Unsupervised Domain Adaptation for Multi-Stain Cell Detection in Breast Cancer with Transformers

Oscar Pina
Universitat Politècnica de Catalunya - BarcelonaTech (UPC)
Barcelona, Spain
oscar.pina@upc.edu

Verónica Vilaplana
Universitat Politècnica de Catalunya - BarcelonaTech (UPC)
Barcelona, Spain
veronica.vilaplana@upc.edu

Abstract

The complexity of digital pathology image analysis arises from histopathological slide variability, including tissue specimen differences and stain variations. While publicly available datasets primarily focus on hematoxylin and eosin (H&E) staining, pathologists often require analysis across multiple stains for comprehensive diagnosis. Deep learning pipelines' implementation in clinical settings is hindered by poor cross-stain generalization, necessitating exhaustive annotations for each stain, which are time-consuming to obtain. In this work, we address these challenges by focusing on breast cancer analysis across four crucial stains: ER, PR, HER2, and Ki-67. Given the necessity of cell-level information for diagnosis, we concentrate on cell detection tasks with detection transformers. Leveraging unsupervised domain adaptation techniques, we bridge the gap between publicly available, annotated H&E datasets and unlabeled data in other stains. We demonstrate the superiority of adversarial feature learning over source-only and image-level generative methods. Our work contributes to improving digital pathology image analysis by enabling robust and efficient computer-aided diagnosis pipelines across multiple stains, thereby improving diagnostic accuracy in practical settings. The code can be found at https://github.com/oscar97pina/stain-celldetr.

1. Introduction

Gaining insights into cellular interactions and the distribution of subpopulations is critical for supporting pathologists in their diagnostic endeavors. Breast cancer, accounting for 30% of new cases among women, necessitates the analysis of tumoral morphological traits through hematoxylin and eosin (H&E), as well as four immunohistochemical (IHC) stains that are crucial for identifying specific biomarkers associated with breast cancer subtypes and assessing tumor aggressiveness: Estrogen Receptor (ER), Progesterone Receptor (PR), Ki-67, and Human Epidermal Growth Factor Receptor 2 (HER2).

Traditionally, this information was obtained through manual cell counting under a microscope. However, with the digitization of histopathological slides, pathologists now employ computers to streamline this process. Currently, given the success of computer vision applications, there is a significant interest in developing computer-aided automatic pipelines that can efficiently extract this essential cell-level information.

The clinical significance of this pursuit has led to a proliferation of datasets tailored for digital pathology image analysis, complete with meticulous annotations at the cell level [2, 5, 8]. These datasets typically comprise images stained with H&E, accompanied by exhaustive annotations delineating and characterizing nuclei based on their respective types. Leveraging these expansive annotated datasets has facilitated the integration of deep learning techniques, known for their need for substantial labeled data, in the creation of cutting-edge cell nuclei identification pipelines, culminating in state-of-the-art performance across numerous benchmarks.

Despite the comprehensive insight H&E staining provides into cell subtype populations and their morphological traits, analysis of IHC stains such as ER, PR, Ki-67, and HER2 becomes imperative for specific diagnoses in breast cancer. However, the availability of annotated datasets for these stains is limited, posing a significant challenge. Manual curation of large and precise datasets for training deep learning models is time-consuming and arduous, presenting a significant obstacle to the development of automated computer-aided diagnosis pipelines beyond H&E staining.

While the main cost lies in obtaining annotations, acquiring unlabeled images from IHC slides is relatively straightforward for medical institutions. Given the giga-scale of Whole Slide Images (WSIs), hundreds of unlabeled patches can be extracted from each slide at a magnification that allows for the visualization and identification of cell nuclei. Therefore, the challenge is how to effectively leverage both the annotated datasets in H&E for cell-level tasks and the unlabeled IHC images to transfer the knowledge gained in H&E to IHC stained images.

In this work, we utilize unsupervised domain adaptation (UDA) techniques to adapt cell detection transformers (Cell-DETR) [11, 13] from H&E to IHC. Specifically, our focus is on the four imperative stains for breast cancer diagnosis: ER, PR, Ki-67, and HER2. Although cell segmentation [5] is the standard approach for obtaining cell-level information, it poses significant computational challenges when processing large histopathological slides, whereas the actual contour of cell nuclei is often ignored [5]. Recently, Cell-DETRs [13] have emerged as a promising alternative, achieving state-of-the-art performance in cell detection and classification, along with significantly faster inference times. We adopt adversarial feature alignment via query token [6], a methodology specifically tailored for UDA with detection transformers.

