# A Comparative Study of Unsupervised Adversarial Domain Adaptation Strategies in Multiple-instance Learning Frameworks for Digital Pathology

**Javier Garcia-Baroja**[1] JAVIER.G.BAROJA@GMAIL.COM
**Samaneh Abbasi-Sureshjani**[2] SAMANEH.ABBASI@ROCHE.COM
**Nazim Shaikh**[2] NAZIM.SHAIKH@ROCHE.COM
**Konstanty Korski**[2] KONSTANTY.KORSKI@ROCHE.COM
[1] *Swiss Federal Institute of Technology,Rämistrasse 101, 8092 Zürich*
[2] *F. Hoffmann-La Roche AG, Grenzacherstrasse 124, 4070 Basel, Switzerland*

**Editors:** Accepted for publication at MIDL 2023

## Abstract

Performance of state-of-the-art deep learning methods is often impacted when evaluated on data coming from unseen acquisition settings, hindering their approval by the regulatory agencies and incorporation to the clinic. In recent years, several techniques have been proposed for improving the generalizability of models by using the target data and their corresponding ground truths. Some of those approaches have been adopted in histopathology, however they either focus on pixel-level predictions or simple tile level classification tasks with or without target labels. In this work, we investigate adversarial strategies in weakly supervised learning frameworks in digital pathology domain without access to the target labels, thereby strengthening the generalizability to unlabeled target domains. We evaluate several strategies on Camelyon dataset for metastatic tumor detection tasks and show that some methods can improve the average F1-score over 10% for the target domain.

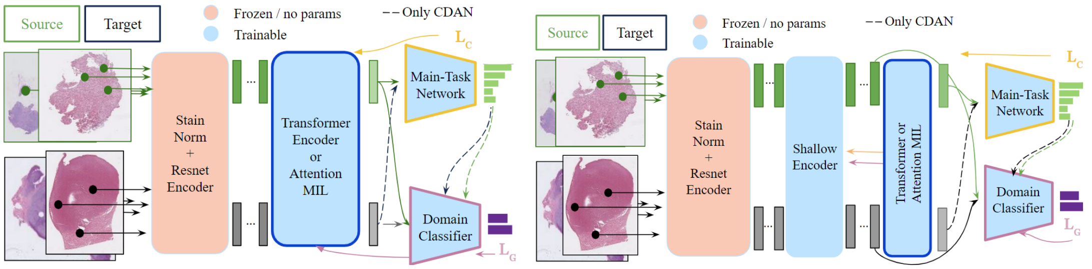
## 1. Introduction

Despite the popularity of computer aided diagnosis tools for Digital Pathology (DP), widespread use of these algorithms is hampered by the inherent variation between images of diverse origin (Howard et al., 2021) (in staining, thickness, patient demographics, etc.) known as *domain shift*. Therefore, a strategy that enables us to build more generalizable models is desired. Among different methods, Marini et al. (2022) propose a Domain Adversarial Neural Network (DANN) (Ganin et al., 2015) to tackle stain heterogeneity with an understanding of domain rooted in Whole Slide Image (WSI) coloring. The Conditional Domain Adversarial Network (CDAN) proposed by Long et al. (2017) also facilitates domain alignment by utilizing the discriminative information offered by main-task classifier predictions.

This paper focuses on Unsupervised Domain Adaptation (UDA) and addresses domain shift caused by scanner variations in weakly supervised metastatic tumor detection. We explore DANN, CDAN, and the impact of changing the position of the domain discriminator in attention MIL and TransMIL (Shao et al., 2021) networks.

## 2. Methods and Experimentations

We propose to adapt MIL models by combining the discriminators ($\mathcal{G}$) in DANN and CDAN at two locations: 1) after a shallow encoder ($loc_i$), where $\mathcal{G}$ would receive *instance-level* samples. In this way, the feature alignment between domains will be provided by a shallow encoder that maps the embeddings from the frozen encoder into an overlapping latent space; 2) $\mathcal{G}$ is positioned after the embedding aggregation step ($loc_b$), that is, after the attention mechanism in attention-MIL or mean pooling of patch tokens after the last transformer layer in TransMIL. This ensures domain alignment on the *aggregated instances* that are forwarded to the final slide-level classifier. The adapted MIL pipelines for DANN and CDAN are depicted in Figure 1($a$) and 1($b$), illustrating each integration location.



($a$) ($b$)

Figure 1: Overview of the two UDA approaches. a) shows $loc_b$ and b) $loc_i$.

### 2.1. Experimental Setup

The experiments used a combination of the publicly available Camelyon16 and Camelyon17 datasets (Litjens et al., 2018), which contains 1399 WSI of lymph nodes (metastatic and healthy) stained with Hematoxylin and Eosin, from three different scanners, five hospitals.

Scanner 1 (S1) data (from three different medical centers digitized by the same scanner) was used as the source ($N_s = 544$), while Scanner 2 (S2) data (from two hospitals) and Scanner 3 (S3) data (from one facility) comprised the target dataset ($N_{t_2} = 253$ and $N_{t_3} = 100$), on which the model is to improve its generalizability. The source dataset was split into 5 non-overlapping subsets (each 20%) stratified by medical center and tumor label, for 5-fold Cross-Validation (CV). The UDA training had S2 as target and was evaluated on S3.

