CENTER-PRIMED MULTIPLE INSTANCE LEARNING FOR REDUCING SHORTCUTS IN CROSS-CENTER WHOLE-SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised multiple instance learning (MIL) requires only slide-level labels with low annotation labor costs and is widely used for whole-slide image (WSI) classification. However, inter-hospital differences in staining, scanning, and patient composition may cause models to rely on center-related cues that are spuriously correlated with disease labels, leading to reduced cross-center classification performance. To meet this challenge, this study proposes CenPrime-MIL, which uses fixed center Primes and zero–real Prime dual-path learning with a shared task head to establish an auxiliary pathway for center-related information that can be removed at inference. A Prime-difference center identification constraint directs center-related changes into this auxiliary pathway, while class-conditional center adversarial regularization suppresses center-predictive information in bag-level representations while also preserving pathological discriminative information. Evaluation on three multicenter datasets of breast cancer and skin diseases shows that CenPrime-MIL consistently improves the classification performance of the base MIL models, with average gains of 3.86% in AUC and 3.79% in ACC. Conditional representation analysis further shows that CenPrime-MIL more effectively disentangles center-related encodings ($\Delta\mathrm{AUC} = -10.50\%$) and diagnostic encodings ($\Delta\mathrm{AUC} = +4.88\%$), supporting its ability to enhance pathological discriminative while reducing the influence of center information. Overall, CenPrime-MIL provides an effective approach to improving the reliability of cross-center WSI classification.

1 INTRODUCTION

Histopathological examination provides morphological evidence for tumor diagnosis, subtyping, and risk assessment, while digitized whole-slide images (WSIs) enable computational analysis of tissue architecture and disease-related phenotypes at high resolution (Bera et al., 2019; Niazi et al., 2019). WSI classification identifies diagnostically informative regions to predict slide- or case-level categories, supporting efficient and consistent computer-assisted pathological interpretation (Litjens et al., 2016). However, the billions of pixels in a WSI impose substantial computational and fine-grained annotation costs on direct end-to-end supervised training (Lu et al., 2021). Existing methods therefore partition each slide into patches, represent them as a bag of instances, and use multiple instance learning (MIL) for slide-level prediction (Li et al., 2021; Ilse et al., 2018). Since clinical labels are generally available at the slide or case level rather than for individual patches, weakly supervised MIL has become an important approach to WSI classification (Campanella et al., 2019). Attention-based MIL assigns nonuniform weights to patches, emphasizing diagnostically informative regions when constructing bag-level representations without exhaustive patch annotations (Ilse et al., 2018). By exploiting readily available coarse-grained labels, this learning paradigm reduces annotation costs and supports large-scale pathological modeling and the clinical translation of computational pathology (Campanella et al., 2019; Lu et al., 2021).

High classification performance under a random patient-level split does not necessarily indicate that a model will generalize to new hospitals, scanning workflows, or patient populations (Howard et al., 2021). Multi-center pathology data commonly differ in hospital source, scanner model, resolution, staining batch, laboratory workflow, tissue background, and patient composition (Howard et al., 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

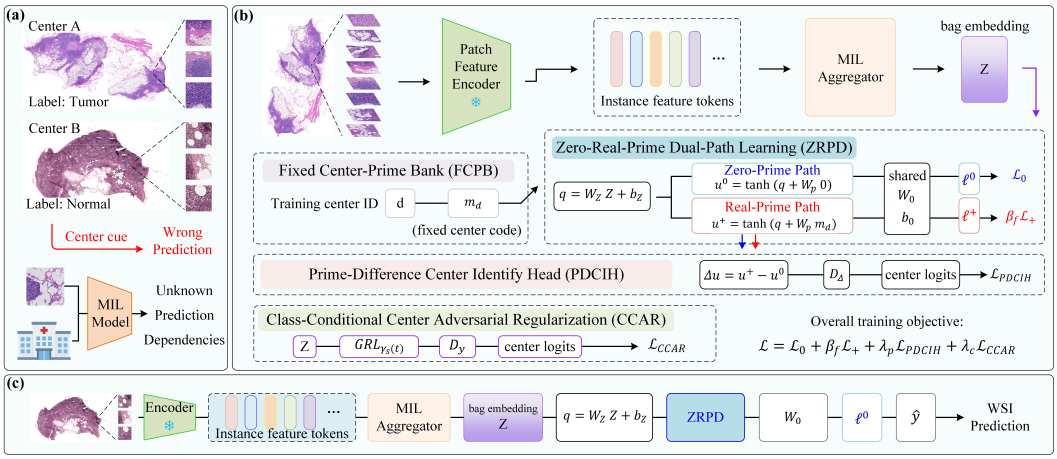

Figure 1: **Motivation and architecture of CenPrime-MIL. (a)** Center-specific shortcuts can undermine cross-center WSI classification. **(b)** Center-guided dual-path learning regularizes the MIL representation during training. **(c)** Inference uses only the zero-Prime path without target-center identity.

When these factors are spuriously correlated with disease labels in the training data, a model may achieve a low training risk by learning center characteristics. For example, if one center contains a higher proportion of positive cases, the model may infer disease labels from color, background texture, or scanning style without adequately recognizing the lesion itself. Such patterns may remain predictive under a random split within the same centers, yet fail when test samples originate from an unseen center and the original statistical associations change (Howard et al., 2021; Geirhos et al., 2020). Domain-generalization research has shown that evaluation protocols, pretraining information, and model-selection procedures can all affect the assessment of genuine out-of-distribution generalization. Cross-center evaluation should therefore strictly isolate the target domain and avoid using test-domain information for model selection (Yu et al., 2024).