Our results demonstrate that the proposed methodologies outperform both the source-only approach and an image-level generative model utilizing Cycle-GAN [19]. This work makes a significant contribution to the development of computer-aided diagnosis pipelines based on digital pathology image analysis beyond H&E. By overcoming the common scenario of scarce available annotations in digital pathology and leveraging large amounts of unlabeled data through unsupervised learning, we pave the way for more effective and scalable diagnostic solutions.

2. Related Work

2.1. Domain adaptation for object detection

Adversarial learning, initially developed for unsupervised domain adaptation in image classification tasks [3], has been extended to object detection models such as Fast-RCNN [4]. The approach applies adversarial learning at different levels to exploit the multi-scale nature inherent in object detection tasks [12]. The goal is to encourage the learning of domain-invariant features while effectively solving the task in the annotated source domain.

With the rise of detection transformers (DETR) [1, 20], recent research has focused on developing adaptation techniques tailored specifically for transformer-based detectors, as methods designed for convolutional neural networks may not yield optimal results when applied to transformer architectures. To address this, researchers have introduced domain queries within the attention mechanism, enabling transformers to adapt effectively to different domains while maintaining high performance [6, 17]. In addition to adversarial learning, alternative methodologies such as mean teacher have also been explored for object detection tasks with transformer architectures [18].

2.2. Unsupervised domain adaptation in digital pathology

The high variability seen in histopathological slides, stemming from differences between tissue samples and variations in staining methods, presents a challenge to the applicability of deep learning methods in the field of digital pathology. This challenge is compounded by the general scarcity of fine-grained available annotations. Consequently, there has been a growing interest in implementing unsupervised domain adaptation (UDA) techniques for digital pathology image analysis [7, 14–16].

A common UDA approach involves using image-level techniques with generative models such as Cycle-GAN. These models serve a role akin to color deconvolution, aiming to minimize variations between source and target data. They have proven effective in accounting for differences between H&E cohorts and in translating images from one stain to another. The applications of such models are diverse, spanning from whole slide images (WSI) [14] to cell-level tasks [16].

3. Materials and Methods

In this section, we detail the data used in our experiments and describe the methodologies used, namely detection transformers and adversarial query token for unsupervised domain adaptation.

3.1. Datasets

The datasets utilized in our experiments comprise a combination of publicly available and private data. We focus on domain adaptation from H&E to four distinct IHC stains: ER, PR, HER2, and Ki-67. Our setup encompasses an annotated source dataset in H&E, a collection of unlabeled WSIs for each stain, and a small set of IHC image patches containing cell annotations for evaluation purposes.

Source dataset (H&E) The PanNuke dataset [2] comprises 7,904 patches, each sized 256×256, extracted from WSIs in The Cancer Genome Atlas (TCGA) dataset, representing 19 diverse tissue types at a magnification of 40x.

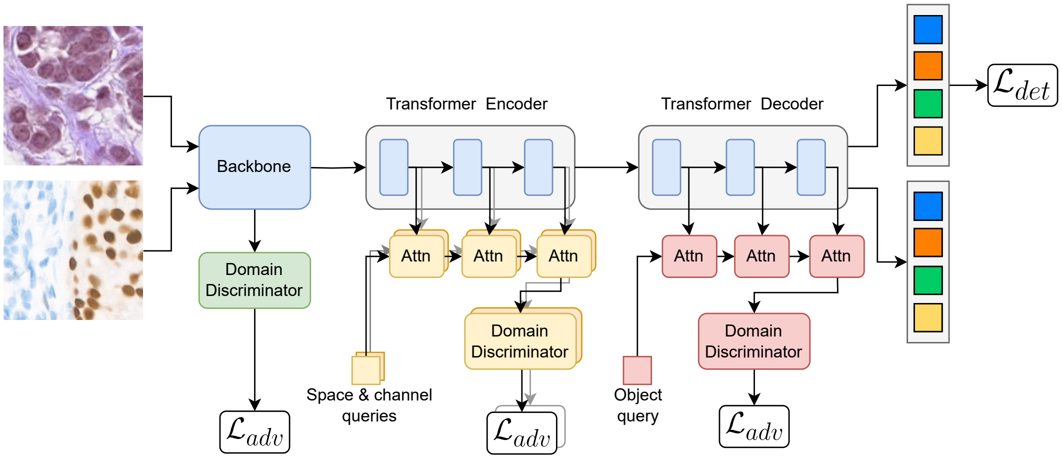


Figure 1. Unsupervised domain adaptation for cell detection with Cell-DETR [13] and AQT [6]
The model receives input from both H&E (source domain) and IHC (target domain) stained images. Adversarial loss occurs in the backbone in a token-wise manner, employing a Feed-Forward Network (FFN) independently for each token. Spatial and channel-wise AQT is applied to the output of the encoder layers, with additional application to the decoder to ensure domain-independent object queries. Supervised detection is conducted solely on the source image output, as annotations are unavailable for the target images.

