consist of: GAN-based [10, 11, 12] and Diffusion-based [13, 14, 15] methods. To our knowledge, CycleGAN-turbo[16] is considered as the SOTA for style transferring, which improves CycleGAN with Diffusion.

However, it has mainly been tested on common datasets and not yet in medical stained slices. This study will explore the feasibility of applying this model to medical staining slice conversion tasks.

**Metric-based Model Selection.** For medical slice translation, besides expert evaluation, several criteria are used for final model selection, including SSIM[17], MS-SSIM[18], and FID[8]. While expert evaluation can yield good results, it also has significant limitations because the availability of experts is not guaranteed, and not all tasks have access to them. Moreover, SSIM, and MS-SSIM all require paired images for calculation, making them unsuitable for translation tasks that use unpaired datasets.

Therefore, this study has designed a new metric for model selection in unpaired medical staining slice translation tasks using GAN methods.

## 3. METHODOLOGY

### 3.1. Basic Model

The specific method used for the conversion of medical stained slices is CycleGAN-turbo. It builds its overall architecture based on a pre-trained single-step text-to-image model. Its main components include a generator and two discriminators. Both discriminators use the CLIP model as a backbone, following the recommendations of Vision-Aided GAN. The generator is based on a model framework proposed by the authors, which includes an encoder, a UNet network with LoRA[19] adapters added, a decoder, and a text encoder. The specific architecture and workflow can be seen in Figure 2.

### 3.2. Traditional Metrics

After the model is trained, model selection is necessary. For such generative models, existing evaluation metrics include FID, SSIM, MS-SSIM, and DINO-Struct-Dist (DINO)[9], among others. The FID metric is a commonly used indicator in generative models, which assesses whether the distribution of generated results is close to the target distribution. However, it mainly considers the statistical characteristics of the image and lacks attention to visual effects and texture details. The SSIM, and MS-SSIM metrics are used for image quality evaluation, but their calculations generally require corresponding matched images for the results to be valid. For tasks lacking matched datasets, such as the unpaired medical image translation task in this paper, the applicability of these metrics is limited. The DINO-Struct-Dist metric is used to evaluate the structural similarity of image content by combining attention mechanisms for feature extraction from two images, and then calculating MSE for the final result.

### 3.3. Tailored Metric-based Model Selection

This task belongs to the image translation category, and therefore, the final model needs to achieve two goals. **First**, the distribution of the generated results must be sufficiently close to the distribution of the target domain images. **Second**, the spatial structure of the generated images must be sufficiently similar to the input images to ensure that the characteristic information of the input images is preserved. For the first goal, the traditional FID metric can be used for evaluation. For the second goal, since the dataset used in this experiment consists of unpaired images, metrics such as SSIM and MS-SSIM are not applicable. Instead, DINO-Struct-Dist is used to assess the structural similarity of the images.For the calculation of FID and DINO, see Eq(1) and Eq(2).

$$\mathrm{FID} = |\mu - \mu_{\mathrm{w}}|^2 + \mathrm{tr}\left(\sum + \sum_{w} - 2 * \mathrm{sqrt}\left(\sum * \sum_{w}\right)\right) \tag{1}$$

where $\mu$ and $\mu_w$ are the mean and covariance of the real sample features, while $\sum$ and $\sum_w$ are the mean and covariance of the generated sample features.

$$\mathrm{DINO} = \mathrm{MSE}\left(S^L(I_s) - S^L(I_o)\right) \tag{2}$$

where $S^L(I)$ represents the cosine similarity between keys in ViT (Vision Transformer)[20]. $I$ stands for the image, and $MSE$ denotes the calculation of Mean Squared Error.

At the beginning of the metric design, FID and DINO were simply combined, as $\alpha * \mathrm{FID} + \beta * \mathrm{DINO}$.

However, preliminary experiments show that models selected using this method do not perform well. This is because the DINO value stabilizes shortly after model training begins, which has little impact on the overall model selection. As a result, the final selected model does not significantly differ from the one selected using only unidirectional FID, failing to meet the goal of considering both objectives simultaneously. To address this, this study attempt to design two methods for improvement.