Existing approaches to cross-center robustness and debiasing in computational pathology include stain style transfer (Shaban et al., 2019), bag-level confounder modeling (Lin et al., 2023), counterfactual sub-bags and instance selection (Lin et al., 2024), and structured causal representation learning (Song et al., 2025). These strategies provide foundations for reducing spurious correlations but introduce additional modeling requirements. Style transfer uses generators and discriminators (Shaban et al., 2019), while causal approaches incorporate mediators, graph modules, memory mechanisms, or auxiliary objectives (Chen et al., 2024; Cui et al., 2025; Song et al., 2025). More importantly, removing all center information may be inappropriate: center differences can reflect genuine variation in patient composition and disease distributions as well as technical artifacts (Howard et al., 2021). Unconditional domain alignment may therefore compromise task-discriminative information, particularly when label distributions differ across domains (Zhao et al., 2019).

To tackle the cross-center shortcut problem illustrated in Figure 1(a), this study proposes CenPrime-MIL, which establishes an auxiliary pathway for center-associated information that is disabled during inference, rather than unconditionally removing all center information from WSI representations (Figure 1(b)). The Fixed Center-Prime Bank (FCPB) assigns a fixed structured vector to each training center, providing an explicit and controllable center encoding. Zero-Real-Prime Dual-Path Learning (ZRPD) computes the hidden representations of both paths in parallel under a shared task head, using center information to learn center-associated changes while retaining a prediction path independent of center identity. The Prime-Difference Center Identify Head (PDCIH) predicts center identity solely from the difference between the hidden representations of the two paths, directing identifiable center effects toward the Prime auxiliary pathway rather than the shared bag-level representation. Class-Conditional Center Adversarial Regularization (CCAR) further suppresses redundant center-predictive information in the bag-level representation while seeking to preserve genuine pathological variation that may accompany center differences. During inference, only the zero-Prime path is retained and the center-related auxiliary heads are bypassed, eliminating the need for target-center identity (Figure 1(c)). The framework preserves the original patch feature extraction and bag-level

aggregation pipeline and promotes the learning of cross-center-stable pathological representations through lightweight training-time constraints. The complete code has been uploaded.

The results show that CenPrime-MIL consistently improves the performance of the base MIL models on WSI classification tasks for breast cancer (CAMELYON17 and CAMELYON16) and skin diseases (Skin-PEAN), with average gains of 3.86% in AUC and 3.79% in ACC. On CAMELYON17, which comprises data from five centers, CenPrime-MIL achieved an AUC improvement of up to 5.5%, strongly demonstrating its ability to generalize to unseen centers. Compared with the base MIL models, CenPrime-MIL more effectively disentangles center-related encodings ($\Delta\mathrm{AUC} = -10.50\%$) and diagnostic encodings ($\Delta\mathrm{AUC} = +4.88\%$), exhibiting a trend in which reduced linear decodability of center information coexists with enhanced pathological discriminative ability. These findings provide empirical evidence for jointly suppressing center dependence and preserving task-relevant information, while offering a methodological basis for cross-center computational pathology models to learn more reliable pathological representations under distribution shifts.

2 METHODS

2.1 PROBLEM FORMULATION FOR WEAKLY SUPERVISED WHOLE-SLIDE CLASSIFICATION

A WSI contains a large number of spatially and morphologically heterogeneous tissue regions. Because pixel- and patch-level annotations are costly, weakly supervised WSI classification uses slide-level pathological labels alone. For the jth WSI, the tissue region is partitioned into N_j patches and mapped by a feature encoder E to an instance bag

$$\mathcal{B}_j = \{\boldsymbol{h}_{j1}, \ldots, \boldsymbol{h}_{jN_j}\}, \qquad \boldsymbol{h}_{ji} = E(\boldsymbol{x}_{ji}) \in \mathbb{R}^{d_h}, \tag{1}$$

where $\boldsymbol{x}_{ji}$ denotes the ith patch. During training, only the bag-level label $y_j \in \{1, \ldots, C\}$ is observed; no ground-truth instance labels are available. The model learns a permutation-invariant mapping

$$F : \mathcal{B}_j \longmapsto \widehat{\boldsymbol{p}}_j, \qquad \widehat{\boldsymbol{p}}_j \in [0,1]^C, \qquad \sum_{c=1}^{C} \widehat{p}_{jc} = 1, \tag{2}$$

and predicts $\widehat{y}_j = \arg\max_c \widehat{p}_{jc}$.

Each training bag in a multi-center dataset additionally has a center label $d_j \in \{1, \ldots, M\}$ identifying its hospital, scanning workflow, or data source. This label is used only for center-related training objectives. Prediction at test time remains an estimate of $p(y \mid \mathcal{B})$ and does not require the center identity of the test slide.

