

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

From H&E to IHC: AI-driven Virtual Immuno Staining

José Miguel Pires Teixeira



Mestrado em Bioengenharia

Supervisor: Professor Jaime dos Santos Cardoso

Co-Supervisor: Sara Pires de Oliveira, PhD

July 31, 2025

From H&E to IHC: AI-driven Virtual Immuno Staining

José Miguel Pires Teixeira

Mestrado em Bioengenharia

Faculdade de Engenharia da Universidade do Porto

July 31, 2025

Resumo

Câncer da mama é o câncer mais letal em mulheres e a detecção precoce é crucial para solucionar este problema. Em patologia, as biópsias tumorais são avaliadas não somente com hematoxilina e eosina (H&E), mas também com coloração Imunohistoquímica (IHQ), cuja avaliação desempenha um papel fulcral no processo de diagnóstico. Contudo, esta técnica é morosa e dispendiosa. Com o advento da era digital, as tradicionais amostras em lâminas de vidro são agora digitalizadas em ficheiros de imagem de alta resolução, facilitando abordagens de análise computacional recorrendo à IA, nomeadamente coloração virtual. Utilizando modelos generativos, esta técnica simula coloração histológica ao converter entre domínios de imagem. Reduzindo custos e trabalho laboratorial, esta técnica promete a diminuição do tempo de diagnóstico.

Investigação em coloração virtual de H&E para IHQ em biomarcadores mamários é um campo de estudo recente e emergente, com 64% dos estudos existentes publicados em 2024. Por conseguinte, o principal objetivo desta dissertação é estudar o atual estado da arte (EDA) e abordar alguns dos problemas que têm impedido o seu desenvolvimento. Esta dissertação apresenta o primeiro estudo abrangente em coloração virtual de H&E para IHQ em biomarcadores clínicos de câncer da mama, incluindo uma revisão aprofundada do EDA e uma avaliação comparativa de modelos representativos do campo de estudo.

Em coloração virtual assumimos ser possível prever IHQ ao nível celular através da morfologia da H&E. Ademais, a maioria dos estudos utiliza conjuntos de dados de pares H&E-IHQ de secções contíguas de tecido, assumindo ainda que estas partilham a mesma informação patológica. Deste modo, impõem-se dificuldades de treino, para as quais os autores desenvolvem modelos complexos usando GANs, mas ignorando a importância da componente adversarial para qualidade das imagens geradas. Além disso, alegam melhoria da EDA com base em métricas clássicas de análise de imagem, as quais arguimos serem limitadas para a avaliação da qualidade das imagens de coloração virtual. Por conseguinte, utilizamos HER2match, um conjunto de dados de pares H&E-IHQ (HER2) da mesma secção de tecido, para o desenvolvimento do modelo CSSP2P GAN. Ao longo do seu desenvolvimento, analisamos o impacto das estruturas da GAN, em particular da componente adversarial, sobre a qualidade das imagens geradas. Ao comparar a qualidade do nosso modelo com modelos do EDA, demonstramos a sua superioridade e sublinhamos as limitações das métricas atualmente usadas. Ademais, a avaliação patológica das imagens geradas pelo nosso modelo afirma a sua autenticidade e consistência patológica. Por fim, reiteramos o desalinhamento ao nível do píxel entre pares de imagens como um dos maiores desafios a ultrapassar e sugerimos potenciais direções de pesquisa futura.

Citando Max Planck, “Conhecimento deve preceder aplicação”. Por conseguinte, pretendemos que esta tese contribua para uma investigação futura fundada, a fim de ultrapassar os desafios da atualidade e responder à questão crítica: É possível prever biomarcador ao nível celular a partir da morfologia da H&E?

Palavras-chave: Patologia Computacional, Coloração Virtual, Modelos Generativos, Câncer da Mama, HER2

Abstract

Breast cancer (BC) is the leading cause of cancer-related deaths among women, and early detection plays a crucial role in addressing this issue. In addition to evaluating tumor biopsies using hematoxylin and eosin (H&E) staining, the traditional pathology workflow uses Immunohistochemistry (IHC), and its evaluation is a critical part of the diagnostic process. However, IHC staining is a laborious and expensive technique. With the advent of digital pathology, traditional glass slides are now digitized into high-resolution whole-slide images, facilitating computational analysis approaches. In particular, virtual staining is an AI-driven image-to-image translation task based on deep generative models that simulates histological staining by converting from one image domain (*e.g.*, H&E) to another (*e.g.*, IHC). Reducing laboratory workload and costs, this technique promises to reduce the turnaround time for breast cancer diagnosis.

Virtual staining from H&E to IHC breast biomarkers is a recent and emerging research field, with 64% of published studies released in 2024. The primary objective of this thesis is to examine the current state-of-the-art (SOTA) on this timely topic and to address some of the issues that have hindered further progress. We provide the first comprehensive overview of virtual staining from H&E to IHC for clinical biomarkers in breast cancer, including a review of the SOTA and a benchmarking study with a unified performance comparison of representative models.

When training a virtual staining model, we assume that cell biomarker expression can be predicted from H&E morphology. However, most studies rely on datasets of H&E-IHC pairs from consecutive tissue sections, further assuming that contiguous tissue sections retain the same diagnostic information. Recognizing the training challenges, most authors develop complex models based on conditional Generative Adversarial Networks, but ignore the importance of adversarial loss in improving the quality of virtual staining. Furthermore, overlooking the issues of model evaluation, they claim improved performance based on classic image metrics, which we argue are not sufficiently robust to evaluate the quality of virtual staining. Therefore, we use the HER2match dataset, the first high-quality dataset of H&E-IHC (HER2) pairs from the same tissue section, to develop CSSP2P GAN. Through its iterative development, we analyze various aspects of the GAN framework and underscore the importance of adversarial loss in improving the quality of virtual staining. While comparing CSSP2P GAN with representative models in the field, we underscore the limitations of current evaluation metrics and demonstrate its superior performance. A qualitative evaluation further reinforces the realism and fidelity of our generated images. Finally, we identify pixel-level misalignment as a major challenge in both training and evaluation, suggesting potential directions for future research to overcome these issues.

As Max Planck aptly said, “Insight must precede application.” Therefore, we aimed for this thesis to contribute valuable insights to future research, advancing the field of virtual staining and addressing the critical question: Can cell biomarker expression be predicted from H&E morphology?

Keywords: Computational Pathology, Virtual Staining, Deep Generative Models, Breast Cancer, HER2

UN Sustainable Development Goals

SGD	Target	Contribution	Performance Indicators and Metrics
3	3.4	Virtual staining improves breast cancer diagnostics, enabling earlier detection and reducing the burden of breast cancer-related deaths	Reduction in breast cancer diagnostic turnaround time.
	3.8	Virtual staining reduces breast cancer diagnosis cost and time, improving accessibility in countries with limited healthcare resources. It supports equitable access to essential health services, contributing to universal health coverage.	The number of health centers adopting virtual staining technology, and the resulting increase in the percentage of patients receiving timely breast cancer diagnosis and treatment.
12	12.4	Virtual staining reduces the need for expensive chemicals and minimizes chemical waste compared to traditional immunohistochemistry	Reduction in chemical usage per diagnostic test and decreased hazardous waste generated by pathology laboratories.
17	17.6	Virtual staining promotes collaboration between AI, medical research, and healthcare sectors, fostering partnerships for the development and dissemination of innovative diagnostic tools.	Number of cross-sector partnerships (between research institutions, healthcare providers, and tech companies) formed to advance virtual staining technology and its adoption.

Agradecimentos

Tenho muito a agradecer à Sara e ao Pascal, sem os quais este trabalho não teria sido possível, ao professor Jaime e à equipa que me acompanhou. Integrar esta equipa proporcionou-me desafios que me permitiram aprender, crescer e atingir objetivos que acredito que irão contribuir significativamente para a comunidade científica.

Não menos importante, foi o apoio dos meus pais e da Adriana, que tanto me suportaram e alegraram, tornando possível esta aventura.

Dedicação e trabalho não teriam sido suficientes não fosse o esforço conjunto destas pessoas.

José Teixeira

“Insight must precede application.”

Max Planck

Contents

1	Introduction	1
1.1	Context and Background	1
1.2	Motivation and Objectives	2
1.3	List of Publications	2
1.4	Document Structure	3
2	Breast Cancer and Virtual Staining	4
2.1	Computational Pathology	4
2.2	Breast cancer and clinical IHC biomarkers	5
2.3	Virtual Staining	7
2.4	Summary	7
3	Deep Generative Models	9
3.1	Deep Generative Models	9
3.1.1	Generative Adversarial Networks (GANs)	9
3.1.2	Contrastive Learning	12
3.1.3	Diffusion models	13
3.2	Model Evaluation	14
3.2.1	Image Metrics	14
3.2.2	Pathology metrics and downstream tasks	18
3.2.3	Qualitative evaluations	19
3.3	Summary	19
4	State-of-the-Art Review	20
4.1	State-of-the-Art	20
4.2	Benchmarking	35
4.3	Limitations in the State-of-the-Art	40
4.4	Summary	42
5	Model Development	43
5.1	Implementation Details	43
5.1.1	Discriminator Architecture	43
5.1.2	Generator Architecture	44
5.1.3	Training Methodology	45
5.1.4	Model Checkpoint and Evaluation	48
5.1.5	Dataset	50
5.2	BCE GAN	50
5.2.1	Training Methodology	50
5.2.2	Adversarial Learning in improving Perceptual Quality	51
5.2.3	Effect of conditioning the Discriminator on the H&E	54
5.3	CSS GAN	54
5.3.1	Training Methodology	55
5.3.2	CSS for Morphology Preservation	56
5.4	CSSP2P GAN	59

5.4.1	Training Methodology	59
5.4.2	Results on the Pyramidal Strategy	60
5.4.3	Efficacy of CSS Loss	60
5.4.4	Experiments in Adaptive Loss Weighting	64
5.5	OkLAB GAN	66
5.5.1	Training Methodology	66
5.5.2	OkLAB Loss for Accurate Positive DAB Staining	72
5.6	Summary	75
6	Model Evaluation	76
6.1	Model Comparison	76
6.1.1	Methodology	76
6.1.2	Results and Discussion	76
6.2	Qualitative Evaluation	79
6.2.1	Methodology	79
6.2.2	Results and Discussion	80
6.3	Summary of Results	81
7	Conclusion	82
7.1	Future Work and Final Remarks	82
	References	85
8	APPENDIX	97
8.1	Summary of the Model Development	97
8.2	Architectures	100
8.2.1	Discriminator Architecture	100
8.2.2	Generator Architecture	101

List of Figures

2.1	Interpretation of commonly used immunohistochemistry (IHC) markers in breast cancer.	6
3.1	Generative adversarial networks basic structures.	10
3.2	Illustration of CycleGAN framework.	12
3.3	Illustration of contrastive unpaired translation framework.	13
3.4	Illustration of diffusion model fundamental concept.	14
4.1	PRISMA flowchart for the selection process of studies included in the review.	21
4.2	Benchmarking results examples from the MIST dataset.	38
4.3	Benchmarking generalization results examples from the BCI dataset.	39
5.1	Illustration of the training step.	47
5.2	Loss and accuracy curves during GAN training.	49
5.3	Example tile from HER2match dataset	50
5.4	Comparison between BCE ncGAN and U-Net on a high resolution example output.	51
5.5	Qualitative assessment of the BCE ncGAN, BCE GAN, and U-net models on the validation set of the HER2match dataset.	52
5.6	Qualitative comparison between the BCE ncGAN, BCE GAN, and U-Net on cases with a 3+ HER2-score and intense positive DAB stain.	53
5.7	Qualitative comparison between the BCE ncGAN, BCE GAN, and U-Net on cases without positive DAB stain.	53
5.8	Example outputs of the BCE ncGAN and BCE GAN models for the 1+ HER2-score image example from the validation set.	54
5.9	Illustration of the integration of CSS loss in the BCE GAN framework.	55
5.10	Heatmap of the grayscale CSS map.	55
5.11	Qualitative assessment of the BCE GAN and CSS GAN models on the validation set of the HER2match dataset.	57
5.12	Qualitative comparison between the BCE GAN and CSS GAN on cases with a 3+ HER2-score and intense positive DAB stain.	58
5.13	Qualitative comparison between the BCE GAN and CSS GAN on cases without positive DAB stain.	58
5.14	Illustration of the Pyramid $\mathcal{L}_1$ loss function.	59
5.15	Qualitative comparison between the CSSP2P GAN and P2P GAN on cases with a 3+ HER2-score and intense positive DAB stain.	62
5.16	Qualitative comparison between the CSSP2P GAN and P2P GAN on cases without positive DAB stain.	62
5.17	Qualitative assessment of the BCE GAN, CSS GAN, and CSSP2P GAN models	63
5.18	Evolution of the adversarial loss component during training for SoftAdapt and NoSoftAdapt.	66
5.19	Example of an estimated mask for an image with a 3+ HER2-score.	67
5.20	Illustration of the DAB channel histograms for different HER2-scores.	68
5.21	Illustration of the OkLAB loss (methodology).	71
5.22	Qualitative assessment of the CSSP2P GAN, OkLAB GAN, masked CSSP2P GAN, and masked OkLAB GAN on cases with a 3+ HER2-score and intense positive DAB staining.	73

5.23	Qualitative comparison between the CSSP2P GAN, OkLAB GAN, masked CSSP2P GAN, and masked OkLAB GAN performance on cases without positive DAB stain.	74
6.1	Qualitative assessment of pyramid Pix2Pix, ASP, and CSSP2P GAN models on the test set of the HER2match dataset.	78
6.2	Example figure of the display window presented to pathologists during the qualitative evaluation.	79
8.1	Summary of the discriminator architecture.	100
8.2	Summary of the generator architecture.	101

List of Tables

4.1	Summary of the state-of-the-art models on virtual staining (H&E to IHC) of the clinical biomarker panel for breast cancer (from Klockner et al. [27]). Q.E.: qualitative evaluation.	22
4.2	Performance metrics on the MIST test set (used for validation) comparing Pyramid Pix2Pix, ASP, and PSPStain.	37
4.3	Performance metrics for generalization on the BCI test set (HER2 staining) comparing Pyramid Pix2Pix, ASP, and PSPStain.	37
5.1	Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN, BCE ncGAN and U-Net models.	51
5.2	Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN and CSS GAN models.	57
5.3	Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN, CSS GAN, and P2P GAN models.	60
5.4	Performance metrics on the validation set of the HER2match dataset comparing the CSSP2P GAN and P2P GAN models.	61
5.5	Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN, CSS GAN, and CSSP2P GAN models.	63
5.6	Performance metrics on the validation set of the HER2match dataset comparing different SoftAdapt implementation method	65
5.7	Performance metrics on the validation set of the HER2match dataset for the CSSP2P GAN, OkLAB GAN, masked CSSP2P GAN, and masked OkLAB GAN models.	73
6.1	Performance metrics on the test set of the HER2match dataset for the Pyramid Pix2Pix, ASP, and CSSP2P GAN models.	77
6.2	Qualitative evaluation consensus results.	80
8.1	Summary of the Model Development	97

List of Acronyms

ASP	Adaptive Supervised PatchNCE
BC	Breast Cancer
BCI	Breast Cancer Immunohistochemical dataset
CSS	Contrast-Structure Similarity
cGAN	conditional GAN
CUT	Contrastive Unpaired Translation
DAB	3,3'-Diaminobenzidine
DGMs	Deep Generative Models
DMs	Diffusion Models
ER	Estrogen Receptor
FID	Fréchet Inception Distance
GANs	Generative Adversarial Networks
H&E	Hematoxylin and Eosin
HED	Hematoxylin, Eosin, DAB
HER2	Human Epidermal Growth Factor Receptor 2
HR	Hormone Receptor
IHC	Immunohistochemistry
ISH	<i>in situ</i> Hybridization
JSD	Jensen-Shannon Divergence
KID	Kernel Inception Distance
LPIPS	Learned Perceptual Image Patch Similarity
MAE	Mean Absolute Error
MIST	Multi-IHC Stain Translation dataset
MMD	Maximum-Mean Discrepancy
MSE	Mean Squared Error
MS-SSIM	Multi-Scale Structural Similarity Index Measure
OD	Optical Density
PR	Progesterone Receptor
PRN	Pathological Representation Network
SP	Supervised PatchNCE
SOTA	State-of-the-Art
SSIM	Structural Similarity Index Measure
WHO	World Health Organization
WSI	Whole Slide Image

List of Notation

G	Generator
D	Discriminator
S	Source domain
T	Target domain
B	Batch dimension
C	Channel dimension
$H \times W$	Height and Width
$\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$	Input image
$\mathbf{Y} \in \mathbb{R}^{B \times C \times H \times W}$	Ground truth image
$\hat{\mathbf{Y}} \in \mathbb{R}^{B \times C \times H \times W}$	Predicted image
$\hat{\mathbf{Y}}_D \in [0, 1]^{B \times I \times H \times W}$	Discriminator's predicted probability
$\mathbf{Y}_D \in \{0, 1\}^{B \times I \times H \times W}$	Discriminator's label
η_t	Learning rate at epoch t
$\odot$	Hadamard (element-wise) product
$\times$	Matrix cross product
$\cdot$	Scalar, Vector, and Matrix multiplication
$*$	Convolution

Chapter 1

Introduction

1.1 Context and Background

Breast cancer (BC) is a heterogeneous disease with variability in presentation, morphology, molecular characteristics, clinical behavior, and therapeutic response [1], [2]. In 2022, 2.3 million women were diagnosed with BC, and there were 670,000 breast cancer-related deaths worldwide [3]. As the leading cause of cancer-related death among women, BC presents an increasing global challenge, intersecting with health, gender, socioeconomic, and equity concerns [4].

Pathology, the branch of medical science focused on understanding the causes, development, structural, and functional changes of diseases [5], is pivotal in the evaluation of BC cases. In the pathological workflow, contiguous breast tumor sections are cut and analyzed using hematoxylin and eosin (H&E) staining and Immunohistochemistry (IHC). H&E staining allows for the examination of the general characteristics of tumor tissue and is commonly used in routine practice, while IHC is crucial for diagnosis and prognosis. Consequently, a combination of IHC stains, including those for estrogen and progesterone receptors (ER, PR), Human Epidermal Growth Factor Receptor 2 (HER2), and Ki67, is commonly used. Evaluation of the expression of these markers is a crucial component of the diagnostic process for all BC patients, in particular HER2, which numerous studies have shown to hold prognostic and predictive significance [6]. However, IHC staining is a labor-intensive and expensive technique.

With the advent of digital pathology, traditional glass slides are now scanned into high-resolution whole-slide images (WSI). This shift to the digital era has become the new standard of care, creating opportunities to develop AI-based computational approaches for automatic image analysis, including tasks such as tumor detection, grading, segmentation, survival prediction, and the discovery of novel diagnostic biomarkers [7], [8].

In particular, virtual staining offers a promising solution for simulating histological staining on screen. This method is an AI-driven image-to-image translation approach based on deep generative models that map one image domain (the source, *e.g.* H&E) to another (the target, *e.g.* IHC). By enabling the faster generation of (digital) IHC from H&E slides, this technique reduces laboratory workload, costs, and reagent consumption, while preserving tissue specimens and improving

overall resource efficiency. More importantly, virtual staining promises to reduce the turnaround time for BC diagnosis [9].

1.2 Motivation and Objectives

Virtual staining promises to revolutionize the current standard of care in BC and beyond. Although we recognize the significant progress that has been made, this is a recent field of investigation, and we still observe a general lack of insight that hinders the clinical application of virtual staining. Therefore, while studying the current state-of-the-art (SOTA) in virtual staining for BC and addressing some of its limitations, this thesis aims to expand knowledge and contribute to this timely field.

As a result, this thesis can be viewed as a two-part study. The first part provides a broad overview of the current SOTA, introducing the commonly used deep generative frameworks and evaluation methodologies in virtual staining. Subsequently, we offer a systematic review of the current SOTA and a benchmarking study, presenting a unified performance comparison of representative models in the field using two widely used datasets. We then summarize the accomplishments that have been achieved so far and discuss the current limitations. Therefore, it provides a valuable resource for researchers and clinicians working on virtual staining technology for BC diagnostics.

Acknowledging the limitations of the current SOTA, the second part of this thesis focuses on developing our virtual staining model, CSSP2P GAN, using the recently released HER2match dataset [10]. We compare its performance with representative models and conduct a blind qualitative evaluation of its results. As we iteratively develop our framework, we explore the impact of each component, with particular emphasis on the role of adversarial learning in improving the quality of virtually stained images. Additionally, we highlight the existing gap in the SOTA regarding the development of a more robust, pathologically oriented metric for virtual staining, as well as the need for standardization of evaluation methods. As a result, in addition to developing a pathologically consistent virtual staining model, we aim to build knowledge and contribute to the future, insightful development of the field.

1.3 List of Publications

The conducted study in this thesis contributed to the following publications:

- Article in International Conference Proceedings:
Teixeira J., Klöckner P., Montezuma D., Fraga J., Horlings H.M., Cardoso J.S. and Oliveira S.P. “Leveraging Adversarial Learning for Pathological Fidelity in Virtual Staining.” *In Deep Generative Models: 5th MICCAI Workshop - DGM4MICCAI*, 2025. (under review)
- Article in International Journal
Klöckner P., Teixeira J., Montezuma D., Fraga J., Horlings H.M., Cardoso J.S. and Oliveira

S.P. “H&E to IHC virtual staining methods in breast cancer: An overview and benchmarking.” *npj Digital Medicine*, 2025.

- Abstract in International Conference

Teixeira J., Klöckner P., Montezuma D., Horlings H.M., Cardoso J.S. and Oliveira S.P. “H&E to IHC virtual staining in breast cancer: methods, benchmarking and challenges.” *In 21st European Congress on Digital Pathology (ECDP)*, 2025.

1.4 Document Structure

In addition to this introduction, which describes the subject under study and summarizes the objectives and contributions of the present dissertation, the remainder of this document is composed of six chapters.

Chapter 2 provides an overview of breast cancer and introduces virtual staining, emphasizing its potential to reduce diagnostic turnaround time.

Chapter 3 presents the deep generative frameworks and evaluation methods commonly used in virtual staining.

Chapter 4 provides a systematic review and benchmarking of seminal models in virtual staining for H&E to IHC breast cancer biomarkers.

Chapter 5 details the development of our proposed GAN model on the HER2match dataset, evaluating the role of adversarial loss in image quality and discussing the limitations of existing evaluation metrics.

Chapter 6 compares the performance of the developed model with re-trained baselines on the same dataset and provides a qualitative evaluation of our model.

Chapter 7 concludes the dissertation by summarizing the main findings and contributions, discussing ongoing challenges, and outlining directions for future work.

Chapter 2

Breast Cancer and Virtual Staining

In this chapter, we begin with an introduction to pathology, followed by a description of the traditional pathology workflow and an introduction to the concept of computational pathology. Next, we discuss the background of breast cancer and the clinical importance of immunohistochemistry biomarkers. Finally, we introduce the virtual staining method.

2.1 Computational Pathology

Pathology is the branch of medical science that focuses on understanding the causes, development, structural, and functional changes of diseases [5]. Cancer is a heterogeneous disease with significant variability in its biological and clinical manifestations; therefore, pathology plays a central role in its diagnosis by conducting detailed assessments of tissue morphology and evaluating molecular and immunohistochemical markers [11]. These procedures are also essential for patient care, providing a comprehensive analysis of biopsy and surgical specimens to determine tumor subtypes, grades, and severity. The traditional pathology workflow involves cutting, staining, and mounting thin tissue sections of tumor samples onto glass slides. Staining allows for highlighting the structures of interest [11], with the standard technique using a combination of hematoxylin and eosin (H&E) to stain the nuclei and cytoplasm of cells, respectively. Immunohistochemistry (IHC) is a more advanced staining technique to detect specific tissue proteins through a series of antibody and enzymatic reactions. A primary antibody binds to the target antigen, followed by a secondary antibody conjugated to an enzyme that reacts with a chromogenic substrate, such as 3,3'-Diaminobenzidine (DAB), forming an insoluble precipitate (brown in the case of DAB). The high contrast between the precipitate and a counter-stain, usually hematoxylin, emphasizes the location of the target antigen [12], [13]. It is also important to note that IHC is typically performed on consecutive tissue sections, following H&E staining.

Digital pathology refers to the practice of pathology through digital imaging, where traditional glass slides are scanned into high-resolution whole-slide images (WSI) [14]. This transition to the digital era has become the new standard of care, creating the opportunity to develop AI-driven computational analysis, such as tumor detection, grading, segmentation, survival prediction, and

biomarker discovery [7], [8], [14]. By integrating AI and WSI analysis techniques, computational pathology has been transforming pathology practice, and is expected to play a central role in the future of pathology [8], [14].

2.2 Breast cancer and clinical IHC biomarkers

BC presents an increasing global challenge, intersecting with health, gender, socioeconomic, and equity concerns [4]. In 2022, 2.3 million women were diagnosed with BC, and there were 670,000 breast cancer-related deaths worldwide [3]. It remains the leading cause of cancer-related deaths among women, with the burden of preventable deaths disproportionately impacting low- and middle-income countries. BC is a heterogeneous disease with variability in presentation, morphology, molecular characteristics, clinical behavior, and therapeutic response [1], [2]. Tumor grading evaluates differentiation and proliferative activity, while staging determines the extent of disease spread. Both are essential for case evaluation, prognosis, and treatment planning. Additionally, invasive BC is broadly classified into four molecular subtypes: luminal A, luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, and basal-like [1], [2]. Further research has identified additional subclasses within these groups [2], [15]. Nonetheless, the determination of molecular subtypes is not currently considered critical in routine clinical practice, as treatment decisions primarily rely on clinico-pathological prognostic and predictive factors, with gene expression signature profiling used only in selected cases [16]–[19]. Specifically, the immuno-expression profile plays a key role in the clinical classification of invasive carcinoma subtypes. Consequently, a combination of IHC stains is routinely employed, including those for estrogen and progesterone receptors (ER, PR), HER2, and, in some cases, Ki67. The various combinations of immuno-expression help identify clinical subgroups. Evaluating the expression of these stains is a crucial component of the diagnostic process for all BC patients and follows standardized protocols (Figure 2.1) and established guidelines [20]–[23]. Additionally, multi-gene prognostic tests, such as OncotypeDX, MammaPrint, and Prosigna, can provide risk stratification in this context [24], [25].

Numerous studies have shown that hormone receptors (HR) and HER2 hold prognostic and predictive significance [6]. In assessing adjuvant therapy for BC, a crucial factor is determining the potential response of the tumor to endocrine treatment or HER2-targeted therapy. ER positivity, observed in 75-80% of patients, is a key predictor for hormone therapy response, while ER-negative tumors do not benefit from this treatment [28], [29]. PgR also provides additional prognostic and predictive value in BC [30], with its expression profile being more variable than ER, and it helps stratify ER-positive cases into different prognostic categories [6]. A 1% cut-off is commonly used to define HR positivity for hormone therapy eligibility, though expression intensity and percentage can vary. BC tumors with strong, diffuse nuclear HR expression generally respond better to endocrine therapy and show more favorable outcomes. HER2 expression, evaluated through IHC and/or *in situ* hybridization (ISH), is crucial in determining the effectiveness of anti-HER2 therapies [31]. While HER2 over-expression was once linked to poor prognosis, the development of targeted therapies has notably improved survival rates in HER2-positive BC [31].

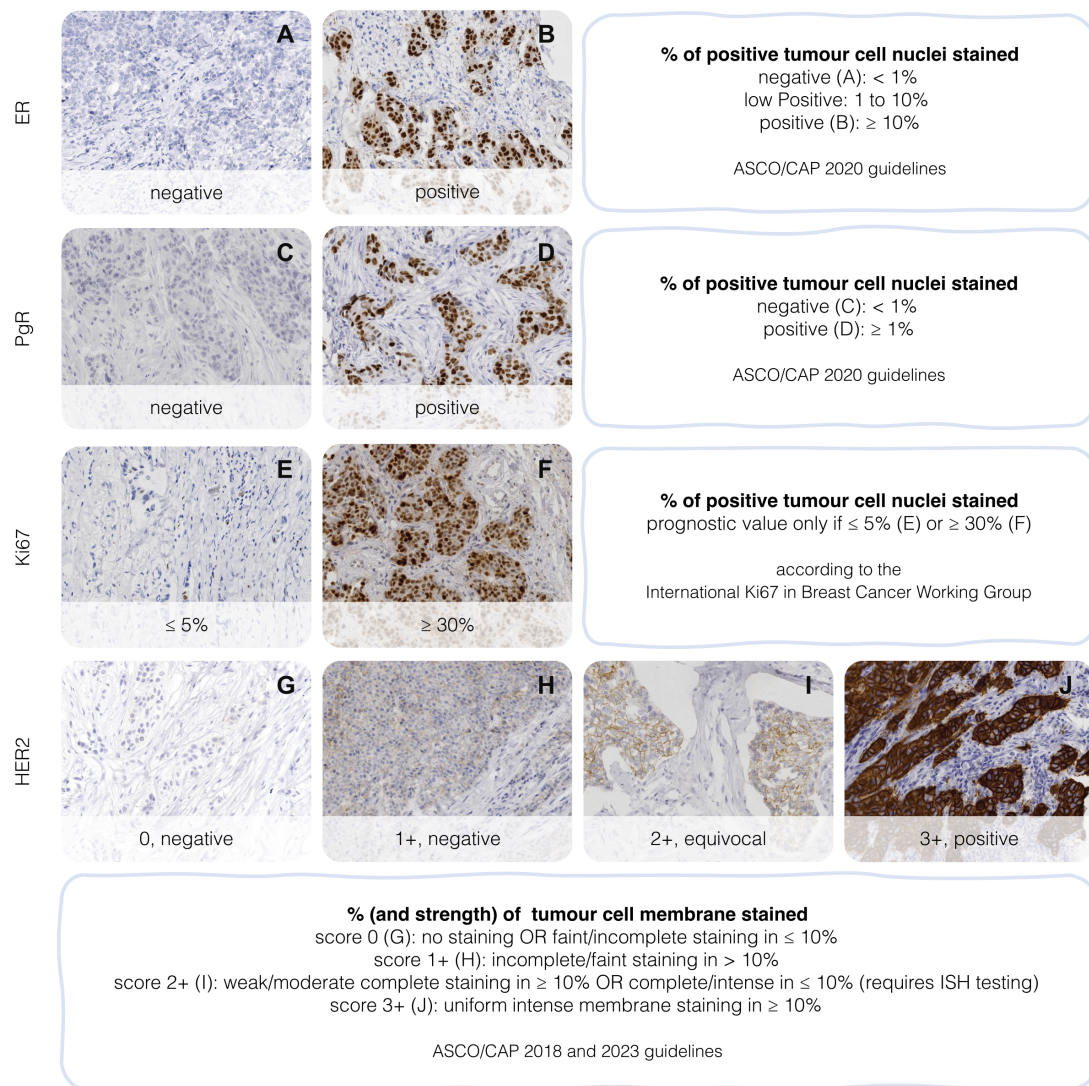


Figure 2.1: Interpretation of commonly used immunohistochemistry (IHC) markers in breast cancer: **A, B** Estrogen receptor (ER); **C, D** Progesterone receptor (PgR); **E, F** Ki67; **G-J** Human epidermal growth factor receptor 2 (HER2). For HER2, an IHC 2+ score requires confirmation by *in situ* hybridisation (ISH), and an IHC 1+ or 2+ may be reported as “HER2 Low” and are eligible for clinically appropriate HER2-targeted therapy, according to ASCO/CAP 2023 Guideline update. Images adapted from the MIST test set [26]. (from Klöckner *et al.* [27])

Ki67, a biomarker of cell proliferation, is recognized by the Breast Cancer Working Group as a prognostic marker for BC. However, its clinical utility is primarily limited to ER+/HER2- patients, where it aids in stratifying those who would benefit from adjuvant chemotherapy [32]. Despite this, the validity of Ki67 remains uncertain due to inconsistent scoring methods, unclear cut-off thresholds, and concerns about the reliability of different antibodies [22].

Finally, it is important to note that the prognostic evaluation of BC extends beyond IHC/molecular panels. Factors such as histologic subtype, tumor stage, lymphovascular invasion, margin status, and other clinico-pathologic characteristics are crucial in guiding patient management and treatment.

2.3 Virtual Staining

Traditional pathology practice has been transformed by the advent of computational pathology with the integration of AI and WSI analysis techniques [8]. In particular, virtual staining offers a promising approach for simulating histological staining on screen [9]. It reduces laboratory workload, costs, and reagent consumption, while preserving tissue specimens and improving overall resource efficiency. According to the World Health Organization (WHO), tackling the three fundamental pillars of the Global Breast Cancer Initiative—early detection, prompt diagnosis, and comprehensive BC management—is estimated to save 2.5 million lives over 20 years [33]. In this context, virtual staining, by enabling rapid and precise tissue analysis without traditional IHC processes [34], holds significant potential for improving the prompt diagnosis of BC, thus supporting the goals of WHO’s Global Breast Cancer Initiative. Additionally, virtual staining enables the customization of staining patterns, allowing for the visualization of multiple biomarkers within the same tissue sample and revealing details that may be difficult to detect with traditional staining methods [35]. It is expected to play an increasingly significant role in research, diagnostics, and personalized medicine, ultimately enhancing patient care, particularly in resource-limited settings. From an AI perspective, virtual staining can be viewed as an image-to-image translation task, where the objective is to map from one image domain (source, *e.g.*, H&E) into another (target, *e.g.*, IHC). Deep generative models are commonly used to address this complex task. Specifically, these algorithms learn the underlying structure of the training data, enabling them to generate new samples that closely resemble the original data distribution. It is worth mentioning that these models can be trained on both paired and unpaired data. While paired data requires corresponding samples from both domains, unpaired training only requires a dataset from each domain. Although paired approaches typically offer better performance, unpaired techniques may be more suitable for medical images, where paired data is often not readily available [36].

2.4 Summary

Breast cancer (BC) is the leading cause of cancer-related deaths among women. In the traditional pathology workflow, thin BC tissue sections are cut, stained, and mounted onto glass slides. While

H&E staining is regularly used to evaluate general characteristics of the tissues, IHC staining plays a crucial role in the prognosis and diagnosis of the disease. However, this technique is laborious and expensive.

Computational pathology has been transforming traditional pathology practice by integrating AI and WSI analysis techniques. In particular, virtual staining is an image-to-image translation task that aims to map one domain to another using deep generative models. This process allows for the generation of digital IHC from H&E slides, offering the potential to reduce laboratory workload, cost, and reagent consumption, while preserving tissue specimens. More importantly, virtual staining holds the promise of reducing turnaround time for BC diagnosis.

Chapter 3

Deep Generative Models

In this chapter, we explore the commonly used frameworks for transforming H&E to IHC images in the context of BC biomarkers. Section 3.1 introduces and explores three commonly used deep generative frameworks: GANs, Contrastive Learning, and Diffusion Models. Section 3.2 describes the methods used to evaluate the quality of the generated images, including image metrics, pathology-specific metrics, downstream tasks, and qualitative evaluation.

3.1 Deep Generative Models

The exploration of deep generative models (DGMs) for imaging is a rapidly evolving field in AI research. With new models and methodologies emerging each year, generative image models have matured and are now widely used, including in applications like AI art generators such as DALL-E [37] and Stable Diffusion [38].

In the context of virtual staining, three key model architectures are commonly used: Generative Adversarial Networks (GANs) [39], contrastive learning, and Diffusion Models (DMs) [40]. GANs are the most widely used class of models and can be further categorized into CycleGAN [36] and conditional GAN (cGAN) [41]. Although contrastive learning is not a model architecture per se, approaches combining GANs with contrastive learning strategies are commonly used, such as Contrastive Unpaired Translation (CUT) [42]. Finally, DMs, which have shown superiority over GANs in some fields, are gaining popularity for virtual staining tasks and are expected to see broader adoption in the near future.

3.1.1 Generative Adversarial Networks (GANs)

GANs are an adversarial framework used to train generative models (Figure 3.1a). This framework consists of a generator model G , which learns the data distribution, and a discriminator model D , which estimates the probability that a sample originates from the training data rather than from G [41]. In other words, the generator aims to create realistic images, while the discriminator's task is to distinguish between real training images and those generated by G . Both G and D are trained simultaneously using the following optimization criterion:

$$\min_G \max_D V(D, G) = \mathbb{E}_{y \sim p_{\text{data}}(y)} [\log D(y)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (3.1)$$

where y is a real image of the target distribution, and z is a noise vector.

Essentially, this approach can be understood as a two-player min-max game, often illustrated using the metaphor of an art forger and an appraiser. The forger (generator) strives to create flawless counterfeit artworks, while the appraiser (discriminator) works to distinguish between genuine and forged pieces. After each round, both the forger and the appraiser receive feedback, enabling the forger to refine their counterfeiting skills and the appraiser to enhance their detection methods for subsequent rounds. Through this competitive strategy, the generator can produce highly realistic results, as long as both the generator and discriminator continue to improve throughout the training process and learn at a similar pace.