This dataset encompasses 189,744 labeled nuclei categorized into five clinically significant classes: neoplastic, inflammatory, connective, necrosis, and epithelial.

Target datasets (IHC) The target dataset consists of 159 unannotated IHC-stained WSIs. Regions from the slides were manually selected to create a comprehensive dataset of image patches for each stain. Efforts were made to ensure heterogeneity within the datasets, encompassing stromal and epithelial regions as well as regions exhibiting positivity and negativity for the corresponding receptors of the stains. A summary of the tailored datasets can be found in Table 2.

Evaluation target datasets (IHC) To assess detection accuracy, an additional small set of image patches for each stain with cell annotations was utilized. These annotations include the position of cell nuclei as well as their reaction to the receptor. The statistics of the available annotated datasets can be found in Table 1.

In our setup, both target and source datasets are utilized for training purposes. PanNuke is divided into three folds, with the first fold used for training, the second for validation, and the third for hold-out test data. All patches extracted from the IHC slides are employed for training in unsupervised domain adaptation. Finally, annotated patches in IHC are split into validation and test sets.

Stain	Num. patches	Patch size	Num. nuclei
H&E	$7,904$	256×256	$189,744$
RE	30	1024×1024	$11,968$
RP	38	1024×1024	$14,631$
Ki-67	52	1024×1024	$20,315$
HER2	39	1024×1024	$16,414$

Table 1. Annotated source and target datasets

Stain	Num. WSIs	Num. Patches	Patch size
RE	20	$2,530$	1024×1024
RP	22	$5,392$	1024×1024
Ki-67	21	$3,252$	1024×1024
HER2	30	$3,501$	1024×1024

Table 2. Unannotated target datasets

3.2. Cell Detection with Transformers

Cell segmentation is the standard approach to identify cell nuclei from digital pathology slides. Although the exact cell contour information is usually ignored in downstream applications, the dense dense prediction format offers a solution to challenges like the potential overlap between cells and their small size. This comes at expense of higher computational demands during both training and inference. However, recent advancements in Cell Detection Transformers (Cell-DETR) [13] have demonstrated superior performance in both cell detection and classification, alongside significantly reduced inference times compared to conventional segmentation algorithms.

The Cell-DETR consists of a hierarchical Swin Transformer [9] backbone, and a Deformable Transformer encoder-decoder [1, 20]. For a given input image I, the backbone outputs a 4-level feature pyramid $\{\mathbf{x}^{(l)}\}_{l=1}^{l=4}$ via shifted window attention (sMHA), layer normalization (LN), skip connection and feed-forward networks (FFN):

$$\begin{aligned} \hat{\mathbf{z}}^{(\mathbf{l})} &= \text{S-MHA}\left[\text{LN}\left(\mathbf{z}^{(\mathbf{l-1})}\right)\right] + \mathbf{z}^{(\mathbf{l-1})}, \\ \mathbf{z}^{(\mathbf{l})} &= \text{FFN}\left[\text{LN}\left(\hat{\mathbf{z}}^{(\mathbf{l})}\right)\right] + \hat{\mathbf{z}}^{(\mathbf{l})} \end{aligned} \tag{1}$$

Specifically, the output resolutions of the feature maps are $1/4$, $1/8$, $1/16$ and $1/32$. These output features are enhanced with a deformable transformer encoder based on multi-scale deformable attention. The computation is similar to a standard transformer encoder [1], but the self-attention module is replaced with a deformable attention layer. Multi-scale deformable attention only attends to a subset of keypoints around a reference point rather than attending to the entire image:

$$\begin{aligned} \text{MSDeformAttn}\left(\mathbf{z}_q, \mathbf{p}_q, \{\mathbf{z}^{(l)}\}_{l=1}^{l=4}\right) &= \sum_{m=1}^{M} \mathbf{W}_m \left[\hat{\mathbf{h}}_m\right], \\ \hat{\mathbf{h}}_m = \sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} \cdot \mathbf{W}'_m \mathbf{z}^{(l)}(\phi_{(l)}(\mathbf{p}_q) &+ \Delta\mathbf{p}_{mlqk}) \end{aligned} \tag{2}$$

