Out-of-Distribution Classification for Histopathology Patches

Dumb and Dumber Team

Amine Maazizi AMINE.MAAZIZI@ENSTA.FR
MVA Master's Program, ENS Paris-Saclay

Nemanja Vujadinović NEMANJA.VUJADINOVIC@ENS-PARIS-SACLAY.FR
MVA Master's Program, ENS Paris-Saclay

Abstract

We study out-of-distribution (OOD) generalization for binary histopathology patch classification, where acquisition-center differences create substantial domain shift between training and evaluation data. Our main study is built around a pretrained DINOv2 backbone, on top of which we analyze linear probing, higher input resolution, stain-aware augmentation, parameter-efficient fine-tuning, color alignment with Reinhard normalization, and feature alignment with CORAL. This DINOv2-based progression reaches 0.96907 accuracy on the public Kaggle leaderboard, substantially improving over frozen-backbone baselines. As a late comparative experiment, we also evaluate UNI, a pathology-specific foundation model, under the same training protocol; a frozen UNI probing model reaches 0.97107 public accuracy, outperforming our best DINOv2 variant. Overall, we find that larger inputs and domain-aware augmentations consistently improve robustness, while explicit alignment methods such as Reinhard and CORAL do not help in this setting. Code is available at: https://github.com/amine-maazizi/histopathology-OOD-classification.

Keywords: histopathology, out-of-distribution generalization, domain shift, medical imaging, deep learning

1. Introduction

Histopathology image analysis is a core task in computational pathology, with applications ranging from diagnosis to treatment planning. In practice, models trained on data from a limited number of medical centers often fail to generalize when deployed on slides acquired elsewhere (Stacke et al., 2019). Differences in staining protocols, tissue preparation, scanner characteristics, and local preprocessing pipelines induce substantial distribution shifts that alter image appearance without necessarily changing the underlying tissue morphology. Convolutional architectures trained end-to-end on such data tend to overfit to center-specific color statistics, motivating the use of strong pretrained Vision Transformers whose self-supervised representations are known to capture more transferable structural features (Oquab et al., 2024).

The present challenge focuses on binary classification of histopathology patches under precisely this setting. Training images come from three centers, while validation and test patches originate from unseen centers. The goal is therefore not only to learn a discriminative classifier, but to do so in a way that remains robust under clinically realistic domain shift.

Our approach is centered on a pretrained visual backbone and a sequence of controlled modifications designed to improve cross-center generalization. The main body of the study

focuses on DINOv2 (Oquab et al., 2024), which we test to measure how far a strong natural-image foundation model can be pushed in histopathology without domain-specific pretraining. Starting from frozen linear probing, we progressively evaluate higher input resolution, stain-aware augmentation, and lightweight adaptation with LoRA, while also testing alignment-based alternatives such as Reinhard normalization and CORAL. As a late comparative experiment, we additionally evaluate UNI (Chen et al., 2023), a foundation model pretrained specifically for computational pathology, to measure the remaining gap between generic visual pretraining and pathology-specific representation learning. This progression allows us to identify which interventions genuinely improve robustness under OOD shift.

2. Architecture and Methodological Components

2.1. Motivation and Intuition

The main difficulty of this challenge is acquisition-center shift rather than class imbalance or label noise. In histopathology, such shifts primarily affect low-level image statistics, especially stain appearance and contrast, while many diagnostically relevant structures remain morphologically stable (Stacke et al., 2019). A robust solution should therefore preserve tissue structure while reducing reliance on center-specific appearance cues.

This motivates the use of a strong pretrained backbone. Rather than training from scratch on a relatively small multi-center dataset, we build on DINOv2, a self-supervised Vision Transformer with transferable visual representations (Oquab et al., 2024). Our strategy is to begin with linear probing as a controlled baseline, then introduce only targeted modifications aimed at likely failure modes of the task: insufficient spatial detail, sensitivity to stain variation, and over-adaptation to source centers. In practice, our experiments show that higher resolution, realistic stain-aware augmentation, and lightweight adaptation are more effective than explicit alignment objectives such as Reinhard normalization or CORAL.