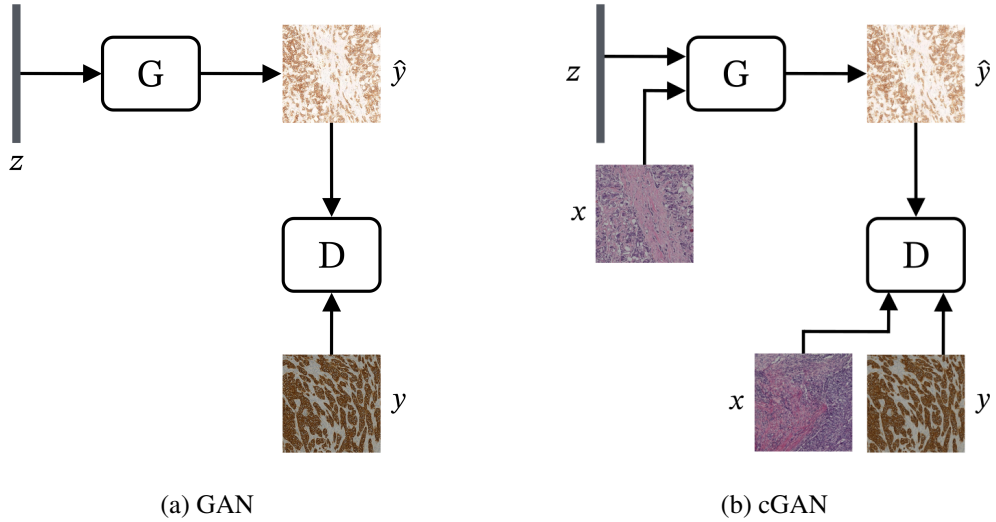


Figure 3.1: Generative adversarial networks basic structures, where z is a noise vector, x is a real image from the source distribution, y is a real image from the target distribution, $\hat{y}$ is the generated image, D is the discriminator and G is the generator. Example tiles adapted from the BCI dataset [43]. (from Klöckner *et al.* [27])

The original GAN framework generates new data from noise. However, in image-to-image translation, the goal is to transform an image from a source domain (*e.g.*, H&E images) to a target domain (*e.g.*, IHC images), rendering the original GAN framework unsuitable for this task. Consequently, alternative GAN-based frameworks accept an image as input to the generator, which learns to map it to the target domain. The frameworks most commonly used for this purpose are based on conditional GANs (cGAN) [41] and CycleGAN [36].

The cGAN (Figure 3.1b) extends the GAN framework by adding conditional information, such as the input image, x , to the training objective:

$$\min_G \max_D V(D, G) = \mathbb{E}_{y, x \sim p_{\text{data}}(y, x)} [\log D(y, x)] + \mathbb{E}_{z \sim p_z(z), x \sim p_x} [\log (1 - D(G(z, x), x))] \quad (3.2)$$

where, in the context of paired images (for example well registered H&E and IHC images from consecutive tissue sections) the conditional information could be, for example, the respective source H&E image, x , that is presented to the discriminator alongside the fake, $G(x, z)$, and real, y , IHC images.

One of the most seminal cGAN frameworks for image-to-image translation is Pix2Pix, introduced by Isola *et al.* [44]. In their work, they emphasize the importance of the discriminator loss, which enables the generator to produce sharp and realistic images, as opposed to relying on MSE or $\mathcal{L}_1$ loss as training criterion. While models trained with these losses capture low frequencies, they fail to promote high-frequency sharpness, resulting in blurry and grayish images. Additionally, they introduce the PatchGAN discriminator, which, unlike previous networks that assess whether the entire image is real or fake, classifies each $N \times N$ patch within the image as real or fake. This discriminator models the image as a Markov random field, assuming independence between pixels separated by more than the patch diameter, which has been empirically proven to be more effective than classifying the entire image. Finally, they discuss the significance of conditioning the discriminator to prevent mismatches between the input and output. This approach also helps avoid mode collapse, which occurs when the generator produces the same or similar outputs regardless of the input image. It is also important to note that, unlike previous GAN frameworks, the authors do not pass the noise vector to the discriminator.

A downside of the Pix2Pix framework in the context of medical imaging is that it assumes a proper pairing exists between images of the source and target domains, which is often not the case. Therefore, a model architecture capable of learning from unpaired data, such as CycleGAN (Figure 3.2), is needed in such cases. CycleGAN employs two generators G_{ST} that maps from the source to the target distribution (*e.g.* H&E to IHC) and G_{TS} that maps from the target to the source distribution (*e.g.* IHC to H&E). Additionally, it also utilizes the idea of cycle-consistency loss [36], which ensures that an image from the source domain that has been translated by G_{ST} to the target domain and then translated back to the source domain via G_{TS} (from H&E to IHC and back to H&E), is almost identical to the original input image. By analogy to language translation, where converting an English sentence to German and then back to English should ideally result in the original sentence, CycleGAN enforces this consistency. Since some meaning may be lost in translation, it is logical to incorporate this consistency into the training objective.

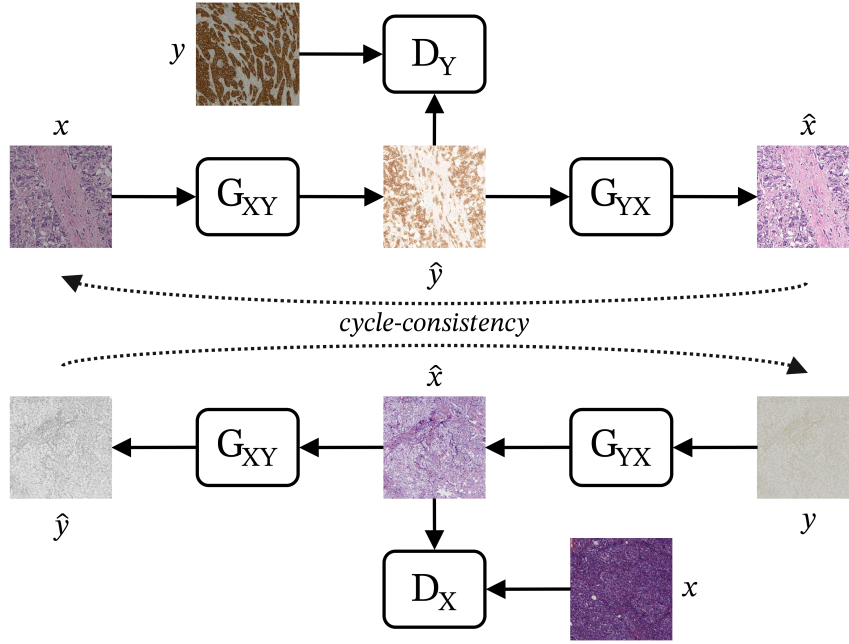


Figure 3.2: CycleGAN framework, where x and y are real images from the source and target distributions, respectively, $\hat{x}$ and $\hat{y}$ are generated images, G_{XY} is the generator from the source to the target domain, G_{YX} the generator from the target to the source domain, D_X the discriminator for the source domain and D_Y the discriminator for the target domain. The cycle-consistency concept ensures that a generated image can be translated back to the source domain via the opposite generator. Example tiles adapted from the BCI dataset [43]. (from Klöckner *et al.* [27])

Although cGANs and CycleGANs describe the high-level architecture, they consist of several distinct modules, including the generator and the discriminator. For the discriminator, the most currently used architecture is PatchGAN, while for the generator, it is common to use a CNN-based model, such as U-Net [45] or ResNet [46], which are typically pre-trained on existing image databases such as ImageNet [47].

3.1.2 Contrastive Learning

Contrastive learning is a discriminative representation learning framework designed to learn meaningful representations by contrasting positive and negative pairs of instances. It typically involves three types of samples: positive examples, anchor examples, and negative examples. The core idea of contrastive learning is to minimize the distance between positive and anchor examples, while maximizing the distance between negative and anchor examples [48]. As a result, contrastive learning transforms the feature space into a metric space where the distance between instances reflects their semantic similarity: similar or matched inputs are placed close together, while dissimilar or unmatched inputs are placed farther apart. This approach enables models to capture relevant features without relying on explicit labels. A common model for image-to-image translation using contrastive learning is the Contrastive Unpaired Translation (CUT) [42] framework. In this approach, a patch from the generated image (anchor) is compared with its corresponding

patch in the input image (positive) and other patches from the same input image (negatives) by computing the distance between the embeddings at different levels of the generator’s encoder. By maximizing the distance between the anchor and the negatives, while minimizing the distance between the anchor and the corresponding positive patch, the model learns to preserve the content of the input image (Figure 3.3). To achieve this, the authors use the PatchNCE loss, which can be combined with the standard adversarial GAN loss to generate images that retain the style of the target domain while preserving the content of the input image.

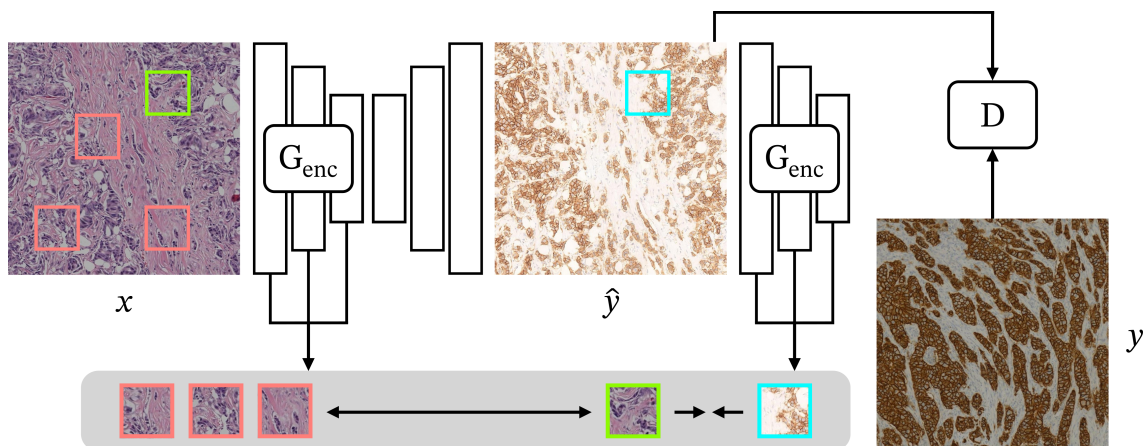


Figure 3.3: Contrastive unpaired translation (CUT) framework, where x and y are real images from the source and target distributions, respectively, $\hat{y}$ is the generated image, G_{enc} is the generator encoder and D is the target discriminator. (from Klöckner *et al.* [27])

An extension of the PatchNCE loss is the Adaptive Supervised PatchNCE (ASP) loss, developed by Li *et al.* [26]. This loss function was specifically designed to train a GAN model for virtual staining of H&E to IHC breast tissue tiles. Unlike PatchNCE, which selects positive and negative examples from the input H&E image, ASP chooses embeddings from the ground-truth IHC image to calculate the supervised PatchNCE (SP) loss. This loss is then weighted by the cosine similarity between the embeddings of each anchor and its respective positive patch in both the generated and real IHC images, resulting in the ASP loss. This weighting strategy aims to reduce the importance of regions where the ground-truth and generated IHC images exhibit significant low consistency. It is important to note that this loss is introduced only at later stages of model training, when a proper distinction between consistent and inconsistent patch locations can be made.

3.1.3 Diffusion models

DMs (Figure 3.4) are a class of latent variable models [49] inspired by diffusion processes in thermodynamics [50]. In the forward diffusion process, an input image is progressively corrupted with Gaussian noise. During the reverse (denoising) process, a neural network (typically U-Net) predicts and removes the added noise, gradually reconstructing the image. This is done by predicting the noise added at each time step and subtracting it from the image at that time step. By learning this reverse process, the model can generate new images from random noise, producing

samples that match the original data distribution. Additionally, the network can be conditioned on modalities such as text prompts or images. In addition to generating images from noise, DMs can also be used for tasks such as super-resolution [51] and style transfer [52]. Recently, these models have gained popularity due to their stability and easier training compared to GANs, while still producing high-quality images [53], [54].

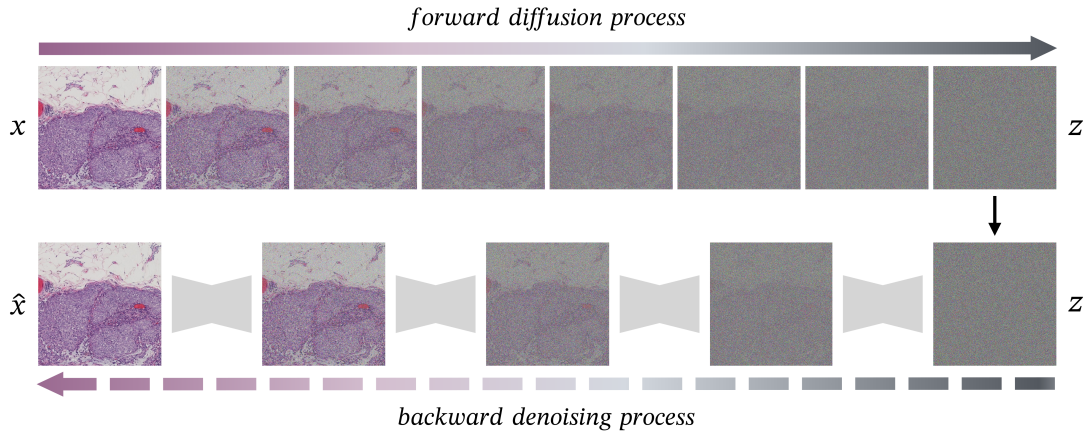


Figure 3.4: Diffusion models’ basic concept, where x is the input image and z is Gaussian noise. The diffusion process iteratively adds noise, while the denoising process aims to reconstruct the original image. Example tile adapted from the BCI dataset [43]. (from Klöckner *et al.* [27])

3.2 Model Evaluation

3.2.1 Image Metrics

The most straightforward way to evaluate the similarity between two images is to compare them pixel by pixel. This can be done by subtracting the pixel values of the generated image from those of the ground-truth image and summing the absolute values. This sum is referred to as $\mathcal{L}_1$ (Equation 3.3), and when divided by the total number of pixels, it becomes the mean absolute error (MAE), defined in Equation 3.5. If, instead of summing the absolute values, we sum the square roots of the squared differences between pixels, we obtain the $\mathcal{L}_2$ measure, also known as the Euclidean distance (Equation 3.4). When averaged across all squared distances, this becomes the mean squared error (MSE), as in Equation 3.6.

$$L_1(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{i=1}^N |\mathbf{y}_i - \hat{\mathbf{y}}_i| \quad (3.3)$$

$$L_2(\mathbf{Y}, \hat{\mathbf{Y}}) = \sqrt{\sum_{i=1}^N (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2} \quad (3.4)$$

$$\text{MAE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (3.5)$$

$$\text{MSE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (3.6)$$

Peak Signal-to-Noise Ratio (PSNR) is a metric used to quantify the ratio between the maximum possible power of a signal (the original image) and the power of corrupting noise (the error between the generated and reference images). It is expressed in decibels and is mathematically derived from the MSE as follows:

$$\text{PSNR}(\mathbf{Y}, \hat{\mathbf{Y}}) = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \quad (3.7)$$

where MAX is the maximum value a pixel can attain (*e.g.* 255 in an 8-bit image). A higher PSNR value indicates a better correspondence between the generated and reference images, approaching infinity as the MSE approaches 0.

MAE, MSE, and PSNR are easy-to-compute metrics; however, they do not always align with human visual perception. As a result, higher PSNR values and lower MAE and MSE values do not always correlate with perceived image quality. Additionally, these metrics are highly sensitive to pixel-level differences, making them less suitable for evaluating scenarios like virtual staining, where the ground truth and source images often lack strong pixel-to-pixel correspondence. To address these issues, Wang *et al.* [55] introduced the Structural Similarity Index Measure (SSIM), a perception-based metric that evaluates image degradation based on perceived changes in luminance (l), contrast (c), and structure (s), as in Equation 3.8.

$$\text{SSIM}(\mathbf{Y}, \hat{\mathbf{Y}}) = [l(\mathbf{Y}, \hat{\mathbf{Y}})]^\alpha \cdot [c(\mathbf{Y}, \hat{\mathbf{Y}})]^\beta \cdot [s(\mathbf{Y}, \hat{\mathbf{Y}})]^\gamma \quad (3.8)$$

$$l(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{2\mu_{\mathbf{Y}}\mu_{\hat{\mathbf{Y}}} + C_1}{\mu_{\mathbf{Y}}^2 + \mu_{\hat{\mathbf{Y}}}^2 + C_1} \quad (3.9)$$

$$c(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{2\sigma_{\mathbf{Y}}\sigma_{\hat{\mathbf{Y}}} + C_2}{\sigma_{\mathbf{Y}}^2 + \sigma_{\hat{\mathbf{Y}}}^2 + C_2} \quad (3.10)$$

$$s(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{\sigma_{\mathbf{Y}\hat{\mathbf{Y}}} + C_3}{\sigma_{\mathbf{Y}}\sigma_{\hat{\mathbf{Y}}} + C_3}, \quad (3.11)$$

In equations from 3.9 to 3.11, $\mu_{\mathbf{Y}}$ and $\mu_{\hat{\mathbf{Y}}}$ are the mean intensities, $\sigma_{\mathbf{Y}}$ and $\sigma_{\hat{\mathbf{Y}}}$ are the standard deviations, $\sigma_{\mathbf{Y}\hat{\mathbf{Y}}}$ is the covariance, α , β , and γ determine the weighting of each component. C_1 , C_2 , and C_3 are constants added to avoid instability.

However, the commonly used expression is a simplified form suggested by the authors, where $\alpha = \beta = \gamma$ and $C_3 = C_2/2$ as depicted in Equation 3.12.

$$\text{SSIM}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{(2\mu_{\mathbf{Y}}\mu_{\hat{\mathbf{Y}}} + C_1)(2\sigma_{\mathbf{Y}\hat{\mathbf{Y}}} + C_2)}{(\mu_{\mathbf{Y}}^2 + \mu_{\hat{\mathbf{Y}}}^2 + C_1)(\sigma_{\mathbf{Y}}^2 + \sigma_{\hat{\mathbf{Y}}}^2 + C_2)} \quad (3.12)$$

This expression is computed for corresponding local patches of two images, with the final result being an average across the patches. According to the definition, the final value lies in the range $[-1, 1]$, where 0 indicates no similarity, while -1 and 1 represent perfect anti-correlation and perfect correlation, respectively.

SSIM is more robust than the aforementioned metrics to small distortions and generally demonstrates higher concordance with human perception. The Multi-Scale SSIM (MS-SSIM) [56] extends the SSIM index by combining measurements across multiple scales of the image, thereby providing greater robustness to distortions and variations in viewing conditions.

As an image-to-image translation task, virtual staining aims to transform the input image from one staining modality (source domain) to another (target domain), while preserving the structure of the input image. SSIM and its variants assess the similarity between the generated and target images within the target domain, but they do not consider the preservation of the input image's content. To address this issue, Liu *et al.* [57] introduced the Contrast Structure Similarity (CSS) metric, which evaluates the preservation of morphology in the input H&E image during virtual immuno-staining. As shown in Equation 3.13, CSS can be considered a variant of SSIM, focusing primarily on the similarity of samples in terms of local contrast and structural similarity, while excluding the absolute intensity (luminance).

$$\text{CSS}(\mathbf{X}, \hat{\mathbf{Y}}) = \frac{2\sigma_{\mathbf{X}\hat{\mathbf{Y}}} + C}{\sigma_{\mathbf{X}}^2 + \sigma_{\hat{\mathbf{Y}}}^2 + C} \quad (3.13)$$

Although straightforward and relatively inexpensive to compute, the SSIM metric and its variants have limitations in capturing the subtleties of human perception. Zhang *et al.* [58] demonstrated that networks trained for visual prediction tasks inherently learn representations that correlate strongly with human perceptual judgment. Notably, the more effective a feature set is at tasks such as classification, the more robust it becomes as a model of perceptual similarity. This insight led to the development of network-based similarity metrics that use either intermediate features or embeddings to assess the similarity between two images. The choice between using intermediate features or embeddings depends on the aspect we wish to evaluate. Embeddings represent the final output of an image produced by the model and encode high-level semantic information, serving as a global representation that summarizes the entire input. In contrast, intermediate features refer to activations within the model before the final embeddings are computed. These features capture localized, low-level information, such as edges, textures, and various patterns [59]. In the field of image-to-image translation, the most prominent network-based metrics are Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID).

Introduced by Zhang *et al.* [58], the definition of LPIPS relies on the intermediate features of a network trained on human similarity judgments. The metric between two paired images (x and x_0) is calculated by computing $\mathcal{L}_2$ distance of channel-normalized activations from different layers

($\hat{\mathbf{y}}_{1hw}^l$ and $\hat{\mathbf{y}}_{2hw}^l$), multiplying them by a weight vector, averaging them spatially, and summing them channel wisely (Equation 3.14).

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\hat{\mathbf{y}}_{hw}^l - \hat{\mathbf{y}}_{0hw}^l) \right\|_2^2 \quad (3.14)$$

On the other hand, the FID definition is based on the embeddings of an InceptionV3 [60] network. Specifically introduced by Heusel *et al.* [61] to evaluate the performance of GANs, FID measures the Wasserstein distance between the embeddings of the generated and real image distributions, assuming both follow a Gaussian distribution. As a result, it is a metric that highlights how well the model has captured the real distribution, rather than just pairwise concordance. After generating the embeddings of real and generated images, the score is obtained by computing the Euclidean distance between the means of the feature vectors (μ_y , $\mu_{\hat{y}}$) of both sets and adding the trace (Tr) of a combination of the covariance matrices (Σ_y , $\Sigma_{\hat{y}}$), as shown in Equation (3.15). The more similar the two distributions are, *i.e.*, the more similar the generated images are to the real images in the embedding space of the Inception model, the smaller the FID score will be. This score represents the state-of-the-art (SOTA) in evaluating image-to-image models and can be applied to unpaired datasets and unsupervised methods.

$$\text{FID}(\mathbf{Y}, \hat{\mathbf{Y}}) = \|\mu_{\mathbf{Y}} - \mu_{\hat{\mathbf{Y}}}\|_2^2 + \text{Tr} \left(\Sigma_{\mathbf{Y}} + \Sigma_{\hat{\mathbf{Y}}} - 2(\Sigma_{\mathbf{Y}} \cdot \Sigma_{\hat{\mathbf{Y}}})^{1/2} \right) \quad (3.15)$$

However, it should be noted that FID is affected by the number of examples which compose each distribution [62]. To address this issue, Binkowski *et al.* [62] introduce the Kernel Inception Distance (KID) which evaluates the quality of generative models by measuring the squared Maximum Mean Discrepancy (MMD) between the distributions of InceptionV3 embeddings of real and generated images. It employs a polynomial kernel of degree 3 to compute the MMD, providing an unbiased estimate of the true distance between these distributions. This metric proves particularly useful for small sample sizes, as it does not assume a Gaussian distribution for the embeddings and offers a more reliable assessment compared to FID [62], [63].

The advantage of using the aforementioned metrics is that they provide a numerical, objective assessment of the apparent quality of generated images from an image-to-image translation perspective. However, the main objective of virtual staining is to generate images that can be useful later in clinical practice. Therefore, the success of a virtual staining model should consider its ability to preserve pathological information and the quality of its insight for accurate clinical decisions. To this end, pathology-specific metrics have been developed, alongside downstream task evaluations to verify the usefulness of the generated images. Additionally, expert qualitative evaluations are commonly performed. In the following subsections, we will explore some pathology-specific metrics (Subsection 3.2.2) and the qualitative expert assessments performed on the models (Subsection 3.2.3).

3.2.2 Pathology metrics and downstream tasks

While models that perform well on image-to-image metrics are valuable, they must also preserve and emphasize the pathological information in the slides. The concentration of chromogen in an image correlates with the protein expression of the biomarker of interest. Consequently, some authors estimate the concentration of chromogen by calculating the concentration of DAB and applying dedicated metrics based on it. For example, a distance measure such as the Jensen-Shannon divergence (JSD) between the DAB histograms of the real and generated images can provide meaningful insight in this context.

To estimate the DAB concentration, most authors build upon the work of Ruifrok *et al.* [64], who demonstrated that the relative intensity in each of the three RGB channels depends non-linearly on stain concentration; however, the optical density (OD) for each channel varies linearly with the concentration of the absorbing material. This linearity enables the separation of contributions from multiple stains within a specimen. As a result, the method of Ruifrok *et al.* [64] enables the conversion from the RGB color space to the HED (Hematoxylin, Eosin, DAB) concentration/color space by first transforming the RGB values into their OD values, as shown in Equation 3.16. Subsequently, color deconvolution is performed using the 3 by 3 normalized optical density matrix, which encodes the absorption factors for each individual stain, as illustrated in Equation 3.17.

$$\text{RGB}_{OD} = -\log_{10}(\text{RGB}) \quad (3.16)$$

$$\text{HED} = \text{RGB}_{OD} \cdot \mathbf{M}_{\text{RGB} \rightarrow \text{HED}} \quad (3.17)$$

Based on the DAB channel, it is also possible to compare the number of positively stained cells in the real versus the generated image by applying an intensity threshold, as demonstrated by Martino *et al.* [65]. Furthermore, using a threshold established by clinical guidelines, which classifies images as positive or negative based on the proportion of stained cells, allows assessment of agreement with categorizations derived from the ground truth. The DAB channel can also be used to create segmentation maps for training or evaluation purposes. Alternatively, the performance of a virtual staining model can be assessed by applying its output to downstream tasks, such as biomarker expression classification. For example, in the study by Peng *et al.* [66], a classifier was trained on both generated and real IHC images to predict HER2-score, with performance comparisons made between the two. Assessing virtual staining models through downstream tasks is crucial because, in clinical practice, the key concern is not just the visual quality of the images but the insights they offer about biomarker expression, which is critical for informing clinical decisions. Although evaluating downstream tasks is valuable for determining practical applicability, most methods lack such assessments, possibly because many virtual staining approaches have not yet achieved sufficient quality.

3.2.3 Qualitative evaluations

To assess the usability of developed models, some researchers conduct qualitative evaluations by presenting generated images to pathologists. These assessments typically focus on whether pathologists can (1) differentiate between generated and real IHC images and (2) reach the same conclusions, such as determining biomarker expression, using the generated slides. For example, Zeng *et al.* [67] presented PgR-positive and PgR-negative slides to a pathologist to evaluate the agreement between the positivity expression of real and generated slides. In another study, Martino *et al.* [65] showed 30 synthetic IHC images alongside real ones to two pathologists, who were asked to identify whether each image was real or generated.

3.3 Summary

GANs involve a competitive training process between a generator and a discriminator, where the discriminator differentiates between real and fake images, while the generator learns to deceive it, capturing the underlying data distribution. The two most commonly used GAN frameworks in virtual staining are Pix2Pix and CycleGAN, depending on whether the data is paired or unpaired, respectively.

In virtual staining, GANs are commonly extended with a contrastive learning objective. In such approaches, the generator is guided to preserve the content of a template image by maximizing the distance between the anchor and negative samples, while minimizing the distance between the anchor and the corresponding positive patch in the feature space.

As an alternative to GANs, some researchers utilize DMs, a class of latent variable models inspired by diffusion processes in thermodynamics. By learning to denoise, the generator can create images that match the original data distribution.

PSNR and SSIM are classical evaluation metrics used in the field, but have been shown to correlate poorly with human perception of similarity. Consequently, network-based metrics, such as FID and LPIPS, are gaining popularity. However, the true value of an image lies in its applicability, making pathology-oriented metrics, downstream tasks, and qualitative evaluations essential for assessing the quality of a virtually stained image.

Chapter 4

State-of-the-Art Review

This chapter provides a comprehensive overview of virtual staining from H&E to IHC of the clinical biomarker panel in BC, including a systematic review of the SOTA, and a benchmarking study with a unified performance comparison of representative models in the field. Finally, we conclude this chapter by summarizing the review and discussing the results obtained in the benchmarking, as well as the gaps in the current SOTA. As a result, this chapter provides a valuable resource for researchers and clinicians advancing virtual staining technology for BC diagnostics.

4.1 State-of-the-Art

We have identified 28 publications since 2021 that focused on virtual staining from H&E to IHC for the clinical panel in BC, as described in Chapter 2. This list results from a bibliographic search in Google Scholar using the following prompt:

```
( "virtual stainin" OR  
  "virtual histological staining" OR  
  "computational staining" OR  
  "digital staining" OR  
  "synthetic staining" ) AND  
( "HER2" OR "PR" OR "PgR" OR "ER" OR "KI67" )
```

We also examined citations of key papers and manually reviewed relevant conference proceedings and publications from authors with prior work in the field (Figure 4.1). This process yielded 1,116 potential articles, which were carefully analyzed, leading to the inclusion of 28 relevant works, all summarized in Table 4.1. These publications were categorized based on model architectures, distinguishing whether they use a GAN or DM framework. Additionally, we further categorized models that incorporate contrastive learning, as we find them distinct enough to group them separately, despite often combining contrastive learning with GAN.

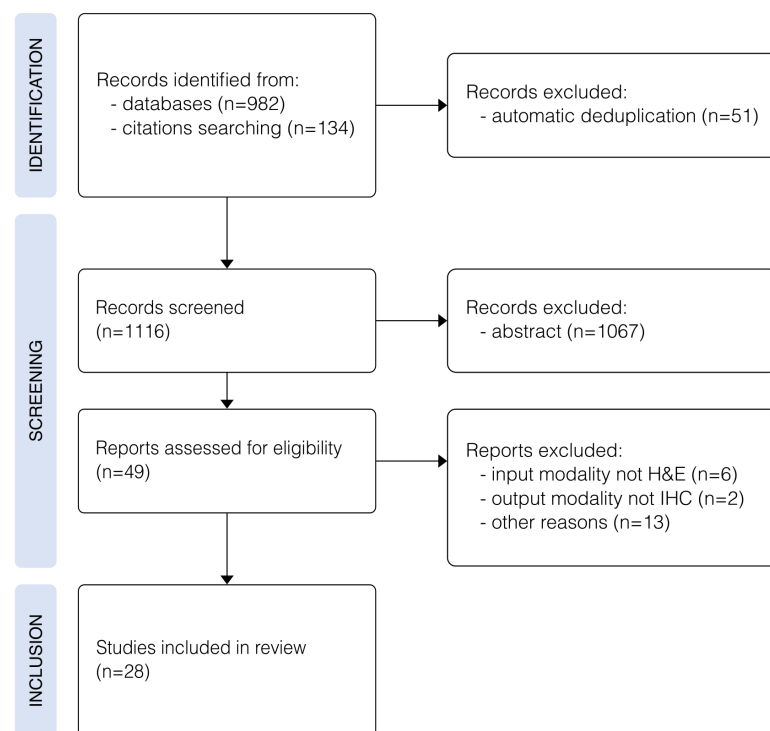


Figure 4.1: PRISMA flowchart illustrating the selection process for the studies included in the review, identified via databases and other sources (from Klöckner *et al.* [27]).

Table 4.1: Summary of the state-of-the-art models on virtual staining (H&E to IHC) of the clinical biomarker panel for breast cancer (from Klockner et al. [27]). Q.E.: qualitative evaluation. (Continued from previous page)

Framework	Author/Model	Year	HER2	ER	PgR	Ki67	Datasets	Details	Q.E.
	Huang <i>et al.</i> [70] TC-cycleGAN	2023	x	-	-	-	BCI	CycleGAN that includes a loss function based on Gabor filters.	-
	Baldeon-C. <i>et al.</i> [71] DeepSIT	2023	x	-	-	-	BCI	U-Net generator with four PatchGAN discriminators on the outputs of different depths of the U-Net decoder.	-
	Liu <i>et al.</i> [72] MGGAN	2023	x	-	-	-	BCI	GAN with a generator consisting of 2 U-Nets that extract features at different scales (3x3 or 7x7 kernels in the first convolutional layer).	-
	Martino <i>et al.</i> [65]	2024	-	-	-	x	In-house: oral	Standard Pix2Pix model without any added loss functions.	x
	Ma <i>et al.</i> [73] DSFF-GAN	2024	x	x	-	x	MIST In-house: endometrial	CycleGAN that generates positive region masks which are compared against masks based on the DAB channel of the ground-truth image. Additionally, a latent space loss is added.	x
	Wu <i>et al.</i> [74] HcGAN	2024	x	x	x	x	BCI MIST	Conditional GAN with a U-Net generator and harmonic convolutions based on discrete cosine transforms.	-

Continued on next page

Table 4.1: Summary of the state-of-the-art models on virtual staining (H&E to IHC) of the clinical biomarker panel for breast cancer (from Klockner et al. [27]). Q.E.: qualitative evaluation. (Continued from previous page)

Framework	Author/Model	Year	HER2	ER	PgR	Ki67	Datasets	Details	Q.E.
GANs	Peng <i>et al.</i> [66] TDKstain	2024	x	-	-	-	In-house: breast	Conditional GAN with a regularised training based on generated/real nuclei density maps and L1 loss on the generated DAB channel.	-
	Wei <i>et al.</i> [75] DeReStainer	2024	x	-	-	-	BCI	A GAN model that processes the shared haematoxylin channels of the H&E and IHC images to constrain the training and uses a HER2-classifier to regularise the decoder.	-
	Guan <i>et al.</i> [63] GramGAN	2024	-	-	-	x	In-house: kidney + lung	A multi-domain stain-transfer model based on cycleGAN, using a style-encoding dictionary and auxiliary stain-type classifiers for regularisation.	-
	Qu <i>et al.</i> [76] BCISainer	2024	x	-	-	-	BCI	A GAN with parameter-free attention layers in the generator and two classifiers to predict HER2-expression levels on both the H&E and IHC images.	-

Continued on next page

Continued on next page

Table 4.1: Summary of the state-of-the-art models on virtual staining (H&E to IHC) of the clinical biomarker panel for breast cancer (from Klockner et al. [27]). Q.E.: qualitative evaluation. (Continued from previous page)

Framework	Author/Model	Year	HER2	ER	PgR	Ki67	Datasets	Details	Q.E.
	Li <i>et al.</i> [80] WPCC	2024	x	x	x	x	BCI MIST	CUT-based model that calculates the Jensen-Shannon divergence (JSD) between the distributions of contrastive losses calculated on the real and the generated image.	-
Contrastive learning	Wang <i>et al.</i> [81] MDCL	2024	x	x	x	x	BCI MIST	GAN framework with two siamese networks with the generated IHC and H&E (or real IHC) as input. Corresponding patch embeddings of H&E, real and generated IHC are pulled together while spatially non-corresponding patches are pushed apart.	-
	Zhang <i>et al.</i> [82] PPT	2024	x	-	-	-	BCI MIST	GAN that expands the PatchNCE loss by using the focal loss on all three combinations of real/fake IHC and the H&E input.	x

Continued on next page

Continued on next page

Table 4.1: Summary of the state-of-the-art models on virtual staining (H&E to IHC) of the clinical biomarker panel for breast cancer (from Klockner et al. [27]). Q.E.: qualitative evaluation. (Continued from previous page)

Framework	Author/Model	Year	HER2	ER	PgR	Ki67	Datasets	Details	Q.E.
	Li <i>et al.</i> [83] Confusion-GAN	2024	x	x	-	-	BCI MIST In-house: liver	A cycleGAN with a multi-branch generator and discriminator to handle features at multiple scales, an auxiliary classifier based on stained-region percentage to capture pathological information and a contrastive loss to between positive/negative real/generated patches.	x
	Hu <i>et al.</i> [84] ULViT-GAN	2024	x	-	-	-	BCI	Based on cycleGAN, the generator has a U-Net-like structure with additional Harr-wavelet downsampling and a vision transformer module on the output of the encoder.	-
Diffusion models	He <i>et al.</i> [85] PST-DIFF	2024	x	x	-	x	BCI ACROBAT	Two diffusion models that learn to denoise their respective domains. The latent space is connected via the dual diffusion implicit bridge method, and structural consistency is learned in the de-noising model via conditioning on frequency information.	-

Continued on next page

Table 4.1: Summary of the state-of-the-art models on virtual staining (H&E to IHC) of the clinical biomarker panel for breast cancer (from Klockner et al. [27]). Q.E.: qualitative evaluation. (Continued from previous page)

Framework	Author/Model	Year	HER2	ER	PgR	Ki67	Datasets	Details	Q.E.
	Li <i>et al.</i> [86] His-MMDM	2024	-	-	x	x	In-house: brain	A diffusion model conditioned on source and target domain that can also be conditioned on mutations or tissue type.	-
Other models	Jia <i>et al.</i> [87] DTNet	2024	x	x	x	x	In-house: breast	U-Net architecture with content and style encoders, transformer skip connections and a pyramidal loss.	-
	Ji <i>et al.</i> [88]	2024	-	-	-	x	In-house: uterus	A U-Net with an H&E image converted to optical density as input.	-

4.1.0.1 GANs

Liu *et al.* [57] published the first study on virtual staining of the BC IHC biomarker Ki67 from H&E-stained samples in 2021. Building upon the cycleGAN framework, the authors introduced a pathological representation network (PRN) within each generator. These networks predicted tumor cell areas, which constrained the generators' learning by penalizing deviations between the predicted tumor regions and the ground-truth heatmaps, as well as discrepancies between predicted heatmaps across generators, using the L1 loss. This methodology aimed to enforce the preservation of structural and pathological content during image translation, thereby ensuring accurate pathological representation. Furthermore, the authors proposed an SSIM-cycle loss and a base-space loss designed to align the latent spaces of the two generators. According to the authors, the cycleGAN framework lacks constraints between the generated and target images, which can result in inconsistencies in structural and pathological content and limit its usefulness in virtual staining. The addition of the two PRN and loss functions addresses this issue, advancing cycleGAN's applicability in the field of virtual staining. The proposed framework was trained on in-house datasets of neuroendocrine and breast tissue, the latter comprising 160 paired H&E and Ki67-stained WSIs. These were tiled into 288×288 pixel images, yielding 57,000 tiles per pair. The experimental results showed that their model, PC-StainGAN, outperformed cycleGAN by achieving better MAE, SSIM, MS-SSIM, and PSNR between the H&E source and reconstructed images, better CSS scores between the H&E source and generated IHC images, and a higher Pearson correlation for the positively and negatively stained regions when comparing the generated and reference IHC images.