(1) Since the primary issue is that DINO's impact is minimal and the model is mainly influenced by FID, the simplest solution is to increase the influence of DINO. A straightforward approach is to increase the weight of DINO in the calculation process. However, this effectively means considering the importance of structural preservation as higher, which does not align with the basic human understanding of this task. In human cognition, both aspects should be equally important, or even closer to the target distribution might be more important. Therefore, in the experiments, FID was replaced with related loss parameters. The FID value represents the closeness of the feature distribution between generated images and target images. During model training, the constraints on generated images are mainly achieved using generator loss (gan_loss) and discriminator loss (disc_loss). These two values are related to FID and are used to replace it, naturally increasing the relative weight of DINO.The calculations for gan_loss and disc_loss are:

$$\text{gan\_loss} = \mathrm{E}_z[log(1 - \mathrm{D}(\mathrm{G}(z, \mathrm{C}_z)))] \quad (3)$$

$$\text{disc\_loss} = \mathrm{E}_x[log\mathrm{D}(x)] + \mathrm{E}_y[log(1 - \mathrm{D}(\mathrm{G}(y, \mathrm{C}_x)))] \quad (4)$$

where $z$ is the input image, $\mathrm{C}_z$ is the text prompt corresponding to the image. $x$ is the real image, and $\mathrm{C}_x$ is the text prompt corresponding to the real image. Overall, there is no difference in the loss from CycleGAN, except that text prompts are added. E represents Mean Squared Error. Then, we can get the new evaluation metric:

$$\textsc{Metric}_1 = \alpha_1 * (\alpha_2 * (\text{gan\_loss+disc\_loss})) + \beta^* \text{DINO} \quad (5)$$

(2) Since the DINO value does not change significantly and is mainly influenced by FID, the cycle_loss, which is also related to preserving features and shows more noticeable variation, is added to increase variability.The calculation for cycle_loss is shown in Eq.(6). The calculation of the second new evaluation metric is shown in Eq.(7).

$$\begin{aligned}\text{cycle\_loss} = \mathrm{E}_x[\mathrm{L}_{\text{rec}}(\mathrm{G}(\mathrm{G}(x, \mathrm{C}_y), \mathrm{C}_x), x)] \\ +\mathrm{E}_y[\mathrm{L}_{\text{rec}}(\mathrm{G}(\mathrm{G}(y, \mathrm{C}_x), \mathrm{C}_y), y)]\end{aligned} \quad (6)$$

$\mathrm{L}_{\text{rec}}$[16] is a combination of L1 loss and LPIPS[21]. $x$ represents an image from domain X, and $\mathrm{C}_y$ is the text prompt from domain Y. The same applies for subsequent terms.

$$\textsc{Metric}_2 = \alpha * \text{FID} + \beta_1 * (\beta_2 * (\text{cycle\_loss+DINO})) \quad (7)$$

In addition to the above metrics, this study also replaces both FID and DINO with loss parameters, using a combination of pure loss functions as the new evaluation metric.The calculation of this metric is shown in Eq.(8).

$$\textsc{Metric}_3 = \alpha_1 * (\alpha_2 * (\text{gan\_loss+disc\_loss})) + \beta * \text{cycle\_loss} \quad (8)$$

In constructing the new metric, we considered that the model shares a single generator during training. The generator can transform from direction a to b, i.e., convert H&E images into IHC images, and also from direction b to a, i.e., convert IHC images into H&E images, with both directions being related to each other. Therefore, in addition to using single-direction evaluation metrics and loss parameters, both directions were ultimately considered together. However, since the primary goal of this model is to convert slices that are easier to obtain into slices that are harder to acquire, even if both directions are considered, their importance differs, and different weights should be assigned. The direction corresponding to the primary goal should be given more weight.

## 4. EXPERIMENTATION

The research objects in the experiment are H&E stained slices, which are relatively easy to obtain, and IHC stained slices, which are more difficult to obtain. First, the CycleGAN-turbo model was used to translate medical stained slices. Next, based on the results, the newly designed metrics were used to evaluate and select the model. During the experiment, three different metrics were designed and compared. In addition, the experiment also compared the model selection results of bidirectional metrics and unidirectional metrics.

### 4.1. Experimental Settings

**Datasets.** The dataset used is the BCI (Breast Cancer Immunohistochemical Image)[22] dataset. The training set includes 3,396 H&E-stained slice images and 3,396 IHC-stained slice images, while the test set contains 500 H&E-stained slice images and 500 IHC-stained slice images. Each image has a size of 1024 x 1204 pixels. During training, the images are resized to 286 x 286 pixels and then randomly cropped to 256 x 256 pixels for training. During inference, the images are not cropped or adjusted.

**Implementation Details.** The entire experiment was conducted on a single NVIDIA 4090 GPU. The GPU was used for training with a single training step set to 25,000 iterations, and the training duration was approximately 50 hours.

### 4.2. Experimental Results

**Model Selection based on Unidirectional Loss.** We first focus on the main objective of the model, which is to translate from domain A to domain B. The relevant parameters selected are: FID_a2b, DINO_a2b, gan_a_loss, disc_a_loss and cycle_a_loss. When calculating the new metrics, normalization is first performed, followed by the calculation of $\textsc{Metric}_1$, $\textsc{Metric}_2$ and $\textsc{Metric}_3$.In this experiment, set $\alpha$, $\alpha_1$, $\alpha_2$, $\beta$, $\beta_1$ and $\beta_2$ in Eq.(5), Eq.(7), Eq.(8) all to 0.5. Specific experimental results are shown in Table 1.