where $\mathbf{z}_q$ and $\mathbf{p}_q$ are the query vector and position, and *M, L* and *K* are the number of heads, layers in the input tokens and sampling point locations, respectively. The attention scores A_{mlqk}, as well as the attention location offsets $\Delta\mathbf{p}_{mlqk}$ are obtained via a learnable linear projection. Finally, the decoder takes as input the output of the encoder as a memory, and decodes the representations for a set of learnable object queries. The architecture follows a transformer decoder layer, but replaces the cross-attention module with deformable multi-scale self-attention. The representations of the object queries, which have a reference point associated, are updated by attending to a subset of points around that reference. The output object queries are then utilized to predict the bounding box and the class of each object in the image.

Although the target and source image patches have a different size, we adopt the *window detection* procedure defined in [13] and train with crops of size 256×256 while in-device splitting the image into overlapped windows for evaluation and inference.

3.3. Adversarial UDA for Cell-DETRs

Adversarial Query Token (AQT) [6] introduces an attention-based domain discriminator to overcome the limitations of domain adversarial learning for object detection with DETRs. The module takes as input a learnable adversarial token $\mathbf{q}$, as well as a set of multi-layer content tokens $\{\mathbf{z}_u^{(i)}\}$ for every layer $i = 1...N$. Note that there are multiple content tokens for each layer. The query token is iteratively updated with multi-head attention:

$$\mathbf{q}^{(i+1)} = \text{Linear}\left(\text{MHA}\left(\mathbf{q}^{(i)}, \{\mathbf{z}_u^{(i)}\}\right)\right) \tag{3}$$

Then, a domain discriminator D takes the output queries as input to predict whether the content tokens were from the source or the target domains via adversarial learning:

$$\mathcal{L}_{adv} = \sum_{i=1}^{N} -d\, log D(\mathbf{q}^{(i+1)}) - (1-d)\, log(1 - D(\mathbf{q}^{(i+1)})) \tag{4}$$

This mechanism is applied at three different stages, each with its own query token and domain discriminator: (i) space-level $(\mathbf{q}_s, D_s)$, (ii) channel-level $(\mathbf{q}_c, D_c)$ and (ii) object-level $(\mathbf{q}_o, D_o)$. The space-level alignment is based on the output of each deformable encoder layer, so that the content tokens are directly the tokens output by each encoder layer. Given that the deformable attention is local and sparse, the channel-level stage is included in the encoder for a global feature alignment. It also takes place at the output of each encoder layer, but the tokens are constructed as the channels of the layer's output. Finally, the object alignment takes place in the decoder, utilizing the object query representations as content tokens. An additional domain discriminator is applied to the output of the backbone, however, it consists of a FFN-based discriminator applied token-wise, rather than employing the attention-based aggregation and the learnable query vector.

Figure 1 shows a diagram of the architecture. The model is fed with images from both source and target domains. The supervised loss for cell detection is only applied to the source images, as no annotations are available for the target domain, whereas the adversarial learning utilizes the images and representations from both domains. The supervised detection loss guides the model to learn features that are useful for cell detection in H&E, whereas the domain adversarial loss forces those feature to be stain invariant, and consequently also useful for cell detection in other stains.

4. Experiments and results

In this section, we conduct both quantitative and qualitative evaluations to assess the performance of unsupervised domain adaptation for cell detection from H&E to IHC with Cell-DETR.

4.1. Experiments

The experiments conducted in this work utilize a pre-trained Cell-DETR model for cell detection on H&E, specifically

Method	H&E			ER			PR			Ki-67			HER2		
	P_{det}	R_{det}	F_{det}	P_{det}	R_{det}	F_{det}	P_{det}	R_{det}	F_{det}	P_{det}	R_{det}	F_{det}	P_{det}	R_{det}	F_{det}
Source-only Cell-DETR	0.82	0.87	**0.84**	0.73	0.34	0.46	0.77	0.33	0.46	0.76	0.29	0.42	0.13	0.06	0.09
Cycle-GAN + Cell-DETR	0.82	0.87	**0.84**	0.70	0.63	0.66	0.80	0.57	0.66	0.58	0.36	0.45	0.57	0.31	0.40
AQT + Cell-DETR	0.81	0.85	0.83	0.85	0.67	0.75	0.89	0.73	**0.80**	0.87	0.63	0.73	0.53	0.32	0.40
AQT$_{all}$ + Cell-DETR	0.80	0.85	0.83	0.89	0.73	**0.80**	0.88	0.74	**0.80**	0.88	0.64	**0.74**	0.60	0.31	**0.41**