In 2022, Zeng *et al.* [67] developed a semi-supervised method for virtual staining of PgR in breast tissue. They noted that the methodology employed by Liu *et al.* was highly dependent on expert annotations. To overcome this limitation, after registering and tiling the consecutive H&E and IHC slides, they applied thresholding on the DAB channel of the IHC images to label each pair as positive or negative. Therefore, the cycleGAN framework could be improved by incorporating a classification loss that enforces consistency between input images and their generated counterparts based on their labels. In addition, inspired by Zhou *et al.* [89] and Kim *et al.* [90], the authors introduced an auxiliary classification loss, class activation map loss, to guide the generator in focusing on discriminative areas. This study used an in-house BC dataset consisting of 22 training slide pairs (58,942 H&E and 62,218 PgR overlapping patches of 256×256 pixels) and 8 pairs for evaluation. The proposed model outperformed cycleGAN, MUNIT [91], and UGATIT [90] on image quality metrics and demonstrated strong agreement with biomarker expression as assessed by a pathologist.

In the same year, Liu *et al.* [43] released the Breast Cancer Immunohistochemical (BCI) dataset, the first publicly available collection of paired H&E and IHC slides. It includes 4,873 pairs of registered 1024×1024 pixel tiles from consecutive H&E and HER2 IHC-stained WSIs. The WSIs were initially manually aligned via projection transformation and subsequently further registered using the Elastix toolbox [92]. In addition to releasing the dataset, the authors intro-

duced the pyramid Pix2Pix model, an improvement over the classic Pix2Pix model by Isola *et al.* [44]. To address the issue of imperfect pixel alignment in consecutive tissue samples, their method employs a pyramid framework that applies Gaussian smoothing and downsampling, simulating various octave levels, to both generated and ground truth images. This approach enhances consistency between image pairs across multiple scales. The standard GAN and $\mathcal{L}_1$ losses of the vanilla Pix2Pix generator are supplemented with a multi-scale L1 loss applied to the downsampled and smoothed image pairs. Employing ResNet9-blocks as the generator backbone and PatchGAN for the discriminator architecture, their framework achieved better PSNR and SSIM than other Pix2Pix models and CycleGAN. This paper marks an important milestone in virtual staining for BC immuno markers, as the BCI dataset has become widely adopted, and the pyramid Pix2Pix model serves as a foundational framework for many subsequent works. However, considerable improvement is needed before these models can be clinically applied, as evaluations by two pathologists showed that fewer than 40% of the generated images matched the HER2 expression of their corresponding real images.

Although most GAN frameworks use a CNN-based model as the core component, Ma *et al.* [69] employed the Swin-Transformer [93] as the backbone model in 2023, to balance local and global feature learning. They also implemented a pre-training method based on masked image modeling (MIM) [94], [95] to improve the performance of the model and ensure it learns meaningful representations from a labeled dataset. The pre-training phase was carried out on the Insillico Labeling (ISL) dataset [96], which includes 148,454 pairs of unstained/stained microscopy images of 256×256 pixels, before proceeding to a fine-tuning phase on the BCI dataset. During fine-tuning, the authors used the PatchGAN [97] discriminator and the Charbonier loss function [98] (a smoothed version of MAE loss), achieving the highest PSNR and second-highest SSIM values compared to Pix2Pix, Pyramid Pix2Pix, and CycleGAN frameworks.

In the same year, Baldeon-Calisto *et al.* [71] trained a U-Net generator using four PatchGAN discriminators at different depths of the U-Net decoder on the BCI dataset. The authors trained their framework using a combination of adversarial and L1 losses, calculated between the generated outputs at various depths and the ground truth. To match the dimensions of the real IHC images, intermediate outputs were resized using bilinear interpolation. The authors reported superior SSIM values compared to other models, including Pix2Pix, Pix2PixHD [99], Pyramid Pix2Pix, and CycleGAN. Notably, their model ranked first in the post-challenge phase leaderboard of the BCI GrandChallenge [76]. Additionally, a qualitative evaluation of the generated outputs showed optimal performance for cases with 0 and 1+ HER2-scores, but reduced performance for cases with 3+ HER2-score, aligning with findings from previous studies.

In 2024, building on PC-StainGAN, Ma *et al.* [73] eliminated the need for expert-annotated segmentation masks by converting IHC-stained RGB images to the HED color space and then thresholding in the DAB channel. They introduced the Dual Scale Feature Fusion (DSFF) block in the generator decoder, which integrates upsampling with Pathological Feature Attention (PFA) blocks. The PFA sub-block calculates attention weights across the channel, width, and height dimensions to improve the coloring ability of the network. Furthermore, they use a modified

version of the SSIM loss to enhance the color fidelity of virtually stained images, penalizing discrepancies in pixel values in the XYZ color space. The authors trained and evaluated their model on an internal dataset of Ki67-stained endometrial samples, demonstrating superior SSIM, PSNR, and CSS compared to variations of the Pyramid Pix2Pix and CycleGAN models. For Ki67, two pathologists quantified the positive cell ratio per patch and compared it to the ratio in the corresponding real patches. Their model showed the smallest deviation and lowest variation among the evaluated models. Further testing on the MIST dataset reinforced the effectiveness of the model in SSIM and CSS for ER and HER2 stains.

In the same year, Peng *et al.* [66] introduced a different strategy to achieve accurate pathological representation without expert annotations. To address this, the authors proposed the TDKstain model, which incorporates task-specific domain knowledge by extracting stained membrane regions from the DAB channel and nuclei density maps from the hematoxylin channel. The model combines a GAN with a density estimation network, both alternately trained, to generate IHC staining and predict the DAB channel from the H&E input and its corresponding membrane staining intensity mask. The authors used an internal dataset of breast tissue containing well-registered H&E- and HER2 IHC-stained non-overlapping images, each with a fixed size of 1024×1024 pixels. The model achieved the highest values for FID and Deep Image Structure and Texture Similarity (DISTS) [100] among the SOTA models, although it did not achieve the highest SSIM score. Additionally, it performed exceptionally well in task-specific evaluations, such as nuclei density estimation and membrane staining intensity. Unlike many other studies, the authors tested their method on a downstream task, reporting comparable performance for a HER2-grade classification model trained on generated IHC-stained images and the same model trained on real IHC-stained images.

Wei *et al.* [75] introduce the DeReStainer model, which exploits the shared hematoxylin channel between H&E- and IHC-stained images. Their generator comprises a DeStainer encoder, which extracts and encodes the hematoxylin channel, and a symmetric ReStainer decoder that applies the virtual DAB stain to generate the IHC-stained image. Throughout training, both the H&E and IHC hematoxylin channels are encoded by the DeStainer module. Afterwards, these inputs are aligned and the HER2-score is estimated from the H&E hematoxylin channel. The alignment task and HER2-score estimation are optimized using cosine similarity loss and Focal loss [101], respectively. The ReStainer then predicts the DAB channel, and $\mathcal{L}_1$ loss is applied between the virtual and ground-truth IHC DAB channels. A comparator module assesses the similarity between virtual and real IHC images using a combination of SSIM, MAE, and cosine similarity. When evaluated on the BCI benchmark, the model achieved top SSIM results but did not outperform SOTA in PSNR. Furthermore, the authors trained a k-Nearest Neighbor (kNN) classifier on the generated images. Although the accuracy (60%) was not optimal, the model trained on their images outperformed models trained on images generated by ASP and Pyramid Pix2Pix.

Guan *et al.* [63], building on CycleGAN [36], introduced a multi-domain stain transfer method (GramGAN), which enables stain transfer across multiple styles using a single network. This model incorporates style-guided (SG) blocks in the generator bottleneck, facilitating the progres-

sive transfer of the stain from the original to the target style using a style encoding dictionary (SED) to gradually adjust the input feature map towards the target style. The discriminator architecture includes an auxiliary classifier, an auxiliary discriminator, and a target stain classifier. After passing through the downsampling module, the auxiliary classifier and discriminator generate attention weights that guide the respective components. The authors trained and evaluated GramGAN on kidney and lung datasets from ANHIR [102], which include 8,961 H&E and 8,419 Ki67 tiles from three lung tumor samples. Their model outperformed several SOTA models in CSS, FID, and KID metrics, as well as in the downstream task of glomeruli segmentation and detection.

4.1.0.2 Contrastive Learning

In 2022, Zhang *et al.* [77] introduced the first network developed to generate multiple IHC stains from H&E, including HER2, PgR, ER, and Ki67. The Multiple Virtual Functional Stain (MVF-Stain) model generator comprises style and content encoders, along with a single decoder that merges the source/target content and style to produce a generated image. The content encoder uses pixel/channel-wise attention, while the style encoder applies KL loss to constrain feature distributions to a Gaussian distribution. The decoder outputs three images: (a) a reconstruction of the source image, (b) the virtually stained image, and (c) a virtually stained image with a Gaussian-style feature vector as input. Images a and c are used for the reconstruction and cycle losses, respectively, while images b and c serve as inputs for a style discriminator. Additionally, a histogram-based loss, inspired by Ustinova *et al.* [103], is applied to differentiate style feature distributions between inter- and intra-domain pairs. A second discriminator is applied to the content features. The authors used four datasets with various tissue types and staining, including ER, PgR, and HER2 for breast internal samples. For this application, they used 5,242 tiles of 256×256 pixels extracted from WSIs from three individuals. The reported results showed higher PSNR and HTI for HER2 compared to a vanilla CycleGAN, as well as better PSNR but slightly lower HTI for ER and PgR. Furthermore, the average optical density (mIOD) [77] for HER2 was closer to the ground truth in CycleGAN, while MVFStain performed better on ER and PgR. Moreover, two pathologists blindly graded both virtually- and manually-stained tissue sections on a scale of 1 to 3 for stain quality, with results showing that virtual images were nearly indistinguishable from the real ones.

In the following year, Li *et al.* [26] released the Multi-IHC Stain Translation (MIST) dataset, which, along with BCI, has become one of the most widely used datasets to date. The dataset consists of paired 1024×1024 pixel tiles of consecutive tissue sections, including H&E and HER2 ($n = 5,642$ from 64 WSIs), ER ($n = 5,153$ from 56 WSIs), PgR ($n = 5,139$ from 56 WSIs), and Ki67 ($n = 5,361$ from 56 WSIs) stainings. Furthermore, recognizing that contrastive learning enhances model robustness against label noise [104]–[106], the authors enhanced the Pyramid Pix2Pix model by incorporating PatchNCE from Park *et al.* [42] and Adaptive Supervised PatchNCE (ASP) into the generator loss function. Both loss functions aim to maximize mutual information in the embedding space of the network using a patch-based InfoNCE loss [107]. However,

while the former penalizes discrepancies between the input H&E and the generated image, the latter penalizes the difference between the generated image and the target IHC. Additionally, ASP introduces an adaptive weighting scheme, based on cosine similarity and the training epoch, to reduce the impact of inconsistent patch locations between the generated and real IHC images in the embedding space, thereby addressing the issue of pixel-level misalignment.

In 2024, Qu *et al.* [78] introduced a model that employs a multi-magnification processing strategy and an attention module. The original images were downsampled, and regions were randomly cropped and zoomed in, producing images with $2\times$ or $4\times$ magnification. Using a super-resolution model [108], these processed images were resized to the same dimensions. This preprocessing step was designed to expose the model to multiple scales, helping it learn to extract features at various magnifications. The overall structure of the model is based on the ASP model, with the addition of the attention module. This module connects both downsampling and upsampling blocks at the same level and concatenates the output of the preceding layer. As a result, in addition to focusing attention on specific spatial regions and feature channels, it also serves as a skip connection. Evaluation on the MIST dataset demonstrated that the model performed slightly better than the baseline ASP [26] model.

Chen *et al.* [79] extended the work of Park *et al.* [42] and Liu *et al.* [43] by addressing the challenge of incorrect spatial alignment and the preservation of pathologically relevant information through their Pathological Semantics-Preserving Learning (PSPStain) framework. PSPStain incorporates two primary strategies: the Protein-Aware Learning Strategy (PALS) and the Prototype-Consistent Learning Strategy (PCLS). In PALS, Focal Optical Density (FOD) maps are generated by isolating the DAB channel. Using the FOD maps of the real and generated images, the authors apply three DAB-specific losses to preserve the biomarker expression: average difference, histogram difference, and average differences across 16 smaller patches. In PCLS, the authors utilize the output from the penultimate layer of a pre-trained U-Net, along with class-label probability maps, to create prototypes for both the real and generated IHC images. They compute the cosine similarity between the generated IHC feature maps and the ground-truth IHC prototypes, and vice versa, to derive a cross-image probability prediction. This prediction is compared to the respective generated mask, resulting in a cross-image tumour prototype consistency (CTPC) loss. Despite having lower scores on traditional image quality metrics, such as PSNR and SSIM, compared to CUT [42], ASP, Pix2Pix, and Pyramid Pix2Pix models, the proposed model achieved top results in pathology-related metrics such as mIOD, cumulative optical density (IOD), and the Pearson correlation coefficient on the MIST and BCI datasets.

To address the scarcity of pixel-wise paired datasets, Li *et al.* [80] built on the work of Park *et al.* [42] by introducing the weakly pathological consistency constraint (WPCC), which acts on multiple layers of the generator. WPCC begins by extracting embedding vectors from both the real and generated IHC images at various layers. It then computes the distribution of patch embedding similarities in each layer and calculates the JSD between the real and generated distributions. The Pearson correlation coefficient is used to assess the degree of alignment between consecutive H&E and IHC image tiles. This alignment is then used to compute a weighted sum of the distances pre-

viously calculated, with a higher degree of alignment resulting in a larger weight. In addition to WPCC, the authors introduced a discriminator regularization loss function to stabilize the discriminator training. The model was trained on the BCI dataset, achieving improved PSNR, FID, KID, and SSIM scores compared to CycleGAN, CUT, Pyramid Pix2Pix, and ASP.

Wang *et al.* [81] introduce Mix-Domain Contrastive Learning (MDCL), which uses Siamese networks to predict the correlation between the anchor and both inter- and intra-domain patches. In their approach, upon generation of the IHC images, some patches are randomly selected from the virtual IHC and input H&E to form the anchor and positive sets, respectively, which serve as input for one of the Siamese networks. The input of the other network consists of the virtual IHC anchor set and the corresponding patches from the ground-truth IHC. However, since these are not aligned at the pixel-level, an adaptive weighting factor is introduced [26]. The contrastive loss function is a modified version of the one used by Park *et al.* [42] to measure similarity in both the inter- and intra-domain. The authors trained and tested their model on the MIST [26] and BCI [43] datasets, demonstrating superior performance on FID, KID, and PHV metrics compared to CycleGAN, CUT, Pix2Pix, Pyramid Pix2Pix, and ASP.

Zhang *et al.* [82] introduced the Patch Alignment-based Paired (PPT) model, which uses bidirectional contrastive learning to address the misalignment issue. Extending the work of Li *et al.* [26], the authors use FocalNCE loss between the input and output images, as well as between the output and ground-truth images, in a bidirectional manner. FocalNCE loss incorporates Focal loss [101] to handle class imbalance over the PatchNCE loss function. Moreover, the training is constrained by consistency loss, which includes patch loss, content loss, and frequency loss. Patch loss computes the mean difference between the output and target images at the patch level, rather than at the pixel level, further addressing the misalignment issue. Content loss assesses consistency at the feature level using a VGG-19 [109] network. Lastly, the Pyramid loss [43] maintains consistency across different levels. The authors trained and evaluated the model on three datasets: an internal dataset of canine lymphoma, BCI [43], and MIST [26]. They achieved improved PSNR, FID, and LPIPS scores across all three datasets compared to ASP, Pix2PixHD, Pyramid Pix2Pix, CycleGAN, and CUT. Evaluation of the CD3 and PAX5 stained internal dataset by two pathologists revealed a good similarity between the generated and real images.

Li *et al.* [83] introduced Confusion-GAN, a CycleGAN-based model that integrates a patch-level pathology information extractor (PPIE) for H&E image classification, along with a multi-branch discriminator (MBD). An IHC classifier is trained using positive patch-level labels (greater than 1% stained patch area) from the generated IHC images, and the loss, combined with the PPIE, produces a term that captures the pathological information of the input patch. PPIE consists of an encoder that maximizes the feature space distance between the top-k H&E tiles (selected via attention from an ABMIL [110]) of a positive WSI and negative tiles transformed by a BBDM [111] into positive examples. Based on these embeddings, a classifier is trained to classify positive and negative classes, which is then combined with the loss of the IHC classifier. Each generator is associated with two MBDs that use a discriminator to distinguish between real and generated images, while a secondary classifier identifies the generated image by mixing embeddings at a

random channel position. In addition to training on an internal hepatocellular carcinoma dataset, the authors retrained and evaluated their model on the BCI and MIST datasets. Their experiments demonstrated improved FID, LPIPS, and SSIM values compared to Pix2Pix variations and ASP on the MIST dataset. Since these datasets provide tiles instead of WSIs, the authors applied IHC labels using the previously mentioned staining cut-off (or the BCI labels) and transferred them to the corresponding H&E images for use with the PPIE. It should also be noted that the authors performed a qualitative assessment of the generated images, but this was only performed on the internal dataset (for hepatocellular carcinoma).

4.1.0.3 Diffusion models

In 2024, He *et al.* [85] proposed the Pathological Stain Transfer Diffusion (PST-DIFF) model, which enables bidirectional translation between the source and target staining domains. The model couples the latent representations of both modalities, leveraging the Dual Diffusion Implicit Bridges (DDIBs) [112] method. The authors introduce a Latent Transfer (LT) module to enforce pathological consistency across domains by minimizing the distance between the latent representations of paired H&E and IHC image pairs. An Asymmetric Attention Mechanism (AAM) is also incorporated, employing multi-head attention for the source-to-latent conversion and global attention for the latent-to-target conversion. To enforce structural consistency, a Conditional Frequency Guidance (CFG) module [113] guides the denoising network with low-frequency information in early stages and high-frequency details in later stages of the reverse process. The authors trained and evaluated their model on the BCI dataset. Despite achieving lower performance on PSNR and SSIM compared to GAN models, their method demonstrated superior performance on luminance and contrast similarity (LC), source-structural similarity (SS), visual information fidelity (VIF), and histogram correlation metrics. Additionally, the model demonstrated good generalizability when evaluated on the ACROBAT-HER2 dataset [114]. Stain quantification into four categories using ImageJ showed the highest accuracy (72%) for PST-DIFF compared to other models.

Similarly, Li *et al.* [86] introduced the Histopathological Image Multi-Domain Multi-Omics Translation with Diffusion Model (His-MMDM), which enables multi-domain translation within a unified framework by conditioning on the H&E and multiple target IHC stains. In addition to IHC stains, the model allows conditioning on mutations, tissue types, and other class-conditional inputs. The authors train their model using 27,530 H&E tiles and 13 IHC stains from an internal brain tumor dataset, including PgR (2,862 tiles) and Ki67 (4,846 tiles). Although they do not perform a direct evaluation of virtual IHC staining, they report FID-based metrics for cryosection-to-FFPE translation that are comparable to those achieved by the specialized model from Ozyoruk *et al.* [115].

4.2 Benchmarking

Beyond reviewing the SOTA, in this section, we conducted a benchmarking assessment of representative models in the field. In selecting our models, we prioritized representativeness, focusing

on those trained on MIST or BCI datasets and with publicly available code. As a result, we selected Pyramid Pix2Pix, ASP, and PSPStain. Pyramid Pix2Pix and ASP are seminal models in virtual staining, forming the foundation for subsequent work, while PSPStain exemplifies a model that directly addresses the issue of pathological consistency. We trained and validated our models using the MIST training and test sets, respectively, as no validation partition or information regarding the corresponding slide for each tile was provided. The best model checkpoint was selected on the basis of the average SSIM metric across the images. Furthermore, we used the test set from the BCI dataset to assess the generalizability of the models trained to predict HER2 staining. All hyperparameters were chosen based on the values reported in the respective articles, when available, or the default values provided in the code. Since the pyramid Pix2Pix was not originally trained on MIST, we adjusted its number of epochs to ensure the total number of training steps was equal to those in the original work. Our evaluation was based on SSIM, PSNR and FID, as these are the most reported metrics. Additionally, we utilized LPIPS with a VGG16 [116] network pre-trained on ImageNet [47], as this metric aligns with human perception of similarity. Moreover, since it evaluates the distance at the network’s activation level, we expect this metric to be more robust to the unpaired pixel-to-pixel nature of the data. As discussed in Chapter 3, the concentration of chromogen in an image correlates with the protein expression of the biomarker of interest. Therefore, we computed the JSD between the generated and real DAB channels to evaluate the consistency of the pathological information. To ensure a fair comparison, all random processes were initialized using the same seed, and all models were trained on an NVIDIA A100 GPU. Table 4.2 and Table 4.3 present the quantitative results comparing the models on the MIST test and BCI test sets, respectively, while Figure 4.2 and Figure 4.3 illustrate example images of the models in both cases.

Table 4.2: Performance metrics on the MIST test set (used for validation) comparing Pyramid Pix2Pix, ASP, and PSPStain. For metrics with $\uparrow$, higher values are better, while for metrics with $\downarrow$, lower values are better. The standard deviation reflects the variability of the metric values across all evaluated images. The best metrics for each IHC marker are highlighted in bold.

Marker	Model	Human Perception		Image Similarity	Noise	DAB
		LPIPS $\downarrow$	FID $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
ER	Pyramid Pix2Pix	0.600 ± 0.040	<u>150</u>	0.380 ± 0.123	<u>15.5 ± 3.6</u>	0.900 ± 0.173
	ASP	<u>0.589 ± 0.033</u>	226	<u>0.340 ± 0.108</u>	15.8 ± 3.1	<u>0.884 ± 0.150</u>
	PSPStain	0.551 ± 0.032	60	0.317 ± 0.113	14.7 ± 3.5	0.748 ± 0.270
PgR	Pyramid Pix2Pix	0.599 ± 0.035	<u>143</u>	0.379 ± 0.127	<u>15.5 ± 4.0</u>	<u>0.878 ± 0.221</u>
	ASP	<u>0.586 ± 0.031</u>	201	<u>0.366 ± 0.107</u>	15.6 ± 3.7	0.901 ± 0.144
	PSPStain	0.549 ± 0.030	67	0.304 ± 0.115	15.1 ± 3.5	0.822 ± 0.218
Ki-67	Pyramid Pix2Pix	0.598 ± 0.035	<u>145</u>	<u>0.377 ± 0.115</u>	15.3 ± 2.7	0.764 ± 0.277
	ASP	<u>0.583 ± 0.033</u>	227	0.384 ± 0.114	16.5 ± 2.2	0.965 ± 0.069
	PSPStain	0.548 ± 0.029	47	0.340 ± 0.098	<u>15.6 ± 2.3</u>	<u>0.788 ± 0.249</u>
HER2	Pyramid Pix2Pix	0.609 ± 0.027	<u>154</u>	0.349 ± 0.107	<u>15.3 ± 3.0</u>	<u>0.774 ± 0.221</u>
	ASP	<u>0.580 ± 0.032</u>	177	<u>0.304 ± 0.098</u>	15.8 ± 2.8	0.827 ± 0.187
	PSPStain	0.556 ± 0.027	54	0.263 ± 0.096	14.7 ± 2.7	0.746 ± 0.254

Table 4.3: Performance metrics for generalization on the BCI test set (HER2 staining) comparing Pyramid Pix2Pix, ASP, and PSPStain. For metrics with $\uparrow$, higher values are better, while for metrics with $\downarrow$, lower values are better. The standard deviation reflects the variability of the metric values across all evaluated images. The best and second-best metric values are highlighted in bold and underlined.

Model	Human Perception		Image Similarity	Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
Pyramid Pix2Pix	0.558 ± 0.058	316	0.624 ± 0.155	15.0 ± 5.1	<u>0.748 ± 0.234</u>
ASP	<u>0.543 ± 0.034</u>	<u>205</u>	<u>0.510 ± 0.144</u>	16.2 ± 3.3	0.732 ± 0.237
PSPStain	0.540 ± 0.046	177	0.462 ± 0.163	<u>15.6 ± 3.1</u>	0.971 ± 0.068

Upon analyzing Table 4.2, we observe that pyramid Pix2Pix generally achieves the highest SSIM, indicating a closer similarity between its outputs and ground truth images. However, a closer inspection of the example images in Figure 4.2 reveals that the pyramid Pix2Pix outputs exhibit significant smoothing, with minimal preservation of the H&E structure and morphology. This tendency is further confirmed on the BCI test set, where the SSIM metric indicates that pyramid Pix2Pix is the best-performing model. This result can be attributed to the unpaired pixel-to-pixel nature of the data, where the structures of the H&E images do not align with those of the IHC pairs. As a result, the SSIM metric may be biased toward highly blurred outputs.

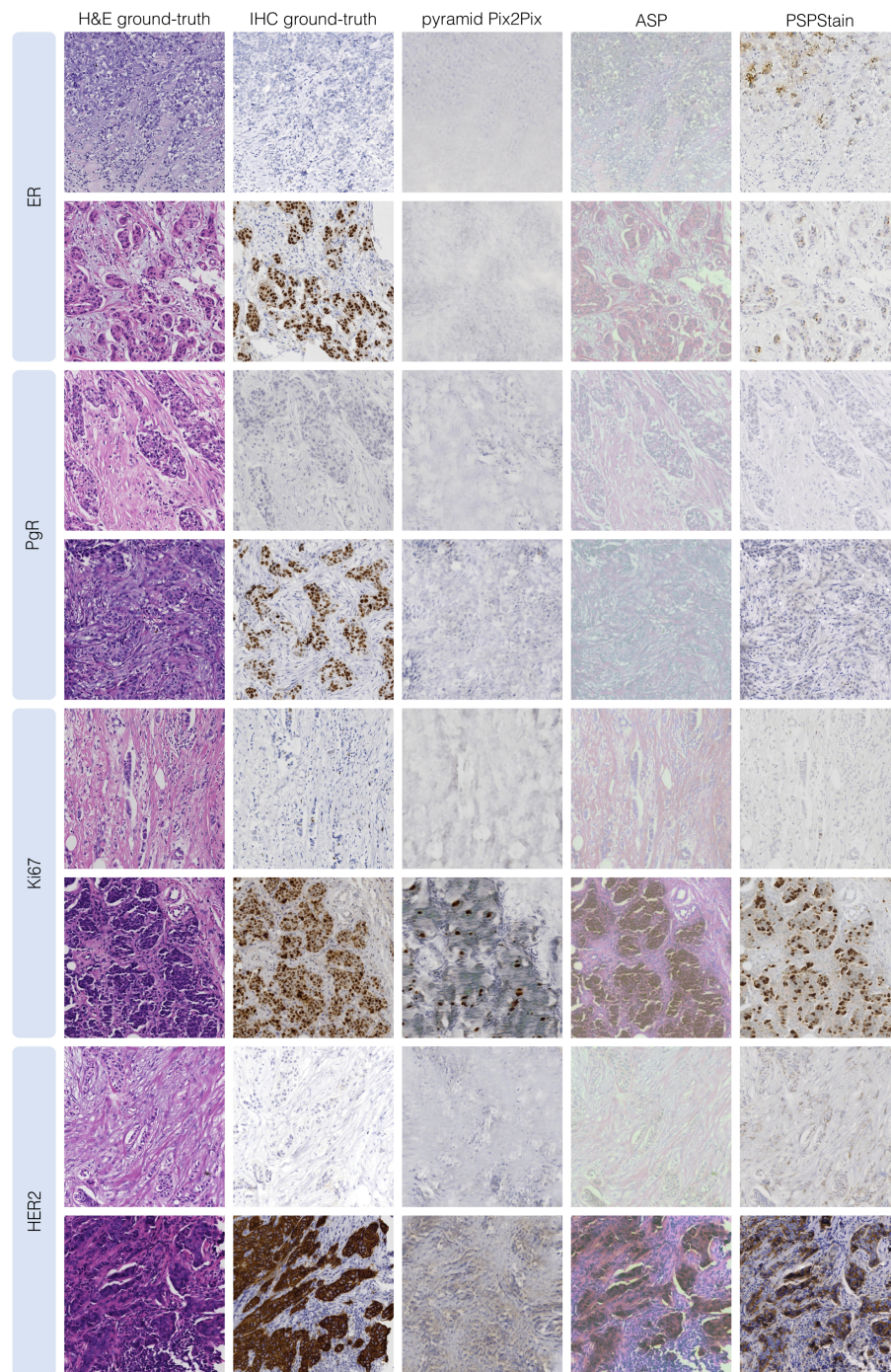


Figure 4.2: Benchmarking results examples from the MIST dataset, for ER, PgR, Ki67 and HER2 stainings.

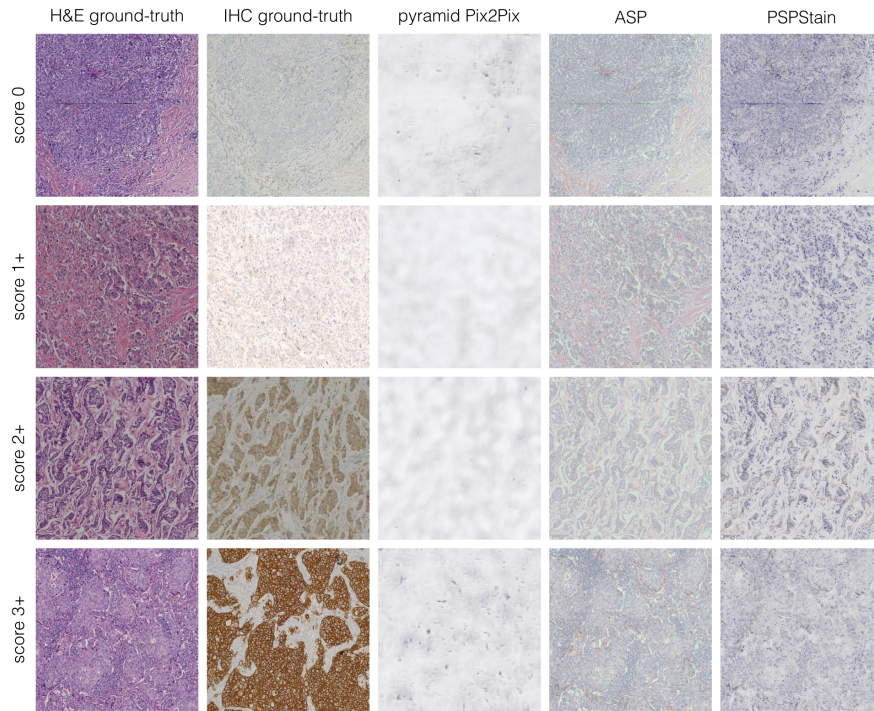


Figure 4.3: Benchmarking generalization results on HER2 examples from the BCI dataset.

In contrast, PSPStain achieves the best LPIPS and FID values on both the MIST test and BCI test sets, which aligns with the perceived quality of its output images and their similarity to the ground truth. As mentioned in Chapter 3, LPIPS and FID measure similarity by evaluating the distance between intermediate network activations and network embeddings, respectively. Consequently, it was expected that these metrics would be more robust to the unpaired pixel-to-pixel nature of the data, and this result confirms that expectation. Furthermore, since embeddings capture high-level semantic information, measuring the distance between embeddings can be useful for assessing the consistency of pathological information. An analysis of the PSPStain output on the MIST test set, along with its generally better performance on JSD computed in the DAB channel, supports this claim. However, this result does not extend to the BCI test set, where PSPStain achieves the best FID performance but the worst JSD computed in the DAB channel. However, it is important to note that FID uses InceptionV3 pre-trained on ImageNet, meaning it cannot be guaranteed that the extracted embeddings contain meaningful pathological information, which helps explain the observed results. Similarly, this reasoning can be applied to LPIPS, which may explain why ASP performs second-best, despite its color characteristics being closer to the input H&E. Nonetheless, it is still possible to distinguish the positively DAB-stained regions.

From the MIST test to the BCI test sets, we observe a similar performance trend across the models, with the exception of the JSD computed in the DAB channel. On the MIST test set, PSPStain performs best, which aligns with the inclusion of the dedicated pathology consistency methodology used by the authors. However, this performance does not extend to out-of-domain

(OOD) distributions, highlighting the need to address the generalization problem in virtual staining.

4.3 Limitations in the State-of-the-Art

The SOTA analysis demonstrated how virtual staining from H&E to IHC for breast biomarkers is an emerging area of research, characterized by unique approaches and challenges within the field of image-to-image translation. Beginning with the work of Liu *et al.* [57] in 2021, the field has grown rapidly, with 64% of existing models just published in 2024.

The majority of models are based on the GAN framework, with a particular emphasis on CycleGAN and Pix2Pix. Notably, the work of Liu *et al.* [43] and Li *et al.* [26] introduced the two most seminal models in the field: Pyramid Pix2Pix and ASP, respectively. In addition to developing these models, the authors introduced the BCI and MIST datasets, the first being publicly available. As a result, their work laid the foundation for subsequent research and now serves as the standard benchmark for model comparison.

Although widespread use of the BCI and MIST datasets facilitates comparison between developed models, the fact that these datasets consist of registered pairs of H&E and IHC from consecutive tissue sections presents significant challenges in training, particularly in achieving pathological consistency. In this context, in addition to assuming that the H&E morphology encodes the necessary information to predict IHC expression, we also assume that the contiguous tissue section retains the same diagnostic information. To address this issue, most authors increase the complexity of their model architectures and training methodologies. However, it is important to note that, in this context, cells present in one tissue section may not even be present in the consecutive one. As a result, this not only poses difficulties in training but also precludes a robust evaluation of the developed models.

In addition to the SOTA study, we conducted a benchmark comparison between Pyramid Pix2Pix, ASP, and PSPStain. While Pyramid Pix2Pix and ASP are foundational models in the field, PSPStain builds upon Pyramid Pix2Pix to directly address the issue of preserving pathological information, allowing us to explore how pathological-dedicated loss functions contribute to achieving pathological consistency. We trained and validated the models using the MIST train and test sets, respectively, and further used the BCI test set to assess the generalizability of the models trained to predict HER2 staining.

Our results revealed that Pyramid Pix2Pix performed best in SSIM and ranked second for PSNR. However, its outputs exhibited significant blurring, thus lacking cellular resolution and preservation of the input H&E structures. This finding highlights a critical limitation of classical image metrics in the evaluation of virtual staining models within the context of BCI and MIST datasets, serving as a critique of the current SOTA, as many studies still rely on these metrics to claim improved performance. On the other hand, despite PSPStain performing worst on SSIM and PSNR, it achieved the best LPIPS and FID metric values. An inspection of its output revealed

that these two metrics align more closely with perceived quality, highlighting how network-based metrics can be more robust than traditional image metrics.

The study of PSPStain further demonstrated how its pathology-dedicated framework enabled it to achieve the best performance on the JSD computed in the DAB channel, positioning it as the leading model for pathological consistency. However, when tested on the BCI dataset, it performed the worst, highlighting a critical issue of generalizability. This challenge has often been overlooked, with many studies opting to retrain their models when using external datasets. Factors such as differences in scanners, study sites, tissue types, and diverse demographics are common in clinical settings, which makes robustness to OOD distributions crucial for successful virtual staining. Therefore, focusing on robustness and generalizability, future work could involve training models on multiple heterogeneous datasets to account for both H&E and IHC staining variance. Similarly, approaches such as training the same model on different stains [117] might achieve similar results in learning more robust features. However, there is a lack of established best practices for virtual staining, common in computational pathology, such as using stain augmentations and stain normalization methods [118]. More in-depth discussions of other challenges for clinical practice can be found in the comprehensive reviews on virtual staining by Bai *et al.* [35] and Latonen *et al.* [9].

Additionally, it is important to note that current studies are based on the assumption that, although not possible for a pathologist, deep generative learning models can predict biomarker expression at the cellular level by analyzing the morphology and general patterns in the H&E-stained image. Future research should focus on model interpretability and explainability, as this assumption may not be true. The learning of non-causal features in the in-domain distribution could explain why a virtual staining model fails to generalize to OOD distributions.