2.2 BAG-LEVEL REPRESENTATION LEARNING

For the instance bag in Eq. (1), a bag-level aggregator A_θ summarizes the patch representations as

$$\boldsymbol{z}_j = A_\theta(\mathcal{B}_j) \in \mathbb{R}^{d_z}. \tag{3}$$

The aggregator may use attention pooling, attention with instance-level constraints, or token-interaction layers. Under attention pooling (Ilse et al., 2018), for example,

$$e_{ji} = a_\theta(\boldsymbol{h}_{ji}), \qquad \alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{k=1}^{N_j} \exp(e_{jk})}, \qquad \boldsymbol{z}_j = \sum_{i=1}^{N_j} \alpha_{ji} \boldsymbol{h}_{ji}. \tag{4}$$

The output dimension d_z depends on the aggregator. A learned linear map subsequently projects $\boldsymbol{z}_j$ from $\mathbb{R}^{d_z}$ to a common task space $\mathbb{R}^{d_u}$. The patch-encoder output dimension and the bag-representation dimension therefore need not be identical.

2.3 FIXED CENTER-PRIME BANK

For the M centers observed during training, FCPB constructs a fixed simplex equiangular tight frame (ETF) codebook (Wang et al., 2024)

$$\mathcal{P} = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_M\}, \qquad \boldsymbol{m}_d \in \mathbb{R}^r, \tag{5}$$

such that

$$\|\boldsymbol{m}_d\|_2 = 1, \qquad \boldsymbol{m}_d^\top \boldsymbol{m}_{d'} = -\frac{1}{M-1} \quad (d \neq d'). \tag{6}$$

The codebook is obtained by applying QR decomposition to the $(M-1)$-dimensional subspace spanned by a centered identity matrix. Exact satisfaction of Eq. (6) requires $r \geq M-1$; when $r > M-1$, the remaining dimensions are padded with zeros. The codebook remains fixed throughout training. For bag j, the center label indexes $\boldsymbol{m}_{d_j}$, whereas the zero-Prime path receives the same-dimensional null vector $\boldsymbol{m}_0 = \mathbf{0}$. The fixed geometry provides an addressable center code that can be removed at inference and prevents the center vectors from drifting under the task loss.

2.4 Zero-Real-Prime Dual-Path Learning

The bag representation is first mapped by the shared linear layer

$$\boldsymbol{q}_j = \boldsymbol{W}_z \boldsymbol{z}_j + \boldsymbol{b}_z. \tag{7}$$

The null code $\boldsymbol{m}_0$ and the center code $\boldsymbol{m}_{d_j}$ are then projected by the bias-free map $\boldsymbol{W}_p$ into the same hidden space. The two paths compute

$$\boldsymbol{u}_j^0 = \tanh(\boldsymbol{q}_j + \boldsymbol{W}_p \boldsymbol{m}_0) = \tanh(\boldsymbol{q}_j), \qquad \boldsymbol{u}_j^+ = \tanh(\boldsymbol{q}_j + \boldsymbol{W}_p \boldsymbol{m}_{d_j}), \tag{8}$$

$$\boldsymbol{\ell}_j^0 = \boldsymbol{W}_o \boldsymbol{u}_j^0 + \boldsymbol{b}_o, \qquad \boldsymbol{\ell}_j^+ = \boldsymbol{W}_o \boldsymbol{u}_j^+ + \boldsymbol{b}_o. \tag{9}$$

Both paths share $\boldsymbol{W}_z$, $\boldsymbol{W}_p$, and the task classifier $(\boldsymbol{W}_o, \boldsymbol{b}_o)$, and the real-Prime path does not repeat patch encoding or bag aggregation. The model always returns $\boldsymbol{\ell}_j^0$ as its primary task output, so the deployment path receives pathological supervision from the beginning of training. The real-Prime logits $\boldsymbol{\ell}_j^+$ are generated only during training, with their pathological supervision controlled by β_f. When $\beta_f = 0$, $\boldsymbol{u}_j^+$ remains active for the Prime-difference constraint described below.

2.5 Prime-Difference Center Identify Head

The hidden change induced by the real Prime is

$$\Delta \boldsymbol{u}_j = \boldsymbol{u}_j^+ - \boldsymbol{u}_j^0. \tag{10}$$

PDCIH takes $\Delta \boldsymbol{u}_j$ as its sole input and predicts the training center:

$$\boldsymbol{\ell}_j^{\mathrm{PDCIH}} = D_\Delta(\Delta \boldsymbol{u}_j), \qquad \mathcal{L}_{\mathrm{PDCIH}} = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\ell}_j^{\mathrm{PDCIH}}, d_j). \tag{11}$$

The head does not directly observe $\boldsymbol{z}_j$ or $\boldsymbol{m}_{d_j}$. Its gradient requires the Prime projection to produce a center-identifiable change in the shared task space, preventing $\boldsymbol{W}_p \boldsymbol{m}_{d_j}$ from being ignored by the task head. Because $\Delta \boldsymbol{u}_j$ jointly depends on the bag representation, Prime projection, and nonlinear activation, it describes only the hidden change induced by Prime injection and is not interpreted as an independent center representation or a clinical counterfactual.