$10,000$ tiles at $40\times$ magnification were extracted from each WSI. Image patches were stain normalized by Tellez et al. (2019) to account for stain variation from multiple acquisition centers. A ResNet-50 (He et al., 2016) pre-trained on DP images via BYOL self-supervised learning strategy (Grill et al., 2020; Abbasi-Sureshjani et al., 2021) was used to extract intermediate features and the backbone weights remained frozen for computational efficiency. The attention network in attention MIL had 5 fully connected layers, followed by batch normalization and dropout (p=0.5). The transformer used Nystrom approximation (Xiong et al., 2021) for Self-Attention (SA) with 3 layers and 8 heads in each multi-head SA block. The main-task classifier had 2 fully connected layers. The discriminator had 3 layers in CDAN and 2 in DANN. ReLU is used as activation function. The Adam optimizer with a learning rate of $10^{-3}$ was used. The adversarial contribution to the updates of the network parameters preceding the domain discriminator $\mathcal{G}$ was defined as $\lambda = \frac{2}{1+\exp(-\gamma p)} - 1$, with $p \in (0, 1]$ the relative progress of the training.

The UDA strategies were compared with three baselines: *source only* (S1), *target only* (S2), and *balanced data* (combining source and target data, with new stratification). Target labels were only used to settle the baseline and evaluating the adapted models (never for UDA). Model selection relied on macro-average validation F1-score. The performance on S3 was obtained using the model with the closest average F1-score to S2 in the CV experiments.

## 3. Results and Conclusion

The results in Table 1 show UDA improves the performance on the target, indicating higher retention of domain agnostic features. The F1-score gap for S2 is reduced by at least 10%, while the models still generalize to S3. The more severe gap for S2 than S3, beyond persisting staining differences after stain normalization, could be attributed to the slide thickness as explained by our pathologist. Moreover, the attention heatmaps showed the effectiveness of UDA to reduce bias towards light coloring that may be irrelevant to the network outcome.

No UDA method clearly outperformed the rest, possibly due to limited bandwidth for domain alignment with a frozen backbone. More complex methods with additional hyperparameters may be required. UDA led to a slight decline in source domain performance that can be addressed by continual learning methods such as Bándi et al. (2022).

## Acknowledgments

The authors thank the Roche Personalized Healthcare Digital Pathology Program for sponsoring project resourcing. The authors declare the following competing interests: S.A., N.S. and K.K. are Roche employees and J.G. was employed by Roche at the time of this work.

Table 1: Results of different UDA strategies on CAMELYON dataset, 5-fold CV [a]

| **Method** | | **S1** | **S1 → S2**[b] | | **S3** |
|---|---|---|---|---|---|
| **MIL** | **Experiment** | **avg. F1** | **avg. F1** | $\mathbf{F1}_{S1} - \mathbf{F1}_{S2}(\downarrow)$ | **avg. F1** |
| Attention MIL | *Balanced data* | 85.2(1.0) | 80.4(3.2) | - | 87.8(2.2) |
| | *Source only* | 86.9(3.6) | 68.4(2.1) | 19.2(3.9) | 83.0 |
| | DANN @ $loc_i$ | 83.6(2.6) | 74.0(6.3) | 9.6(2.0) | 82.6 |
| | CDAN @ $loc_b$ | 86.2(6.2) | 80.2(3.0) | 6.0(3.2) | 85.2 |
| | DANN @ $loc_b$ | **87.6(4.3)** | **81.4(3.2)** | 6.2(1.1) | 86.0 |
| | CDAN @ $loc_i$ | 83.0(3.2) | 79.4(3.0) | 3.6(1.2) | 83.4 |
| | *Target only (S2)* | - | - | 88.4(6.1) | - |
| TransMIL | *Balanced data* | 88.5 | 85.1 | - | 92.9 |
| | *Source only* | 86.3(4.2) | 67.6(7.8) | 18.7(3.6) | 84.0 |
| | DANN @ $loc_b$ | 85.6(3.6) | 79.3(2.2) | 6.3(1.5) | 85.8 |
| | CDAN @ $loc_b$ | 86.4(2.1) | 79.0(2.5) | 7.0(0.4) | 85.0 |
| | *Target only (S2)* | - | 90.5(5.3) | - | - |

[a] Percentages with standard deviation. Best in bold, second underlined; [b] Arrow for adaptation direction.

## References

Samaneh Abbasi-Sureshjani, Anıl Yüce, Simon Schönenberger, Maris Skujevskis, Uwe Schalles, Fabien Gaire, and Konstanty Korski. Molecular subtype prediction for breast cancer using h&e specialized backbone. In *MICCAI Workshop on Computational Pathology*, pages 1–9. PMLR, 2021.

Péter Bándi, Maschenka C. A. Balkenhol, Marcory van Dijk, Bram van Ginneken, Jeroen van der Laak, and Geert J. S. Litjens. Domain adaptation strategies for cancer-independent detection of lymph node metastases. *ArXiv*, abs/2207.06193, 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Advances in Computer Vision and Pattern Recognition*, 17:189–209, 5 2015. ISSN 21916594. doi: 10.48550/arxiv.1505.07818.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

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

Frederick M. Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman, and Alexander T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12(1), July 2021. doi: 10.1038/s41467-021-24698-1.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), May 2018. doi: 10.1093/gigascience/giy065.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation, 2017.

Niccoì O Marini, Manfredo Atzori, Sebastian Otálora, Stephane Marchand-Maillet, and Henning Müller. He-adversarial network: a convolutional neural network to learn stain-invariant features through hematoxylin  eosin regression. 1 2022. doi: 10.48550/arxiv.2201.06329.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and yongbing zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and

J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021.

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.101544.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nystöm-based algorithm for approximating self-attention. *Proc. Conf. AAAI Artif. Intell.*, 35(16):14138–14148, May 2021.