Table 3. Performance unsupervised cell detection

on the PanNuke dataset. The original Cell-DETR architecture has been adapted to focus solely on detection, omitting the classification aspect. Although the PanNuke dataset includes nuclei classification annotations, aligning the available labels for IHC (positive vs negative nuclei) with cell types is not straightforward. Thus, our evaluation primarily centers on assessing detection capabilities. The main modification in the detection-only model is the assignment of a single class (i.e., *nuclei*) to each instance. Therefore, the models assignns a score to each object queries (i.e., the probability of being a cell nuclei), rather than a score for each nuclei type (*neoplastic*, *inflammatory*, *connective*, *necrosis*, and *epithelial*).

Source-only The first baseline we employ is the source-only model, which entails performing cell detection on the various IHC stains using the model trained on H&E. Assessing the performance of the source-only model provides insights into the (dis)similarity between domains and aids in drawing more accurate conclusions regarding the efficacy of unsupervised domain adaptation techniques.

Generative image-level UDA In our experiments, we incorporate an image-based generative unsupervised domain adaptation model, specifically utilizing CycleGAN [19]. By leveraging the unannotated patches extracted from the IHC WSIs alongside the PanNuke dataset, we train independent CycleGAN models for each stain. These models facilitate the generation of corresponding IHC images based on input H&E images and vice versa. The comprehensive approach involves training the generative CycleGAN and converting evaluation IHC images to H&E. Subsequently, the Cell-DETR trained on H&E can be fed with these images to make corresponding predictions.

Adversarial feature-level UDA We combine Cell-DETR and AQT for feature-level domain adaptation, as outlined in Section 3. Our approach incorporates space, channel, and instance feature alignment with the attention-based query token. Additionally, backbone alignment is performed token-wise with a FFN discriminator. We train the *AQT+Cell-DETR* model for each IHC stain with H&E. Furthermore, we train another model by merging all IHC images, denoted as *AQT$_{all}$+Cell-DETR*. By training the model with all available source and target domains (H&E, ER, PR, Ki-67, and HER2), we obtain a multi-stain model for cell detection in histopathological slides.

For a fairer comparison, all models utilize the same Cell-DETR architecture, which has achieved state-of-the-art performance for cell detection and classification in H&E. As a result, the downstream performance comparison is decoupled from the detection architecture itself and specifically focuses on evaluating the contribution of the domain adaptation techniques.

4.2. Results

The numerical results for cell detection *precision* (P_{det}), *recall* (R_{det}), and *F1-Score* (F_{det}) across each target IHC domain and H&E are presented in Table 3. Adversarial feature alignment demonstrates superior performance across all target domains. It is important to note that the metrics for the Cycle-GAN model in the source domain are extracted using the source-only model and original images, rather than artificially generated ones. The performances in H&E for source-only and Cycle-GAN models are slightly better compared to AQT, as the weights have not been modified, whereas the adversarial models have undergone training with the other stains.

Figure 2 shows the ground truth targets and predictions with the distinct models in the IHC stained image patches. Among the methods investigated, adversarial feature-level alignment consistently outperforms both the source-only and generative image-level domain adaptation approaches when employing the same detection architecture.

The challenges inherent in utilizing the source-only model become apparent when considering the differences between H&E and IHC stained slides, resulting in suboptimal detection performance. As depicted in Figure 2b, qualitative analysis of images and model predictions reveals that the source-only model tends to prioritize IHC-positive (brown) nuclei while overlooking negative and stromal cells, which blend with the background due to similar coloration.