2.6 Class-Conditional Center Adversarial Regularization

The bag representation $\boldsymbol{z}_j$ is also passed through a gradient reversal layer (GRL) (Ganin et al., 2016) to a class-conditional center head. A linear map D_c is assigned to each pathological class c. Bag j activates only the head D_{y_j} associated with its observed pathological label, provided that the class occurs in at least two training centers. The center prediction is

$$\boldsymbol{\ell}_j^{\mathrm{CCAR}} = D_{y_j}\left(\mathrm{GRL}_{\gamma s(t)}(\boldsymbol{z}_j)\right), \qquad \mathcal{L}_{\mathrm{CCAR}} = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\ell}_j^{\mathrm{CCAR}}, d_j). \tag{12}$$

The GRL is the identity during the forward pass, allowing D_{y_j} to learn center classification normally. During backpropagation, the gradient passed to A_θ is multiplied by $-\gamma s(t)$. The center head thus improves its prediction of d_j, while the aggregator is updated in the opposite direction to reduce center predictability conditional on the pathological label.

2.7 Joint Optimization Objective

The pathological classification losses for the zero-Prime and real-Prime paths are

$$\mathcal{L}_0 = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\ell}_j^0, y_j), \qquad \mathcal{L}_+ = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\ell}_j^+, y_j). \tag{13}$$

Their bag-level task term is

$$\mathcal{L}_{\mathrm{ZRPD}} = \mathcal{L}_0 + \beta_f \mathcal{L}_+, \tag{14}$$

where $\mathcal{L}_0$ is active in every configuration and β_f controls whether and to what extent the real-Prime path receives pathological supervision. FCPB has no trainable parameters and contributes no separate loss.

To prevent label-center combinations with more samples from dominating the two center objectives, we first compute

$$\widetilde{w}_{y,d} = \sqrt{\frac{n_y}{M n_{y,d}}}, \tag{15}$$

where $n_{y,d}$ is the number of training bags with label y from center d, and $n_y = \sum_d n_{y,d}$. The final weight is

$$w_{y,d} = \mathrm{clip}(c\widetilde{w}_{y,d}, 0.5, 2.0)\,, \tag{16}$$

where c is solved numerically so that $N^{-1}\sum_{y,d} n_{y,d} w_{y,d} = 1$. These weights are applied only to PDCIH and CCAR and do not alter the sampling distribution of the pathological classification losses.

The center constraints follow a delayed schedule so that the bag-level task can first establish a stable representation. Let T denote the total number of training epochs, $T_s = \min(T, 20)$ the effective schedule length, and $T_w = \lceil 0.2 T_s \rceil$ the number of warm-up epochs. We define

$$s(t) = \begin{cases} 0, & t \le T_w, \\ \dfrac{2}{1 + \exp\left[-10\frac{t - T_w}{T_s - T_w}\right]} - 1, & T_w < t < T_s, \\ 1, & t \ge T_s. \end{cases} \tag{17}$$

During warm-up, PDCIH contributes no gradient. The CCAR heads continue to minimize their center-classification losses, but the GRL scale passed to the bag-level aggregator is zero. For a single training bag, the objective is

$$\mathcal{L}_{\mathrm{CP}}(t) = \mathcal{L}_{\mathrm{ZRPD}} + w_{y_j,d_j}\lambda_p s(t)\mathcal{L}_{\mathrm{PDCIH}} + w_{y_j,d_j}\lambda_c \mathcal{L}_{\mathrm{CCAR}}, \tag{18}$$

where the GRL scale in Eq. (12) is $\gamma s(t)$. If A_θ also produces an instance-level training loss $\mathcal{L}_{\mathrm{ins}}$, the full objective becomes

$$\mathcal{L} = \rho \mathcal{L}_{\mathrm{CP}} + (1 - \rho)\mathcal{L}_{\mathrm{ins}}, \tag{19}$$

where $\rho \in [0, 1]$ is the bag-level loss weight. For an aggregator without an instance-level constraint, $\mathcal{L} = \mathcal{L}_{\mathrm{CP}}$.

2.8 Zero-Prime Inference

At inference, the same aggregator A_θ feeds only the zero-Prime path:

$$\widehat{\boldsymbol{p}} = \mathrm{softmax}(\boldsymbol{W}_o \tanh(\boldsymbol{W}_z A_\theta(\mathcal{B}) + \boldsymbol{b}_z) + \boldsymbol{b}_o)\,. \tag{20}$$

Neither center labels nor FCPB are accessed; the real-Prime path, D_Δ, and D_y are bypassed. Predictions depend solely on the zero-Prime logits optimized throughout training.

3 Results

3.1 Datasets and Experimental Details

The CAMELYON17 (Bandi et al., 2019) experiments used the 500 WSIs in the official training set, covering five centers. All 100 slides from center 1 were designated as the test set, while the 400 slides from the remaining four centers constituted the training set. This split enables evaluation of model generalization to a center not observed during training: center 1 provides a relatively large number of pixel-level annotations and a comparatively balanced class distribution, making it suitable both for evaluating slide-level classification performance and for lesion-localization quantification and heatmap analysis.