**Model Selection based on Bidirectional Loss.** We also consider all parameters for both directions during the model training process. In the calculation, the metrics for the two directions are combined with different weights. The formula used is: $\textsc{Metric}_{[i]} = \alpha * \textsc{Metric}_{[i]\_a2b} + \beta * \textsc{Metric}_{[i]\_b2a}$.

**Table 1**. Results of training with uni-direction. a2b denotes H&E translated to IHC.

| Selection Criteria | DINO_a2b | FID_a2b | FID_b2a | DINO_b2a |
|---|---|---|---|---|
| FID_a2b | 0.0196 | 90.3117 | 110.4120 | 0.0186 |
| DINO_a2b | 0.0114 | 113.3111 | 88.5869 | 0.0155 |
| METRIC$_1$ | 0.0139 | 117.0864 | 93.7011 | 0.0150 |
| METRIC$_2$ | 0.0166 | 97.6371 | 92.8660 | 0.0183 |
| METRIC$_3$ | 0.0139 | 117.0864 | 93.7011 | 0.0150 |

**Table 2**. Results of Bidirectional

| Selection Criteria | DINO_a2b | FID_a2b | FID_b2a | DINO_b2a |
|---|---|---|---|---|
| FID_Bidirectional | 0.0162 | 98.6232 | 88.5616 | 0.0190 |
| DINO_Bidirectional | 0.0125 | 120.1047 | 103.3174 | 0.0140 |
| METRIC$_1$ | 0.0139 | 117.0864 | 93.7011 | 0.0150 |
| METRIC$_2$ | 0.0154 | 99.7883 | 88.9877 | 0.0174 |
| METRIC$_3$ | 0.0139 | 117.0864 | 93.7011 | 0.0150 |

In this experiment, set $\alpha$ to 0.7, $\beta$ to 0.3. Specific experimental results are shown in Table 2.

Based on the single-direction experiments, the model selected with METRIC$_1$, which replaces FID with the loss parameters gan_loss and disc_loss and produced different results compared to the model selected using FID, and the DINO value showed improvement.Despite these changes, the final model still faced limitations. GANs often prioritize deceiving the discriminator, potentially overlooking feature generation, which can lead to higher FID values. Additionally, these loss values fluctuate minimally after training, making it challenging to assess training progress directly. However, using loss parameters related to both directions improves the FID score in the other direction, balancing performance.

By adding cycle_loss to increase the consideration of structure preservation during model selection, the METRIC$_2$ selected a better model compared to the first replacement method. Like METRIC$_1$, cycle_loss, being relevant to both directions, resulted in better FID scores in both directions. The only downside was a slight decrease in the DINO score, though the overall performance remained strong.

Models selected using METRIC$_3$, similar to those chosen with METRIC$_1$, also faced issues with less-than-ideal FID scores.

Considering both directions, the models selected with the involvement of loss parameters showed relatively small variations in results, while the others exhibited some changes. Notably, the final model selected using FID outperformed the one chosen by considering only one direction, nearly matching the performance of the final model selected using METRIC$_2$, highlighting the importance of considering both directions in model selection.

Based on the experimental results and the evaluation metrics, the two best metrics for model selection are the FID and METRIC$_2$ metrics, both considering both directions. Overall, the METRIC$_2$ metric proposed in this paper is superior. This is because FID requires consideration of both directions

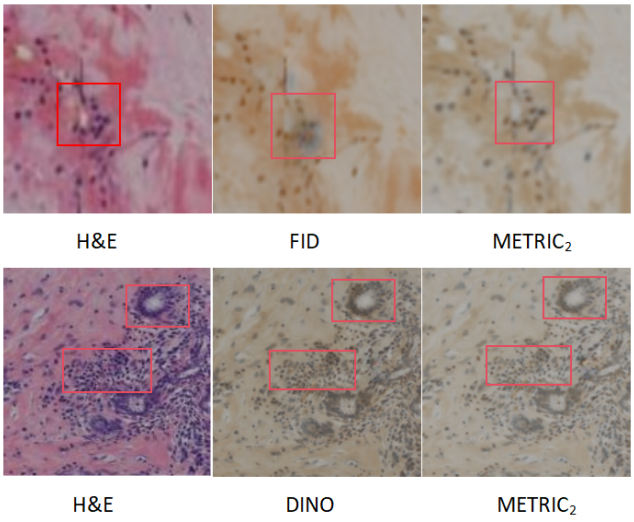



**Fig. 3**. Comparison between traditional metrics and METRIC$_2$.

to achieve optimal results, whereas METRIC$_2$ performs well even when focusing on the model's primary objective. Additionally, METRIC$_2$ achieves better DINO values.