In our benchmarking, we used the JSD computed in the DAB channel to assess pathological consistency, as chromogen staining reveals the presence of the target antigen, thus encoding crucial pathological information. While most authors rely on the DAB channel for training and evaluation, there is a lack of consensus or a gold standard metric to assess the pathological quality of output images. Furthermore, we emphasize that the value of a virtual staining model lies in its ability to enable the same diagnostic conclusions from both generated and real IHC images. Although these metrics constitute a reference point in achieving this objective, the efficacy of a virtual staining model can only be fully validated through downstream tasks and/or expert evaluation. Most studies overlook the importance of these assessments; however, we recognize that many of the models developed so far have not achieved sufficient performance to be tested at this level, as demonstrated in our benchmarking. Nonetheless, studies that do evaluate the quality of their models at this level often choose expert evaluation. These assessments may be less reproducible, but can be more pathologically meaningful than downstream tasks, especially for HER2. Until now, the only downstream task used for this biomarker was conducted by Peng *et al.* [66], which involved assessing the difference in performance of a classification model trained on both generated and real IHC images to predict HER2-score. However, since tile labels were derived from the slide, this task can be considered noisy and less representative, as slide-level labels, particularly for 3+

HER2-scores, do not always align with the information in the tile.

4.4 Summary

This chapter comprehensively reviews the SOTA methods for virtual staining of H&E to IHC breast biomarkers. This is a timely and emerging topic, with 64% of the models published in 2024. Among the works studied, we highlight those of Liu *et al.* [43] and Li *et al.* [26], who not only laid the groundwork for subsequent models with pyramid Pix2Pix and ASP but also contributed with BCI and MIST datasets, which constitute the standard benchmark for model comparison to date. However, these datasets consist of registered pairs of H&E and IHC images from consecutive tissue sections, which, more importantly than posing a challenge to training, hinder a robust and meaningful evaluation. When using these datasets, in addition to assuming that it is possible to predict cell biomarker expression from H&E morphology, we also assume that pairs of images from contiguous tissue sections share the same diagnostic information, even though they do not share the same structures and might even present different cells. Overlooking this issue, most authors have developed complex frameworks and training methodologies based on cGAN, treating the adversarial loss component as an accessory and not recognizing its impact on the quality of the virtually generated images. Furthermore, in their pursuit of outperforming previous models, most authors report improved performance based on SSIM and PSNR, which, as discussed in the Benchmarking 4.2 section, are not robust enough to evaluate the quality of virtual staining models.

Chapter 5

Model Development

Considering the gaps identified in the current SOTA in the previous chapter, we develop our proposed model on the HER2match dataset [10]. We start by describing the implementation details, namely the discriminator and generator architectures, the training methodology, and the dataset. The subsequent sections outline each step of the model development, with each section representing one iteration toward our final virtual staining model. As we progress, we evaluate the impact of each component, particularly the adversarial loss, on the quality of the generated images, while also discussing the strengths and limitations of the existing evaluation metrics.

5.1 Implementation Details

5.1.1 Discriminator Architecture

Building upon the work of Li *et al.* [26], we employed a PatchGAN architecture for the discriminator. As described in Chapter 3, this architecture classifies non-overlapping image patches as real or fake, producing a probability map where each pixel value indicates the likelihood of the corresponding patch being real. The network consists of six layers, with the first four blocks comprising a 2D convolutional layer, a BatchNorm [119] layer, and a LeakyReLU [120] activation. After each block, we use a custom separable convolution for efficient and smooth downsampling. It applies a Pascal’s triangle-based smoothing filter in two steps: a vertical 1D convolution followed by a horizontal 1D convolution, both with a stride of two. Reflection padding is applied before convolution to maintain dimensional integrity. Using this methodology, we prevent aliasing artifacts and reduce computation compared to standard 2D convolutions. The final two layers do not involve downsampling and, together with the final sigmoid activation, produce a single-channel prediction probability map.

Training instability is a major challenge in GANs. Therefore, we carefully selected BatchNorm as the intermediate normalization layer, as it was originally developed to accelerate and stabilize the training of discriminative networks by reducing the dependence of gradients on parameter scale and initial values [119], [121]. This choice is strongly motivated by the issue of

internal covariate shift, where the input distribution to each layer changes as preceding layers update during training. Such a shift slows down training by necessitating lower learning rates and careful parameter initialization. Based on empirical evidence that networks converge faster when inputs are whitened, meaning linearly transformed to have zero mean and unit variance, Ioffe and Szegedy [119] designed BatchNorm, as described in Equation 5.1.

$$\text{BN}(x) = \gamma \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \beta \quad (5.1)$$

In this equation, $\mu(x)$ and $\sigma(x)$ denote the mean and standard deviation computed across the batch size and spatial dimensions, separately for each feature channel. The parameters γ and β are learnable affine coefficients. This formulation not only normalizes the input of each layer but also allows the model to learn the most appropriate distribution for each layer. Additionally, BatchNorm uses mini-batch statistics during training and replaces them with population statistics during inference.

In addition to utilizing BatchNorm, we also employed spectral normalization. This technique was specifically designed to stabilize GAN training by ensuring that the discriminator is 1-Lipschitz continuous, thus preventing issues such as exploding gradients and mode collapse [122]. Furthermore, empirical evidence shows that training GANs with spectral normalization results in higher quality generated images and faster convergence [122]. This is achieved by normalizing the weight matrices of each layer in the discriminator according to their spectral norm, as illustrated in Equation 5.2, where $\sigma(W)$ represents the matrix spectral norm. To integrate spectral normalization into PatchGAN, we used the native PyTorch implementation [123], which employs power iteration to compute the right and left singular vectors, thus approximating $\sigma(W) \approx \tilde{u}^T W \tilde{v}$, as directly computing the singular value decomposition is computationally expensive.

$$\overline{W}_{SN} = \frac{W}{\sigma(W)} \quad (5.2)$$

5.1.2 Generator Architecture

We used a U-Net architecture for our generator based on the work of Ronneberger *et al.* [45]. This architecture differs from standard encoder-decoder designs by incorporating skip connections between the encoder and decoder. These connections enable the decoder to access high-resolution, low-level features from the encoder, preserving fine-grained details and propagating contextual information to higher-resolution layers. Additionally, skip connections provide early feedback during backpropagation, helping to mitigate vanishing gradients and improve convergence. The success of this network architecture extends beyond semantic segmentation. According to Isola *et al.* [44], U-Net outperforms general encoder-decoder architectures in various tasks within cGANs. Moreover, the authors demonstrate that the U-Net architecture remains effective even when the predicted pixels are not geometrically aligned with the information in the input.

Contrary to the methodology employed by Ronneberger *et al.* [45], we do not crop the incoming feature map from the skip connection before concatenating it with the upsampled feature map

in the decoder. Instead, we pad the upsampled feature map with zeros to match the dimensions of the skip connection feature map. Furthermore, we replace the transposed convolution with a sequence of bilinear interpolation followed by a 1×1 convolution, which reduces the number of learnable parameters and helps prevent checkerboard artifacts [124]. At the output of our decoder, we use a sigmoid activation to ensure that the pixel values lie within the $[0, 1]$ range [125].

Due to its efficacy and widespread use in computational pathology, we replaced the original U-Net encoder with a ResNet-50 backbone [126]. This architecture employs residual connections that facilitate gradient flow and mitigate the problem of vanishing gradient, thus helping to stabilize GAN training. For implementation, we used the ResNet-50 model with pre-trained ImageNet weights [47] from PyTorch’s torchvision library [127], inserting skip connections between the encoder and the decoder immediately after the initial input transformation layer and following each main sequence of bottleneck blocks. To avoid channel mismatches between encoder and decoder feature maps, we apply a 1×1 convolution at the skip connection level.

Regarding the normalization layer, we should note that the original U-Net architecture does not use any form of layer normalization, whereas ResNet-50 employs BatchNorm. However, for our generator, we found InstanceNorm to be more appropriate, as it has been empirically shown to perform better in feedforward-style transfer tasks [121], [128]. InstanceNorm follows the same mathematical formulation as BatchNorm, as demonstrated in Equation 5.1; however, $\mu(x)$ and $\sigma(x)$ are computed independently across spatial dimensions for each channel and each sample, unlike BatchNorm, which computes these statistics across the batch dimension. Additionally, BatchNorm replaces mini-batch statistics with population statistics during inference, whereas InstanceNorm layers are applied unchanged at inference time. As demonstrated by Huang *et al.* [121], models trained with BatchNorm on style-normalized images can converge as quickly as models trained with InstanceNorm on original images, suggesting that InstanceNorm effectively performs style normalization. Based on this evidence, the authors argue that the success of InstanceNorm stems from its ability to normalize the individual style of each sample within a batch, allowing the rest of the network to focus on content manipulation. In contrast, when using BatchNorm, samples within the same batch may retain different styles. Although network layers can learn to compensate for these intra-batch style variations, this adds complexity and challenges to training. Therefore, we replaced BatchNorm in ResNet-50 with InstanceNorm and applied InstanceNorm before ReLU activations [120], [125] throughout the rest of the network, to further stabilize the GAN training and enhance the generator performance.

5.1.3 Training Methodology

In our training methodology, we alternately optimize the discriminator, D , and the generator, G , to prevent performance imbalance between the two models and to keep D near its optimal solution, providing meaningful feedback for updating G [39]. Recalling the cGAN optimization criterion presented in Equation 3.2 of Chapter 3, we define the loss functions for D and G as outlined in Equation 5.3 and Equation 5.4, respectively. Here, $\mathbf{1}$ represents a tensor of ones, while B , H , and

W denote the batch size, height, and width dimensions of the prediction output map.

$$\mathcal{L}_D(\mathbf{X}, \mathbf{Y}, \mathbf{Y}_D) = -\frac{1}{BHW} \mathbf{1}_{BHW}^\top \cdot \left(\mathbf{Y}_D \odot \log(D(\mathbf{Y}, \mathbf{X})) + (\mathbf{1}_{BHW} - \mathbf{Y}_D) \odot \log(\mathbf{1}_{BHW} - D(G(\mathbf{X}), \mathbf{X})) \right) \cdot \mathbf{1}_{BHW} \quad (5.3)$$

$$\mathcal{L}_G(\mathbf{X}) = \frac{1}{BHW} \mathbf{1}_{BHW}^\top \cdot \left(\log(\mathbf{1}_{BHW} - D(G(\mathbf{X}), \mathbf{X})) \right) \cdot \mathbf{1}_{BHW} \quad (5.4)$$

Using these loss functions, we train D to maximize the correct classification of both real and generated samples, while training G to minimize the probability that its generated images are classified as fake by D . However, in practice, the formulation of Equation 5.4 may easily saturate, especially early in training when G is weak and D confidently rejects generated samples. To address this, we train G to maximize $\log(D(G(\mathbf{X}), \mathbf{X}))$ instead. This alternative objective shares the same fixed point in the dynamics of G and D but provides stronger gradients during early learning [39]. Essentially, this re-frames the problem: rather than minimizing the probability that generated images are classified as fake, G seeks to maximize the probability that its images are classified as real [129].

$$\mathcal{L}_G(\mathbf{X}) = -\frac{1}{BHW} \mathbf{1}_{BHW}^\top \cdot \left(\log(D(G(\mathbf{X}), \mathbf{X})) \right) \cdot \mathbf{1}_{BHW} \quad (5.5)$$

Two main modifications distinguish our cGAN implementation from the original. First, we do not input a noise vector $\mathbf{Z}$ to G . In many real-world scenarios, there is not just a single correct output: for example, translating a sketch to a photo can have multiple valid interpretations [44]. In such cases, $\mathbf{Z}$ introduces stochasticity into the generation process. However, we want to achieve an accurate and consistent mapping from H&E to IHC images. Therefore, we omit $\mathbf{Z}$ from our image generation. Second, we concatenate the corresponding H&E pair to the IHC image before inputting it into D . This approach helps prevent mode collapse during training [44], which manifests when G outputs the same or similar output regardless the input. More importantly, in addition to leveraging D to ensure that the generated IHC image appears realistic, this methodology preserves the corresponding H&E information during image generation. We highlight that this approach is innovative compared to the studied SOTA models and assume that the pathological information present in the H&E image is consistent with its IHC counterpart, since both originate from the same tissue section.

The training step begins with the generation of fake IHC images (Figure 5.1). We then optimize D by feeding it real and fake examples. Since we focus solely on updating D at this stage, the generated images are detached from the computational graph to prevent gradients from flowing through G . Finally, we optimize G using the generated fake images, with D serving as the network loss function.

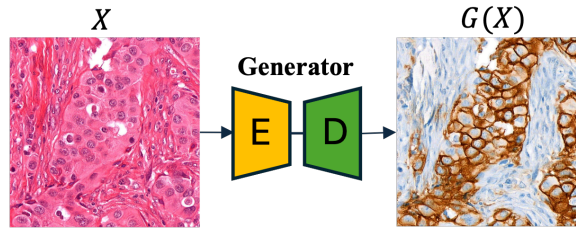
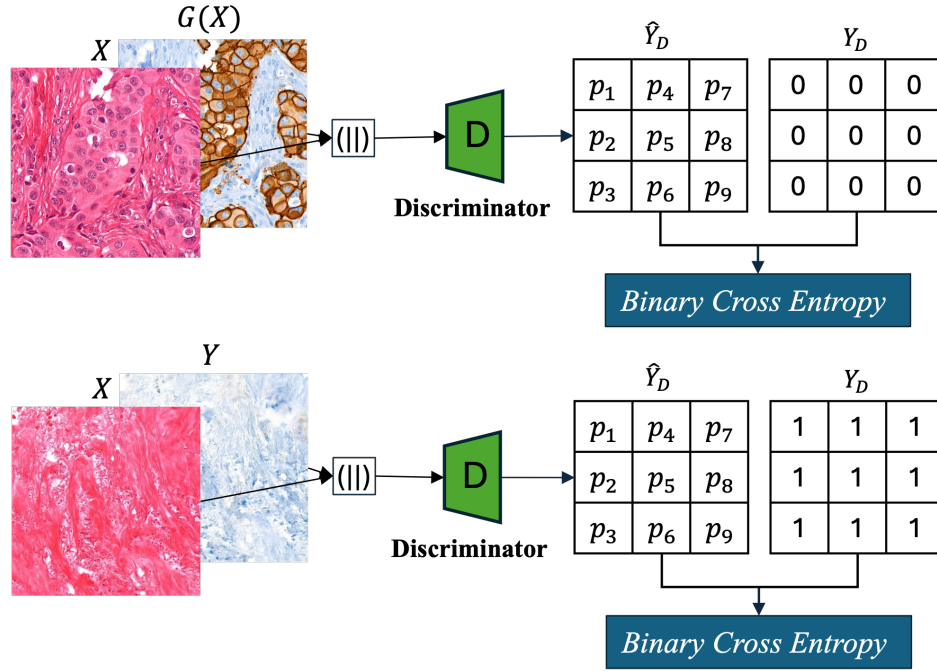
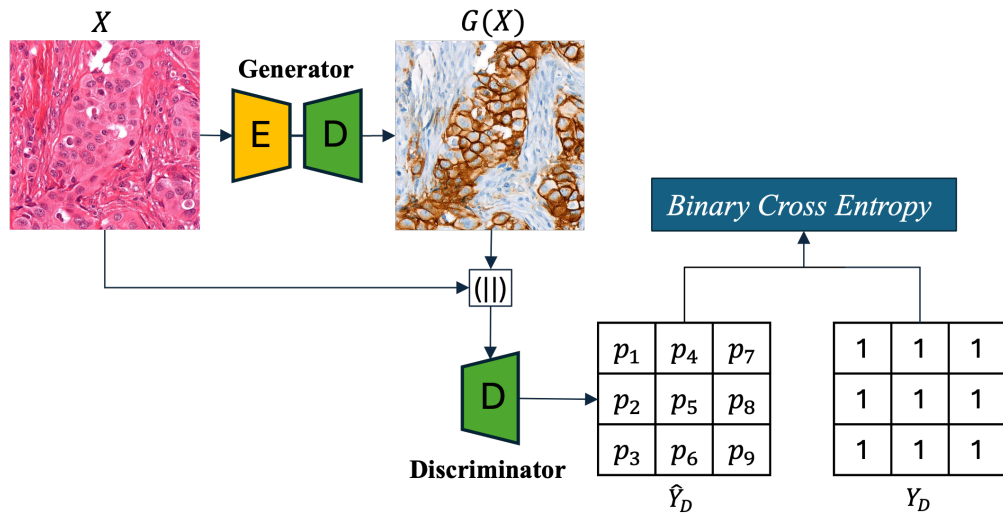
1. Generate Images2. Optimize Discriminator3. Optimize Generator

Figure 5.1: Illustration of the training step.

Since maintaining stability between D and G is paramount when training GANs, we recorded the loss values of each network at every training step, averaged them at the end of the training epoch, and logged these averages to monitor stability. Similarly, we used the same routine to monitor the accuracy of D . However, since PatchGAN outputs the likelihood of each patch being real rather than the entire image, computing accuracy is not straightforward. Therefore, we followed the procedure outlined in Equations 5.6–5.10, where $\mathbf{1}_{\text{HW}}$ is a tensor of ones for summation, and $\mathbf{1}(\cdot)$ is the indicator function returning 1 if the condition is true and 0 otherwise.

$$1. \text{ Computing Votes: } \mathbf{V} = \hat{\mathbf{Y}}_D > \theta \in \{0, 1\}^{B \times 1 \times H \times W} \quad (5.6)$$

$$2. \text{ Calculating Mean Vote per Example: } \bar{\mathbf{V}} = \frac{1}{IHW} \mathbf{1}_{\text{IHW}}^\top \cdot \mathbf{V} \cdot \mathbf{1}_{\text{IHW}} \in [0, 1]^B \quad (5.7)$$

$$3. \text{ Decision per Example: } \hat{\mathbf{y}} = \bar{\mathbf{V}} > \theta \in \{0, 1\}^B \quad (5.8)$$

$$4. \text{ Reducing Ground Truth Labels: } \mathbf{y}_D = \max_{h,w} \mathbf{Y}_D \in \{0, 1\}^{B \times 1} \quad (5.9)$$

$$5. \text{ Computing Accuracy: } \text{Accuracy} = \frac{1}{B} \mathbf{1}_B^\top \mathbf{1}(\hat{\mathbf{y}} = \mathbf{y}_D) \quad (5.10)$$

Throughout our experiments, we verified that the GAN training remained stable as long as the accuracy of D remained close to 0.50 and the loss values of both networks were comparable (Figure 5.2). Additionally, we observed the expected behavior of D outperforming G early in the training, as evidenced by high accuracy and low loss values.

All experiments were run on an NVIDIA A100 GPU, with all random processes initialized using a common seed to ensure reproducibility. During training, we used bfloat16 mixed-precision to optimize memory usage and computational efficiency without a significant loss of model accuracy [130], while using float32 during inference for higher numerical precision. We conducted all experiments using 200 epochs and a batch size of 8. For both G and D , we utilized the AdamW optimizer [131], reducing the default β_1 to 0.5, since the use of a more aggressive modifier may cause convergence failure [132], and applied a weight decay of 1×10^{-5} . Furthermore, high learning rates also lead to non-convergence and mode collapse; therefore, we chose an initial learning rate of 5×10^{-5} and employed an exponential learning rate decay scheduler as defined in Equation 5.11, where $\gamma = 0.995$.

$$\eta_t = \eta_0 \times \gamma^t \quad (5.11)$$

Since the ResNet-50 backbone was initialized with ImageNet pre-trained weights, we normalized the H&E images using ImageNet’s mean and standard deviation values before inputting them to G . During training, we further randomly cropped patches of size 512×512 pixels and applied random horizontal and vertical flips, with a probability of 0.5 each, to the corresponding H&E and IHC images as a data augmentation step.

5.1.4 Model Checkpoint and Evaluation

Classic per-pixel measures assume pixel-wise independence and are therefore insufficient for assessing structured outputs such as images. Since PSNR is based on MSE, it inherits the same

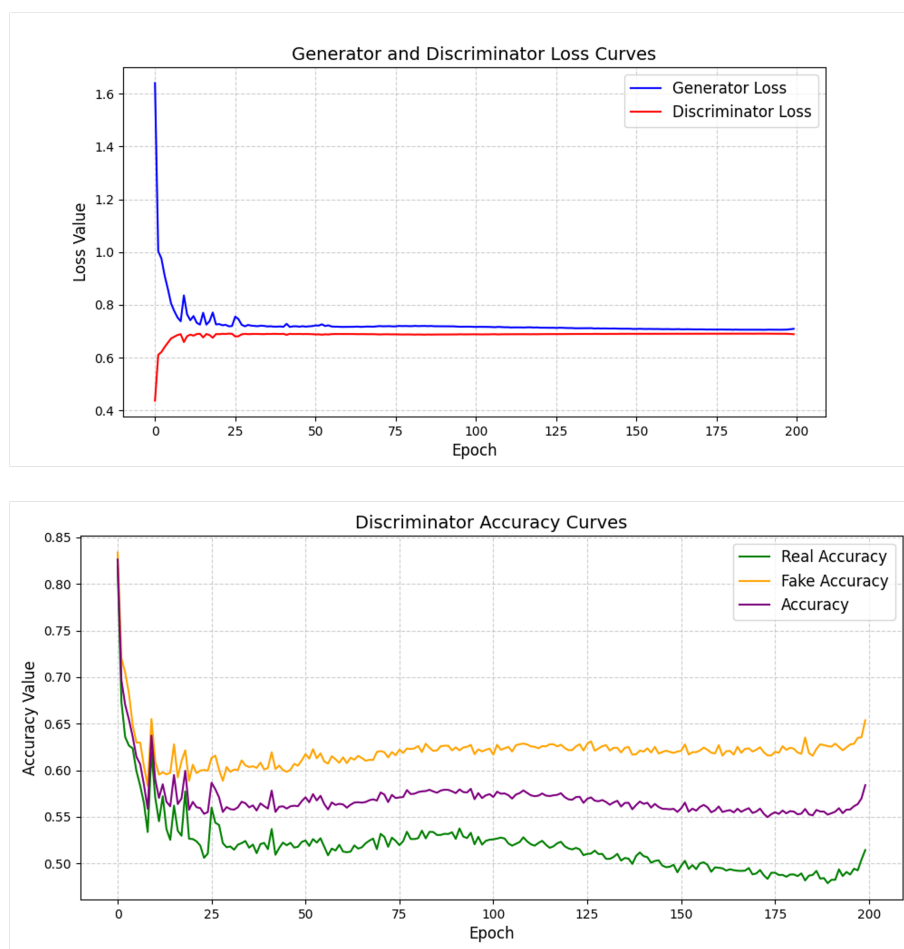


Figure 5.2: Training loss curves and D accuracy for the BCE GAN baseline model.

limitations, making it less suitable as an evaluation metric. Although more robust than per-pixel measures and simple to compute, SSIM and its variants fail to capture many nuances of human perception of similarity. Furthermore, as highlighted by Zhang *et al.* [58] and confirmed in the Benchmarking Section 4.2 of Chapter 4, these metrics tend to favor blurred outputs and are not designed to handle cases with significant geometric distortions. Supported by evidence from Zhang *et al.* [58], which shows that LPIPS correlates well with human perception of similarity, and considering that LPIPS values generally aligned with the visual quality of generated images in our benchmarking, we selected LPIPS as the evaluation metric to choose the best model in the validation set (excluding the first epoch) and as the primary reference metric in our experiments. FID and KID are also robust to pixel-level misalignment; however, since they evaluate differences between the distributions of embeddings from real and generated images, their assessment focuses more on preserved semantic information than on the visual aspect of the images. Additionally, these metrics rely on pre-trained ImageNet weights; thus, they are not guaranteed to correlate with the pathological information in the images. Furthermore, the results of these metrics are biased by the number of examples, making comparisons between validation and test sets less reliable due to the large difference in sample sizes. In addition to metric evaluations, we asked pathologists to

select one representative image per HER2-score from the validation and test sets, allowing us to assess how the metric values align with the visual quality of the generated images.

5.1.5 Dataset

The HER2match [10] dataset includes 17 WSIs of breast tissue stained with H&E, followed by de-staining and re-staining with HER2 IHC, resulting in H&E-IHC sample pairs from the same tissue section (Figure 5.3). The slides were curated, registered, and tiled into 1024×1024 pixel tiles at the highest resolution of $0.25 \mu\text{m}/\text{px}$, resulting in 21,172 tile pairs. Subsequently, the authors split the dataset at the per-patient level, ensuring that each set contained at least one case for each HER2 score, resulting in 11,610 tiles for training, 3,582 for validation, and 5,980 for testing.

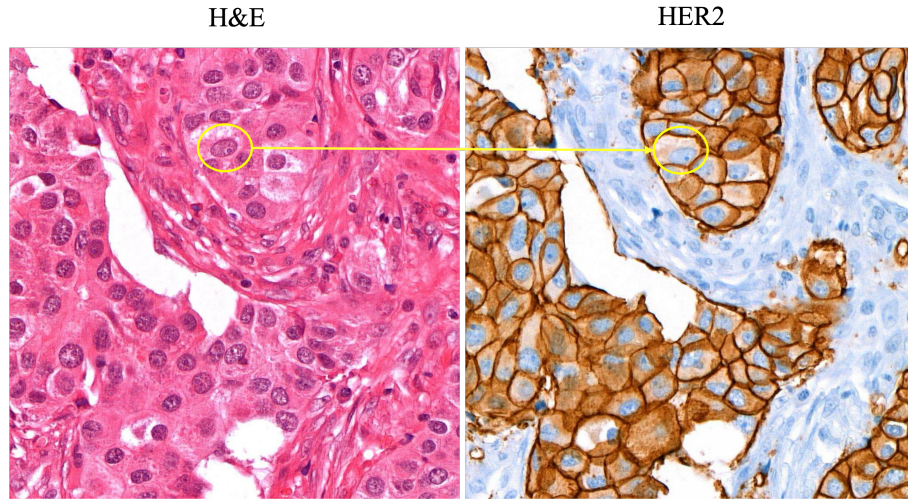


Figure 5.3: Example tile from HER2match dataset demonstrating the high degree of alignment between the H&E and IHC pairs.

5.2 BCE GAN

Following the training methodology outlined earlier, this section presents the results of our baseline BCE GAN model. We also investigated the role of adversarial loss in enhancing the perceptual quality of the generated images and preserving semantic information.

5.2.1 Training Methodology

To simplify the study of the impact of adversarial loss, we first trained G using MSE loss, as defined in Equation 5.12. This choice was motivated by MSE’s widespread use, simplicity, efficiency, and its status as a standard benchmark for general image-to-image translation tasks [133]. Next, we trained our baseline model (BCE GAN) and a modified version in which D is not conditioned on the corresponding H&E pair (BCE ncGAN).

$$\mathcal{L}_G = \frac{1}{BCHW} \mathbf{1}_{BCHW}^\top \cdot (\mathbf{Y} - G(\mathbf{X}))^2 \cdot \mathbf{1}_{BCHW} \quad (5.12)$$

5.2.2 Adversarial Learning in improving Perceptual Quality

According to Lee *et al.* [134], since per-pixel distance functions such as $\mathcal{L}_1$ and $\mathcal{L}_2$ are applied independently to each pixel, their use as a loss function guides the network toward estimating the mean RGB value for each pixel. As a result, these functions enable the model to capture low-frequency features but fail to encourage high-frequency details, often producing overly smoothed and non-realistic images [44]. We observe that the U-Net results preserve only the global H&E morphology and structure, lacking cell-level resolution due to significant smoothing (Figure 5.4-5.5). As presented in Table 5.1, this results in a notable decrease in performance in human perception metrics and CSS compared to BCE ncGAN (Table 5.1). For reference, we also present the results of the Identity model, which outputs the exact same input H&E (resulting in a perfect CSS metric value).

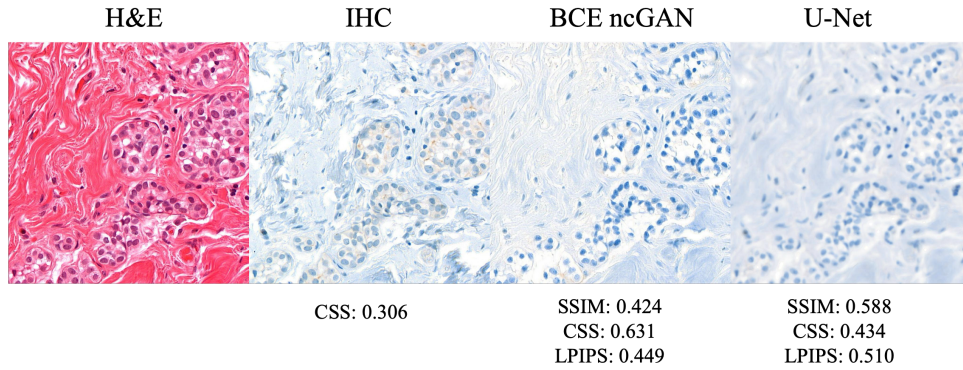


Figure 5.4: Comparison between BCE ncGAN and U-Net on a high resolution example output from the validation set with a 1+ HER2-score.

Table 5.1: Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN, BCE ncGAN and U-Net experiments. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best and second-best metric values are highlighted in bold and underlined.

Model	Human Perception			Image Similarity			Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	CSS $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
Identity	0.664 ± 0.079	252	26.09 ± 1.67	0.341 ± 0.106	0.310 ± 0.098	1.00	9.85	—
U-Net	0.500 ± 0.070	102	8.52 ± 0.86	0.606 ± 0.111	0.582 ± 0.108	0.485 ± 0.104	22.2 ± 3.7	0.879 ± 0.173
BCE ncGAN	<u>0.466 ± 0.071</u>	<u>73</u>	<u>4.46 ± 0.80</u>	0.472 ± 0.122	0.434 ± 0.122	0.654 ± 0.068	18.9 ± 4.1	0.860 ± 0.197
BCE GAN	0.458 ± 0.067	34	1.17 ± 0.56	0.503 ± 0.120	0.464 ± 0.120	0.414 ± 0.104	<u>19.8 ± 3.9</u>	0.890 ± 0.193

In contrast, U-Net achieves the highest SSIM, MS-SSIM, and PSNR metric values among the three models, providing empirical evidence that even with a higher-quality dataset of pairs sampled

from the same tissue section, SSIM and its variants, as well as PSNR, tend to favor blurred images. However, we note that even when sampled from the same tissue section, structural dissimilarities can still arise between H&E-IHC pairs due to mechanical and chemical treatments during re-staining, which can deform the tissue [135]. As a result, pixel-level alignment is not always guaranteed, as demonstrated by the low CSS value of the ground-truth IHC image (Figure 5.4). This highlights the limitations of these metrics in evaluating the quality of virtually generated images, as previously discussed in Chapter 4.

On the other hand, it appears that U-Net performs better in capturing chromogen staining (Figure 5.5). In the image example with a 3+ HER2-score, the U-Net captures the characteristic dark brown of positive DAB regions, while BCE ncGAN and BCE GAN fail to do so.

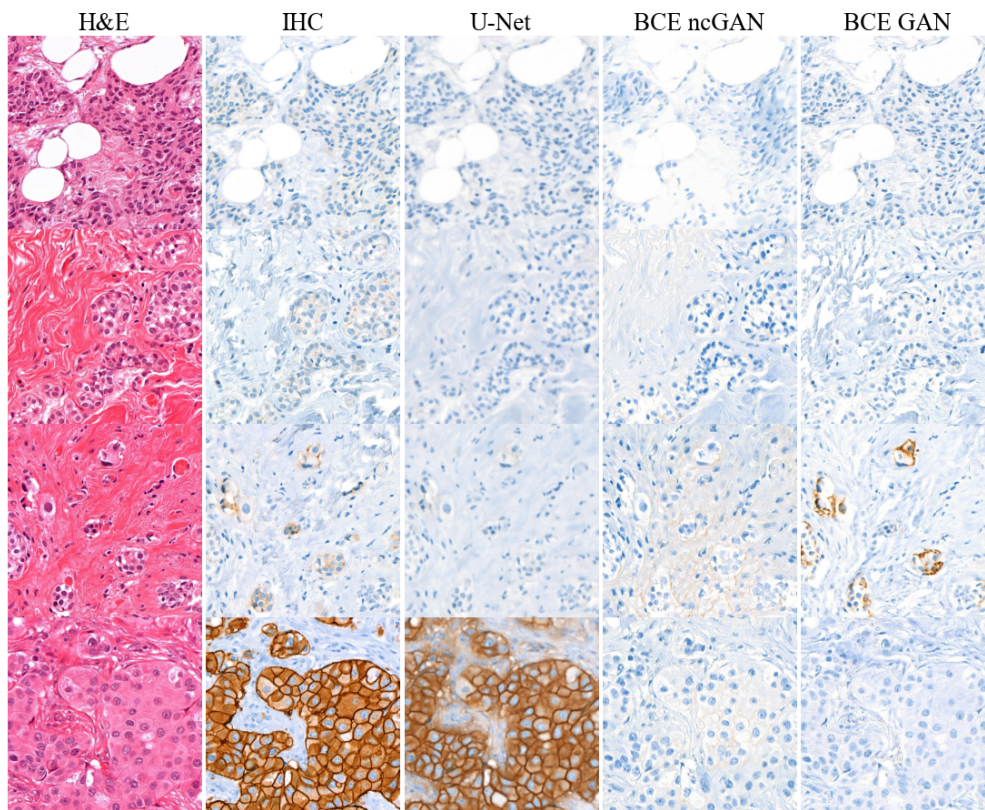


Figure 5.5: Qualitative assessment of the BCE ncGAN, BCE GAN, and U-net models on the validation set of the HER2match dataset.

Based on this observation, we selected two additional subsets from the validation set: one containing cases of 3+ HER2-score and intense positive DAB staining (Figure 5.6), and another without such staining (Figure 5.7). We observe that despite demonstrating higher accuracy in identifying cases with a 3+ HER2-score, the U-Net also tends to overstain. In contrast, BCE ncGAN missed positive DAB staining in cases with a 3+ HER2-score. On the other hand, although BCE GAN demonstrates lower accuracy in distinguishing cases with a 3+ HER2 score compared to U-Net, it localizes positive DAB regions with greater precision. In addition, it produces more

realistic and higher-resolution images, making it the most promising model. However, it is important to note that false-positive cases stained with this model often present chromogen staining along the edges, particularly at the interface between tissue and background. Although HER2 is a membrane target protein, and this feature could be correlated, it is certainly not causal for these regions to be represented as positive DAB staining.

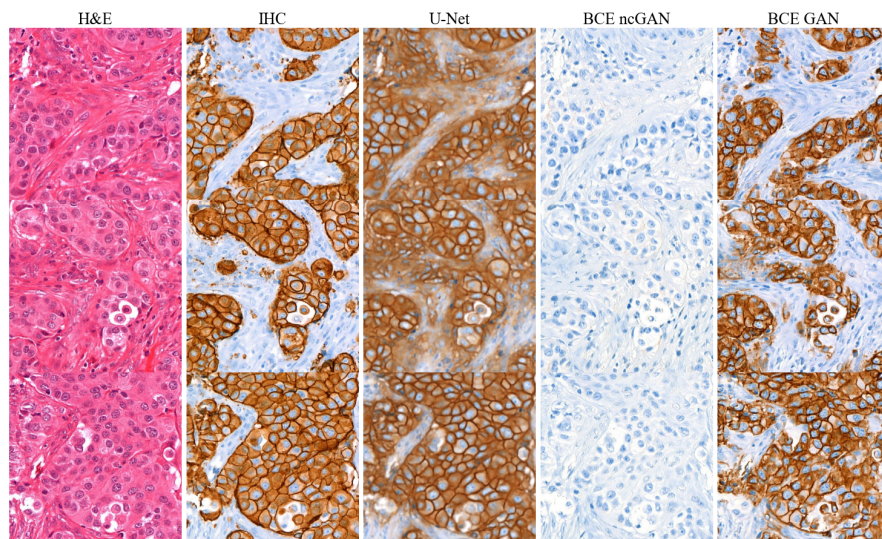


Figure 5.6: Qualitative comparison between the BCE ncGAN, BCE GAN, and U-Net performances on the selected image examples from the validation set with a 3+ HER2-score and intense positive DAB stain.