Table 1: Cross-Center WSI Classification

Method	ACC	AUC	Recall	Precision	F1
CAMELYON17					
IBMIL-AB	0.676 ± 0.006	0.662 ± 0.007	0.444 ± 0.006	0.481 ± 0.138	0.388 ± 0.012
IBMIL-CLAM	0.664 ± 0.011	0.704 ± 0.013	0.416 ± 0.014	0.426 ± 0.016	0.396 ± 0.018
IBMIL-Trans	0.664 ± 0.039	0.748 ± 0.030	0.461 ± 0.012	0.445 ± 0.040	0.445 ± 0.022
ABMIL	0.676 ± 0.025	0.723 ± 0.007	0.464 ± 0.037	0.517 ± 0.049	0.465 ± 0.035
ABMIL + CenPrime-MIL	$0.690 \pm 0.007 \uparrow 0.014$	$0.738 \pm 0.018 \uparrow 0.015$	0.435 ± 0.009	0.463 ± 0.016	0.433 ± 0.011
CLAM	0.682 ± 0.008	0.702 ± 0.018	0.438 ± 0.018	0.450 ± 0.020	0.425 ± 0.015
CLAM + CenPrime-MIL	$0.700 \pm 0.016 \uparrow 0.018$	$0.708 \pm 0.006 \uparrow 0.006$	0.449 ± 0.015	0.496 ± 0.018	0.442 ± 0.025
TransMIL	0.690 ± 0.029	0.695 ± 0.047	0.467 ± 0.047	0.534 ± 0.060	0.456 ± 0.071
TransMIL + CenPrime-MIL	$0.694 \pm 0.035 \uparrow 0.004$	$0.750 \pm 0.043 \uparrow 0.055$	0.480 ± 0.042	0.594 ± 0.054	0.503 ± 0.040
CAMELYON16					
IBMIL-AB	0.724 ± 0.024	0.717 ± 0.058	0.678 ± 0.033	0.729 ± 0.033	0.682 ± 0.032
IBMIL-CLAM	0.739 ± 0.041	0.701 ± 0.026	0.692 ± 0.027	0.758 ± 0.074	0.706 ± 0.034
IBMIL-Trans	0.755 ± 0.040	0.729 ± 0.056	0.746 ± 0.021	0.747 ± 0.034	0.741 ± 0.032
ABMIL	0.642 ± 0.118	0.616 ± 0.094	0.500 ± 0.147	0.607 ± 0.227	0.512 ± 0.082
ABMIL + CenPrime-MIL	$0.671 \pm 0.104 \uparrow 0.029$	$0.675 \pm 0.111 \uparrow 0.059$	0.496 ± 0.081	0.611 ± 0.208	0.536 ± 0.104
CLAM	0.744 ± 0.033	0.747 ± 0.062	0.608 ± 0.084	0.685 ± 0.074	0.638 ± 0.038
CLAM + CenPrime-MIL	$0.757 \pm 0.119 \uparrow 0.013$	$0.761 \pm 0.107 \uparrow 0.014$	0.617 ± 0.161	0.776 ± 0.188	0.656 ± 0.118
TransMIL	0.701 ± 0.082	0.689 ± 0.081	0.554 ± 0.070	0.624 ± 0.154	0.583 ± 0.097
TransMIL + CenPrime-MIL	$0.777 \pm 0.022 \uparrow 0.076$	$0.709 \pm 0.032 \uparrow 0.020$	0.529 ± 0.067	0.818 ± 0.082	0.637 ± 0.037
Skin-PEAN					
IBMIL-AB	0.737 ± 0.030	0.917 ± 0.006	0.700 ± 0.024	0.646 ± 0.027	0.645 ± 0.023
IBMIL-CLAM	0.802 ± 0.018	0.946 ± 0.004	0.756 ± 0.014	0.731 ± 0.028	0.735 ± 0.015
IBMIL-Trans	0.748 ± 0.043	0.905 ± 0.011	0.719 ± 0.035	0.653 ± 0.045	0.670 ± 0.046
ABMIL	0.851 ± 0.031	0.971 ± 0.005	0.788 ± 0.053	0.823 ± 0.031	0.794 ± 0.038
ABMIL + CenPrime-MIL	$0.851 \pm 0.026\ 0.000$	$0.981 \pm 0.004 \uparrow 0.010$	0.809 ± 0.041	0.806 ± 0.037	0.798 ± 0.036
CLAM	0.813 ± 0.012	0.950 ± 0.002	0.764 ± 0.012	0.728 ± 0.005	0.740 ± 0.004
CLAM + CenPrime-MIL	$0.826 \pm 0.026 \uparrow 0.013$	$0.955 \pm 0.011 \uparrow 0.005$	0.771 ± 0.048	0.758 ± 0.037	0.755 ± 0.029
TransMIL	0.600 ± 0.126	0.751 ± 0.104	0.398 ± 0.127	0.397 ± 0.174	0.360 ± 0.141
TransMIL + CenPrime-MIL	$0.774 \pm 0.013 \uparrow 0.174$	$0.914 \pm 0.004 \uparrow 0.163$	0.611 ± 0.088	0.647 ± 0.053	0.599 ± 0.107

The CAMELYON16 (Ehteshami Bejnordi et al., 2017) data followed a bidirectional cross-center evaluation protocol. Because CenPrime-MIL requires data from at least two centers during training, approximately 10% of the labeled samples from the target center were included in the training data, whereas the formal test slides were excluded from both model training and weight selection. Across the two directions, the test sets comprised 129 nonoverlapping WSIs, including 48 positive and 81 negative slides. Final reporting aggregates the results across multiple random-seed splits and the two transfer directions.

The Skin-PEAN dataset (Nan et al., 2025) contains five-class WSIs from two centers. A total of 198 slides from the external test cohort of the second center and 1,468 slides from the first center jointly formed a training set of 1,666 WSIs. The 2,431 slides from the first center were used as the validation set, whereas the remaining 1,784 slides from the second center constituted the test set.