### 4.3. Case Study

To visually demonstrate the effectiveness of the method proposed in this paper, several case examples are provided for explanation.The specific comparison can be seen in Figure 3.

**VS. FID.** Compared to the final model selected using only the FID metric, the model chosen based on the METRIC$_2$ indicator show better performance in generating finer details. The FID-selected model tend to struggle in areas with shadows or densely packed cell nuclei, where the generation quality is subpar.

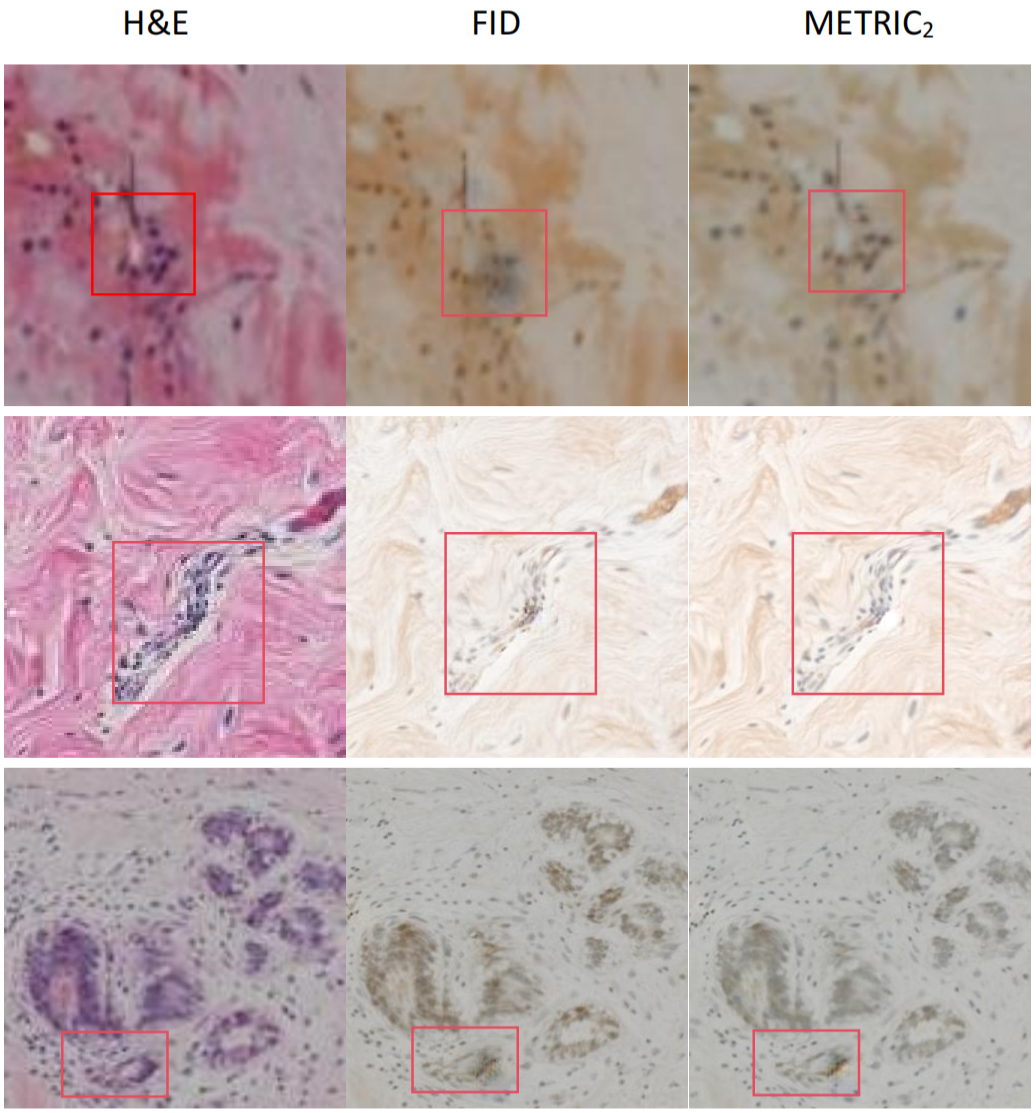
**VS. DINO.** Compared to the final model selected using METRIC$_2$, the model selected using DINO shows a relatively larger discrepancy between the result distribution and the target distribution. The final image appears darker overall, leading to a less clean visual effect.