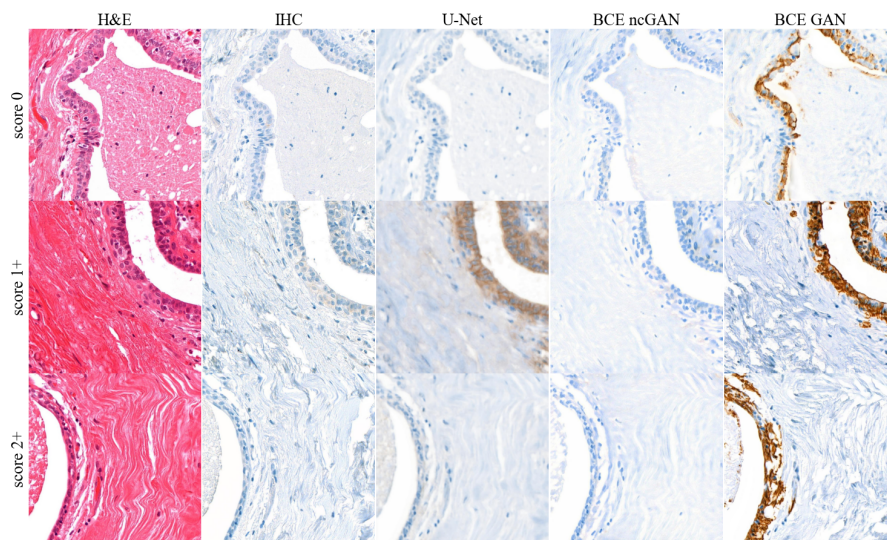


Figure 5.7: Qualitative comparison between the BCE ncGAN, BCE GAN, and U-Net performances on the selected image examples from the validation set without positive DAB staining.

5.2.3 Effect of conditioning the Discriminator on the H&E

By conditioning D on the paired H&E image, we aimed to preserve the information contained in the H&E during the transformation to IHC. Although further tests are needed to confirm that pathological information in the H&E image matches its IHC counterpart and that the preserved information is indeed meaningful, the significantly lower FID and KID scores achieved by conditioning on the H&E (Table 5.1), along with improved visual quality of positive DAB staining, and increased accuracy in identifying cases with 3+ HER2-score (Figures 5.6-5.7), suggest progress in this direction. However, this conditioning comes at the cost of reduced performance in the CSS metric (Table 5.1 and Figure 5.8), indicating that BCE GAN is less effective at preserving H&E morphology. An ideal virtual staining model should be capable of predicting how an H&E

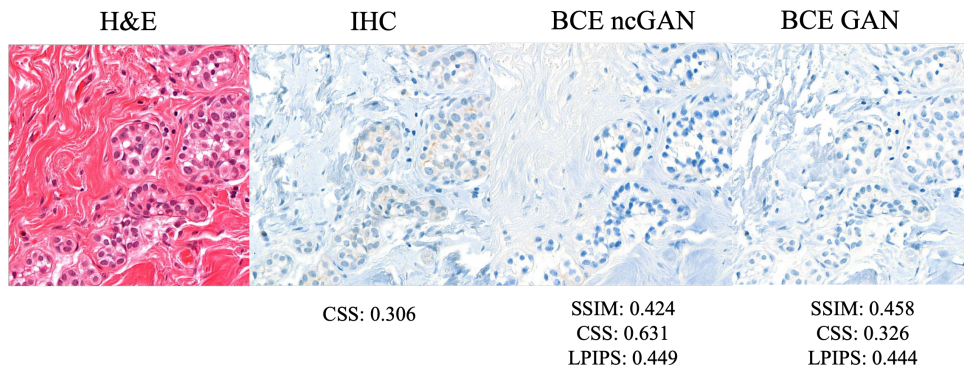


Figure 5.8: Example outputs of the BCE ncGAN and BCE GAN models for the 1+ HER2-score image example from the validation set.

image would appear if it had been subjected instead to IHC staining, making the preservation of H&E structures crucial for this task. Furthermore, we observe missing tissue regions (blank spaces) in the BCE GAN outputs that are absent in both the H&E and IHC images, suggesting a form of GAN hallucination. Such artifacts are particularly concerning, as they can undermine trust in AI models and hinder their clinical adoption. In addition to pixel-level misalignment, we verify that re-staining can also lead to loss of connective tissue regions (Figure 5.8), resulting in instances where connective tissue regions are missing in the IHC image but present in the H&E. Consequently, one can reason that in the BCE GAN setting, G may exploit this discrepancy by hallucinating random blank regions to fool D , reflecting the characteristics of the dataset.

5.3 CSS GAN

Despite producing highly realistic results, BCE GAN failed to adequately preserve the local H&E structures, with notable hallucinations of missing connective tissue regions. Furthermore, we confirmed that this model learned non-causal features of HER2 staining and demonstrated low

accuracy in identifying cases with a 3+ HER2 score. In this section, we introduce CSS loss to penalize structural differences between the input H&E and generated IHC images, thus regularizing the training process.

5.3.1 Training Methodology

Building on the BCE GAN framework, we extend $\mathcal{L}_G$ by adding a term that penalizes structural differences between the input H&E and the generated IHC images. Using the CSS metric definition, we compute a CSS map between the generated IHC and input H&E images, highlighting areas where their structures differ (Figure 5.9). For clarity, this can be visualized using a heatmap of the grayscale CSS map (Figure 5.10). In the heatmap, lighter regions represent areas where the structures of both images align, while darker regions indicate structural discrepancies, suggesting that the model should be updated. As a result, our CSS loss function can be defined as outlined

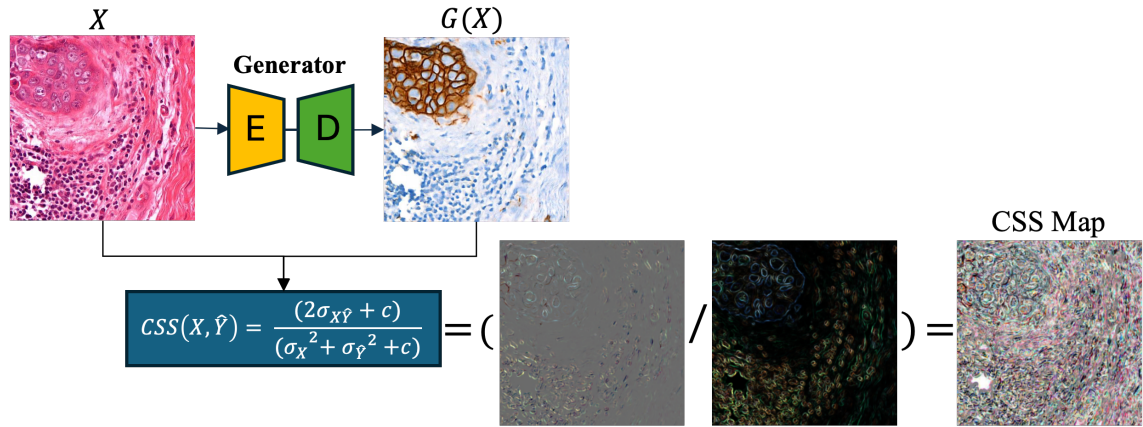


Figure 5.9: Illustration of the integration of CSS loss in the BCE GAN framework.

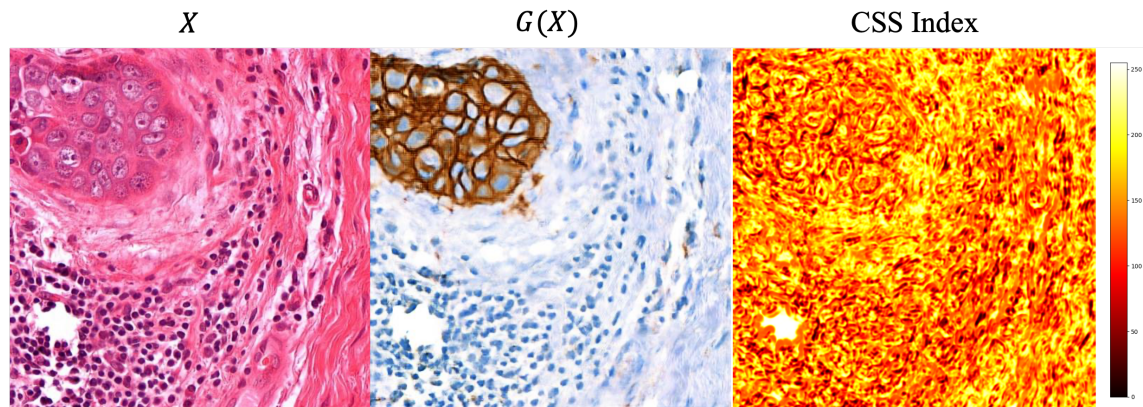


Figure 5.10: Heatmap of the grayscale CSS map highlights regions in the generated image where the structure does not align with that of the H&E input image. In this figure, lighter regions represent areas where the structure in the generated image aligns with the H&E image, while darker regions indicate areas where the structure differs, suggesting that the model should be updated accordingly.

in Equations 5.13–5.15. We begin by computing the CSS Map using Equation 5.13, where C is a stabilization constant and $\sigma_{\mathbf{X}\hat{\mathbf{Y}}}$ is the covariance between $\mathbf{X}$ and $\hat{\mathbf{Y}}$. Next, using Equation 5.14, we normalize to the range $[0, 1]$ and average it spatially and across channels to obtain a structure similarity score for each instance in the mini-batch. Finally, in Equation 5.15, we apply the negative logarithm and average the result across the batch. The logarithm is used at this stage because it is a monotonically increasing function and helps provide a smooth optimization landscape. Therefore, the loss function of G in the CSS GAN framework can be expressed as shown in Equation 5.16.

$$\text{CSS}(\mathbf{X}, \hat{\mathbf{Y}}) = \frac{2\sigma_{\mathbf{X}\hat{\mathbf{Y}}} + C}{\sigma_{\mathbf{X}}^2 + \sigma_{\hat{\mathbf{Y}}}^2 + C} \in [-1, 1]^{B \times C \times H \times W} \quad (5.13)$$

$$\overline{\text{CSS}}(\mathbf{X}, \hat{\mathbf{Y}}) = \frac{1}{CHW} \mathbf{1}_{CHW}^\top \cdot \left(\frac{\text{CSS}(\mathbf{X}, \hat{\mathbf{Y}}) + 1}{2} \right) \cdot \mathbf{1}_{CHW} \in [0, 1]^B \quad (5.14)$$

$$\mathcal{L}_{\text{CSS}}(\mathbf{X}, \hat{\mathbf{Y}}) = -\frac{1}{B} \mathbf{1}_B^\top \cdot \log(\overline{\text{CSS}}(\mathbf{X}, \hat{\mathbf{Y}})) \cdot \mathbf{1}_B \quad (5.15)$$

$$\mathcal{L}_G(\mathbf{X}) = -\frac{1}{BHW} \mathbf{1}_{BHW}^\top \cdot \left(\log(D(G(\mathbf{X}), \mathbf{X})) \right) \mathbf{1}_{BHW} + \mathcal{L}_{\text{CSS}}(\mathbf{X}, G(\mathbf{X})) \quad (5.16)$$

5.3.2 CSS for Morphology Preservation

We verify that the CSS GAN examples show improved tissue preservation compared to BCE GAN (Figure 5.11 and Figure 5.13). Specifically, the rate and size of missing tissue areas appear to be smaller in the CSS GAN examples than in those generated by BCE GAN. This would suggest an improvement in the CSS metric; however, its quantitative results indicate a slight decrease (Table 5.2). Although CSS loss did not improve the results on the CSS metric in this experiment, it led to a small improvement in LPIPS, our reference metric, and, more importantly, in JSD applied at the DAB channel, suggesting a positive regularization effect.

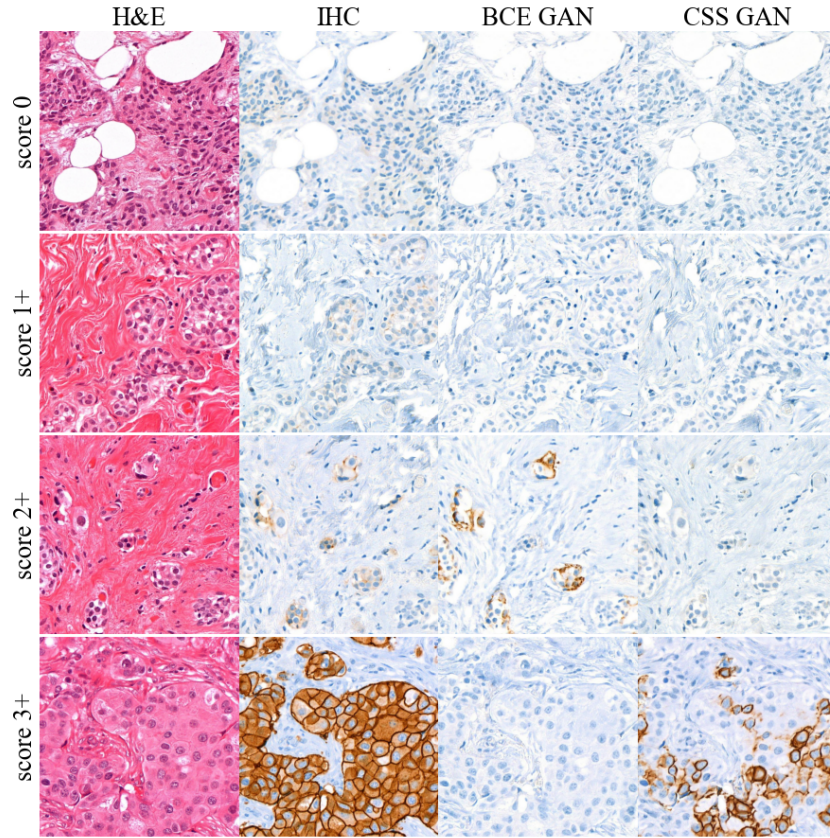


Figure 5.11: Qualitative assessment of the BCE GAN and CSS GAN models on the validation set of the HER2match dataset.

Table 5.2: Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN and CSS GAN experiments. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best metric values are highlighted in bold.

Model	Human Perception			Image Similarity			Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	CSS $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
Identity	0.664 ± 0.079	252	26.09 ± 1.67	0.341 ± 0.106	0.310 ± 0.098	1.00	9.85	—
BCE GAN	0.458 ± 0.067	34	1.17 ± 0.56	0.503 ± 0.120	0.464 ± 0.120	0.414 ± 0.104	19.8 ± 3.9	0.890 ± 0.193
CSS GAN	0.457 ± 0.065	34	1.32 ± 0.42	0.474 ± 0.120	0.459 ± 0.127	0.392 ± 0.101	19.8 ± 3.8	0.876 ± 0.209

In addition, we found that CSS GAN is more accurate in distinguishing cases with a 3+ HER2 score, showing a significant decrease in false positives (Figure 5.13). This result further demonstrates that by enforcing the preservation of H&E structures, this loss function acted as a regularizer, encouraging G to focus on important image features that can deceive D , rather than generating random instances of missing tissue areas. However, it is important to note that, in addition to not improving the CSS metric, the CSS loss also led to a decrease in SSIM, MS-SSIM, and KID. Given the observed trade-off in metric performance, the CSS loss will be further tested in subsequent experiments.

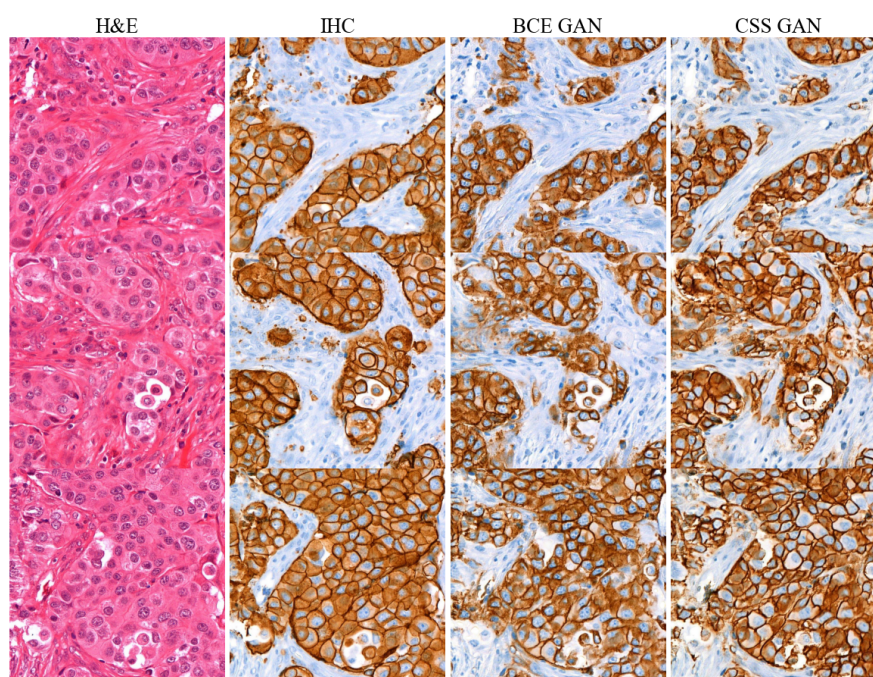


Figure 5.12: Qualitative comparison between the BCE GAN and CSS GAN performances on the selected image examples from the validation set with a 3+ HER2-score and intense positive DAB stain.

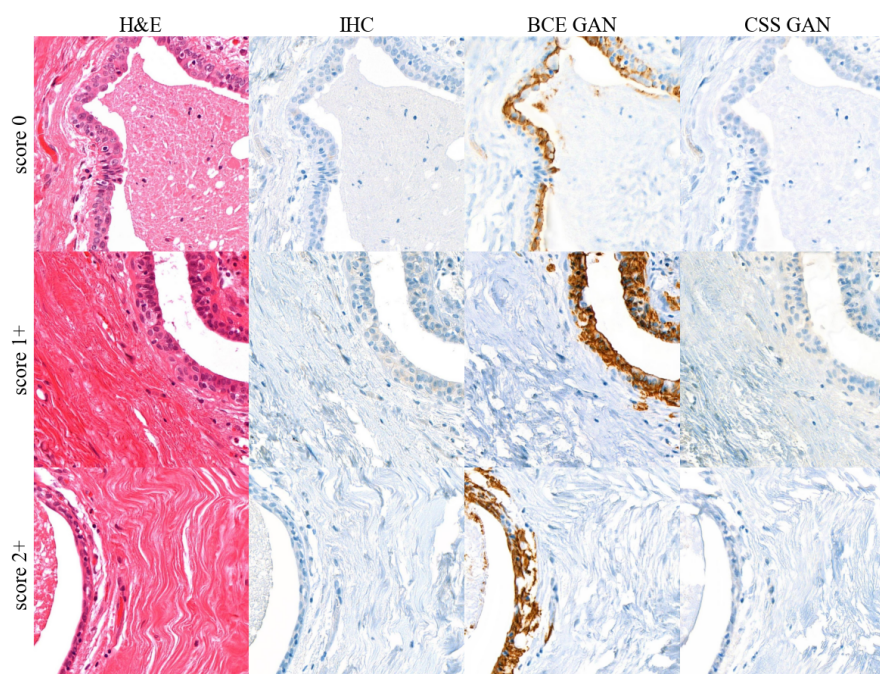


Figure 5.13: Qualitative comparison between the BCE GAN and CSS GAN performances on the selected image examples from the validation set without positive DAB staining.

5.4 CSSP2P GAN

To improve the similarity between generated and real IHC images, we extended the BCE GAN and CSS GAN frameworks by incorporating the Pyramid $\mathcal{L}_1$ loss [43]. In this section, we investigate the impact of this loss function and demonstrate the positive regularizing effect of CSS loss. Additionally, we assess the efficacy of SoftAdapt in addressing the challenges of complex fine-tuning in GANs.

5.4.1 Training Methodology

The intuition behind the pyramid strategy is based on the observation that two similar but distinct objects often appear more alike when viewed from a distance due to the loss of high-frequency details, which are crucial for distinguishing fine features. The pyramid strategy simulates multiple octave levels, where each octave applies four Gaussian filters, acting as low-pass filters, followed by downsampling, effectively mimicking the effect of viewing objects from farther away (Figure 5.14). In our application, as the octave level increases, information in both the real and generated images is progressively lost; however, this also increases their consistency, thereby addressing the problem of imperfect pixel-to-pixel match. The described phenomenon is illustrated in Figure 5.14, where the generated and real IHC images become increasingly similar as the octave level increases.

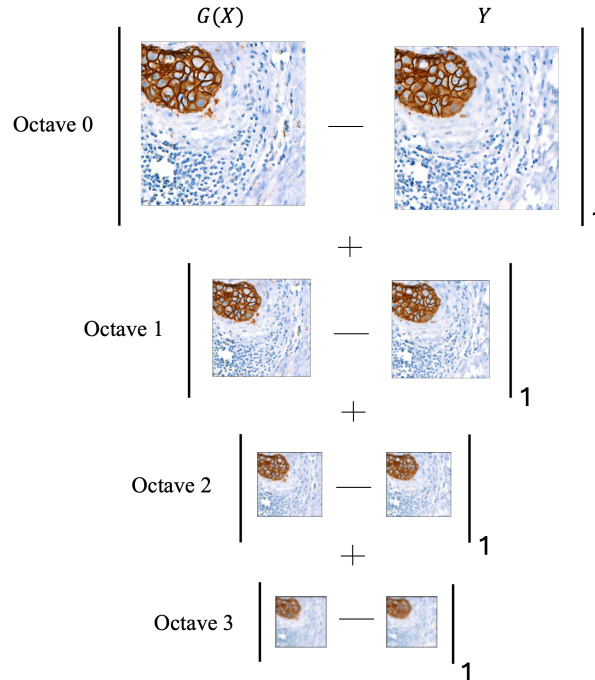


Figure 5.14: Illustration of the Pyramid $\mathcal{L}_1$ loss function. As the octave level increases, the resolution of both images decreases; however, their consistency improves, effectively addressing the problem of imperfect pixel-to-pixel correspondence.

As outlined in Equations 5.17-5.19, our implementation of Pyramid $\mathcal{L}_1$ loss follows that of Liu *et al.* [43]. It recursively applies four identical Gaussian smoothing filters with kernel size 3 and $\sigma = 1$ to the generated, $\hat{\mathbf{Y}}$, and ground truth, $\mathbf{Y}$, images, followed by downsampling via average pooling (kernel size 3, stride 2, padding 1) at each level l (Equation 5.18). This produces progressively downsampled tensors. At every scale, the loss computes the $\mathcal{L}_1$ distance between the two images. The losses on all four scales are weighted by coefficients λ_l (set to 1), summed, and the final loss is computed as the average across spatial, channel, and batch dimensions (Equation 5.19). Using the Pyramid $\mathcal{L}_1$ loss, the loss function of G can be described as in Equation 5.20.

$$\hat{\mathbf{Y}}^{(0)} = \hat{\mathbf{Y}} \quad (5.17)$$

$$\hat{\mathbf{Y}}^{(l)} = \text{AvgPool} \left(\underbrace{\mathcal{G}_\sigma * \mathcal{G}_\sigma * \mathcal{G}_\sigma * \mathcal{G}_\sigma}_{4 \text{ times}} * \hat{\mathbf{Y}}^{(l-1)} \right), \quad l = 1, 2, 3, 4 \quad (5.18)$$

$$\mathcal{L}_{\text{pyramid}(\mathcal{L}_1)}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{BCHW} \mathbf{1}_{BCHW}^\top \cdot \left(\sum_{l=1}^4 \lambda_l \mathcal{L}_1(\mathbf{Y}^{(l)}, \hat{\mathbf{Y}}^{(l)}) \right) \cdot \mathbf{1}_{BCHW} \quad (5.19)$$

$$\mathcal{L}_G(\mathbf{X}) = -\frac{1}{BHW} \mathbf{1}_{BHW}^\top \cdot \left(\log(D(G(\mathbf{X}), \mathbf{X})) \right) \cdot \mathbf{1}_{BHW} + \mathcal{L}_{\text{pyramid}(\mathcal{L}_1)}(\mathbf{Y}, G(\mathbf{X})) \quad (5.20)$$

5.4.2 Results on the Pyramidal Strategy

We observe that P2P GAN matches BCE GAN in image similarity metrics while achieving a better LPIPS score than both BCE GAN and CSS GAN (Table 5.3). This result aligns with the intended goal of approximating the generated images to the ground truth. Although P2P GAN demonstrates improvements over the other GAN frameworks, it performs worse on FID and KID metrics, suggesting a decrease in the preservation of semantic information.

Table 5.3: Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN, CSS GAN, and P2P GAN experiments. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best and second-best metric values are highlighted in bold and underlined.

Model	Human Perception			Image Similarity			Noise	DAB
	LPIPS ↓	FID ↓	KID ↓	SSIM ↑	MS-SSIM ↑	CSS ↑	PSNR ↑	JSD ↓
Identity	0.664 ± 0.079	252	26.09 ± 1.67	0.341 ± 0.106	0.310 ± 0.098	1.00	9.85	—
BCE GAN	0.458 ± 0.067	34	1.17 ± 0.56	0.503 ± 0.120	<u>0.464 ± 0.120</u>	0.414 ± 0.104	<u>19.8 ± 3.9</u>	0.890 ± 0.193
CSS GAN	<u>0.457 ± 0.065</u>	34	<u>1.32 ± 0.42</u>	0.474 ± 0.120	0.459 ± 0.127	0.392 ± 0.101	19.8 ± 3.8	0.876 ± 0.209
P2P GAN	0.454 ± 0.072	50	3.10 ± 0.84	<u>0.491 ± 0.123</u>	0.466 ± 0.138	<u>0.401 ± 0.102</u>	20.1 ± 4.2	<u>0.885 ± 0.206</u>

5.4.3 Efficacy of CSS Loss

In Section 5.3, we verified that CSS loss reduced the appearance and size of missing tissue areas in the generated IHC images. Furthermore, we observed that the loss function regularized the

training of G , leading to an improved performance in LPIPS and JSD computed in the DAB channel. The improvement in JSD suggested that the model learned to capture the pathological information better, thus improving in distinguishing the cases with a 3+ HER2-score. This finding was further supported by the analysis of the extended group of images selected from the validation set. However, the use of CSS loss unexpectedly resulted in a decrease in performance across all other metrics, particularly the CSS metric. Consequently, this trade-off raised doubts about the usefulness of the CSS loss. Therefore, to further assess its impact, we integrated CSS loss into the P2P GAN training (Equation 5.21), which resulted in the creation of the CSSP2P GAN.

$$\mathcal{L}_G(\mathbf{X}) = -\frac{1}{BHW} \mathbf{1}^\top \left(\log(D(G(\mathbf{X}), \mathbf{X})) \right) \mathbf{1} + \mathcal{L}_{\text{pyramid}(\mathcal{Z}_1)}(\mathbf{Y}, G(\mathbf{X})) + \mathcal{L}_{\text{CSS}}(\mathbf{X}, G(\mathbf{X})) \quad (5.21)$$

We verified that using CSS loss, we were able to improve the performance of the model on all metrics except PSNR, thus confirming the effectiveness of CSS loss (Table 5.4). In addition, a subsequent analysis of the performance in the extended selection of images from the validation set (Figures 5.15–5.16) further demonstrates the regularization effect of CSS loss. Comparing the outputs of the two models, we verify that CSSP2P GAN outperformed P2P GAN in distinguishing the cases with a 3+ HER2-score. This result aligns with the improved JSD computed in the DAB channel for CSSP2P GAN. Although CSS loss was not specifically designed to capture the pathological information of the image, we can once again observe its regularizing effect in this regard.

Table 5.4: Performance metrics on the validation set of the HER2match dataset comparing the CSSP2P GAN and P2P GAN experiments. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best metric values are highlighted in bold.

Model	Human Perception			Image Similarity		CSS $\uparrow$	Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$		PSNR $\uparrow$	JSD $\downarrow$
Identity	0.664 ± 0.079	252	26.09 ± 1.67	0.341 ± 0.106	0.310 ± 0.098	1.00	9.85	—
P2P GAN	0.454 ± 0.072	50	3.10 ± 0.84	0.491 ± 0.123	0.466 ± 0.138	0.401 ± 0.102	20.1 ± 4.2	0.885 ± 0.206
CSSP2P GAN	0.454 ± 0.069	44	2.68 ± 0.59	0.492 ± 0.117	0.475 ± 0.123	0.413 ± 0.097	20.0 ± 3.9	0.871 ± 0.211

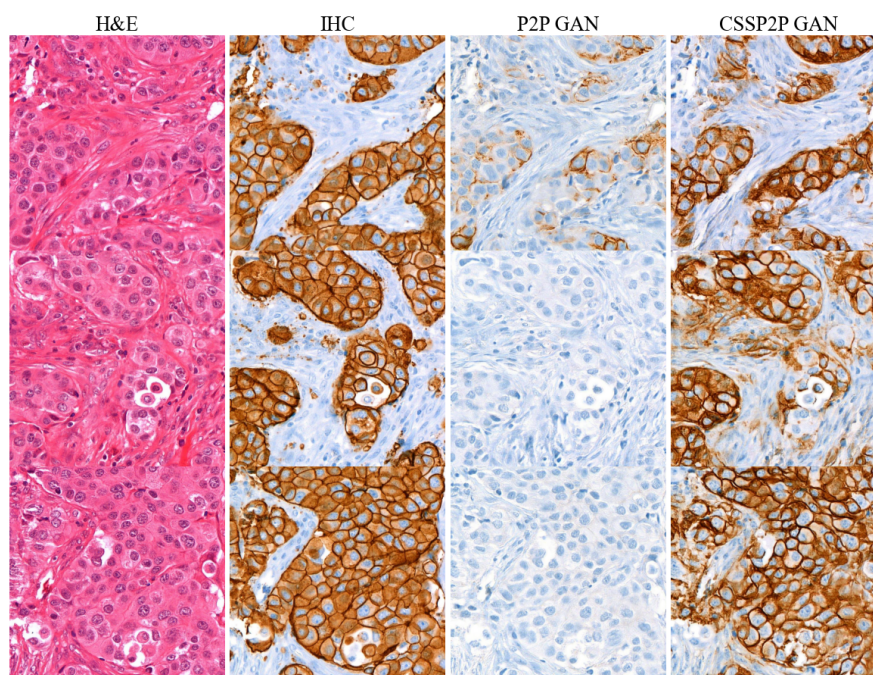


Figure 5.15: Qualitative comparison between the CSSP2P GAN and P2P GAN performance on the selected image examples from the validation set with a 3+ HER2-score and intense positive DAN stain.

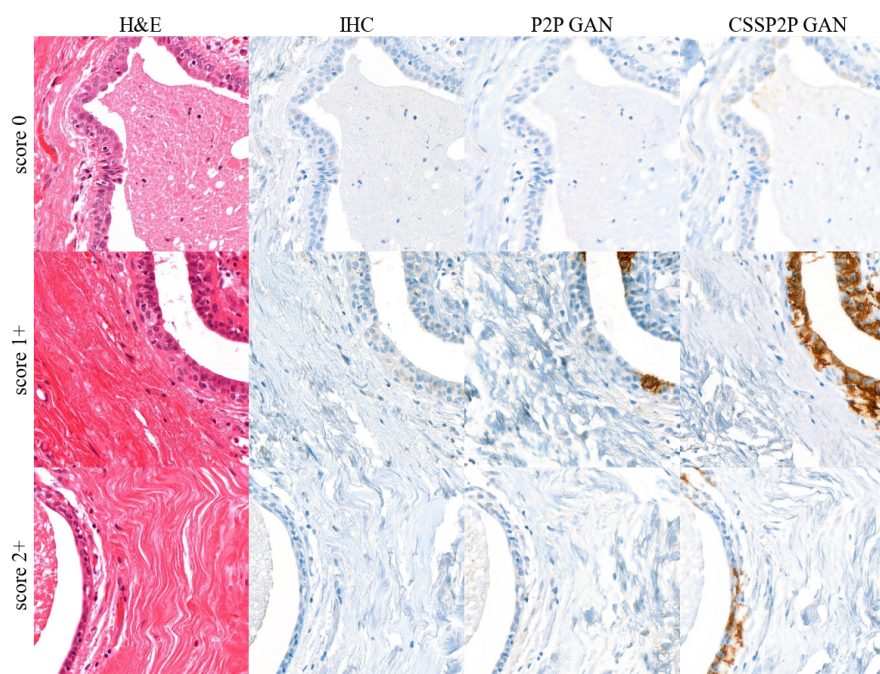


Figure 5.16: Qualitative comparison between the CSSP2P GAN and P2P GAN performance on the selected image examples from the validation set without positive DAB staining.

In conclusion, this experiment demonstrated that CSS loss not only improved performance

across the metrics but also enhanced the ability of the model to capture pathological information. This suggests its positive impact on refining the baseline model towards a more effective virtual staining model, resolving previous doubts about its usefulness. Consequently, CSS P2P GAN was considered the best-performing model (Table 5.5 and Figure 5.17) so far, and CSS loss was incorporated into all subsequent experiments during the training process.

Table 5.5: Performance metrics on the validation set of the HER2match dataset comparing the BCE GAN, CSS GAN, and CSSP2P GAN experiments. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best and second-best metric values are highlighted in bold and underlined.

Model	Human Perception			Image Similarity			Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	CSS $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
Identity	0.664 ± 0.079	252	26.09 ± 1.67	0.341 ± 0.106	0.310 ± 0.098	1.00	9.85	—
BCE GAN	0.458 ± 0.067	34	1.17 ± 0.56	0.503 ± 0.120	<u>0.464 ± 0.120</u>	0.414 ± 0.104	<u>19.8 ± 3.9</u>	0.890 ± 0.193
CSS GAN	<u>0.457 ± 0.065</u>	34	<u>1.32 ± 0.42</u>	0.474 ± 0.120	0.459 ± 0.127	0.392 ± 0.101	19.8 ± 3.8	<u>0.876 ± 0.209</u>
CSSP2P GAN	<u>0.454 ± 0.069</u>	<u>44</u>	2.68 ± 0.59	<u>0.492 ± 0.117</u>	0.475 ± 0.123	<u>0.413 ± 0.097</u>	20.0 ± 3.9	0.871 ± 0.211

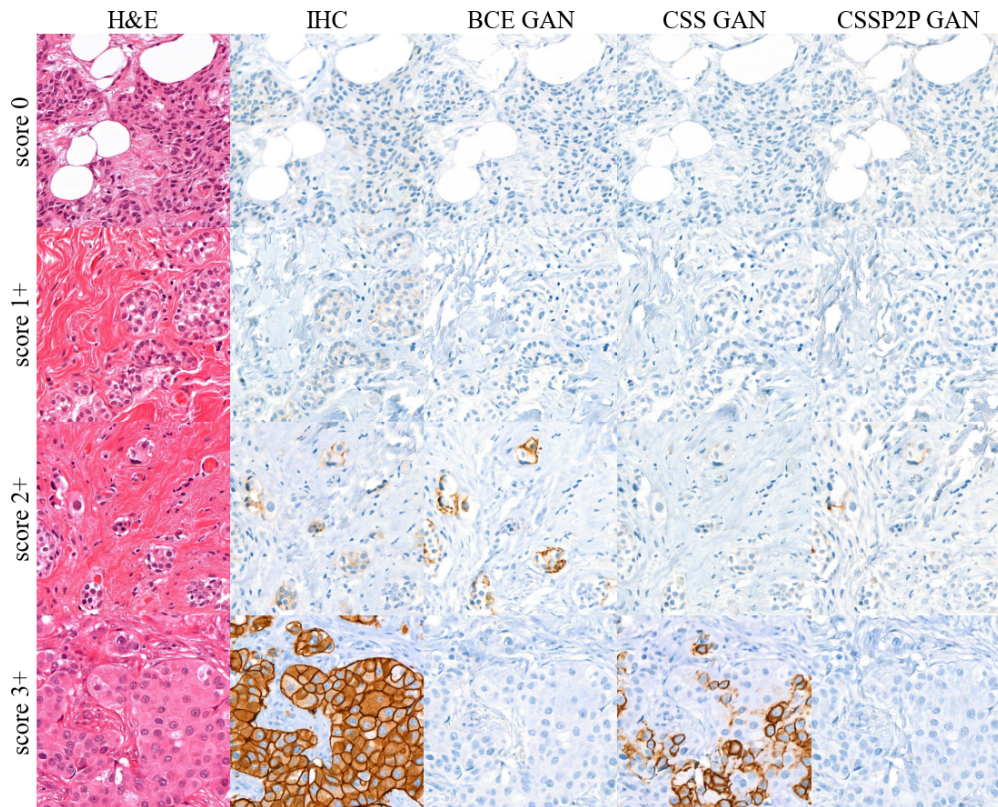


Figure 5.17: Qualitative assessment of the BCE GAN, CSS GAN, and CSSP2P GAN models on the validation set of the HER2match dataset.