All WSIs first underwent foreground-background segmentation and were then divided into nonoverlapping 256×256-pixel patches at $20\times$ magnification. Patch-level features were subsequently extracted using a ResNet-50 (He et al., 2016) image encoder pretrained on ImageNet (Deng et al., 2009). CenPrime-MIL was instantiated with three commonly used bag-level MIL frameworks, namely ABMIL (Ilse et al., 2018), CLAM (Lu et al., 2021), and TransMIL (Shao et al., 2021), to evaluate its applicability to different aggregation mechanisms. The comparison methods included IBMIL-ABMIL, IBMIL-CLAM, and IBMIL-TransMIL (Lin et al., 2023). All models were trained and evaluated using identical data splits and the same five random seeds to ensure fair comparison. For comparison methods with official configurations, the hyperparameters followed the corresponding official settings.

3.2 WSI Classification Performance

To assess whether CenPrime-MIL consistently improves WSI classification across cross-center data settings and bag-level aggregation mechanisms, comparative experiments were conducted on CAMELYON17, CAMELYON16, and Skin-PEAN. All results were obtained using the same five random seeds. As shown in Table 1, CenPrime-MIL achieved consistent overall performance improvements across the three base architectures. Specifically, on CAMELYON17, CenPrime-MIL increased the AUC of TransMIL from 0.695 ± 0.047 to 0.750 ± 0.043 and the Accuracy of CLAM

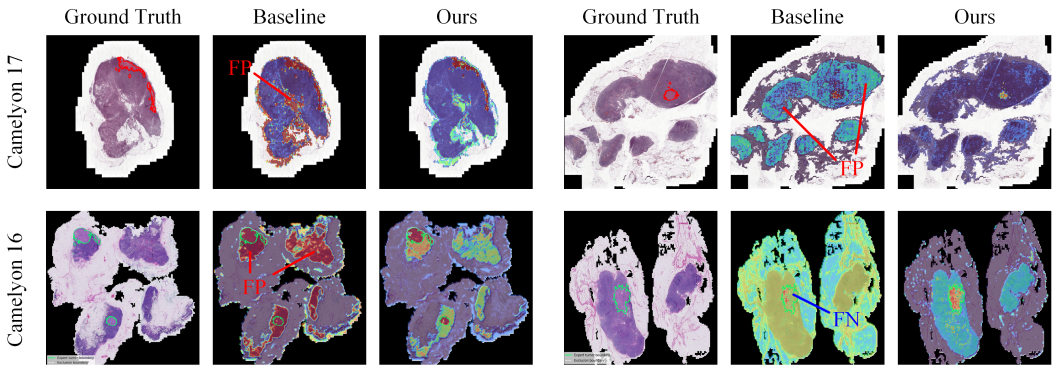

Figure 2: Ground-truth tumor annotations and patch-level heatmaps of the baseline and CenPrime-MIL on CAMELYON17 and CAMELYON16.

from 0.682 ± 0.008 to 0.700 ± 0.016. On CAMELYON16, the AUC gains over ABMIL, CLAM, and TransMIL were 5.9, 1.4, and 2.0 percentage points, respectively, while the corresponding Accuracy gains were 2.9, 1.3, and 7.6 percentage points. On Skin-PEAN, the largest performance gain was observed for TransMIL, whose Accuracy increased from 0.600 ± 0.126 to 0.774 ± 0.013 and whose AUC increased from 0.751 ± 0.104 to 0.914 ± 0.004. Overall, the stable improvements in AUC and Accuracy and the substantial gains over the base models indicate that CenPrime-MIL can enhance the adaptability of bag-level representations to center-specific differences and improve the reliability of cross-center WSI classification across multiple aggregation mechanisms.

3.3 Patch-Level Tumor Recognition

Patch-level tumor recognition directly affects the applicability and interpretability of WSI classification models in medical tasks, because a slide-level class prediction alone cannot establish whether the prediction is based on the actual lesion (Lu et al., 2021; Nan et al., 2025). This experiment was conducted only on CAMELYON16 and CAMELYON17, whose test sets provide complete pixel-level tumor annotations, and evaluated localization performance both qualitatively and quantitatively. As shown in Figure 2, red denotes a higher predicted tumor probability and blue denotes a lower predicted probability; the quantitative evaluation in Table 2 uses WSI-macro patch AUROC, WSI-macro patch AUPRC, and the pointing-game hit rate (Zhang et al., 2016).

Table 2: Patch-Level Tumor Localization

Method	AUROC	AUPRC	Pointing-game hit rate
CAMELYON17			
ABMIL	0.809 ± 0.028	0.111 ± 0.009	0.100 ± 0.105
ABMIL +CP	0.824 ± 0.007	0.132 ± 0.063	0.250 ± 0.000
CLAM	0.777 ± 0.009	0.120 ± 0.046	0.175 ± 0.112
CLAM +CP	$\mathbf{0.890 \pm 0.029}$	$\mathbf{0.366 \pm 0.029}$	$\mathbf{0.450 \pm 0.068}$
TransMIL	0.708 ± 0.111	0.259 ± 0.091	0.325 ± 0.190
TransMIL +CP	0.729 ± 0.133	0.290 ± 0.108	0.400 ± 0.163
CAMELYON16			
ABMIL	0.651 ± 0.098	0.093 ± 0.067	0.179 ± 0.081
ABMIL +CP	0.699 ± 0.100	0.135 ± 0.077	0.304 ± 0.217
CLAM	0.805 ± 0.059	0.278 ± 0.089	0.542 ± 0.130
CLAM +CP	$\mathbf{0.838 \pm 0.041}$	$\mathbf{0.393 \pm 0.114}$	0.621 ± 0.060
TransMIL	0.442 ± 0.001	0.037 ± 0.000	0.000 ± 0.000
TransMIL +CP	0.776 ± 0.106	0.351 ± 0.227	$\mathbf{0.625 \pm 0.127}$

AUROC and AUPRC: WSI macro. +CP: CenPrime-MIL.