Although converting IHC to H&E stain images via Cycle-GAN may seem intuitive, the heterogeneity between

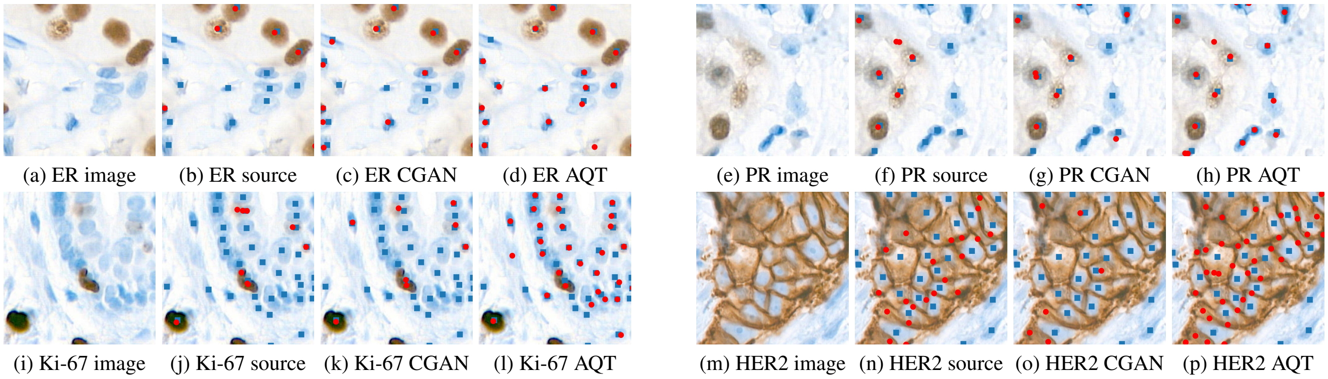


Figure 2. Source-only (source), Cycle-GAN (CGAN) and AQT adapted models' predictions on ER, PR, Ki-67 and HER2 stained patches. Blue squares represent the ground-truth nuclei centroids, while red dots indicate the model predictions. The source-only model effectively detects brown cell nuclei but overlooks negative cells, predicting them as background. Conversely, the AQT-adapted model effectively detects non-positive cell nuclei in ER, PR, and Ki-67 stains. However, the HER2 results highlight morphological differences between stains, as HER2 marks membranes rather than nuclei, resulting in poor performance.

H&E and IHC images, as well as across different stains, poses significant obstacles. The generative model may generate regions interpreted as cell nuclei and incorrectly color some original nuclei as background in H&E. Adding to the complexity, training a Cycle-GAN involves a new model whose performance will influence the overall accuracy, alongside the known challenges of training GANs. Additionally, as the information that can be extracted from IHC and H&E is different, the generative model must approximate or create this information, potentially introducing inconsistencies and damaging the downstream performance.

In contrast, adversarial feature alignment demonstrates superior performance among the methods investigated, emerging as the preferred solution for domain adaptation in cell detection. Despite variations in stains, the morphological traits of cell nuclei remain similar, allowing the model to focus solely on morphological information for cell detection in H&E. This knowledge can then be effectively transferred to the target domains. As illustrated in Figure 2, compared to the source-only model, the adversarial feature alignment approach ensures that even negative (blue) cells are detected, despite their nuclei exhibiting similar coloration to the background. Notably, the best performance across stains is achieved when all target domains are combined. This can be attributed to the increased training data, albeit unannotated, and the model's exposure to greater domain differences, enhancing its generalization capabilities.

While significant improvements over the source-only model are observed across all stains, the performance on the HER2 stain remains poor. This can be attributed to the differing staining information; while the hematoxylin channel, ER, PR, and Ki-67 stains highlight cell nuclei, HER2 focuses on cell membranes, resulting in substantial morphological differences between stains that challenge knowledge transferability. As shown in Figure 2n and Figure 2p, both the source-only and AQT models tend to detect the colored membranes of the cells as nuclei, while disregarding the actual nuclei, which blend with the background due to similar coloration. The adversarial feature alignment approach appears ineffective in addressing these morphological differences.

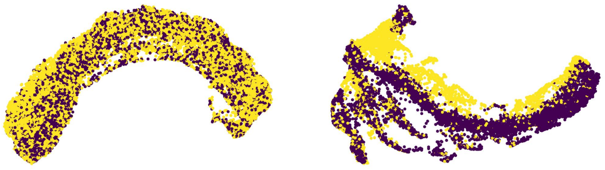
4.3. Implementation Details

The models were implemented in PyTorch and trained using four 16GB GPUs with a batch size of 2, comprising one source and one target domain image per batch. All Cell-DETR models were trained for 100 epochs using the Adam optimizer with a learning rate of 10^{-4}. Hyperparameters for Cell-DETR [13] and AQT [6] related losses were kept consistent with the original authors' code, except for the weight associated with the backbone adversarial loss. We increased this weight to match the weight of the spatial-level adversarial learning, 10^{-1}, as we observed a higher influence of backbone feature alignment for the task at hand.