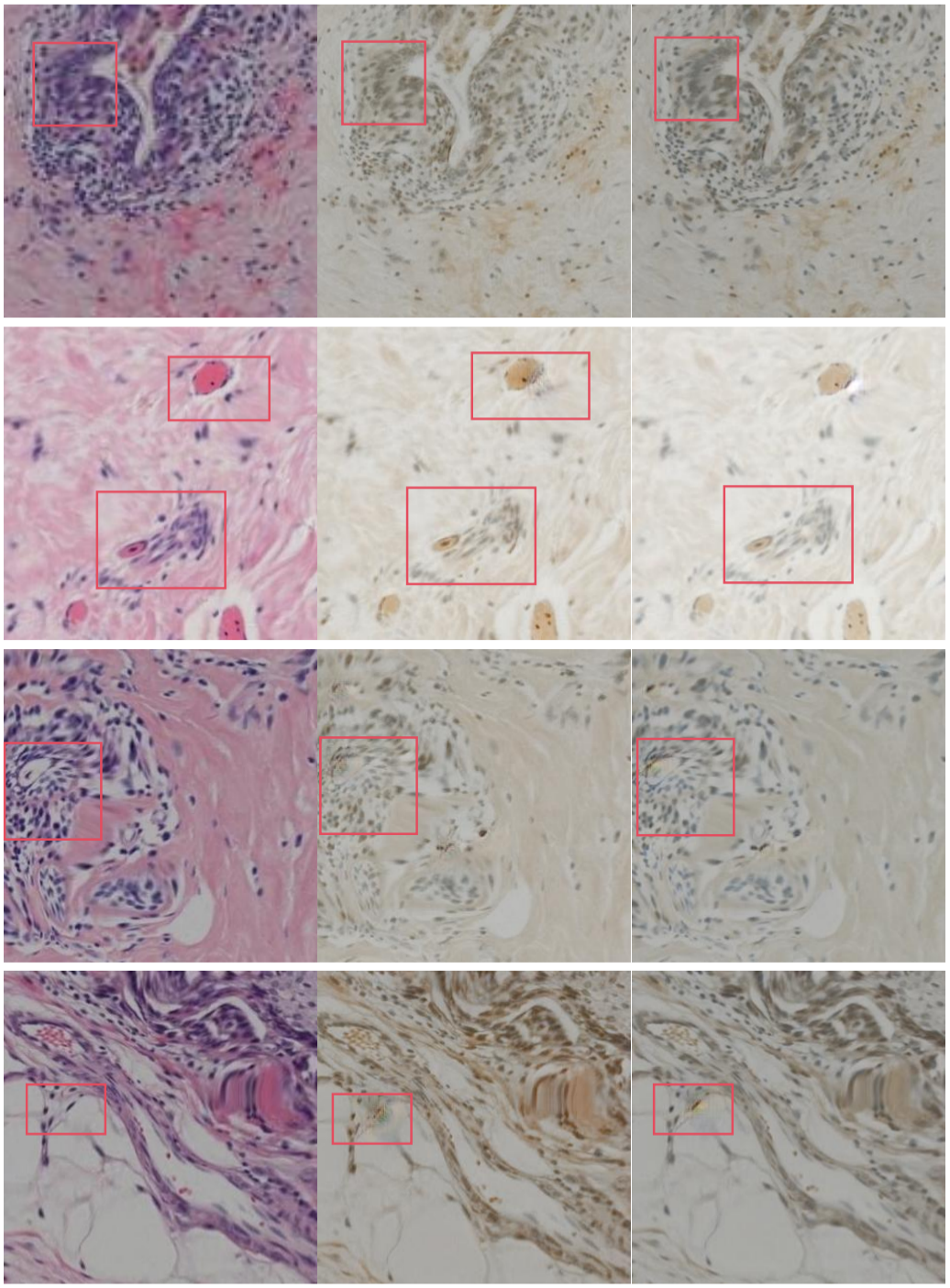
## 5. CONCLUSION

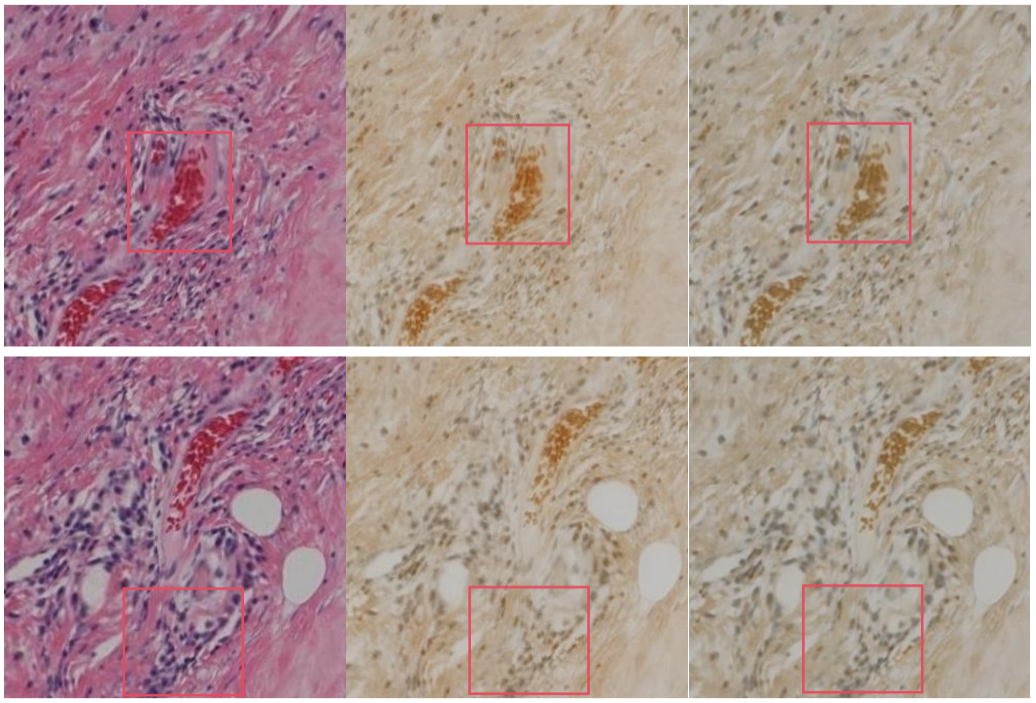
This experiment differs from traditional methods of stained slice translation by employing the new technology, CycleGAN-turbo. Compared to commonly used GAN-based models, this model exhibits superior performance in handling complex images and produces results with enhanced visual effects. Additionally, this study has designed new model selection metrics specifically for such models to identify the best-performing model.