The results show that the baseline heatmaps produced numerous high responses in non-tumor regions while responding insufficiently to parts of the annotated lesions, resulting in evident false positives and false negatives; after introducing CenPrime-MIL, the high-response regions were more spatially concentrated within the pixel-level tumor annotations, and abnormal activations in non-tumor regions were reduced. Across both datasets and all three base architectures, CenPrime-MIL improved patch AUROC, patch AUPRC, and the pointing-game hit rate. On CAMELYON17, the three metrics of CLAM increased from 0.777 ± 0.009, 0.120 ± 0.046, and 0.175 ± 0.112 to 0.890 ± 0.029, 0.366 ± 0.029, and 0.450 ± 0.068, respectively. On CAMELYON16, the corresponding metrics of TransMIL increased from 0.442 ± 0.001, 0.037 ± 0.000, and 0.000 ± 0.000 to 0.776 ± 0.106, 0.351 ± 0.227, and 0.625 ± 0.127. These results indicate that the improvement in slide-level classification performance obtained by CenPrime-MIL does not arise solely from exploiting global statistical correlations; the local evidence attended to by the model is also more consistent with the actual tumor regions, thereby enhancing the pathological interpretability of cross-center WSI classification and its credibility for computer-assisted diagnosis.

Table 3: Ablation Studies

Loss components			ABMIL + CenPrime-MIL		CLAM + CenPrime-MIL		TransMIL + CenPrime-MIL	
Prime task	PDCIH	CCAR	ACC	AUC	ACC	AUC	ACC	AUC
CAMELYON17								
✓	✓	✓	**0.690±0.007**	**0.738±0.018**	**0.700±0.016**	**0.708±0.006**	**0.694±0.035**	**0.750±0.043**
×	✓	✓	0.674±0.006	0.725±0.005	0.686±0.021	0.705±0.016	0.670±0.038	0.726±0.038
✓	×	✓	**0.690±0.012**	0.722±0.003	0.690±0.022	0.707±0.014	0.686±0.038	0.710±0.030
✓	✓	×	0.686±0.009	0.718±0.004	0.684±0.013	0.702±0.019	**0.694±0.044**	0.698±0.042
CAMELYON16								
✓	✓	✓	**0.671±0.104**	**0.675±0.111**	**0.757±0.119**	**0.761±0.107**	**0.777±0.022**	**0.709±0.032**
×	✓	✓	0.648±0.058	0.631±0.070	0.754±0.052	0.754±0.044	0.747±0.037	0.669±0.034
✓	×	✓	0.670±0.114	0.638±0.067	0.726±0.164	0.757±0.105	0.715±0.053	0.650±0.034
✓	✓	×	0.670±0.093	0.660±0.046	0.757±0.119	0.760±0.107	0.719±0.062	0.679±0.062
Skin-PEAN								
✓	✓	✓	**0.851±0.026**	**0.981±0.004**	**0.826±0.026**	**0.955±0.011**	**0.774±0.013**	**0.914±0.004**
×	✓	✓	0.838±0.029	0.970±0.006	0.801±0.034	0.944±0.013	0.752±0.044	0.910±0.004
✓	×	✓	0.851±0.044	0.961±0.005	0.823±0.026	0.952±0.011	0.736±0.016	0.885±0.026
✓	✓	×	0.847±0.030	0.970±0.004	0.798±0.061	0.955±0.010	0.747±0.057	0.907±0.028

3.4 Ablation Studies

To assess the contribution of each training constraint in CenPrime-MIL to cross-center classification performance, ablation experiments were conducted on CAMELYON17, CAMELYON16, and Skin-PEAN using ABMIL, CLAM, and TransMIL as the base architectures. With the full configuration as the reference, the real-Prime task loss, the Prime-difference center identification loss (PDCIH), and the class-conditional center adversarial loss (CCAR) were individually removed to examine their respective roles. As shown in Table 3, the full configuration achieved the highest mean AUC across all three datasets and all three base architectures, while also attaining the highest or joint-highest mean ACC; removing any loss resulted in some degree of performance degradation. These results support the contributions of Prime-branch task supervision, Prime-difference center identification, and class-conditional center adversarial regularization to cross-center classification, indicating that retaining the full training objective helps achieve better classification performance across different datasets and aggregation mechanisms.