An additional question emerged during the project: how much of the remaining performance gap is due to the training strategy itself, and how much is due to the choice of backbone? We deliberately postponed this comparison until late in the study because our primary goal was first to understand how far a natural-image foundation model could be pushed under severe histopathology domain shift. We therefore introduced UNI (Chen et al., 2023) only after establishing a strong DINOv2-based pipeline, using it as a comparative pathology-specific backbone under the same experimental protocol.

2.2. Architecture Design

Our primary architecture uses a pretrained Vision Transformer backbone followed by a classification head for binary tumor-versus-normal prediction. Most of the study is built on DINOv2, which serves as our main generic visual backbone, while a late comparative experiment replaces it with UNI, a pathology-specific foundation model (Chen et al., 2023). In both cases, we first evaluate a frozen-backbone probing setup in which only the classification head is trained, providing a controlled reference point for measuring feature transfer under OOD shift.

To adapt the representation while limiting overfitting, we then use Low-Rank Adaptation (LoRA) (Hu et al., 2021). LoRA updates only a small subset of trainable parameters while

keeping the rest of the pretrained backbone fixed, yielding a more controlled alternative to full fine-tuning. This is particularly appealing in the OOD setting, where aggressive end-to-end adaptation can distort pretrained features and increase sensitivity to source-domain artifacts (Kumar et al., 2022). Additional implementation details are provided in the appendix.

2.3. Methodological Choices and Dataset Handling

Our methodological choices were guided by a simple question: which interventions reduce sensitivity to acquisition-center shift while preserving the morphological information needed for tumor classification? We therefore focused on modifications targeting two likely weaknesses of the frozen baseline: loss of fine spatial detail and over-reliance on center-specific color statistics.

First, we increased the input resolution from 98×98 to 224×224, which improved the public leaderboard score from 0.90560 to 0.92539. This indicates that preserving finer histological detail is beneficial even without adapting the backbone. Second, we evaluated generic RGB color jitter, but this did not improve performance, suggesting that naive photometric perturbations do not adequately model inter-center stain variation. We therefore introduced stain-aware augmentation in HED space (Tellez et al., 2018; Gao et al., 2024), which proved substantially more effective. Exact augmentation mechanics and parameter settings are deferred to the appendix.

All experiments follow the same dataset protocol : training uses only source centers, while evaluation is performed on held-out centers provided by the challenge. Keeping this protocol fixed across experiments ensures that performance differences can be attributed to methodological changes rather than to changes in the evaluation setting

To isolate the effect of the pretrained representation itself, the late UNI experiment was run under the same data protocol, input resolution, and augmentation framework as the strongest DINOv2 variants. This makes the comparison primarily a backbone comparison rather than a change in training setup.

2.4. Exploratory Experiments and Discarded Alternatives

We also evaluated two explicit alignment strategies that were ultimately not retained in the final system: Reinhard normalization for color-distribution alignment and CORAL for feature-statistics alignment. A short summary is given here, while technical details and theoretical background are deferred to Appendix B.

Reinhard normalization was tested as a preprocessing step on top of our strongest augmentation-based model. Although it achieved a competitive public leaderboard score of 0.96627, it did not improve over the final model and increased preprocessing and training cost. We therefore did not retain it in the final pipeline.

We also evaluated CORAL as an explicit domain-alignment strategy. In our setting, it consistently underperformed simpler augmentation-based approaches: the initial class-conditional CORAL model dropped to 0.74998 public accuracy, and a tuned version recovered only to 0.91459. These results suggest that, for this task, global feature-statistics alignment is less effective than preserving morphological detail and improving robustness through realistic stain-aware augmentation.

3. Model Tuning and Comparison

3.1. Preprocessing and Training Procedure

All experiments share the same basic preprocessing pipeline. Images are resized before being passed to the backbone, with 224×224 selected as the default resolution after outperforming the smaller baseline setting, and inputs are normalized using the backbone-specific evaluation pipeline. Augmentations are applied only during training. Our final DINOv2 pipeline combines simple geometric transformations with stain-aware color augmentation, while exact augmentation strengths are reported in the appendix.