## Appendix

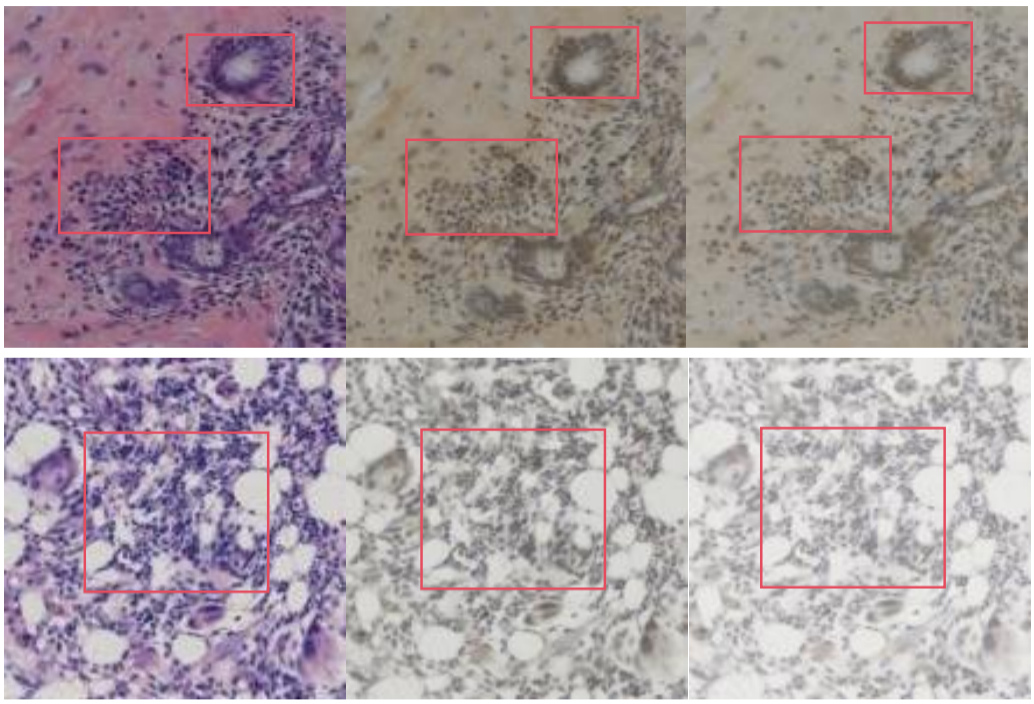
Here, an additional comparison between the generated results of models selected using traditional metrics and the model selected using the new METRIC$_2$ proposed in this paper will be provided. This further reinforces the point made in the paper: compared to FID-selected models, models selected by METRIC$_2$ exhibit superior performance in generating fine details (such as shadowed areas or densely packed cell nuclei) and produce cleaner images compared to the model selected by DINO.