4.4. Interpretability

4.4.1 Stain invariant nuclei representations

Adversarial feature learning enforces representations to be domain invariant, meaning they are invariant to the stain of the input images. Figure 3 illustrates the 2D UMAP [10] embedding of the instance-level representations for the detected nuclei in H&E and RE stains with the adapted (Figure 3a) and the source only models (Figure 3b). Intuitively, the embedding space of the adapted model lacks information about the input stain, and clusters cannot be discerned to recover the input domain. This suggests that the method has effectively aligned the feature spaces. On the contrary, the



(a) AQT adapted Cell-DETR (b) Source-only Cell-DETR

Figure 3. UMAP of source (black) and target (yellow) domain object queries.

output space of the source-only model exhibits clear grouping based on the input domain, indicating the presence of domain-specific information.

4.4.2 Where are AQTs attending at?

The domain discrimination of both the backbone and space-level adversarial token can be visualized in the image. For the backbone discrimination, which consists of a token-wise FFN, we construct a heatmap representing the probability of each token being part of the target domain. As for the space-level discrimination, adversarial query tokens are employed, allowing us to visualize the attention maps of the tokens to distinct regions of the image. These maps provide a measure of relevance for the target domain prediction.

In practice, we observe that token-wise backbone domain discrimination results in nearly uniform prediction maps. However, we found that adversarial tokens focus on distinct parts of the image after each encoder layer. Specifically, the first and second layers tend to focus on wide, complementary parts of the image, whereas subsequent layers become increasingly tailored to specific image regions.

5. Discussion

The ability of deep learning models to detect cell nuclei across various histopathological stains is crucial for comprehensive cancer diagnosis and treatment planning. However, the scarcity of annotated data beyond commonly available stains like H&E poses a significant challenge. Pathologists rely on multiple IHC stains for cancer subtyping and prognosis, necessitating models capable of robust performance across stains. Manual curation of annotated datasets for each stain is impractical, hindering the development of comprehensive computer-aided diagnosis pipelines. Unsupervised domain adaptation (UDA) offers a promising solution by bridging the gap between abundant H&E data and scarce labeled data in other stains, enabling the development of generalized models for digital pathology analysis.

The findings of this study shed light on the effectiveness of UDA techniques in addressing the challenges posed by the variability in staining patterns across different histopathological stains, particularly in the context of cell detection tasks relevant to breast cancer diagnosis.

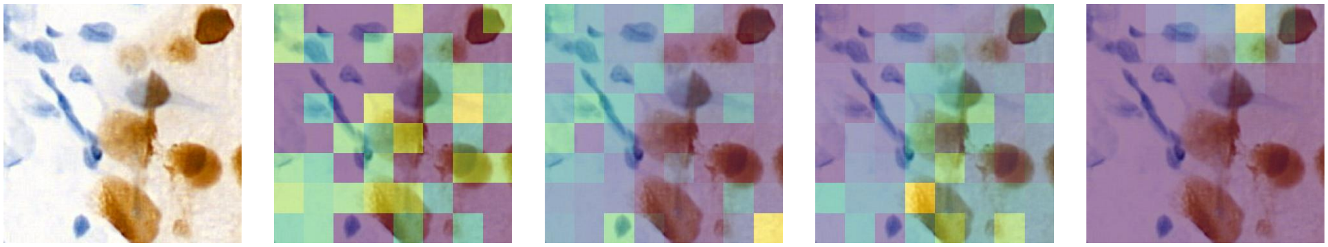
The source-only model, while capable of achieving reasonable performance in detecting cell nuclei in H&E stained images, demonstrates limited generalizability when applied to immunohistochemistry (IHC) stains. This limitation is attributed to the significant differences in staining characteristics between H&E and IHC. The application of adversarial feature alignment demonstrates notable improvements over the source-only model. By leveraging domain adversarial training, the model learns to focus solely on morphological features relevant to cell detection, thereby mitigating the impact of staining variations across different domains. This is evident in Figure 2d, where the adversarial feature alignment model successfully detects negative (blue) cells that were overlooked by the source-only model.

Moreover, the incorporation of all target domains in the training data further enhances the model's generalization capabilities, as evidenced by the superior performance achieved when combining ER, PR, HER2, and Ki-67 stains. This highlights the importance of leveraging diverse unlabeled data sources in UDA tasks, even in the absence of annotated labels, to improve the robustness of deep learning models across different staining modalities.