Training follows a probing-to-adaptation progression. For DINOv2, we first train a head on top of a frozen backbone and then use this solution to initialize LoRA-based adaptation. The late UNI experiment follows the same logic: we first train a frozen UNI probing model and then use its head checkpoint to initialize a second LoRA stage. At the time of writing, the UNI LoRA stage was still training.

3.2. Ablation Study and Component Comparison

We evaluate the pipeline cumulatively to identify which changes improve OOD robustness. The results in Table 1 support three main conclusions. First, increasing input resolution from 98×98 to 224×224 improves the frozen probing baseline, confirming that finer histological detail is beneficial at the feature-extraction stage. Second, generic RGB color jitter does not help, indicating that simple photometric perturbations are a poor proxy for acquisition-center variation, whereas stain-aware augmentation provides a clear gain. Third, lightweight LoRA adaptation produces the strongest DINOv2 results, reaching 0.96907 public accuracy and 0.9553 internal validation accuracy. Late in the project, we evaluated UNI (Chen et al., 2023) under the same protocol; a frozen UNI probing model reaches 0.97107 public and 0.9651 internal validation accuracy, outperforming our best DINOv2 configuration and suggesting that pathology-specific pretraining provides an additional advantage beyond what augmentation and adaptation can recover. Reinhard normalization remains competitive without improving over the strongest models, while class-conditional CORAL degrades performance substantially. Validation accuracy was not systematically logged across all runs; model selection relied on early-stopping checkpoints, and we therefore report internal validation scores only for the two strongest variants. Overall, robustness is better improved by preserving morphology and modeling realistic stain variability than by enforcing explicit global alignment.

3.3. Hyperparameters and Performance

Our strongest DINOv2 model combines a DINOv2 backbone with 224×224 inputs, stain-aware augmentation, and parameter-efficient LoRA adaptation, reaching 0.96907 on the public Kaggle leaderboard. In a late comparative experiment, a frozen UNI probing model trained under the same protocol further improves the public score to 0.97107. For reproducibility, we report the final architecture and optimization hyperparameters together with the corresponding validation and test results in a compact configuration summary. Implementation-level details, including augmentation strengths, LoRA configuration, and UNI-specific settings, are provided in the Appendix A.

Table 1: Ablation on the public Kaggle leaderboard. LP denotes linear probing. Internal val. accuracy is reported only where available.

Variant	**Test acc.**	**Val. acc.**
DINOv2-LP (98×98)	0.90560	—
DINOv2-LP (224×224)	0.92539	—
DINOv2-LP + RGB jitter	0.90419	—
DINOv2-LP + stain augmentation	0.93358	—
DINOv2-LoRA + stain aug. ($r = 8$)	0.96420	—
DINOv2-LoRA + stain aug. ($r = 4$)	0.96484	—
DINOv2-LoRA + stain aug. + Reinhard	0.96627	—
DINOv2-LoRA + stain aug. (best DINOv2)	0.96907	0.9553
DINOv2-LoRA + class-cond. CORAL	*0.74998*	—
DINOv2-LoRA + class-cond. CORAL (tuned)	0.91459	—
UNI-LP (224×224) + stain aug. + Reinhard	**0.97107**	**0.9651**

3.4. Internal Validation Procedure

Following the validation principles discussed in class, our goal was to use an internal protocol that remains unbiased and reflects the true challenge setting, namely generalization to unseen acquisition centers. For this reason, we do not perform random source-only splits, which would likely overestimate performance by validating on data too similar to the training set. Instead, we use the official challenge validation set as our internal validation benchmark, since it comes from centers not seen during training and is therefore a more faithful proxy for the final test distribution.

Concretely, all models are trained only on `train.h5` and validated only on `val.h5`; `test.h5` is never used during development. This separation is explicit in the code through distinct dataset objects and dataloaders for train and validation. Training-time augmentations are applied only to the training set, while validation uses the deterministic evaluation preprocessing pipeline. For frozen-backbone models, precomputed features are also generated separately for each split, preventing leakage. For CORAL-based experiments, center identifiers are used only to form balanced training batches and are never used in validation.

Model selection is performed by early stopping with patience 10 and checkpointing the best validation model. Since Kaggle is evaluated with accuracy, validation accuracy is also our main internal selection metric.