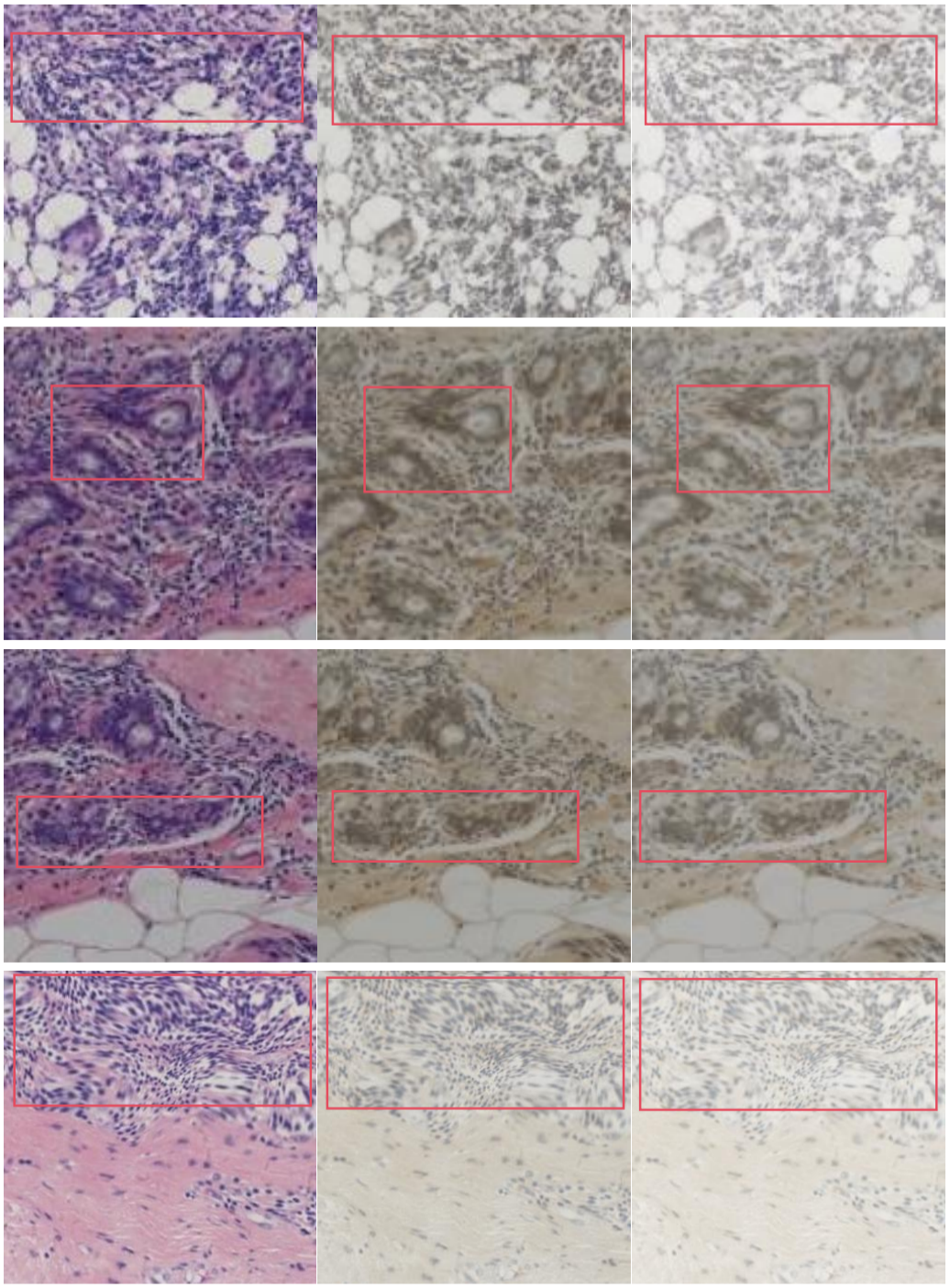


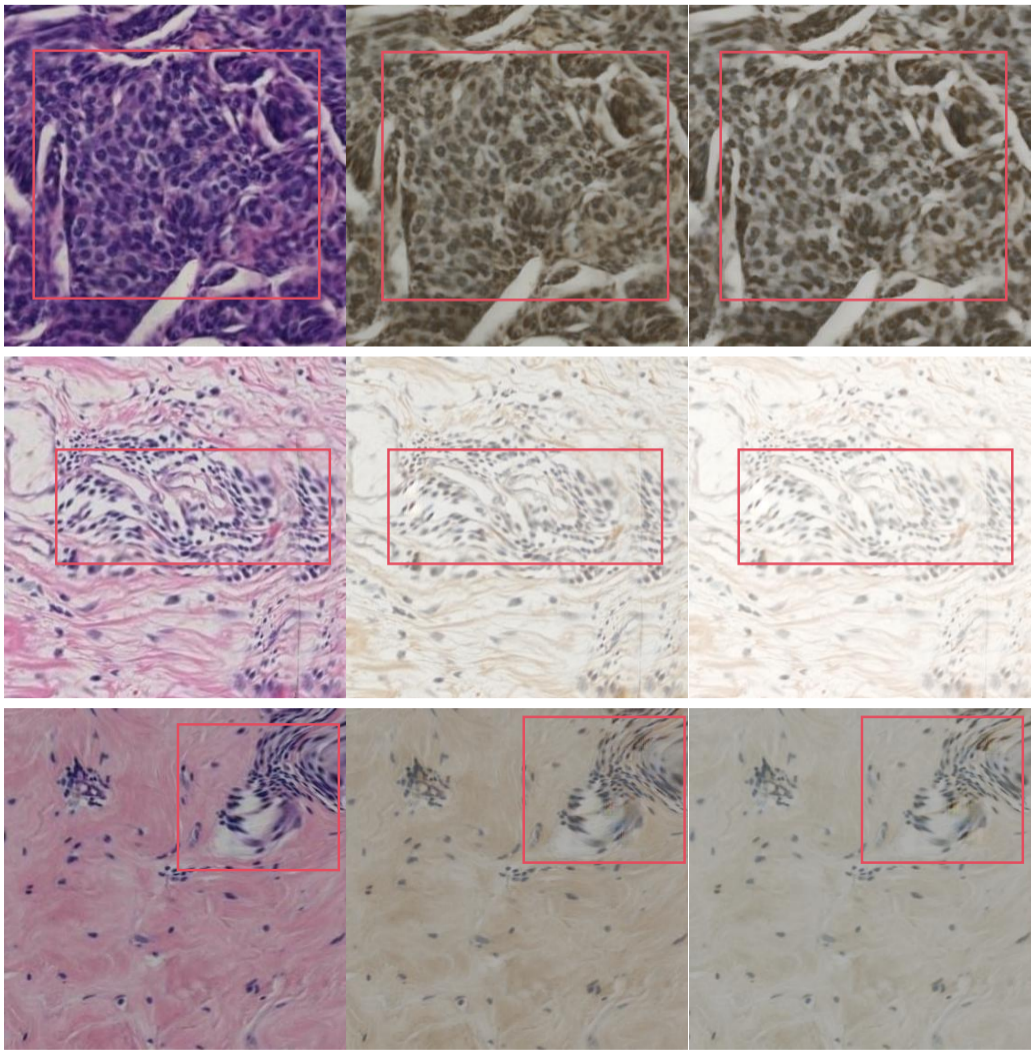




H&E DINO $METRIC_2$







## 6. REFERENCES


[1] Song Wang, Zhong Zhang, Huan Yan, Ming Xu, and Guanghui Wang, "Mix-domain contrastive learning for unpaired h&e-to-ihc stain translation," *CoRR*, vol. abs/2406.11799, 2024.

[2] Jiahan Li, Jiuyang Dong, Shenjin Huang, Xi Li, Junjun Jiang, Xiaopeng Fan, and Yongbing Zhang, "Virtual immunohistochemistry staining for histological images assisted by weakly-supervised learning," in *CVPR*, 2024, pp. 11259–11268.

[3] Georg Wölflein, In Hwa Um, David J. Harrison, and Ognjen Arandjelovic, "Hoechstgan: Virtual lymphocyte staining using generative adversarial networks," in *WACV*, 2023, pp. 4986–4996.

[4] Bowei Zeng, Yiyang Lin, Yifeng Wang, Yang Chen, Jiuyang Dong, Xi Li, and Yongbing Zhang, "Semi-supervised PR virtual staining for breast histopathological images," in *MICCAI*, 2022, vol. 13432, pp. 232–241.

[5] Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash C. Kak, "Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs," in *MICCAI*, 2023, vol. 14225, pp. 632–641.

[6] 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 *MICCAI*, 2022, vol. 13432, pp. 356–365.

[7] 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 Trans. Medical Imaging*, vol. 40, no. 8, pp. 1977–1989, 2021.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel, "Splicing vit features for semantic appearance transfer," in *CVPR*, 2022, pp. 10748–10757.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[12] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020*, 2020, pp. 319–345.

[13] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[15] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee, "Gligen: Open-set grounded text-to-image generation," in *CVPR*, 2023, pp. 22511–22521.

[16] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu, "One-step image translation with text-to-image models," *arXiv preprint arXiv:2403.12036*, 2024.

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[20] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[22] Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin, "Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix," in *CVPR Workshops*, June 2022, pp. 1815–1824.