5.4.4 Experiments in Adaptive Loss Weighting

One of the difficulties in using GANs lies in the complex fine-tuning of hyperparameters, particularly the weighting of each loss component, to achieve optimal results. Based on the work of Heydari *et al.* [136], we tried to address this problem by using soft adaptation, a technique that dynamically adjusts weights for multi-part loss functions based on live performance statistics of the loss components. Equations 5.22–5.28 describe the implemented SoftAdapt procedure. Initially, all loss components are equally weighted. At each training step, the loss values for each component are stored until a specified history window size is reached. Once fully accumulated, every new loss value replaces the oldest, and SoftAdapt begins to adaptively weight the components of the loss function. As demonstrated in Equation 5.25, where H is the history size and $f_i(t_k)$ represents the value of the i -th loss component at time t_k , it computes the average loss value of each component, capturing its typical magnitude. These averaged losses are then subtracted from the total sum of the averaged components and normalized, as shown in Equation 5.26, where ε is a small constant to prevent division by zero. This computes normalized difference-based loss weights that emphasize components with smaller values, effectively downweighting dominant losses and promoting balanced optimization. Simultaneously, the algorithm calculates and normalizes the rate of change of each loss component to detect which are changing most significantly, as depicted in Equations 5.22–5.23. Afterward, using Equation 5.24, a softmax function is applied to these normalized rates to produce preliminary weights that emphasize components with a higher rate of change. Here, the temperature hyperparameter β controls the softness of the softmax. The higher β , the more weight is assigned to the component with the highest positive rate of change (worst performing component). Finally, the preliminary weights and the normalized difference-based loss weights are combined and re-normalized to yield the final adaptive weights, which are used to compute a weighted sum of the loss components, Equations 5.27–5.28. Using this methodology, instead of having one weight hyperparameter per loss component, we reduce the number of hyperparameters to two: the history size and the temperature parameter β . Additionally, it enables dynamic weighting of multiple loss components, allowing model updates to adapt based on the worst-performing components at each training step.

$$\text{Rate of Change per Loss Component: } s_i = \frac{1}{H-1} \sum_{k=1}^{H-1} \frac{f_i(t_k) - f_i(t_{k-1})}{f_i(t_0) + \varepsilon} \quad (5.22)$$

$$\text{Normalization of Rate of Change: } \tilde{s}_i = \frac{s_i}{\sum_{j=1}^m |s_j| + \varepsilon} \quad (5.23)$$

$$\text{Softmax to Compute Preliminary Weights: } \alpha_i^{(1)} = \frac{\exp(\beta(\tilde{s}_i - \max_j \tilde{s}_j))}{\sum_{k=1}^m \exp(\beta(\tilde{s}_k - \max_j \tilde{s}_j))} \quad (5.24)$$

$$\text{Average Loss per Component: } \bar{f}_i = \frac{1}{H} \sum_{k=0}^{H-1} f_i(t_k) \quad (5.25)$$

$$\text{Normalized Difference-Based Loss Weights: } f'_i = \frac{\sum_{j=1}^m \bar{f}_j + \varepsilon - \bar{f}_i}{\sum_{j=1}^m \bar{f}_j + \varepsilon} \quad (5.26)$$

$$\text{Final Combined Weights: } \alpha_i = \frac{\alpha_i^{(1)} \cdot f'_i}{\sum_{k=1}^m \alpha_k^{(1)} \cdot f'_k} \quad (5.27)$$

$$\text{Weighted Loss: } \mathcal{L} = \sum_{i=1}^m \alpha_i f_i \quad (5.28)$$

In this experiment, we trained CSSP2P GAN using SoftAdapt in two configurations: soft adaptation applied to all loss components of G (SoftAdapt), and soft adaptation applied only to CSS and Pyramid $\mathcal{L}_1$ losses, excluding the adversarial loss (NoAdvSoftAdapt). We compared these two approaches with our current CSSP2P GAN model, where each component of the loss function had been equally weighted.

We observe that the model trained with equal weights across the loss components of G achieved the best performance compared to both SoftAdapt configurations (Table 5.6). We verify that the NoAdvSoftAdapt achieved an intermediate performance between SoftAdapt and NoSoftAdapt, with best-performing metrics, especially FID and KID, mostly attributed to the adversarial loss component. In this setting, the adversarial loss weight was fixed to 1, while the weights of the CSS and Pyramid $\mathcal{L}_1$ loss components were between 0 and 1, summing up to 1. This suggests that using soft adaptation slowed down convergence by scaling loss components with weights smaller than 1.

Table 5.6: Performance metrics on the validation set of the HER2match dataset comparing different SoftAdapt implementation methods. SoftAdapt refers to the experiment where all loss values were weighted using soft adaptation; NoSoftAdapt indicates no use of soft adaptation; and NoAdvSoftAdapt denotes weighting all loss terms except the adversarial loss. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best and second-best metric values are highlighted in bold and underlined.

Model	Human Perception			Image Similarity			Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	CSS $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
SoftAdapt	0.463 $\pm$ 0.069	44	2.27 $\pm$ 0.55	0.468 $\pm$ 0.118	0.443 $\pm$ 0.134	0.373 $\pm$ 0.106	19.7 $\pm$ 4.0	0.903 $\pm$ 0.181
NoSoftAdapt	0.454 $\pm$ 0.069	44	2.68 $\pm$ 0.59	0.492 $\pm$ 0.117	0.475 $\pm$ 0.123	0.413 $\pm$ 0.097	20.0 $\pm$ 3.9	0.871 $\pm$ 0.211
NoAdvSoftAdapt	<u>0.456 $\pm$ 0.067</u>	41	2.08 $\pm$ 0.54	0.476 $\pm$ 0.120	<u>0.464 $\pm$ 0.124</u>	<u>0.396 $\pm$ 0.103</u>	<u>19.9 $\pm$ 4.0</u>	<u>0.890 $\pm$ 0.196</u>

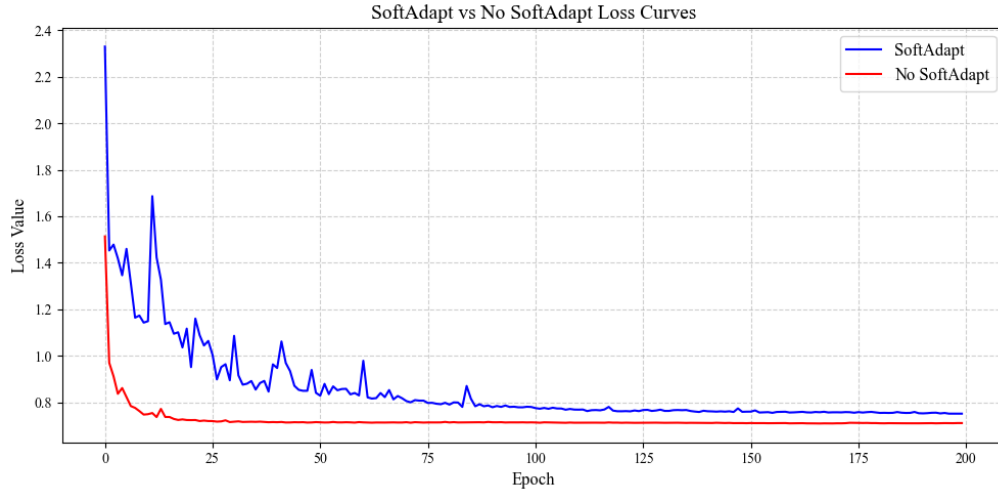


Figure 5.18: Evolution of the adversarial loss component during training for SoftAdapt and NoSoftAdapt.

Figure 5.18 shows the evolution of the adversarial loss component for the two extreme cases: NoSoftAdapt and SoftAdapt. We observe that, although soft adaptation did not destabilize GAN training, it did slow the convergence to equilibrium, corroborating the previous observation that SoftAdapt slows convergence in our use case. This behavior was not unique to the adversarial component; the CSS and Pyramid $\mathcal{L}_1$ loss components exhibited a similar trend. For this reason, and considering the additional computational overhead it introduces, we dropped the use of SoftAdapt. All subsequent results were obtained by assigning equal weights to all loss components.

5.5 OkLAB GAN

The ability to identify cases with a 3+ HER2-score and precisely predict chromogen-stained regions is crucial in virtual staining, as chromogen staining identifies the presence of the target antigen and encodes important pathological information within IHC images. In our case, the chromogen is DAB, which appears brown; the darker the shade, the stronger the expression of HER2 in the tissue. Although CSS loss brought us closer to achieving this objective, its design does not focus on achieving an accurate pathological representation. As a result, this section describes a two-fold methodology designed to address this challenge specifically. First, we utilize a masking strategy to increase the relative importance of pixels within the positive DAB signal region in the loss function. Second, we penalize perceptual color differences between the generated and ground-truth images using the OkLAB loss.

5.5.1 Training Methodology

The developed masking strategy enables differential penalization of regions within the generated image, guiding the model to focus on specific areas of the output. This strategy applies element-wise multiplication ($\odot$) between a 2D mask and the generated and real images, effectively scaling

pixel importance before computing the loss function (Equation 5.29). It is agnostic to the choice of loss function, meaning it can be applied to any type of image loss.

$$\hat{\mathbf{Y}}^{(0)}, \mathbf{Y}^{(0)} = \mathcal{M}(\hat{\mathbf{Y}}, \mathbf{Y}, \mathbf{mask}) = (\hat{\mathbf{Y}} \odot \mathbf{mask}, \mathbf{Y} \odot \mathbf{mask}) \quad (5.29)$$

Figure 5.19 illustrates one of the masks used, where positive DAB regions (brown areas) are assigned a greater weight compared to DAB-negative regions (blue areas), while regions of missing tissue are assigned a value of zero. Using this strategy, we increase the importance of DAB-positive regions relative to DAB-negative ones, while excluding background pixels from contributing to the update of the model.

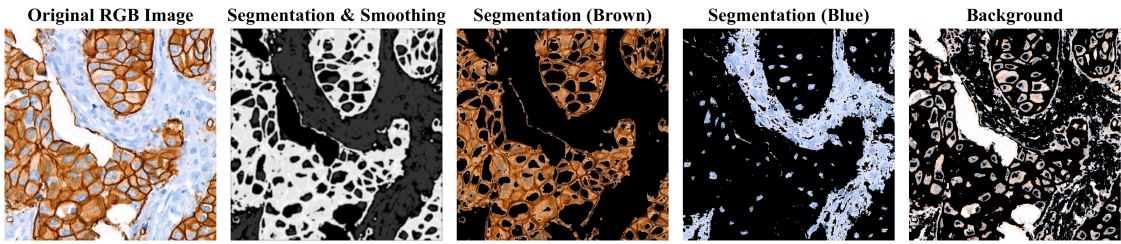


Figure 5.19: Example of an estimated mask for an image with a 3+ HER2-score.

To obtain the masks, we start by converting the IHC image from the RGB color space to the HED concentration (Equations 5.30-5.35). Using Equation 5.32, we first convert the IHC image from RGB color space to OD, where the use of $\max(\mathbf{Y}_{RGB}, \epsilon)$ and the division by $\log(\epsilon)$ prevent infinite negative values, thus enhancing the stability of the procedure. Subsequently, we apply the transformation matrix to convert the IHC image from the OD space to the HED concentration space (Equations 5.33-5.35).

$$\text{Matrix Definition: } \mathbf{M}_{\text{HED} \rightarrow \text{RGB}} = \begin{bmatrix} 0.65 & 0.70 & 0.29 \\ 0.07 & 0.99 & 0.11 \\ 0.27 & 0.57 & 0.78 \end{bmatrix} \quad (5.30)$$

$$\mathbf{M}_{\text{RGB} \rightarrow \text{HED}} = \mathbf{M}_{\text{HED} \rightarrow \text{RGB}}^{-1} \quad (5.31)$$

$$\text{Convert to OD: } \mathbf{Y}_{\log} = \frac{\log(\max(\mathbf{Y}_{\text{RGB}}, \epsilon))}{\log(\epsilon)}, \quad \epsilon = 1 \times 10^{-6} \quad (5.32)$$

$$\text{Computation of HED: } \mathbf{Y}_{\log}^{\text{perm}} = \text{permute}(\mathbf{Y}_{\log}, (0, 2, 3, 1)) \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.33)$$

$$\mathbf{Y}_{\text{HED}}^{\text{perm}} = \mathbf{Y}_{\log}^{\text{perm}} \cdot \mathbf{M}_{\text{RGB} \rightarrow \text{HED}} \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.34)$$

$$\mathbf{Y}_{\text{HED}} = \text{permute}(\mathbf{Y}_{\text{HED}}^{\text{perm}}, (0, 3, 1, 2)) \in \mathbb{R}^{B \times 3 \times H \times W} \quad (5.35)$$

This methodology, based on the work of Ruifrok *et al.* [64] and its implementation by the Scikit-Image Team [137], which we adapted to the PyTorch framework, is commonly used in current SOTA approaches, as mentioned in Chapter 4. However, we omit the clamping of negative values. Although negative concentration values are physically meaningless, they are possible in our case

because the transformation matrix derived by Ruifrok *et al.* [64] is specific to their work. As highlighted by Ji *et al.* [88], variability in chemical stain conditions, staining time, and specimen transmittance causes variations in absorption characteristics. Consequently, the transformation matrix should be recalibrated for each case, where negative values would likely just arise from noise or computational artifacts and should therefore be disregarded. Current SOTA methods often overlook this issue, potentially losing important information in the transformed image. In our approach, we re-implemented the stain deconvolution method by Ruifrok *et al.* [64] using their stain matrix (Equations 5.30-5.35) and verified that clamping negative values in the DAB channel causes loss of blue regions in the IHC image (Figure 5.20).

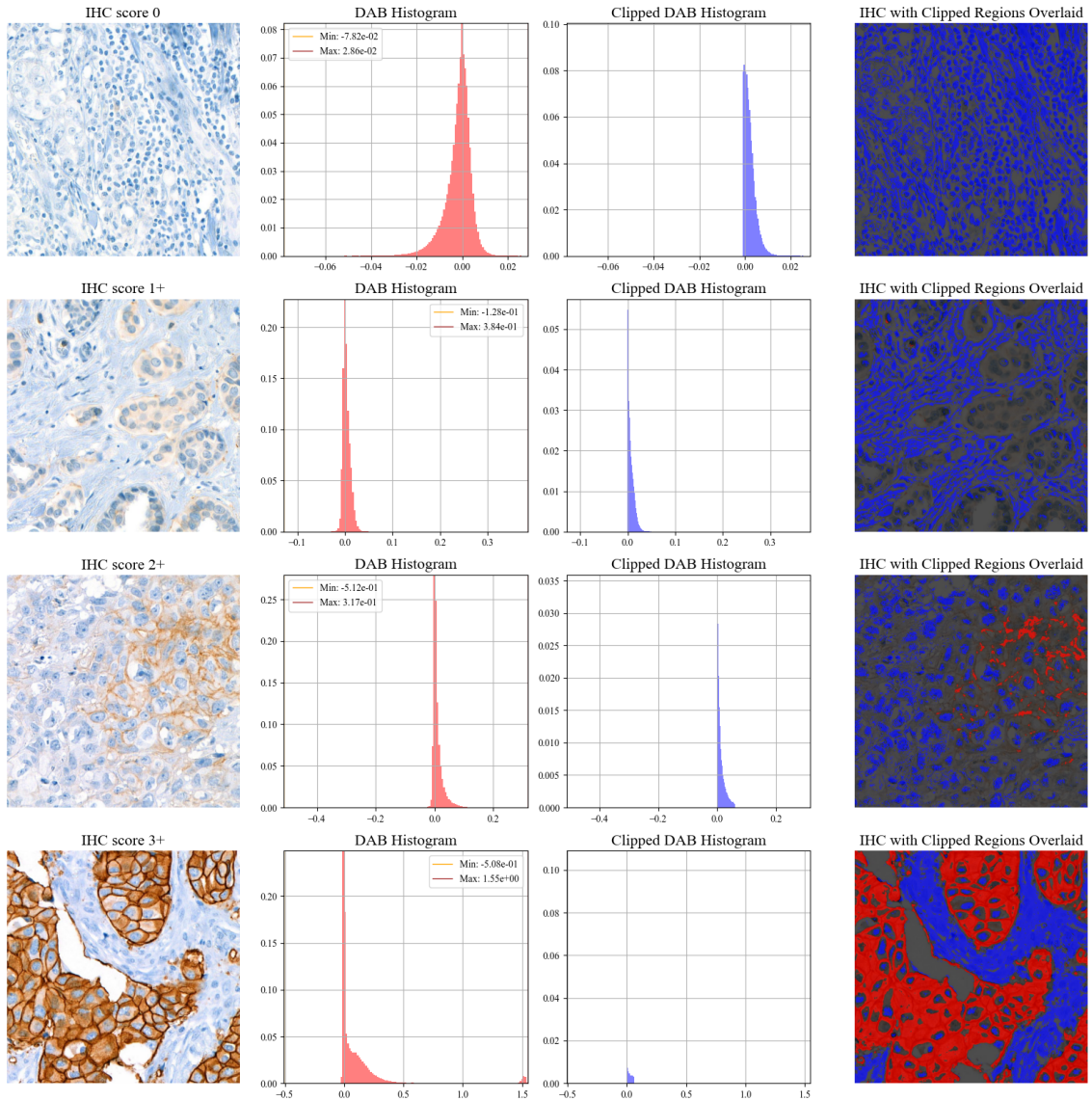


Figure 5.20: Illustration of the DAB channel histograms for different HER2-scores, including their clamped counterparts at the negative extreme -1×10^{-3} and positive extreme 6×10^{-2} , and the resulting loss of content from the original image.

In Equations 5.36–5.42, we outline the procedure to calculate the masks. After estimating the DAB channel (Equation 5.36), we apply thresholding using Equations 5.37–5.38 to obtain the brown and blue regions of interest, with $\tau_{brown}^{(i)}$ and $\tau_{blue}^{(j)}$ representing the respective i -th and j -th values from the list of thresholds, which have sizes N_{brown} and N_{blue} , respectively. In our experiments, we set $\tau_{brown}^{(1)} = 6 \times 10^{-2}$, $\tau_{brown}^{(2)} = 3 \times 10^{-1}$, and $\tau_{blue} = -1 \times 10^{-3}$. We chose two thresholds for the brown regions to prioritize the separation regions between cells, ensuring that these areas were weighted the most. These threshold values were determined by evaluating DAB histograms of different HER2-score examples from the training set (Figure 5.20). Subsequently, we apply smoothing using a Gaussian filter with a kernel of 9 and a σ of 5 to expand the edges of the computed mask and smooth the transition between regions (Equation 5.41). Finally, we normalize the smoothed mask by the maximum pixel value within the mini-batch, resulting in values between 0 and 1 across the batch (Equation 5.42).

$$\text{Extract the DAB channel: } \mathbf{DAB} = \text{ToHED}(\mathbf{Y})[:, 2, :, :] \in \mathbb{R}^{B \times 1 \times H \times W} \quad (5.36)$$

$$\text{Define Indicator Functions: } \mathbf{1}_{\mathbf{DAB} > \tau_{brown}^{(i)}}^{(i)}(h, w) = \begin{cases} 1, & \text{if } \mathbf{DAB}(h, w) > \tau_{brown}^{(i)} \\ 0, & \text{otherwise} \end{cases} \quad (5.37)$$

$$\mathbf{1}_{\mathbf{DAB} < \tau_{blue}^{(j)}}^{(j)}(h, w) = \begin{cases} 1, & \text{if } \mathbf{DAB}(h, w) < \tau_{blue}^{(j)} \\ 0, & \text{otherwise} \end{cases} \quad (5.38)$$

$$\text{Construct Mask: } \mathbf{mask}(h, w) = \sum_{i=1}^{N_{brown}} w_{brown}^{(i)} \cdot \mathbf{1}_{\mathbf{DAB}^{(brown)} > \tau_{brown}^{(i)}}^{(i)}(h, w) \quad (5.39)$$

$$\mathbf{mask}(h, w) = \sum_{j=1}^{N_{blue}} w_{blue}^{(j)} \cdot \mathbf{1}_{\mathbf{DAB} < \tau_{blue}^{(j)}}^{(j)}(h, w) \quad (5.40)$$

$$\text{Apply Smoothing: } \tilde{\mathbf{mask}} = \mathcal{G}_{\sigma} * \mathbf{mask}, \quad \text{kernel} = 9 \quad \text{and} \quad \sigma = 5 \quad (5.41)$$

$$\text{Normalize by the Maximum value: } \hat{\mathbf{mask}} = \frac{\tilde{\mathbf{mask}}}{\max_{b, h, w} \tilde{\mathbf{mask}}(b, h, w)} \quad (5.42)$$

Using the masking strategy, we developed a method to penalize image regions differently, assigning greater importance to positive DAB regions. Building on the premise that the brown color encodes critical pathological information, we also developed a new loss function, OkLAB loss, to directly address perceptual color differences between the generated and ground truth images. By combining OkLAB loss with masking, we guide G to focus on accurately reproducing the positive signal of DAB, the brown-stained regions.

To this end, we begin by introducing the CIELAB color space, a color-opponent model with three dimensions: $\mathbf{L}$, representing the perceived lightness of a color; $\mathbf{a}$, representing the green–red axis; and $\mathbf{b}$, representing the blue–yellow axis, based on the non-linearly compressed CIEXYZ color space coordinates [138]. This color space was designed to be perceptually uniform, meaning that a given numerical change corresponds to a similar perceived change in color, thereby

better approximating human visual experience compared to RGB. However, it was later found that equal distances in CIELAB do not always correspond to equal perceived differences, particularly in the blue region, due to non-uniformities in the color manifold [139]. To address this issue, distance functions, such as ΔE_{2000} [140], were developed to account for these non-uniformities and provide a more faithful measure of perceived color difference. These functions are complex and computationally slow, making them unsuitable for use as loss functions.

Building upon the CIELAB formulation, Ottosson [141] designed the OkLAB color space. This addresses the shortcomings of perceptual uniformity of CIELAB, thus providing a modern refinement for perceptually calibrated color error measurements [139], [142]. Furthermore, OkLAB components are orthogonal, which means that changing one component does not inadvertently affect others, making color adjustments more predictable and tractable [139], [141]. As a result, unlike the CIELAB color space, applying the $\mathcal{L}_2$ distance in OkLAB space is meaningful, as it closely corresponds to perceived color differences [139], [142]. In conclusion, due to its perceptual uniformity and orthogonality, which together enable meaningful use of the $\mathcal{L}_2$ distance to measure perceived color differences, the OkLAB color space emerges as a highly promising manifold to penalize color differences between generated and ground-truth images in a manner that is both computationally efficient and aligned with human perception.

To implement the OkLAB loss, we first convert the RGB image to the OkLAB color space. This requires an intermediate conversion from RGB to XYZ color space coordinates. Based on the implementation of the Scikit-Image Team [143], we adapted this procedure to the PyTorch framework, following the steps demonstrated in Equations 5.43–5.48.

$$\mathbf{M}_{\text{RGB} \rightarrow \text{XYZ}} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \quad (5.43)$$

$$\mathbf{m}(h, w) = \begin{cases} 1, & \text{if } \mathbf{m}(h, w) > 0.04045 \\ 0, & \text{otherwise} \end{cases} \quad (5.44)$$

$$\mathbf{Y}'_{\text{RGB}}(h, w) = \begin{cases} \frac{(\mathbf{Y}_{\text{RGB}} + 0.055)^{2.4}}{1.055}, & \text{if } \mathbf{m}(h, w) = 1 \\ \frac{\mathbf{Y}_{\text{RGB}}}{12.92}, & \text{if } \mathbf{m}(h, w) = 0 \end{cases} \quad (5.45)$$

$$\mathbf{Y}'_{\text{RGB}}^{\text{perm}} = \text{permute}(\mathbf{Y}'_{\text{RGB}}, (0, 2, 3, 1)) \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.46)$$

$$\mathbf{Y}_{\text{XYZ}}^{\text{perm}} = \mathbf{Y}'_{\text{RGB}}^{\text{perm}} \cdot \mathbf{M}_{\text{RGB} \rightarrow \text{XYZ}} \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.47)$$

$$\mathbf{Y}_{\text{XYZ}} = \text{permute}(\mathbf{Y}_{\text{XYZ}}^{\text{perm}}, (0, 3, 1, 2)) \in \mathbb{R}^{B \times 3 \times H \times W} \quad (5.48)$$

After converting the IHC image to XYZ color space coordinates, we follow the procedure described by Ottosson [141] to convert it to OkLAB color space. This procedure was implemented in the PyTorch framework, following the steps outlined in Equations 5.49–5.57. First, we convert the XYZ coordinates to an approximate cone response (Equations 5.51–5.53), and then apply a non-linearity (Equation 5.54). This step is essential to transform luminance, a photometric mea-

sure of luminous intensity, into lightness, which reflects the human perception of how light or dark a color appears [139]. Finally, we convert to the OkLAB color space (Equations 5.55-5.57).

$$\mathbf{M}_1 = \begin{bmatrix} 0.8189330101 & 0.3618667424 & -0.1288597137 \\ 0.0329845436 & 0.9293118715 & 0.0361456387 \\ 0.0482003018 & 0.2643662691 & 0.6338517070 \end{bmatrix} \quad (5.49)$$

$$\mathbf{M}_2 = \begin{bmatrix} 0.2104542553 & 0.7936177850 & -0.0040720468 \\ 1.9779984951 & -2.4285922050 & 0.4505937099 \\ 0.0259040371 & 0.7827717662 & -0.8086757660 \end{bmatrix} \quad (5.50)$$

$$\mathbf{Y}_{XYZ}^{\text{perm}} = \text{permute}(\mathbf{Y}_{XYZ}, (0, 2, 3, 1)) \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.51)$$

$$\mathbf{Y}_{LMS}^{\text{perm}} = \mathbf{Y}_{XYZ}^{\text{perm}} \cdot (\mathbf{M}_1^{-1})^\top \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.52)$$

$$\mathbf{Y}_{LMS} = \text{permute}(\mathbf{Y}_{LMS}^{\text{perm}}, (0, 3, 1, 2)) \in \mathbb{R}^{B \times 3 \times H \times W} \quad (5.53)$$

$$\mathbf{Y}'_{LMS} = \text{sign}(\mathbf{Y}_{LMS}) \odot |\mathbf{Y}_{LMS}|^{\frac{1}{3}} \quad (5.54)$$

$$\mathbf{Y}_{LMS}^{\text{perm}} = \text{permute}(\mathbf{Y}'_{LMS}, (0, 2, 3, 1)) \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.55)$$

$$\mathbf{Y}_{OkLAB}^{\text{perm}} = \mathbf{Y}_{LMS}^{\text{perm}} \cdot (\mathbf{M}_2^{-1})^\top \in \mathbb{R}^{B \times H \times W \times 3} \quad (5.56)$$

$$\mathbf{Y}_{OkLAB} = \text{permute}(\mathbf{Y}_{OkLAB}^{\text{perm}}, (0, 3, 1, 2)) \in \mathbb{R}^{B \times 3 \times H \times W} \quad (5.57)$$

With the generated and real IHC images represented in the OkLAB color space, we can compute the $\mathcal{L}_2$ distance between them to obtain a meaningful measure of the perceptual color error, making this a computationally efficient method with minimal overhead from the OkLAB conversion. However, to leverage the benefits of the pyramidal strategy and to allow the integration of the masking strategy, our loss function is formulated as described in Equations 5.58-5.62 and illustrated in Figure 5.21.

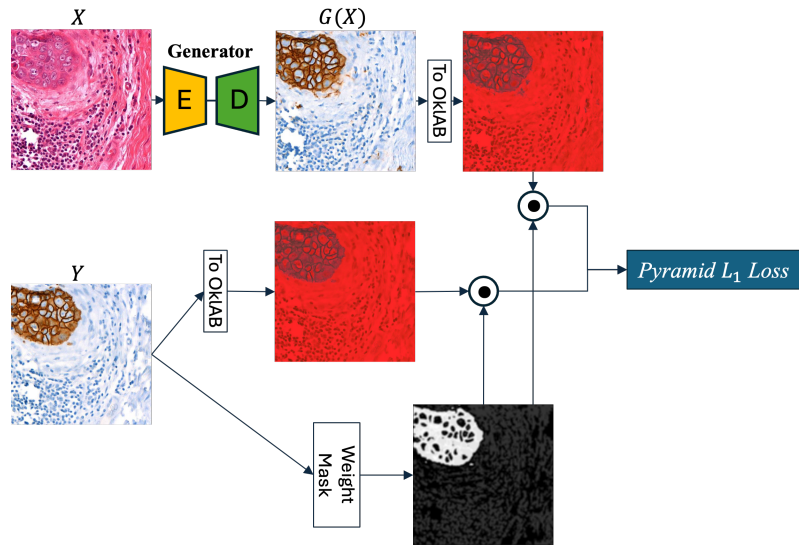


Figure 5.21: Illustration of the OkLAB loss (methodology).

After converting the images to the OkLAB color space, we compute the mask (Equation 5.59). This mask is either a weighted 2D map that matches the image size or a 2D map of ones if the masking strategy is not applied. Subsequently, we use the computed mask to scale the importance of the pixels (Equation 5.60). Finally, we compute the $\mathcal{L}_2$ distance across the octave levels using the pyramidal methodology (Equations 5.61–5.62).

$$\begin{aligned} \hat{\mathbf{Y}}_{\text{OkLAB}}^{(0)'}, \mathbf{Y}_{\text{OkLAB}}^{(0)'} &= \text{ToOkLAB}(\hat{\mathbf{Y}}_{\text{OkLAB}}), \quad \text{ToOkLAB}(\mathbf{Y}_{\text{OkLAB}}) \\ \text{mask} &= \begin{cases} \frac{\mathcal{G}_\sigma * \left(\sum_{i=1}^{N_{\text{brown}}} w_{\text{brown}}^i \cdot \mathbf{1}_{\text{DAB} > \tau_{\text{brown}}^i}(b, h, w) + \sum_{j=1}^{N_{\text{blue}}} w_{\text{blue}}^j \cdot \mathbf{1}_{\text{DAB} < \tau_{\text{blue}}^j}(b, h, w) \right)}{\max_{b, h, w} \left[\mathcal{G}_\sigma * \left(\sum_{i=1}^{N_{\text{brown}}} w_{\text{brown}}^i \cdot \mathbf{1}_{\text{DAB} > \tau_{\text{brown}}^i} + \sum_{j=1}^{N_{\text{blue}}} w_{\text{blue}}^j \cdot \mathbf{1}_{\text{DAB} < \tau_{\text{blue}}^j} \right) \right]} & \text{if masking} \\ 1 & \text{otherwise} \end{cases} \end{aligned} \quad (5.59)$$

$$\hat{\mathbf{Y}}_{\text{OkLAB}}^{(0)}, \mathbf{Y}_{\text{OkLAB}}^{(0)} = \mathcal{M}(\hat{\mathbf{Y}}_{\text{OkLAB}}^{(0)'}, \mathbf{Y}_{\text{OkLAB}}^{(0)'}, \text{mask}) \quad (5.60)$$

$$\hat{\mathbf{Y}}^{(l)} = \text{AvgPool} \left(\underbrace{\mathcal{G}_\sigma * \mathcal{G}_\sigma * \mathcal{G}_\sigma * \mathcal{G}_\sigma}_{4 \text{ times}} * \hat{\mathbf{Y}}^{(l-1)} \right), \quad l = 1, 2, 3, 4 \quad (5.61)$$

$$\mathcal{L}_{\text{OkLAB}}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{BCHW} \mathbf{1}^\top \left(\sum_{l=1}^4 \lambda_l \mathcal{L}_2(\mathbf{Y}_{\text{OkLAB}}^{(l)}, \hat{\mathbf{Y}}_{\text{OkLAB}}^{(l)}) \right) \mathbf{1} \quad (5.62)$$

Therefore, we can define the loss function of G as outlined in Equation 5.63.

$$\mathcal{L}_G(\mathbf{X}) = -\frac{1}{BHW} \mathbf{1}_{\text{BHW}}^\top \cdot \left(\log(D(G(\mathbf{X}), \mathbf{X})) \right) \cdot \mathbf{1}_{\text{BHW}} + \mathcal{L}_{\text{OkLAB}}(\mathbf{Y}, G(\mathbf{X})) + \mathcal{L}_{\text{CSS}}(\mathbf{X}, G(\mathbf{X})) \quad (5.63)$$

5.5.2 OkLAB Loss for Accurate Positive DAB Staining

The OkLAB loss introduces three variations compared to previous experiments: the use of $\mathcal{L}_2$ loss, penalization in the OkLAB color space, and integration of a masking strategy. To study the effect of each component, we conducted four experiments. Specifically, we trained CSSP2P GAN using $\mathcal{L}_2$ loss (CSSP2P(L2) GAN), CSSP2P GAN with masking strategy (mCSSP2P GAN), OkLAB GAN without masking, and OkLAB GAN with masking (mOkLAB GAN).

When evaluating the effect of masking on the CSSP2P GAN framework, we find that it improved performance on FID and KID, and most importantly on the JSD computed in the DAB channel (Table 5.7). However, it led to a decrease in performance in all other metrics. We also notice that the masked CSSP2P GAN overstains 3+ HER2-score cases, leading to an increase in recall of DAB-positive stained pixels but a decrease in precision within the image (Figure 5.22). Furthermore, we observe a reduction in the amount of predicted positive DAB staining regions in the negative cases (Figure 5.23). This reduction decreases the rate of false positives, which further

Table 5.7: Performance metrics on the validation set of the HER2match dataset for the CSSP2P GAN, OkLAB GAN, masked CSSP2P GAN, and masked OkLAB GAN experiments. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best and second-best metric values are highlighted in bold and underlined.

Model	Human Perception			Image Similarity			Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	CSS $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
CSSP2P GAN	<u>0.454 ± 0.069</u>	44	2.68 ± 0.59	<u>0.492 ± 0.117</u>	<u>0.475 ± 0.123</u>	<u>0.413 ± 0.097</u>	<u>20.0 ± 3.9</u>	0.871 ± 0.211
mCSSP2P GAN	<u>0.455 ± 0.067</u>	32	<u>1.37 ± 0.49</u>	0.472 ± 0.121	0.453 ± 0.129	0.387 ± 0.103	19.8 ± 3.9	<u>0.867 ± 0.212</u>
CSSP2P(L2) GAN	0.462 ± 0.070	64	4.63 ± 0.87	0.466 ± 0.124	0.453 ± 0.141	0.365 ± 0.107	<u>20.3 ± 4.1</u>	0.873 ± 0.212
OkLAB GAN	0.457 ± 0.070	45	<u>2.09 ± 0.50</u>	0.475 ± 0.116	0.462 ± 0.132	<u>0.391 ± 0.096</u>	19.9 ± 4.0	0.884 ± 0.200
mOkLAB GAN	0.456 ± 0.067	48	2.63 ± 0.93	0.472 ± 0.118	0.455 ± 0.129	0.394 ± 0.102	19.9 ± 4.0	<u>0.860 ± 0.218</u>

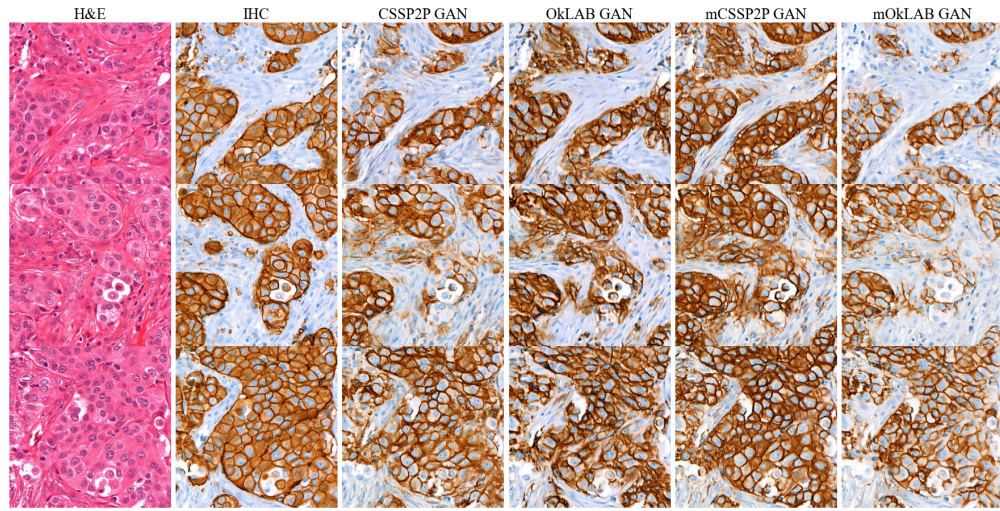


Figure 5.22: Qualitative assessment of the CSSP2P GAN, OkLAB GAN, masked CSSP2P GAN and masked OkLAB GAN on cases with a 3+ HER2-score and intense positive DAB staining.

helps explain the improved performance on the JSD in the DAB channel. These results align with the initial goal of using masking, which was to guide the model to focus on positive DAB-stained regions, thus improving the quality of the pathological representation. To achieve this, the masking strategy increased the contribution of pixels within the positive DAB-stained regions of the ground-truth, compared to non-stained pixels, in the loss function, thus explaining why overstaining occurs in cases with a 3+ HER2-score. Moreover, we observe that the separation between cells is marked by a darker brown stain, which helps to better perceive the boundaries between cells. This aligns with the use of the two brown thresholds, as the pixels in the regions between cells received the highest weight.