Acknowledgments

This work was conducted as part of the MVA DLMI 2026 data challenge.

References

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, Mane Williams, Anurag Vaidya, Sharifa Sahai, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Walt Williams, Long Phi Le, Georg Gerber, and Faisal Mahmood. A general-purpose self-supervised model for computational pathology, 2023. URL https://arxiv.org/abs/2308.15474.

Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations, 2024. URL https://arxiv.org/abs/2302.11861.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022. URL https://arxiv.org/abs/2202.10054.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21:34–41, 10 2001. doi: 10.1109/38.946629.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A closer look at domain shift for deep learning in histopathology. *CoRR*, abs/1909.11575, 2019. URL http://arxiv.org/abs/1909.11575.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation, 2016. URL https://arxiv.org/abs/1607.01719.

David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Whole-slide mitosis detection in

h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136, 2018. doi: 10.1109/TMI.2018.2820199.

Appendix A. Additional Experimental Details

This appendix reports implementation details omitted from the main text for brevity. Unless stated otherwise, the final DINOv2-LoRA model is trained at an input resolution of 224×224 with batch size 16 for up to 100 epochs, using early stopping with patience 10. Optimization uses Adam with cosine annealing, a learning rate of 10^{-3} for the classification head, and 4×10^{-4} for LoRA-adapted backbone parameters.

For parameter-efficient adaptation, we use LoRA with rank 4 and $\alpha = 1.0$, injected into the attention projection layers of the DINOv2 backbone (query, key, value, and output projection). The final augmentation pipeline combines random horizontal and vertical flips, rotations by multiples of 90°, RGB color jitter, and stain-aware color augmentation. For stain augmentation, we use a perturbation strength of $\sigma = 0.1$.

Additional exploratory details, including CORAL formulation and tuning choices, checkpoint-selection rules, and implementation issues encountered during preliminary experiments, are also reported in this appendix where relevant.

Appendix B. Discarded Alignment Strategies

This appendix summarizes the two alignment-based approaches that we explored but did not retain in the final system: Reinhard normalization and CORAL. These experiments were informative for understanding which classes of interventions were less suitable for the challenge setting.

B.1. Reinhard Normalization

Reinhard normalization was evaluated as a color-alignment baseline. Unlike stain-aware augmentation, which increases training-time variability to improve robustness, Reinhard normalization enforces a common color distribution across images by matching color statistics to those of a reference image in CIELAB space (Reinhard et al., 2001). In our implementation, the reference image was selected from the training set and the normalization was applied uniformly to training, validation, and test images.

We integrated this preprocessing step into the strongest augmentation-based model. The resulting public Kaggle score was 0.96627, which was competitive but remained below the final model. In addition, this normalization step increased preprocessing overhead and training time. We therefore did not pursue it further. One possible explanation is that the chosen reference image was not fully representative of the dataset distribution, making the alignment suboptimal.

B.2. CORAL

We also evaluated CORAL as a feature-alignment strategy (Sun and Saenko, 2016). In the original CORAL formulation, the problem is viewed through the lens of unsupervised domain adaptation with shared label space and a mainly covariate-shift assumption,

$$p_S(Y \mid X) = p_T(Y \mid X), \qquad p_S(X) \neq p_T(X),$$

and adaptation is performed by aligning second-order statistics between source and target features. In its deep variant, this is implemented by adding a covariance-matching penalty

to the classification objective,

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \left\| \text{Cov}(\phi(X_S)) - \text{Cov}(\phi(X_T)) \right\|_F^2,$$

where $\phi(\cdot)$ denotes the learned feature representation (Sun and Saenko, 2016).

This hypothesis does not appear to match our setting well. Although the task shares the same label space across centers, the dominant domain gap is better characterized as a structured center-induced appearance shift, closer to

$$p_S(X \mid Y) \neq p_T(X \mid Y),$$

due to stain and scanner variability across institutions (Stacke et al., 2019). Under this type of shift, a single global covariance-alignment objective may be too coarse: it can reduce domain discrepancy in aggregate while still distorting class-discriminative structure.