Despite these advancements, challenges persist, particularly in stains with markedly different staining characteristics such as HER2. The morphological differences between stains present significant obstacles for knowledge transferability, and the limitations of current adversarial feature alignment approaches in addressing these discrepancies underscore the complexity of adapting models to diverse staining patterns.

In future work, it will be crucial to address the challenges observed in adapting the model to the HER2 stain, where significant morphological differences pose obstacles to knowledge transferability. Investigating methods specifically tailored to handle the unique characteristics of HER2 staining could lead to improved performance in this domain. Additionally, while this work focuses on cell detection, incorporating classification tasks, such as identifying positive and negative nuclei, would further enhance the utility of the developed pipelines for comprehensive diagnosis. Exploring approaches to integrate classification alongside detection in a unified framework could be a promising direction for future research.

In summary, this study demonstrates the potential of unsupervised domain adaptation techniques, particularly adversarial feature alignment, in improving the generalizability of deep learning models for cell detection tasks across diverse histopathological stains. By leveraging unlabeled data and domain adversarial training, we pave the way for more robust and efficient computer-aided diagnosis pipelines in digital pathology, ultimately enhancing diag-



(a) ER input image (b) Backbone token (c) Space AQT ($l = 1$) (d) Space AQT ($l = 2$) (e) Space AQT ($l = 4$)

Figure 4. Backbone-level domain discrimination and AQT space-level attention maps.
The backbone domain discrimination is applied token-wise, resulting in nearly uniform probability maps. Adversarial query token focuses on distinct part of the image at each layer.

nostic accuracy and patient outcomes.

6. Conclusions

In this work we have demonstrated the efficacy of unsupervised domain adaptation techniques, particularly adversarial feature alignment, for cell detection across diverse stains in digital pathology. Focusing on breast cancer and crucial stains (ER, PR, HER2, and Ki-67), we bridge the gap between annotated H&E datasets and unlabeled data in other stains, showcasing the potential of domain adaptation in enhancing cancer diagnosis pipelines. Our findings underscore the importance of robust deep learning models capable of generalizing across stains, offering promising avenues for improving cancer diagnosis and treatment.

7. Acknowledgements

This work has been supported by the Spanish Research Agency (AEI) under project PID2020-116907RB-I00 of the call MCIN/ AEI /10.13039/501100011033 and the FI-AGAUR grant funded by Direcció General de Recerca (DGR) of Departament de Recerca i Universitats (REU) of the Generalitat de Catalunya.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 4

[2] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020. 1, 2

[3] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[5] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, page 101563, 2019. 1, 2

[6] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *IJCAI*, pages 972–979, 2022. 2, 3, 4, 6

[7] Ansh Kapil, Armin Meier, Keith Steele, Marlon Rebelatto, Katharina Nekolla, Alexander Haragan, Abraham Silva, Aleksandra Zuraw, Craig Barker, Marietta L Scott, et al. Domain adaptation-based deep learning for automated tumor cell (tc) scoring and survival analysis on pd-l1 stained tissue images. *IEEE Transactions on Medical Imaging*, 40(9): 2513–2523, 2021. 2

[8] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019. 1

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[10] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 6

[11] Ahmad Obeid, Taslim Mahbub, Sajid Javed, Jorge Dias, and Naoufel Werghi. Nucdetr: End-to-end transformer for nucleus detection in histopathology images. In *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pages 47–57. Springer, 2022. 2

[12] Poojan Oza, Vishwanath A Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[13] Oscar Pina, Eduard Dorca, and Veronica Vilaplana. Cell-DETR: Efficient cell detection and classification in WSIs with transformers. In *Submitted to Medical Imaging with Deep Learning*, 2024. under review. 2, 3, 4, 6

[14] Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Unsupervised domain adaptation for classification of histopathology whole-slide images. *Frontiers in bioengineering and biotechnology*, 7:102, 2019. 2

[15] Maximilian E Tschuchnig, Gertie J Oostingh, and Michael Gadermayr. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Patterns*, 1(6), 2020.

[16] Shidan Wang, Ruichen Rong, Zifan Gu, Junya Fujimoto, Xiaowei Zhan, Yang Xie, and Guanghua Xiao. Unsupervised domain adaptation for nuclei segmentation: adapting from hematoxylin & eosin stained slides to immunohistochemistry stained slides using a curriculum approach. *Computer Methods and Programs in Biomedicine*, 241:107768, 2023. 2

[17] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021. 2

[18] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mttrans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, pages 629–645. Springer, 2022. 2

[19] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 5

[20] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 4