When performing the same evaluation for the OkLAB framework, we observe that masking consistently improved performance on the JSD computed in the DAB channel. However, it led to an improvement in LPIPS and a reduction in performance across all other metrics, except CSS and PSNR. A comparison of the positive cases from the extended selection of images from the validation set between the two settings further reveals a trend opposite to that of the CSSP2P framework

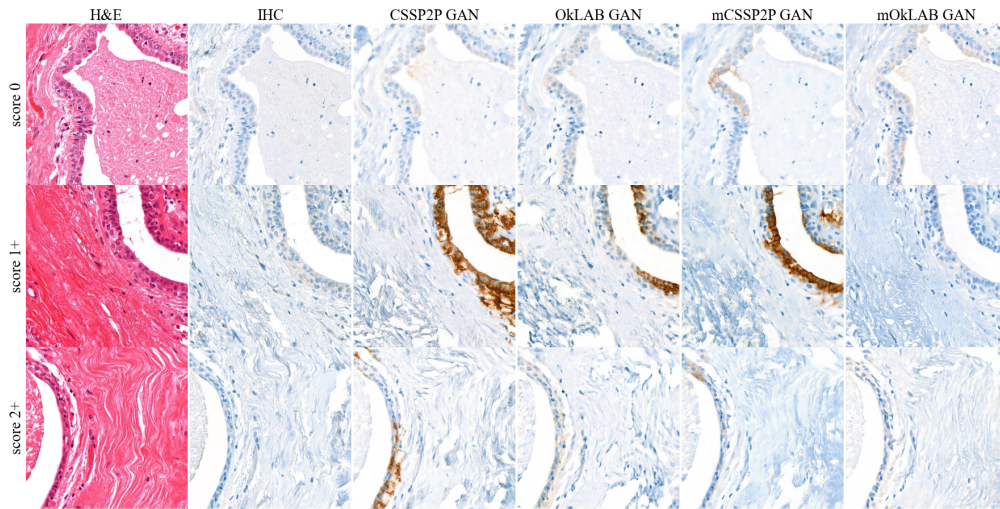


Figure 5.23: Qualitative comparison between the CSSP2P GAN, OkLAB GAN, masked CSSP2P GAN, and masked OkLAB GAN performance on the selected image examples from the validation set without positive DAB staining.

(Figure 5.22 and Figure 5.23). Although we observed overstaining in CSSP2P GAN, in this case, we see a decrease in the size of the positive DAB-stained regions, which improves the staining precision but reduces the recall of positively DAB-stained pixels. Additionally, examination of the negative cases shows that the use of masking effectively decreased the rate of false positives, making this model the best at identifying 3+ HER2-score cases. This may explain why it achieved the best performance on the JSD computed in the DAB channel.

This result further highlights the limitations of using JSD computed in the DAB channel as a comprehensive metric for evaluating the pathological representativeness of the generated images. A lower JSD value indicates a better approximation between the DAB channel histograms, suggesting an improved representativeness of pathological information. However, it does not provide information on how this representation is achieved, as histograms, by definition, do not account for the spatial arrangement of pixels in an image. This explains the discrepancy in the effect of masking between the CSSP2P GAN and OkLAB GAN models. However, as previously studied, a pixel-to-pixel measure would be perturbed by imperfect pixel-level correspondence. While visual inspection remains the most reliable method for evaluating model outputs, it is impractical to assess and compare all images across models. Additionally, it lacks the objectivity of a numerical metric. Therefore, with this result, we emphasize the pressing need for a meaningful and holistic pathological metric to assess the pathological representativeness of the generated images.

When comparing the two unmasked settings of CSSP2P GAN and OkLAB GAN, we observe that the latter exhibits overall reduced performance compared to the former. The penalization in the OkLAB color space was intended to target positive DAB signal regions in conjunction with the masking strategy. Indeed, analysis of the masked OkLAB GAN experiment shows that it achieved the best performance on the JSD computed in the DAB channel, demonstrating the efficacy of this approach. However, some performance loss was observed in other metrics, particularly LPIPS,

our reference metric, which prevents it from being selected as the best model. As a result, we selected CSSP2P GAN as our proposed model.

5.6 Summary

Beginning with BCE GAN, we demonstrated that the importance of adversarial loss goes beyond generating high-frequency-rich and realistic images. By conditioning the discriminator on the input H&E, we enhanced the preservation of semantic information, suggesting improved performance in achieving pathological consistency.

However, this conditioning also led to hallucinations of missing tissue regions, reflecting the characteristics of the dataset. To address this issue, we extended the BCE GAN by incorporating CSS loss, resulting in the CSS GAN. In addition to reducing the rate of hallucination by penalizing structural discrepancies between the H&E and generated images, it also had a positive effect on improving pathological consistency.

Subsequently, we achieved our proposed and best-performing model by adding Pyramid $\mathcal{L}_1$ loss to CSS GAN, resulting in CSSP2P GAN. The combination of these two losses had a positive regularizing effect, encouraging the generator to focus on more relevant image features to deceive the discriminator, rather than allowing the model to exploit missing tissue regions.

Despite improving the pathological consistency of the generated images, neither CSS loss nor Pyramid $\mathcal{L}_1$ loss specifically penalizes inaccurate representations of DAB stain. Therefore, we developed a two-fold methodology that combined masking and $\mathcal{L}_2$ loss in the OkLAB color space to penalize perceptual inaccuracies in positive DAB staining, resulting in OkLAB GAN. Although this strategy proved effective in enhancing DAB representation, the model showed decreased performance across other metrics, leading us to select CSSP2P GAN as our top-performing model.

Lastly, throughout this chapter, we confirmed that even when using pairs from the same tissue section, pixel-level alignment is not always guaranteed, posing challenges both in training and, more importantly, in achieving a robust evaluation with the currently used metrics in the field.

Chapter 6

Model Evaluation

In this chapter, we compare the performance of CSSP2P GAN with Pyramid Pix2Pix and ASP on the test set of the HER2match dataset, emphasizing the importance of using datasets with pairs from the same tissue section and highlighting the limitations of the currently used evaluation metrics. We then conduct a qualitative evaluation of a subset of images generated by CSSP2P GAN from the test set.

6.1 Model Comparison

6.1.1 Methodology

To compare the performance of CSSP2P GAN with that of Pyramid Pix2Pix and ASP on the HER2match dataset, we followed a methodology similar to that described in the Benchmarking Section 4.2 in Chapter 4. All random processes were initialized with the same seed, and all models were trained on an NVIDIA A100 to ensure a fair comparison. We selected the best checkpoint based on LPIPS, excluding the first epoch, as done during our model development. The hyperparameters were chosen based on the values reported in the respective articles, when available, or the default values provided in the code. The only exception was the number of epochs, which we adjusted to the HER2match dataset to ensure the total number of training steps was consistent with the original work. PSPStain was not included in this comparison, as its methodology relies on a pre-trained network specific to each dataset, and the authors do not provide the code or describe the pre-training process.

6.1.2 Results and Discussion

We verify that SSIM and PSNR identify ASP as the best-performing model (Table 6.1). However, ASP did not even achieve the correct IHC style. As previously mentioned, even when using datasets with pairs sampled from the same tissue section, pixel-level alignment cannot always be guaranteed due to mechanical and chemical treatments during re-staining that may deform the

tissue. This explains the limitations of these metrics in evaluating the quality of virtually stained images.

On the other hand, MS-SSIM, by evaluating SSIM at multiple scales, is more robust and faithful to the perceived quality of the generated images than SSIM. Similarly, LPIPS, which measures the distance at the activation layers of a pre-trained network, also demonstrates superior robustness. However, these metrics identified Pyramid Pix2Pix as the best-performing model (Table 6.1). Upon inspecting the generated examples (Figure 6.1), we observe that the Pyramid Pix2Pix outputs are smoother and lack the high-frequency details predicted by CSSP2P GAN. Therefore, while more robust than SSIM and PSNR, these metrics are still limited by imperfect pixel-level alignment, making them less effective in evaluating the similarity between generated and real IHC images, particularly for high-frequency details. Although these details may appear in different locations between the real and generated images, they can still be correct. However, pixel-level misalignment can lead to incorrect penalization by these metrics.

Table 6.1: Performance metrics on the test set of the HER2match dataset for the Pyramid Pix2Pix, ASP, and CSSP2P GAN models. Higher values are better for metrics with $\uparrow$, while lower values are better for metrics with $\downarrow$. The standard deviation reflects the variability of the metric values across all evaluated images. The best and second-best metric values are highlighted in bold and underlined.

Model	Human Perception			Image Similarity			Noise	DAB
	LPIPS $\downarrow$	FID $\downarrow$	KID $\downarrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	CSS $\uparrow$	PSNR $\uparrow$	JSD $\downarrow$
Identity	0.664 ± 0.079	252	26.09 ± 1.67	0.341 ± 0.106	0.310 ± 0.098	1.00	9.85	—
Pyramid Pix2Pix	<u>0.456 ± 0.065</u>	<u>67</u>	<u>5.94 ± 0.908</u>	<u>0.508 ± 0.105</u>	<u>0.533 ± 0.111</u>	<u>0.383 ± 0.105</u>	<u>20.3 ± 3.0</u>	0.947 ± 0.132
ASP	0.472 ± 0.046	102	9.09 ± 0.77	<u>0.513 ± 0.104</u>	<u>0.481 ± 0.106</u>	<u>0.608 ± 0.088</u>	<u>20.7 ± 2.7</u>	<u>0.877 ± 0.180</u>
CSSP2P GAN	<u>0.464 ± 0.070</u>	33	<u>2.53 ± 0.72</u>	0.426 ± 0.128	0.444 ± 0.128	0.354 ± 0.096	18.5 ± 3.8	<u>0.910 ± 0.183</u>

We observe that CSSP2P GAN outputs most closely resemble the nature of the dataset, and therefore, we argue that its outputs are closer to the real IHC images. However, as we have seen previously, faithfully mimicking the nature of the dataset comes with the possibility of generating missing regions of connective tissue, which further explains why LPIPS and MS-SSIM favor Pyramid Pix2Pix instead of CSSP2P GAN. These regions are the most susceptible to being damaged during re-staining, which explains why these tissue regions are sometimes missing in the IHC counterpart. Relying on this characteristic of the dataset and hallucinating missing regions of connective tissue is concerning during training, as it allows the generator to take a shortcut instead of learning meaningful image features to deceive the discriminator. Although undesirable, we argue that this form of hallucination does not compromise the pathological fidelity of CSSP2P GAN.

In addition to not observing any absence of cells in the generated images, we highlight that CSSP2P GAN achieved the best FID and KID metric values. This result demonstrates that the generated images by CSSP2P GAN and the real IHC images share the greatest amount of semantic information, suggesting the best performance in achieving pathological consistency. However,

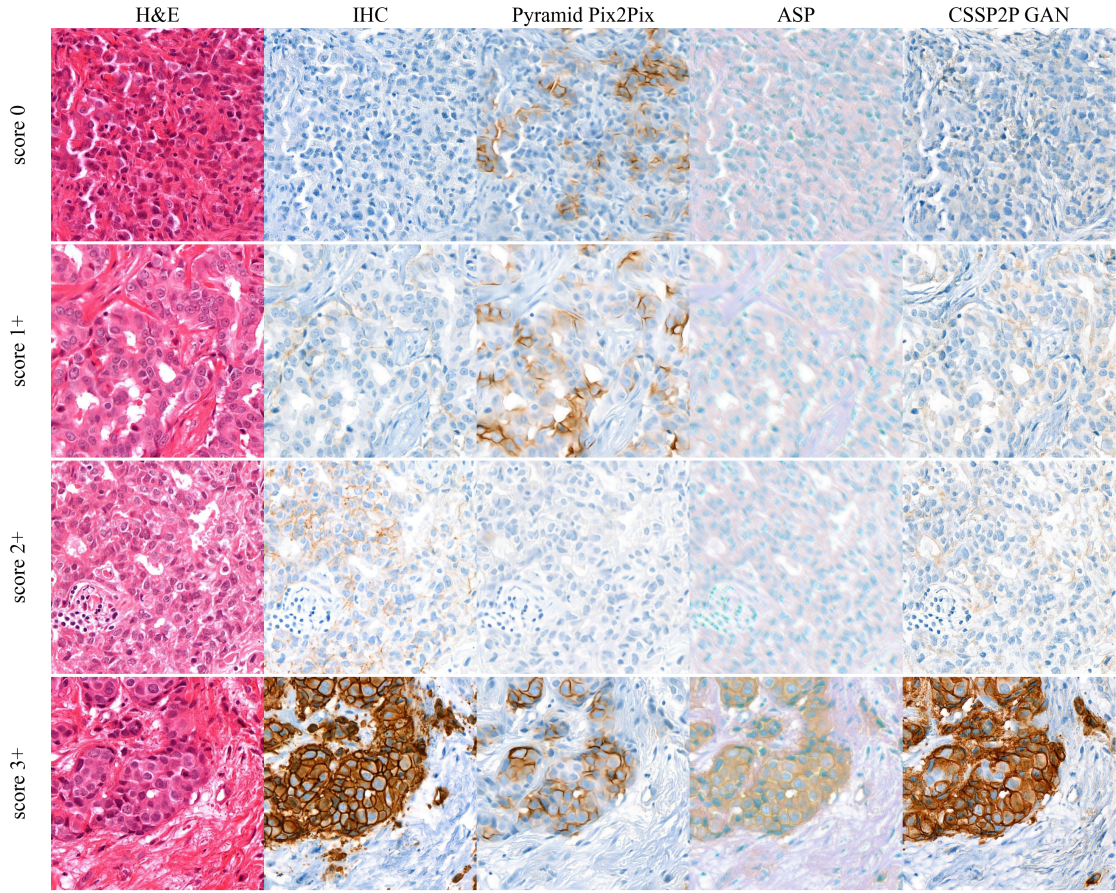


Figure 6.1: Qualitative assessment of pyramid Pix2Pix, ASP, and CSSP2P GAN models on the test set of the HER2match dataset.

these metrics use ImageNet pre-trained weights and cannot be considered definitive indicators of pathological fidelity.

On the other hand, ASP achieved the worst FID and KID values, which we attribute to its inability to achieve the IHC style. This result is sufficient to consider the ASP unsuitable for clinical practice. However, it achieved the best performance on the JSD metric. This constitutes an important finding as it further highlights the limitations of the JSD metric. Specifically, the limitation lies in the use of a non-calibrated deconvolution matrix to convert from the RGB color space to the HED concentration space. As discussed in Chapter 5, this introduces inaccuracies in the conversion, implying that one cannot impose lower and upper bounds on the histogram computation without biasing the result. To address this issue, we applied min-max normalization to the samples before computing the histogram, ensuring that both the generated and real IHC images fell within the range $[0, 1]$. While this approach allows for the computation of an unbounded histogram, it is sensitive to outliers and ignores any systematic offset between the distributions, making the histograms comparable despite potential shifts in pixel intensities.

We attribute the limitation of ASP in learning the IHC style to the possible competition between PatchNCE and ASP loss functions. As described in Chapter 4, while PatchNCE maximizes

the mutual information of a network’s embeddings between the H&E and generated images, ASP maximizes the mutual information of the embeddings of the same network between the generated and IHC images. As a result, although PatchNCE enables ASP to excel in CSS, it can also compete with ASP, thereby explaining the intermediate form between the H&E and IHC styles in the ASP outputs.

As a final remark, we highlight the significantly better quality of the Pyramid Pix2Pix and ASP generated images on the HER2match dataset, compared to BCI and MIST, as shown in the Benchmarking section 4.2. A deep learning-based model is only as good as the data on which it is trained, and this result underscores the critical importance of using high-quality datasets with H&E-IHC pairs sampled from the same tissue section in developing effective virtual staining models.

6.2 Qualitative Evaluation

6.2.1 Methodology

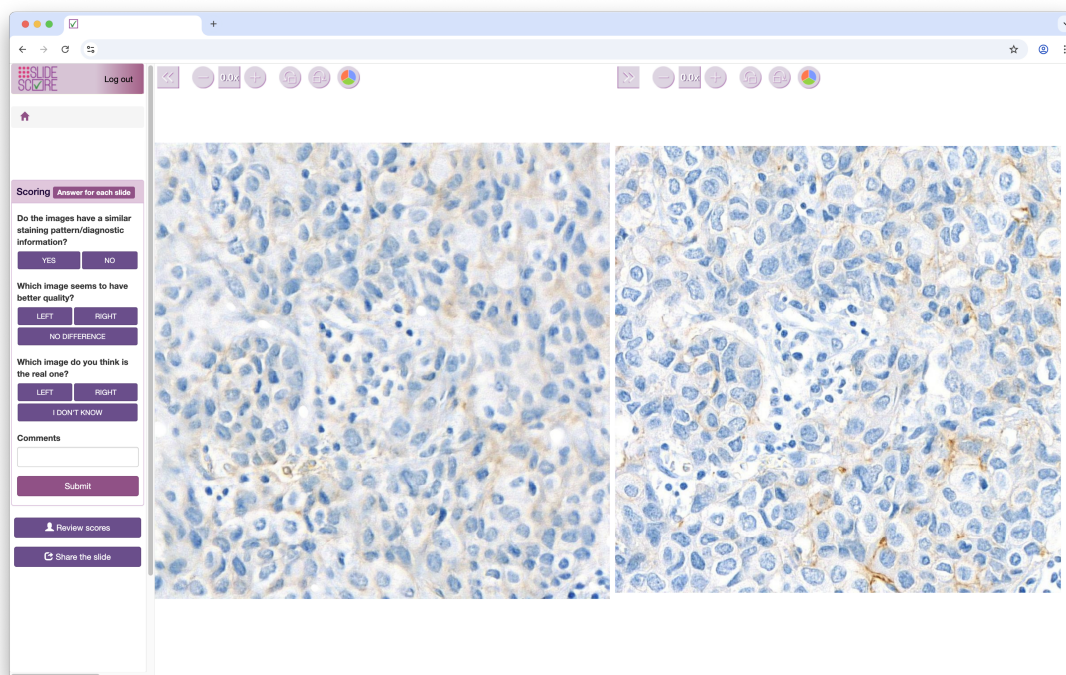


Figure 6.2: Example figure of the display window presented to pathologists during the qualitative evaluation.

In Section 5.2, we observed a significant improvement in FID and KID by conditioning the discriminator on the input H&E. This result indicates better performance in preserving semantic information, suggesting that this straightforward yet innovative approach enhances pathological consistency. When comparing CSSP2P GAN with Pyramid Pix2Pix and ASP, we found that

CSSP2P GAN significantly outperformed both models in these metrics. Additionally, visual inspection of the outputs suggested that CSSP2P GAN represented the DAB stain more accurately, further reinforcing the expectation that conditioning the discriminator on the H&E leads to heightened pathological fidelity. However, we emphasized that FID and KID were computed using pre-trained weights from ImageNet [47]; thus, this result cannot be a definitive indicator of the preservation of pathological information. Therefore, we invited three pathologists to participate in a blinded experiment to assess the quality of our virtually generated images. The pathologists were presented with two images, one generated and one real (Figure 6.2), and tasked with answering the following questions:

- Do the images have similar staining pattern/ diagnostic information?
- Which image seems to be of better quality?
- Which image do you think is the real one?

For this experiment, we randomly selected 500 images from the test set, stratified by HER2-score and slide ID. To minimize observer bias, we randomly duplicated 1% of the images. All images were anonymized, and the positions of real and generated images were randomly assigned (left or right). Pathologists were instructed to perform the evaluation individually, using monitors of the same size to ensure a consistent level of visual detail.

6.2.2 Results and Discussion

To guarantee the reliability of the qualitative evaluation, the reported values are based on a consensus, which means that at least two pathologists must agree (Table 6.2). Despite being rated as lower quality on average, the generated images are frequently mistaken for real ones. This can imply that authenticity is associated with particular features rather than the overall quality of the images. However, one should be aware of the possibility that the two questions are coupled. Therefore, the discrepancy between the quality and authenticity results may reflect their expectations, influencing their judgments.

Table 6.2: Qualitative evaluation consensus results on a selected set of 500 images from the HER2match test set. The reported values are based on a consensus, which means that at least two pathologists must agree. For questions (2) and (3), the values represent the proportion of times the generated image was chosen for each answer.

	HER2-Score				
	0	1+	2+	3+	All
(1) Consistent staining pattern?	0.683	0.726	0.650	0.658	0.676
(2) Which image has better quality?	0.233	0.234	0.228	0.203	0.226
(3) Which one is the real image?	0.683	0.661	0.585	0.602	0.636

In addition to closely resembling authentic IHC images, the similar staining pattern underscores a high level of pathological fidelity, supporting our earlier expectation in the previous section that the enhanced performance in preserving semantic information is connected to greater pathological consistency. Furthermore, the consistent performance across the HER2-scores reinforces the ability of the model to produce images with reliable pathological accuracy.

6.3 Summary of Results

In this chapter, while comparing the outputs of CSSP2P GAN with those of Pyramid Pix2Pix and ASP, we verify that our proposed model achieved the best performance. However, this improvement was not reflected in most evaluation metrics, highlighting the limitations of the currently used metrics in scenarios where pixel-level alignment is not always guaranteed, making them insufficient for evaluating the quality of virtual staining.

Of the metrics used, FID and KID proved to be the most aligned with the perceived quality of the generated images. Our model achieved top performance in these metrics due to the conditioning of the discriminator on the input H&E, which enhanced the preservation of shared semantic information with the ground-truth by leveraging the information encoded in the H&E. The subsequent qualitative evaluation emphasized the authenticity of our generated images and their high pathological fidelity, suggesting that the improved preservation of semantic information is linked to greater pathological consistency.

Chapter 7

Conclusion

In this dissertation, we provided a comprehensive overview of virtual staining from H&E to IHC for the clinical biomarker panel in BC, including a systematic review of the SOTA and a benchmarking study with a unified performance comparison of representative models in the field. This work is pioneering in the field, and we hope it will serve as a valuable resource for researchers and clinicians advancing virtual staining technology for BC diagnostics.

Recognizing the gaps in the current SOTA, we developed the CSSP2P GAN model on the HER2match dataset. Through iterative development, we analyzed various aspects of the GAN framework and their impact on the quality of the virtually stained images, with particular emphasis on the role of adversarial loss. By re-training Pyramid Pix2Pix and ASP on the HER2match dataset, we demonstrated the superior performance of our model and highlighted the limitations of the currently used evaluation metrics. Moreover, a qualitative evaluation further corroborated the realism and pathological fidelity of our generated images.

7.1 Future Work and Final Remarks

In this study, we demonstrate that adversarial loss is paramount in the quality of virtually stained images. In fact, by conditioning the discriminator on the input H&E, CSSP2P GAN achieved highly realistic and pathologically consistent outputs. This finding brings us closer to answering the question of whether cell biomarker expression can be predicted from H&E morphology and suggests that future research should explore further advancements in this topic.

We utilized binary cross-entropy loss to train our model. However, binary cross-entropy is a saturating loss function, meaning that as the discriminator improves, the update gradient becomes more reliable, but its magnitude approaches zero [144], leading to a problem of vanishing gradients. Furthermore, when a generated sample successfully fools the discriminator, it does not contribute to the update of the generator, which also slows convergence and hampers the approximation of the generated data distribution to the real data [145]. Therefore, as a perspective for future research, we propose experimenting with different adversarial losses and evaluating their

impact on the quality of virtually stained images, such as least squares [145] and Wasserstein losses [144].

Additionally, since adversarial loss is a dynamic network loss, we have limited control over the features the generator learns, which can lead to the learning of non-meaningful features. We therefore emphasize that future work should focus on developing strategies to regularize adversarial learning, ensuring that it complements, rather than dominates, the overall training process. Furthermore, we highlight the importance of studying the generalizability and interpretability of virtual staining models, as this will not only improve model robustness but also serve as a necessary step toward clinical application.

To further address the issue of hallucination of missing tissue regions and local structural discrepancies between the H&E and generated IHC, we propose modifying the discriminator architecture to process the H&E and IHC through separate initial layers, followed by a common branch that processes the concatenated latent features. Moreover, we note that due to the complex hyperparameter selection in GANs, we did not fine-tune the weighting of the CSS loss. Furthermore, since CSS loss penalizes contrast and structural discrepancies equally, we expect further refinements to enhance its performance, thus improving the preservation of H&E morphology.

Regarding data imbalance, we highlight that HER2-score labels are derived from the slide, meaning that examples with a 3+ HER2 score may not exhibit positive DAB staining. As a result, an evenly distributed dataset across HER2-scores can still be considered imbalanced. We believe that this imbalance may hinder the performance of the model in accurately representing DAB staining. To address this issue, we propose using common techniques, such as oversampling with augmentation or weighting cases differently during network updates.

Throughout the study, we emphasize that most of the metrics currently used do not adequately assess the quality of virtually stained images. We argue that without robust and representative evaluation methods, efforts in model development will always fall short. Therefore, we emphasize that focusing on the development of robust and pathologically faithful evaluation metrics should be a priority of future research.

Our findings suggest that developing a network-based metric may be the most promising approach. In addition to being more robust to pixel-level misalignment, these metrics evaluate the encoded information in the image, providing a more comprehensive assessment. Despite using ImageNet pre-trained weights, we verified that these metrics closely align with the perceived quality of virtually stained images. Therefore, using these methodologies with pre-trained weights on IHC images could be a first step in developing a higher-quality and pathologically faithful metric. LPIPS, on the other hand, evaluates similarity based on intermediate activations, which demonstrated lower robustness than FID and KID. However, we believe that there is still room for improvement in this type of metric. Enhancements may arise from the choice of architectural type and intermediate activations, with the use of pre-trained weights on IHC images also being a strong starting point.

Evaluation metrics are essential for model development and comparison. However, the true value of a virtually stained image lies in its usability and ability to convey accurate diagnostic

information. To this end, we emphasize the importance of qualitative evaluation, which not only tests the models in scenarios close to real clinical settings but also helps build trust in AI. Therefore, this type of test should be standardized in future research. In addition, although downstream tasks do not substitute for qualitative evaluation, they are easier to conduct, provide a more reproducible basis for model comparison, and offer meaningful insight into the utility and pathological fidelity of the generated images. However, the current SOTA lacks a robust downstream task for evaluating virtual staining of HER2 from H&E, and we anticipate further research in this area.

In conclusion, despite the contributions highlighted previously, we believe that the primary value of this dissertation lies in the constructed knowledge and identification of gaps in the current SOTA that have hindered further progress. Therefore, by taking a step back, we hope that the provided insights and directions for future research have created enough momentum to advance the field of virtual staining.

References

- [1] T. C. G. A. Network., “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012. DOI: [10.1038/nature11412](https://doi.org/10.1038/nature11412).
- [2] A. Goldhirsch, E. P. Winer, A. Coates, *et al.*, “Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013,” *Annals of Oncology*, vol. 24, no. 9, pp. 2206–2223, 2013. DOI: [10.1093/annonc/mdt303](https://doi.org/10.1093/annonc/mdt303).
- [3] World Health Organization, *Breast cancer*, <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>, Accessed: 05-11-2024, 2024.
- [4] C. E. Coles, B. O. Anderson, D. Cameron, *et al.*, “The lancet breast cancer commission: Tackling a global health, gender, and equity challenge,” *The Lancet*, vol. 399, no. 10330, pp. 1101–1103, 2022. DOI: [10.1016/S0140-6736\(22\)00184-2](https://doi.org/10.1016/S0140-6736(22)00184-2).
- [5] W. K. Funkhouser Jr, “Pathology,” in *Molecular Pathology*, Elsevier, 2018, pp. 217–229. DOI: [10.1016/B978-0-12-802761-5.00011-0](https://doi.org/10.1016/B978-0-12-802761-5.00011-0).
- [6] S. M. Hacking, E. Yakirevich, and Y. Wang, “From immunohistochemistry to new digital ecosystems: A state-of-the-art biomarker review for precision breast cancer medicine,” *Cancers*, vol. 14, no. 14, p. 3469, 2022. DOI: [10.3390/cancers14143469](https://doi.org/10.3390/cancers14143469).
- [7] E. Abels, L. Pantanowitz, F. Aeffner, *et al.*, “Computational pathology definitions, best practices, and recommendations for regulatory guidance: A white paper from the digital pathology association,” *The Journal of Pathology*, vol. 249, no. 3, pp. 286–294, 2019. DOI: doi.org/10.1002/path.5331.
- [8] B. Acs, M. Rantalainen, and J. Hartman, “Artificial intelligence as the next step towards precision pathology,” *Journal of Internal Medicine*, vol. 288, no. 1, pp. 62–81, 2020. DOI: [10.1111/joim.13030](https://doi.org/10.1111/joim.13030).
- [9] L. Latonen, S. Koivukoski, U. Khan, and P. Ruusuvaori, “Virtual staining for histology by deep learning,” *Trends in Biotechnology*, vol. 42, no. 9, pp. 1177–1191, Sep. 2024, ISSN: 0167-7799. DOI: [10.1016/j.tibtech.2024.02.009](https://doi.org/10.1016/j.tibtech.2024.02.009).
- [10] P. Klöckner, J. Teixeira, D. Montezuma, J. S. Cardoso, H. M. Horlings, and S. P. Oliveira, *Gans vs. diffusion models for virtual staining with the her2match dataset*, 2025. DOI: [10.48550/arXiv.2506.18484](https://doi.org/10.48550/arXiv.2506.18484). arXiv: 2506.18484 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2506.18484>.
- [11] W. K. Funkhouser, “Pathology: The clinical description of human disease,” in *Essential Concepts in Molecular Pathology (Second Edition)*, W. B. Coleman and G. J. Tsongalis, Eds., Second Edition, Academic Press, 2020, pp. 177–190, ISBN: 978-0-12-813257-9. DOI: [10.1016/B978-0-12-813257-9.00011-5](https://doi.org/10.1016/B978-0-12-813257-9.00011-5).

- [12] S. Magaki, S. Hojat, B. Wei, A. So, and W. Yong, “An introduction to the performance of immunohistochemistry,” English, in *Methods in Molecular Biology* (Methods in Molecular Biology), Methods in Molecular Biology. Humana Press Inc., 2019, pp. 289–298. DOI: [10.1007/978-1-4939-8935-5_25](https://doi.org/10.1007/978-1-4939-8935-5_25).
- [13] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever, “Breast Cancer Histopathology Image Analysis: A Review,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400–1411, 2014. DOI: [10.1109/TBME.2014.2303852](https://doi.org/10.1109/TBME.2014.2303852).
- [14] M. G. Hanna and L. Pantanowitz, “Digital pathology,” in *Encyclopedia of Biomedical Engineering*, R. Narayan, Ed., Oxford: Elsevier, 2019, pp. 524–532, ISBN: 978-0-12-805144-3. DOI: <https://doi.org/10.1016/B978-0-12-801238-3.99958-6>.
- [15] H. G. Russnes, O. C. Lingjærde, A.-L. Børresen-Dale, and C. Caldas, “Breast cancer molecular stratification: From intrinsic subtypes to integrative clusters,” *The American Journal of Pathology*, vol. 187, no. 10, pp. 2152–2162, 2017. DOI: [10.1016/j.ajpath.2017.04.022](https://doi.org/10.1016/j.ajpath.2017.04.022).
- [16] A. Spitale, P. Mazzola, D. Soldini, L. Mazzucchelli, and A. Bordoni, “Breast cancer classification according to immunohistochemical markers: Clinicopathologic features and short-term survival analysis in a population-based study from the south of switzerland,” *Annals of Oncology*, vol. 20, no. 4, pp. 628–635, 2009. DOI: [10.1093/annonc/mdn675](https://doi.org/10.1093/annonc/mdn675).
- [17] A. S. Coates, E. P. Winer, A. Goldhirsch, *et al.*, “Tailoring therapies—improving the management of early breast cancer: St gallen international expert consensus on the primary therapy of early breast cancer 2015,” *Annals of Oncology*, vol. 26, no. 8, pp. 1533–1546, 2015. DOI: [10.1093/annonc/mdv221](https://doi.org/10.1093/annonc/mdv221).
- [18] G. Curigliano, H. J. Burstein, E. P. Winer, *et al.*, “De-escalating and escalating treatments for early-stage breast cancer: The st. gallen international expert consensus conference on the primary therapy of early breast cancer 2017,” *Annals of Oncology*, vol. 28, no. 8, pp. 1700–1712, 2017. DOI: [10.1093/annonc/mdx308](https://doi.org/10.1093/annonc/mdx308).
- [19] L. Collins, *Tumors of the Mammary Gland* (AFIP Atlas of Tumor and Non-Tumor Pathology, Series 5, 18 Series). American Registry of Pathology, 2024, ISBN: 9781933477367. DOI: [10.55418/9781933477367](https://doi.org/10.55418/9781933477367).
- [20] E. A. Rakha, S. E. Pinder, J. M. Bartlett, *et al.*, “Updated uk recommendations for her2 assessment in breast cancer,” *Journal of clinical pathology*, vol. 68, no. 2, pp. 93–99, 2015. DOI: [10.1136/jclinpath-2014-202571](https://doi.org/10.1136/jclinpath-2014-202571).
- [21] K. H. Allison, M. E. H. Hammond, M. Dowsett, *et al.*, “Estrogen and progesterone receptor testing in breast cancer: Asco/cap guideline update,” *Journal of Clinical Oncology*, vol. 38, no. 12, pp. 1346–1366, 2020. DOI: [10.1200/JCO.19.02309](https://doi.org/10.1200/JCO.19.02309).
- [22] E. A. Rakha, E. Chmielik, F. C. Schmitt, P. H. Tan, C. M. Quinn, and G. Gallaghy, “Assessment of predictive biomarkers in breast cancer: Challenges and updates,” *Pathobiology*, vol. 89, no. 5, pp. 263–277, 2022. DOI: [10.1159/000525092](https://doi.org/10.1159/000525092).
- [23] A. C. Wolff, M. R. Somerfield, M. Dowsett, *et al.*, “Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology–college of american pathologists guideline update,” *Archives of Pathology & Laboratory Medicine*, vol. 147, no. 9, pp. 993–1000, 2023. DOI: [10.1200/JCO.22.02864](https://doi.org/10.1200/JCO.22.02864).
- [24] A. F. Vieira and F. Schmitt, “An update on breast cancer multigene prognostic tests—emergent clinical biomarkers,” *Frontiers in medicine*, vol. 5, p. 248, 2018. DOI: [10.3389/fmed.2018.00248](https://doi.org/10.3389/fmed.2018.00248).

- [25] R. Erber, M. Angeloni, R. Stöhr, *et al.*, “Molecular subtyping of invasive breast cancer using a pam50-based multigene expression test-comparison with molecular-like subtyping by tumor grade/immunohistochemistry and influence on oncologist’s decision on systemic therapy in a real-world setting,” *International Journal of Molecular Sciences*, vol. 23, no. 15, p. 8716, 2022. DOI: [10.3390/ijms23158716](https://doi.org/10.3390/ijms23158716).
- [26] F. Li, Z. Hu, W. Chen, and A. Kak, “Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Springer Nature Switzerland, 2023, pp. 632–641. DOI: [10.1007/978-3-031-43987-2_61](https://doi.org/10.1007/978-3-031-43987-2_61).
- [27] P. Klockner, J. Teixeira, D. Montezuma, *et al.*, “H&e to ihc virtual staining methods in breast cancer: An overview and benchmarking,” *npj Digital Medicine (Nature)*, 2025. DOI: [10.1038/s41746-025-01741-9](https://doi.org/10.1038/s41746-025-01741-9).
- [28] E. A. Rakha, G. M. Tse, and C. M. Quinn, “An update on the pathological classification of breast cancer,” *Histopathology*, vol. 82, no. 1, pp. 5–16, 2023. DOI: [10.1111/his.14786](https://doi.org/10.1111/his.14786).
- [29] C. Karaali, M. Emiroğlu, M. Değirmenci, M. Keser, S. Salimoğlu, and C. K. Talu, “The clinical and pathological characteristics that differentiate cases with “low estrogen receptor expression” from triple-negative breast cancer,” *European Journal of Breast Health*, vol. 20, no. 1, p. 19, 2024. DOI: [10.4274/ejbh.galenos.2023.2023-6-3](https://doi.org/10.4274/ejbh.galenos.2023.2023-6-3).
- [30] R. Colomer, M. Beltran, J. Dorcas, *et al.*, “It is not time to stop progesterone receptor testing in breast cancer,” *Journal of clinical oncology*, vol. 23, no. 16, pp. 3868–3869, 2005. DOI: [10.1200/JCO.2005.05.203](https://doi.org/10.1200/JCO.2005.05.203).
- [31] J. Li, X. Wang, S. Wang, *et al.*, “Expert consensus on the clinical diagnosis and targeted therapy of her2 breast cancer (2023 edition),” *Translational Breast Cancer Research*, vol. 3, 2022. DOI: [10.21037/tbcr-22-48](https://doi.org/10.21037/tbcr-22-48).
- [32] T. O. Nielsen, S. C. Y. Leung, D. L. Rimm, *et al.*, “Assessment of ki67 in breast cancer: Updated recommendations from the international ki67 in breast cancer working group,” *JNCI: Journal of the National Cancer Institute*, vol. 113, no. 7, pp. 808–819, 2021. DOI: [10.1093/jnci/djaa201](https://doi.org/10.1093/jnci/djaa201).
- [33] World Health Organization, *Global Breast Cancer Initiative Implementation Framework: assessing, strengthening and scaling-up of services for the early detection and management of breast cancer*, <https://iris.who.int/bitstream/handle/10665/365784/9789240067134-eng.pdf>, Accessed: 05-11-2024, 2023.
- [34] N. Pillar and A. Ozcan, “Virtual tissue staining in pathology using machine learning,” *Expert Review of Molecular Diagnostics*, vol. 22, no. 11, pp. 987–989, 2022. DOI: [10.1080/14737159.2022.2153040](https://doi.org/10.1080/14737159.2022.2153040).
- [35] B. Bai, X. Yang, Y. Li, Y. Zhang, N. Pillar, and A. Ozcan, “Deep learning-enabled virtual histological staining of biological samples,” *Light: Science & Applications*, vol. 12, no. 1, p. 57, 2023. DOI: [10.1038/s41377-023-01104-7](https://doi.org/10.1038/s41377-023-01104-7).
- [36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251. DOI: [10.1109/ICCV.2017.244](https://doi.org/10.1109/ICCV.2017.244).
- [37] A. Ramesh, M. Pavlov, G. Goh, *et al.*, *Zero-shot text-to-image generation*, 2021. DOI: [10.48550/arXiv.2102.12092](https://doi.org/10.48550/arXiv.2102.12092).

- [38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 10 674–10 685. DOI: [10.1109/CVPR52688.2022.01042](https://doi.org/10.1109/CVPR52688.2022.01042).
- [39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014, pp. 2672–2680.
- [40] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, 2015. DOI: [10.48550/arXiv.1503.03585](https://doi.org/10.48550/arXiv.1503.03585).
- [41] M. Mirza and S. Osindero, *Conditional generative adversarial nets*, 2014. DOI: [10.48550/arXiv.1411.1784](https://doi.org/10.48550/arXiv.1411.1784).
- [42] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, “Contrastive learning for unpaired image-to-image translation,” in *European Conference on Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Springer International Publishing, 2020, pp. 319–345, ISBN: 978-3-030-58545-7. DOI: [10.1007/978-3-030-58545-7_19](https://doi.org/10.1007/978-3-030-58545-7_19).
- [43] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin, “Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 1814–1823. DOI: [10.1109/CVPRW56347.2022.00198](https://doi.org/10.1109/CVPRW56347.2022.00198).
- [44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976. DOI: [10.1109/CVPR.2017.632](https://doi.org/10.1109/CVPR.2017.632).
- [45] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 248–255. DOI: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848).
- [48] H. Xiao, T. Liu, Y. Sun, Y. Li, S. Zhao, and A. Avolio, “Remote photoplethysmography for heart rate measurement: A review,” *Biomedical Signal Processing and Control*, vol. 88, p. 105 608, 2024, ISSN: 1746-8094. DOI: <https://doi.org/10.1016/j.bspc.2023.105608>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1746809423010418>.
- [49] J. M. Tomczak, *Deep Generative Modeling*. Springer International Publishing, 2024, ISBN: 9783031640872. DOI: [10.1007/978-3-031-64087-2](https://doi.org/10.1007/978-3-031-64087-2).
- [50] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *arXiv*, 2020. DOI: [10.48550/arXiv.2006.11239](https://doi.org/10.48550/arXiv.2006.11239).

- [51] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image super-resolution via iterative refinement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023. DOI: [10.1109/TPAMI.2022.3204461](https://doi.org/10.1109/TPAMI.2022.3204461).
- [52] Y. Zhang, N. Huang, F. Tang, *et al.*, “Inversion-based style transfer with diffusion models,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 146–10 156. DOI: [10.1109/CVPR52729.2023.00978](https://doi.org/10.1109/CVPR52729.2023.00978).
- [53] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” in *Proceedings of the 35th International Conference on Neural Information Processing Systems – NeurIPS*, Red Hook, NY, USA: Curran Associates Inc., 2021, pp. 8780–8794.
- [54] G. Müller-Franzes, J. M. Niehues, F. Khader, *et al.*, “A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis,” *Sci. Rep.*, vol. 13, no. 1, p. 12 098, 2023. DOI: [10.1038/s41598-023-39278-0](https://doi.org/10.1038/s41598-023-39278-0).
- [55] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861).
- [56] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, 1398–1402 Vol.2. DOI: [10.1109/ACSSC.2003.1292216](https://doi.org/10.1109/ACSSC.2003.1292216).
- [57] S. Liu, B. Zhang, Y. Liu, *et al.*, “Unpaired stain transfer using pathology-consistent constrained generative adversarial networks,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 1977–1989, 2021. DOI: [10.1109/TMI.2021.3069874](https://doi.org/10.1109/TMI.2021.3069874).
- [58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595. DOI: [10.1109/CVPR.2018.00068](https://doi.org/10.1109/CVPR.2018.00068).
- [59] A. Ghildyal, N. Barman, and S. Zadtootaghaj, *Foundation models boost low-level perceptual similarity metrics*, 2025. DOI: [10.48550/arXiv.2409.07650](https://doi.org/10.48550/arXiv.2409.07650). arXiv: 2409.07650 [cs.CV].
- [60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2016, pp. 2818–2826. DOI: [10.1109/CVPR.2016.308](https://doi.org/10.1109/CVPR.2016.308).
- [61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a nash equilibrium,” *arXiv*, 2017. DOI: [10.48550/arXiv.1706.08500](https://doi.org/10.48550/arXiv.1706.08500).
- [62] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, *Demystifying mmd gans*, 2021. DOI: [10.48550/arXiv.1801.01401](https://doi.org/10.48550/arXiv.1801.01401).
- [63] X. Guan, Y. Wang, Y. Lin, X. Li, and Y. Zhang, “Unsupervised multi-domain progressive stain transfer guided by style encoding dictionary,” *IEEE Transactions on Image Processing*, vol. 33, pp. 767–779, 2024. DOI: [10.1109/TIP.2024.3349866](https://doi.org/10.1109/TIP.2024.3349866).
- [64] A. C. Ruifrok and D. A. Johnston, “Quantification of histochemical staining by color deconvolution,” *en, Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, Aug. 2001.

- [65] F. Martino, G. Ilardi, S. Varricchio, *et al.*, “A deep learning model to predict ki-67 positivity in oral squamous cell carcinoma,” *Journal of Pathology Informatics*, vol. 15, p. 100354, 2024, ISSN: 2153-3539. DOI: [10.1016/j.jpi.2023.100354](https://doi.org/10.1016/j.jpi.2023.100354).
- [66] Q. Peng, W. Lin, Y. Hu, *et al.*, “Advancing h&e-to-ihc virtual staining with task-specific domain knowledge for her2 scoring,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru, Q. Dou, A. Feragen, *et al.*, Eds., Cham: Springer Nature Switzerland, 2024, pp. 3–13, ISBN: 978-3-031-72083-3. DOI: [10.1007/978-3-031-72083-3_1](https://doi.org/10.1007/978-3-031-72083-3_1).
- [67] B. Zeng, Y. Lin, Y. Wang, *et al.*, “Semi-supervised pr virtual staining for breast histopathological images,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Springer Nature Switzerland, 2022, pp. 232–241. DOI: [10.1007/978-3-031-16434-7_23](https://doi.org/10.1007/978-3-031-16434-7_23).
- [68] G. Duan, Y. Cao, W. Guo, L. Cui, and Z. Liu, “A virtual staining method for immunohistochemical images of breast cancer,” in *2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2023, pp. 1–5. DOI: [10.1109/CISP-BMEI60920.2023.10373388](https://doi.org/10.1109/CISP-BMEI60920.2023.10373388).
- [69] J. Ma and H. Chen, “Efficient supervised pretraining of swin-transformer for virtual staining of microscopy images,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 4, pp. 1388–1399, 2024. DOI: [10.1109/TMI.2023.3337253](https://doi.org/10.1109/TMI.2023.3337253).
- [70] S. Huang, H. Wang, Y. Hao, S. Guo, Y. Wang, and T. Wang, “Tc-cyclegan: Improved cyclegan with texture constraints for virtual staining of pathological images,” in *Proceedings of the 2023 3rd International Conference on Bioinformatics and Intelligent Computing*, ser. BIC ’23, New York, NY, USA: Association for Computing Machinery, 2023, pp. 147–152, ISBN: 9798400700200. DOI: [10.1145/3592686.3592713](https://doi.org/10.1145/3592686.3592713).
- [71] M. Baldeon-Calisto, F. Rivera-Velastegui, D. Riofrío, R. Flores-Moyano, N. Pérez, and D. Benítez, “Deepsit: Deeply supervised framework for image translation on breast cancer analysis,” in *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, 2023, pp. 1–7. DOI: [10.1109/ICPRS58416.2023.10178999](https://doi.org/10.1109/ICPRS58416.2023.10178999).
- [72] L. Liu, Z. Liu, J. Chang, H. Qiao, T. Sun, and J. Shang, “Mggan: A multi-generator generative adversarial network for breast cancer immunohistochemical image generation,” *Heliyon*, vol. 9, no. 10, e20614, 2023, ISSN: 2405-8440. DOI: [10.1016/j.heliyon.2023.e20614](https://doi.org/10.1016/j.heliyon.2023.e20614).
- [73] Y. Ma, Y. Zhang, Z. Wang, *et al.*, “Dsff-gan: A novel stain transfer network for generating immunohistochemical image of endometrial cancer,” *Computers in Biology and Medicine*, vol. 170, p. 108046, 2024. DOI: [10.1016/j.combiomed.2024.108046](https://doi.org/10.1016/j.combiomed.2024.108046).
- [74] S. Wu and S. Xu, “HcGAN: Harmonic conditional generative adversarial network for efficiently generating high-quality IHC images from H&E,” *Heliyon*, vol. 10, no. 20, e37902, Oct. 2024. DOI: [10.1016/j.heliyon.2024.e37902](https://doi.org/10.1016/j.heliyon.2024.e37902).
- [75] L. Wei, S. Hua, S. Zhang, and X. Zhang, “Derestainer: H&e to ihc pathological image translation via decoupled staining channels,” in *Deep Generative Models*, A. Mukhopadhyay, I. Oksuz, S. Engelhardt, D. Mehrof, and Y. Yuan, Eds., Cham: Springer Nature Switzerland, 2025, pp. 1–10, ISBN: 978-3-031-72744-3. DOI: [10.1007/978-3-031-72744-3_1](https://doi.org/10.1007/978-3-031-72744-3_1).

- [76] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin, *Breast cancer immunohistochemical image generation challenge*, bci.grand-challenge.org, Accessed: 2024-11-20, 2024.
- [77] R. Zhang, Y. Cao, Y. Li, *et al.*, “Mvfstain: Multiple virtual functional stain histopathology images generation based on specific domain mapping,” *Medical Image Analysis*, vol. 80, p. 102 520, 2022, ISSN: 1361-8415. DOI: [10.1016/j.media.2022.102520](https://doi.org/10.1016/j.media.2022.102520).
- [78] L. Qu, C. Zhang, G. Li, H. Zheng, C. Peng, and W. He, “Advancing h&e-to-ihc stain translation in breast cancer: A multi-magnification and attention-based approach,” in *2024 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, 2024, pp. 441–446. DOI: [10.1109/CIS-RAM61939.2024.10673328](https://doi.org/10.1109/CIS-RAM61939.2024.10673328).
- [79] F. Chen, R. Zhang, B. Zheng, Y. Sun, J. He, and W. Qin, “Pathological semantics-preserving learning for h&e-to-ihc virtual staining,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Cham: Springer Nature Switzerland, 2024, pp. 384–394. DOI: [10.1007/978-3-031-72083-3_36](https://doi.org/10.1007/978-3-031-72083-3_36).
- [80] Y. Li, X. Guan, Y. Wang, and Y. Zhang, “Exploiting supervision information in weakly paired images for ihc virtual staining,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru, Q. Dou, A. Feragen, *et al.*, Eds., Springer Nature Switzerland, 2024, pp. 113–122, ISBN: 978-3-031-72083-3. DOI: [10.1007/978-3-031-72083-3_11](https://doi.org/10.1007/978-3-031-72083-3_11).
- [81] S. Wang, Z. Zhang, H. Yan, M. Xu, and G. Wang, *Mix-domain contrastive learning for unpaired h&e-to-ihc stain translation*, 2024. DOI: [10.48550/arXiv.2406.11799](https://doi.org/10.48550/arXiv.2406.11799).
- [82] W. Zhang, T. H. Hui, P. Y. Tse, F. Hill, C. Lau, and X. Li, “High-resolution medical image translation via patch alignment-based bidirectional contrastive learning,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru, Q. Dou, A. Feragen, *et al.*, Eds., Springer Nature Switzerland, 2024, pp. 178–188, ISBN: 978-3-031-72083-3. DOI: [10.1007/978-3-031-72083-3_17](https://doi.org/10.1007/978-3-031-72083-3_17).
- [83] J. Li, J. Dong, S. Huang, *et al.*, “Virtual immunohistochemistry staining for histological images assisted by weakly-supervised learning,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2024, pp. 11 259–11 268. DOI: [10.1109/CVPR52733.2024.01070](https://doi.org/10.1109/CVPR52733.2024.01070).
- [84] J. Hu, G. Sun, Y. Liu, B. Sheng, X. Zou, and L. Xin, “Ulvit-gan: Advancing stain transfer in h&e and ihc pathology images with a unet-like vision-transformer gan,” in *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, ser. CAICE ’24, Xi’ an, China: Association for Computing Machinery, 2024, pp. 524–528, ISBN: 9798400716942. DOI: [10.1145/3672758.3672843](https://doi.org/10.1145/3672758.3672843).
- [85] Y. He, Z. Liu, M. Qi, *et al.*, “Pst-diff: Achieving high-consistency stain transfer by diffusion models with pathological and structural constraints,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 10, pp. 3634–3647, 2024. DOI: [10.1109/TMI.2024.3430825](https://doi.org/10.1109/TMI.2024.3430825).
- [86] Z. Li, T. Su, B. Zhang, *et al.*, “His-mmmdm: Multi-domain and multi-omics translation of histopathology images with diffusion models,” *medRxiv*, 2024. DOI: [10.1101/2024.07.11.24310294](https://doi.org/10.1101/2024.07.11.24310294).

- [87] Y. Jia, G. Duan, Y. Song, L. Ye, and Z. Liu, “Dtnet: Dual-encoder generative adversarial network for generating breast cancer immunohistochemical images,” in *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, 2024, pp. 937–942. DOI: [10.1109/CVIDL62147.2024.10603639](https://doi.org/10.1109/CVIDL62147.2024.10603639).
- [88] C. Ji, K. Oshima, T. Urata, *et al.*, “Transformation from hematoxylin-and-eosin staining to Ki-67 immunohistochemistry digital staining images using deep learning: experimental validation on the labeling index,” *Journal of Medical Imaging*, vol. 11, no. 4, p. 047 501, 2024. DOI: [10.1117/1.JMI.11.4.047501](https://doi.org/10.1117/1.JMI.11.4.047501).
- [89] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929. DOI: [10.1109/CVPR.2016.319](https://doi.org/10.1109/CVPR.2016.319).
- [90] J. Kim, M. Kim, H. Kang, and K. Lee, *U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation*, 2020. DOI: [10.48550/arXiv.1907.10830](https://doi.org/10.48550/arXiv.1907.10830).
- [91] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *European Conference on Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 179–196, ISBN: 978-3-030-01219-9. DOI: [10.1007/978-3-030-01219-9_11](https://doi.org/10.1007/978-3-030-01219-9_11).
- [92] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, “Elastix: A toolbox for intensity-based medical image registration,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010. DOI: [10.1109/TMI.2009.2035616](https://doi.org/10.1109/TMI.2009.2035616).
- [93] Z. Liu, Y. Lin, Y. Cao, *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002. DOI: [10.1109/ICCV48922.2021.00986](https://doi.org/10.1109/ICCV48922.2021.00986).
- [94] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009. DOI: [10.1109/CVPR52688.2022.01553](https://doi.org/10.1109/CVPR52688.2022.01553).
- [95] Z. Xie, Z. Zhang, Y. Cao, *et al.*, “Simmim: A simple framework for masked image modeling,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663. DOI: [10.1109/CVPR52688.2022.00943](https://doi.org/10.1109/CVPR52688.2022.00943).
- [96] E. M. Christiansen, S. J. Yang, D. M. Ando, *et al.*, “In silico labeling: Predicting fluorescent labels in unlabeled images,” *Cell*, vol. 173, no. 3, 792–803.e19, 2018, ISSN: 0092-8674. DOI: [10.1016/j.cell.2018.03.040](https://doi.org/10.1016/j.cell.2018.03.040).
- [97] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 5967–5976. DOI: [10.1109/CVPR.2017.632](https://doi.org/10.1109/CVPR.2017.632).
- [98] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, “Two deterministic half-quadratic regularization algorithms for computed imaging,” in *Proceedings of 1st International Conference on Image Processing*, vol. 2, 1994, 168–172 vol.2. DOI: [10.1109/ICIP.1994.413553](https://doi.org/10.1109/ICIP.1994.413553).

- [99] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition – CVPR*, 2018, pp. 8798–8807. DOI: [10.1109/CVPR.2018.00917](https://doi.org/10.1109/CVPR.2018.00917).
- [100] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022. DOI: [10.1109/TPAMI.2020.3045810](https://doi.org/10.1109/TPAMI.2020.3045810).
- [101] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007. DOI: [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324).
- [102] J. Borovec, J. Kybic, I. Arganda-Carreras, *et al.*, “Anhir: Automatic non-rigid histological image registration challenge,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, pp. 3042–3052, 2020. DOI: [10.1109/TMI.2020.2986331](https://doi.org/10.1109/TMI.2020.2986331).
- [103] E. Ustinova and V. Lempitsky, “Learning deep embeddings with histogram loss,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS’16, Barcelona, Spain: Curran Associates Inc., 2016, pp. 4177–4185, ISBN: 9781510838819.
- [104] A. Ghosh and A. Lan, “Contrastive learning improves model robustness under label noise,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 2697–2702. DOI: [10.1109/CVPRW53098.2021.00304](https://doi.org/10.1109/CVPRW53098.2021.00304).
- [105] Y. Xue, K. Whitecross, and B. Mirzasoleiman, “Investigating why contrastive learning benefits robustness against label noise,” in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 17–23 Jul 2022, pp. 24 851–24 871. [Online]. Available: <https://proceedings.mlr.press/v162/xue22a.html>.
- [106] E. Zheltonozhskii, C. Baskin, A. Mendelson, A. M. Bronstein, and O. Litany, “Contrast to divide: Self-supervised pre-training for learning with noisy labels,” in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 387–397. DOI: [10.1109/WACV51458.2022.00046](https://doi.org/10.1109/WACV51458.2022.00046).
- [107] A. van den Oord, Y. Li, and O. Vinyals, *Representation learning with contrastive predictive coding*, 2019. DOI: [10.48550/arXiv.1807.03748](https://doi.org/10.48550/arXiv.1807.03748). arXiv: [1807.03748](https://arxiv.org/abs/1807.03748) [cs.LG]. [Online]. Available: <https://arxiv.org/abs/1807.03748>.
- [108] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, “Designing a practical degradation model for deep blind image super-resolution,” in *IEEE International Conference on Computer Vision*, 2021, pp. 4791–4800. DOI: [10.1109/ICCV48922.2021.00475](https://doi.org/10.1109/ICCV48922.2021.00475).
- [109] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. DOI: [10.48550/arXiv.1409.1556](https://doi.org/10.48550/arXiv.1409.1556).
- [110] M. Ilse, J. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” in *35th International Conference on Machine Learning – ICML 2018*, J. Dy and A. Krause, Eds., ser. 35th International Conference on Machine Learning, ICML 2018, International Machine Learning Society (IMLS), 2018, pp. 3376–3391.
- [111] B. Li, K. Xue, B. Liu, and Y.-K. Lai, “Bbmd: Image-to-image translation with brownian bridge diffusion models,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1952–1961. DOI: [10.1109/CVPR52729.2023.00194](https://doi.org/10.1109/CVPR52729.2023.00194).

- [112] X. Su, J. Song, C. Meng, and S. Ermon, *Dual diffusion implicit bridges for image-to-image translation*, 2023. DOI: [10.48550/arXiv.2203.08382](https://doi.org/10.48550/arXiv.2203.08382).
- [113] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, “Ilvr: Conditioning method for denoising diffusion probabilistic models,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 347–14 356. DOI: [10.1109/ICCV48922.2021.01410](https://doi.org/10.1109/ICCV48922.2021.01410).
- [114] P. Weitz, M. Valkonen, L. Solorzano, *et al.*, “The acrobat 2022 challenge: Automatic registration of breast cancer tissue,” *Medical Image Analysis*, vol. 97, p. 103 257, 2024, ISSN: 1361-8415. DOI: [10.1016/j.media.2024.103257](https://doi.org/10.1016/j.media.2024.103257).
- [115] K. B. Ozyoruk, S. Can, B. Darbaz, *et al.*, “A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded,” *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1407–1419, Dec. 2022, ISSN: 2157-846X. DOI: [10.1038/s41551-022-00952-9](https://doi.org/10.1038/s41551-022-00952-9).
- [116] S. Liu and W. Deng, “Very deep convolutional neural network based image classification using small training sample size,” in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734. DOI: [10.1109/ACPR.2015.7486599](https://doi.org/10.1109/ACPR.2015.7486599).
- [117] P. Pati, S. Karkampouna, F. Bonollo, *et al.*, “Accelerating histopathology workflows with generative ai-based virtually multiplexed tumour profiling,” *Nature Machine Intelligence*, vol. 6, pp. 1077–1093, Sep. 2024. DOI: [10.1038/s42256-024-00889-5](https://doi.org/10.1038/s42256-024-00889-5).
- [118] D. Tellez, G. Litjens, P. Bándi, *et al.*, “Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology,” *Medical Image Analysis*, vol. 58, p. 101 544, 2019, ISSN: 1361-8415. DOI: <https://doi.org/10.1016/j.media.2019.101544>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1361841519300799>.
- [119] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, 2015. DOI: [10.48550/arXiv.1502.03167](https://doi.org/10.48550/arXiv.1502.03167). arXiv: [1502.03167](https://arxiv.org/abs/1502.03167) [cs.LG]. [Online]. Available: <https://arxiv.org/abs/1502.03167>.
- [120] H. Ide and T. Kurita, “Improvement of learning for cnn with relu activation by sparse regularization,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2684–2691. DOI: [10.1109/IJCNN.2017.7966185](https://doi.org/10.1109/IJCNN.2017.7966185).
- [121] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1510–1519. DOI: [10.1109/ICCV.2017.167](https://doi.org/10.1109/ICCV.2017.167).
- [122] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, *Spectral normalization for generative adversarial networks*, 2018. DOI: [10.48550/arXiv.1802.05957](https://doi.org/10.48550/arXiv.1802.05957). arXiv: [1802.05957](https://arxiv.org/abs/1802.05957) [cs.LG]. [Online]. Available: <https://arxiv.org/abs/1802.05957>.
- [123] P. Foundation, *Spectral normalization*, https://docs.pytorch.org/docs/stable/generated/torch.nn.utils.spectral_norm.html, Accessed: 22-05-2025, 2017.
- [124] A. Odena, V. Dumoulin, and C. Olah, *Deconvolution and checkerboard artifacts*, <https://distill.pub/2016/deconv-checkerboard/>, Accessed: 23-05-2025, 2017.
- [125] Y. Singh, M. Saini, and Savita, “Impact and performance analysis of various activation functions for classification problems,” in *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, vol. 1, 2023, pp. 1–7. DOI: [10.1109/InC457730.2023.10263129](https://doi.org/10.1109/InC457730.2023.10263129).

- [126] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1512.03385>.
- [127] P. Foundation, *Resnet-50*, <https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html>, Accessed: 22-05-2025, 2017.
- [128] D. Ulyanov, A. Vedaldi, and V. Lempitsky, *Instance normalization: The missing ingredient for fast stylization*, 2017. arXiv: 1607.08022 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1607.08022>.
- [129] H. Dwivedi, *Understanding gan loss functions*, <https://neptune.ai/blog/gan-loss-functions>, Accessed: 25-05-2025, 2023.
- [130] S. Wang and P. Kanwar, *Bfloat16: The secret to high performance on cloud tpus*, <https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus>, Accessed: 27-05-2025, 2019.
- [131] P. Foundation, *Adamw*, <https://docs.pytorch.org/docs/stable/generated/torch.optim.AdamW.html>, Accessed: 27-05-2025, 2025.
- [132] T. Agrawal, *Gans failure modes: How to identify and monitor them*, <https://neptune.ai/blog/gan-failure-modes>, Accessed: 27-05-2025, 2025.
- [133] H. Hoyez, C. Schockaert, J. Rambach, B. Mirbach, and D. Stricker, “Unsupervised image-to-image translation: A review,” en, *Sensors (Basel)*, vol. 22, no. 21, p. 8540, Nov. 2022. DOI: 10.3390/s22218540.
- [134] J. Lee, J. Park, S. Baik, and K. M. Lee, *Rethinking rgb color representation for image restoration models*, 2024. DOI: 10.48550/arXiv.2402.03399. arXiv: 2402.03399 [eess.IV]. [Online]. Available: <https://arxiv.org/abs/2402.03399>.
- [135] J. Lotz, N. Weiss, J. van der Laak, and S. Heldmann, “Comparison of consecutive and restained sections for image registration in histopathology,” en, *J. Med. Imaging (Bellingham)*, vol. 10, no. 6, p. 067 501, Nov. 2023. DOI: 10.1117/1.JMI.10.6.067501.
- [136] A. A. Heydari, C. A. Thompson, and A. Mehmood, *Softadapt: Techniques for adaptive loss weighting of neural networks with multi-part loss functions*, 2019. DOI: 10.48550/arXiv.1912.12355. arXiv: 1912.12355 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/1912.12355>.
- [137] S.-I. Team, *Api reference, skimage.color*, <https://scikit-image.org/docs/0.25.x/api/skimage.color.html#skimage.color.rgb2hed>, Accessed: 02-06-2025, 2025.
- [138] R. Warner, “Measurements of water-holding capacity and color: Objective and subjective,” in *Encyclopedia of Meat Sciences (Third Edition)*, M. Dikeman, Ed., Third Edition, Oxford: Elsevier, 2024, pp. 478–492, ISBN: 978-0-323-85198-5. DOI: <https://doi.org/10.1016/B978-0-323-85125-1.00072-7>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780323851251000727>.
- [139] C. Lilley, *Better than lab? gamut reduction cie lab & oklab*, <https://www.w3.org/Graphics/Color/Workshop/slides/talk/lilley>, Accessed: 31-05-2025, 2021.

- [140] G. Sharma, W. Wu, and E. N. Dalal, “The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations,” *Color Research & Application*, vol. 30, no. 1, pp. 21–30, 2005. DOI: <https://doi.org/10.1002/col.20070>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/col.20070>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/col.20070>.
- [141] B. Ottosson, *A perceptual color space for image processing*, <https://bottosson.github.io/posts/oklab/>, Accessed: 31-05-2025, 2020.
- [142] R. Levien, *An interactive review of oklab*, <https://raphlinus.github.io/color/2021/01/18/oklab-critique.html>, Accessed: 31-05-2025, 2021.
- [143] S.-I. Team, *Api reference, skimage.color*, <https://scikit-image.org/docs/0.25.x/api/skimage.color.html#skimage.color.rgb2xyz>, Accessed: 05-06-2025, 2025.
- [144] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML’17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 214–223. DOI: [10.5555/3305381.3305404](https://doi.org/10.5555/3305381.3305404).
- [145] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, “On the effectiveness of least squares generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2947–2960, 2019. DOI: [10.1109/TPAMI.2018.2872043](https://doi.org/10.1109/TPAMI.2018.2872043).

Chapter 8

APPENDIX

8.1 Summary of the Model Development

Table 8.1: Summary of the Model Development

Section	Details
BCE GAN 5.2	Developed Model BCE GAN is a cGAN model that uses a U-Net generator with a ResNet-50 backbone and a 4-layer PatchGAN discriminator. At each training step, the two networks are updated alternately, with binary cross-entropy as the criterion. Unlike common SOTA approaches, the discriminator receives both the IHC and the respective H&E images.
	Strengths • Highly realistic outputs. • High-frequency detail preservation.
	Limitations • Low accuracy in identifying cases with a 3+ HER2-score. • Inadequate preservation of local H&E structures. • Hallucination of missing connective tissue regions.
	Summary of Results • Empirical evidence shows that SSIM, its variants, and PSNR are biased metrics that favor smoothed outputs in virtual staining. • The use of the discriminator network loss allows for the generation of realistic images with high-frequency details.

Continued on next page

Table 8.1: Summary of the Model Development – continued from previous page

Section	Details
	Conditioning the discriminator on the H&E improves visual quality and the preservation of semantic information; however, it reduces the preservation of local H&E structures and encourages the generator to randomly lose some connective tissue regions, thereby reflecting the nature of the dataset.
CSS GAN 5.3	Developed Model An extension of the BCE GAN is achieved by incorporating the CSS loss into the generator’s loss function. Based on the CSS metric definition [57], this loss function penalizes structural discrepancies between the H&E and IHC images. As a result, it helps preserve the H&E morphology while also regularizing the training process.
	Strengths - Decreased rate of hallucination of missing connective tissue regions. - Improved perceptual quality and pathological consistency.
	Limitations Decreased classical image similarity metrics performance.
	Summary of Results - CSS loss helps reduce the rate of hallucination. - The CSS loss helps regularize the training by encouraging the generator to focus on relevant image features, rather than relying on randomly generating missing tissue regions to deceive the discriminator. This regularization positively impacts the improvement of pathological consistency.
CSSP2P GAN 5.4	Developed Model An extension of the CSS GAN is made by incorporating the Pyramid $\mathcal{L}_1$ loss [43]. This loss function penalizes differences between the generated and real IHC images across various octave levels, thereby improving the similarity between the generated images and the ground truth.
	Strengths - Enhanced similarity between generated and real IHC images. - Improved DAB stain representation.
	Limitations Decreased performance in FID and KID, suggesting worse preservation of semantic information.
	Summary of Results Experiments with the CSS loss achieve top performance, further demonstrating its positive regularizing effect.

Continued on next page

Table 8.1: Summary of the Model Development – continued from previous page

Section	Details
	• The inclusion of the SoftAdapt [136] strategy in training aimed to reduce the GAN’s dependence on hyperparameter tuning by dynamically weighting the contributions of each loss function. However, it proved unsuccessful, resulting in poorer performance, likely due to a slower convergence rate. This outcome led to its exclusion from further use.
OkLAB GAN 5.5	Developed Model The DAB stain identifies the target antigen HER2; therefore, the OkLAB GAN extends the CSS GAN by incorporating a penalization on positively DAB-stained regions using the OkLAB loss to achieve a more accurate pathological representation. This two-fold methodology first increases the relative contribution of pixels within positive DAB-stained regions using masks derived from the DAB channel intensities of the real IHC. Next, it applies the Pyramid $\mathcal{L}_2$ loss in the OkLAB color space to penalize perceptual color differences.
	Strengths • Increased DAB stain representation.
	Limitations • Decreased performance in image similarity, particularly in LPIPS.
	Summary of Results • Masking improves DAB stain representation. • Masking and penalization in the OkLAB color space achieves the best DAB stain representation. • The OkLAB GAN achieves the best representation of the DAB stain compared to previously developed models; however, it shows a decrease in performance across other metrics.

8.2 Architectures

8.2.1 Discriminator Architecture

Layer (type:depth-idx)	Output Shape	Param #
NLayerDiscriminator	[1, 1, 30, 30]	--
└Sequential: 1-1	[1, 512, 31, 31]	--
└Conv2d: 2-1	[1, 64, 511, 511]	3,136
└LeakyReLU: 2-2	[1, 64, 511, 511]	--
└Downsample: 2-3	[1, 64, 256, 256]	--
└ReflectionPad2d: 3-1	[1, 64, 513, 513]	--
└Conv2d: 2-4	[1, 128, 255, 255]	131,072
└BatchNorm2d: 2-5	[1, 128, 255, 255]	256
└LeakyReLU: 2-6	[1, 128, 255, 255]	--
└Downsample: 2-7	[1, 128, 128, 128]	--
└ReflectionPad2d: 3-2	[1, 128, 257, 257]	--
└Conv2d: 2-8	[1, 256, 127, 127]	524,288
└BatchNorm2d: 2-9	[1, 256, 127, 127]	512
└LeakyReLU: 2-10	[1, 256, 127, 127]	--
└Downsample: 2-11	[1, 256, 64, 64]	--
└ReflectionPad2d: 3-3	[1, 256, 129, 129]	--
└Conv2d: 2-12	[1, 512, 63, 63]	2,097,152
└BatchNorm2d: 2-13	[1, 512, 63, 63]	1,024
└LeakyReLU: 2-14	[1, 512, 63, 63]	--
└Downsample: 2-15	[1, 512, 32, 32]	--
└ReflectionPad2d: 3-4	[1, 512, 65, 65]	--
└Conv2d: 2-16	[1, 512, 31, 31]	4,194,304
└BatchNorm2d: 2-17	[1, 512, 31, 31]	1,024
└LeakyReLU: 2-18	[1, 512, 31, 31]	--
└Conv2d: 1-2	[1, 1, 30, 30]	8,193
Total params: 6,960,961		
Trainable params: 6,960,961		
Non-trainable params: 0		
Total mult-adds (Units.GIGABYTES): 3.39		
Input size (MB): 3.15		
Forward/backward pass size (MB): 373.32		
Params size (MB): 27.84		
Estimated Total Size (MB): 404.31		

Figure 8.1: Summary of the discriminator architecture.

8.2.2 Generator Architecture

Layer (type:depth-idx)	Output Shape	Param #
ResUNet	[1, 3, 512, 512]	--
└ResNetEncoder: 1-1	[1, 2048, 32, 32]	--
└ResNet: 2-1	--	2,049,000
└Conv2d: 3-1	[1, 64, 512, 512]	9,408
└InstanceNorm2d: 3-2	[1, 64, 512, 512]	--
└ReLU: 3-3	[1, 64, 512, 512]	--
└MaxPool2d: 3-4	[1, 64, 256, 256]	--
└Sequential: 3-5	[1, 256, 256, 256]	--
└Bottleneck: 4-1	[1, 256, 256, 256]	73,728
└Bottleneck: 4-2	[1, 256, 256, 256]	69,632
└Bottleneck: 4-3	[1, 256, 256, 256]	69,632
└Sequential: 3-6	[1, 512, 128, 128]	--
└Bottleneck: 4-4	[1, 512, 128, 128]	376,832
└Bottleneck: 4-5	[1, 512, 128, 128]	278,528
└Bottleneck: 4-6	[1, 512, 128, 128]	278,528
└Bottleneck: 4-7	[1, 512, 128, 128]	278,528
└Sequential: 3-7	[1, 1024, 64, 64]	--
└Bottleneck: 4-8	[1, 1024, 64, 64]	1,507,328
└Bottleneck: 4-9	[1, 1024, 64, 64]	1,114,112
└Bottleneck: 4-10	[1, 1024, 64, 64]	1,114,112
└Bottleneck: 4-11	[1, 1024, 64, 64]	1,114,112
└Bottleneck: 4-12	[1, 1024, 64, 64]	1,114,112
└Bottleneck: 4-13	[1, 1024, 64, 64]	1,114,112
└Sequential: 3-8	[1, 2048, 32, 32]	--
└Bottleneck: 4-14	[1, 2048, 32, 32]	6,029,312
└Bottleneck: 4-15	[1, 2048, 32, 32]	4,456,448
└Bottleneck: 4-16	[1, 2048, 32, 32]	4,456,448
└ResNetDecoder: 1-2	[1, 3, 512, 512]	--
└ModuleList: 2-2	--	--
└Conv2d: 3-9	[1, 1024, 64, 64]	1,049,600
└UpconvBlock: 3-10	[1, 1024, 64, 64]	--
└Sequential: 4-17	[1, 1024, 64, 64]	2,098,176
└DoubleConvBlock: 4-18	[1, 1024, 64, 64]	28,313,600
└Conv2d: 3-11	[1, 512, 128, 128]	262,656
└UpconvBlock: 3-12	[1, 512, 128, 128]	--
└Sequential: 4-19	[1, 512, 128, 128]	524,800
└DoubleConvBlock: 4-20	[1, 512, 128, 128]	7,078,912
└Conv2d: 3-13	[1, 256, 256, 256]	65,792
└UpconvBlock: 3-14	[1, 256, 256, 256]	--
└Sequential: 4-21	[1, 256, 256, 256]	131,328
└DoubleConvBlock: 4-22	[1, 256, 256, 256]	1,769,984
└Conv2d: 3-15	[1, 128, 512, 512]	8,320
└UpconvBlock: 3-16	[1, 64, 512, 512]	--
└Sequential: 4-23	[1, 128, 512, 512]	32,896
└DoubleConvBlock: 4-24	[1, 64, 512, 512]	184,448
└OutConvBlock: 3-17	[1, 3, 512, 512]	--
└Conv2d: 4-25	[1, 3, 512, 512]	195
└Sigmoid: 1-3	[1, 3, 512, 512]	--
Total params: 67,024,619		
Trainable params: 67,024,619		
Non-trainable params: 0		
Total mult-adds (Units.GIGABYTES): 531.28		
Input size (MB): 3.15		
Forward/backward pass size (MB): 3609.20		
Params size (MB): 259.90		
Estimated Total Size (MB): 3872.25		

Figure 8.2: Summary of the generator architecture.