3.5 Conditional Analysis of Center Leakage and Task Fidelity

The center leakage and task fidelity analysis was designed to examine whether CenPrime-MIL can reduce center-related information while retaining pathological discriminative information, using conditional evaluation to reduce the influence of associations between center identity and class distributions on the assessment. The model representations comprise the slide representation $\boldsymbol{z}$ obtained by MIL aggregation, the hidden representation $\boldsymbol{u}^0$ of the zero-Prime path, and the difference representation $\Delta\boldsymbol{u} = \boldsymbol{u}^+ - \boldsymbol{u}^0$ induced by Prime injection. This analysis focuses on $\boldsymbol{u}^0$, which is used at inference, and compares it with the corresponding hidden representation in the classification head of each base model. With the model weights frozen, center identity was predicted within each pathological class, and pathological class was predicted within each center, with results reported as mean AUC. The former evaluates the decodability of residual center information, whereas the latter evaluates the preservation of pathological discriminative information (Alain & Bengio, 2016).

As shown in Table 4, the two panels summarize center prediction conditioned on class and class prediction conditioned on center, respectively. Both report probe AUC for $\boldsymbol{u}^0$ and retain the class and center identifiers of each dataset. CenPrime-MIL achieved a lower center-prediction AUC in 31 of the 33 label-conditioned comparisons and a higher class-prediction AUC in 21 of the 27 center-conditioned comparisons. CenPrime-MIL thus exhibited an overall reduction in center decodability, while task discriminability was preserved or improved in most comparisons. These results support the ability of CenPrime-MIL to reduce linearly decodable center information under most conditions while retaining pathological discriminative information.

3.6 Robustness to Input Perturbations

To assess the ability of CenPrime-MIL to preserve pathological discriminative ability under input perturbations, WSI classification performance was further compared across different perturbation intensities.

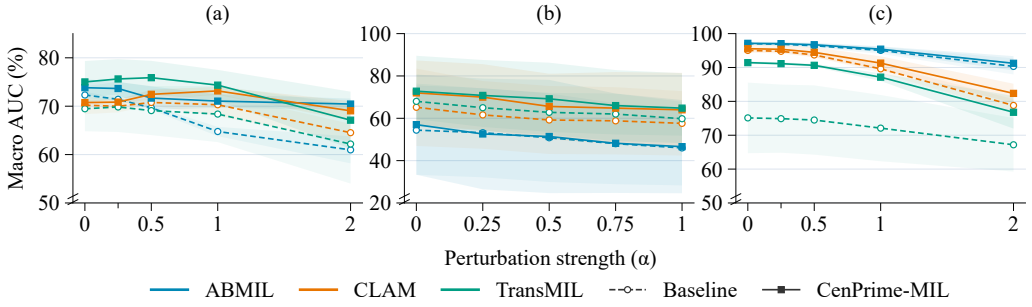

Figure 3: WSI classification under input perturbations on (a) CAMELYON17, (b) CAMELYON16, and (c) Skin-PEAN. CenPrime-MIL maintains relatively high AUC across different perturbation intensities.

Table 4: Conditional Representation Analysis

(a) Center prediction | class

Class	ABMIL		CLAM		TransMIL	
	Base	+CP	Base	+CP	Base	+CP
CAMELYON17						
0	98.77	98.73	98.15	95.38	99.82	72.65
1	86.53	80.12	86.58	79.55	100.00	47.67
2	91.97	89.21	89.20	82.38	99.94	48.62
3	90.69	90.54	86.08	82.92	98.96	61.33
CAMELYON16						
0	97.25	93.55	96.83	95.83	99.83	77.92
1	88.35	88.39	89.20	82.14	99.44	58.75
Skin-PEAN						
0	73.46	71.93	69.54	63.71	62.11	56.06
1	60.86	56.67	58.18	52.96	57.70	57.01
2	74.90	74.90	71.71	69.68	64.53	44.54
3	73.17	72.90	63.20	55.43	68.66	55.08
4	95.22	94.92	85.03	80.14	87.07	84.86

(b) Class prediction | center

Center	ABMIL		CLAM		TransMIL	
	Base	+CP	Base	+CP	Base	+CP
CAMELYON17						
0	95.36	99.93	84.69	88.37	100.00	100.00
1	67.71	65.92	68.04	66.41	69.16	70.62
2	68.82	96.58	71.28	73.21	99.89	100.00
3	94.06	99.41	88.71	92.94	100.00	100.00
4	89.02	99.70	86.29	89.04	100.00	100.00
CAMELYON16						
0	70.58	85.28	84.70	87.34	69.14	88.43
1	68.98	67.37	67.90	69.53	64.54	72.67
Skin-PEAN						
0	95.19	95.59	90.93	92.99	86.51	94.09
1	96.96	97.20	94.15	95.22	75.74	92.17

Mean AUC (%). Base: baseline; +CP: baseline with CenPrime-MIL.

With model weights frozen and no additional training, patches were subjected to 10 repetitions of random H&E staining perturbations of varying intensity ($\alpha = 0$–2), after which features were re-extracted and the models were re-evaluated in terms of AUC.

As shown in Figure 3, CenPrime-MIL achieved higher AUC than the baselines in 35 of the 36 nonzero perturbation conditions, indicating that its performance advantage was maintained under most perturbation conditions. On CAMELYON17 (Figure 3a), CAMELYON16 (Figure 3b), and Skin-PEAN (Figure 3c), CenPrime-MIL consistently maintained a performance advantage over the baselines at the maximum perturbation intensity specified for each dataset. This result indicates that CenPrime-MIL retains strong pathological discriminative ability under the evaluated staining variations.

4 Conclusion

To mitigate reliance on center-specific shortcuts in cross-center WSI classification, this study proposes CenPrime-MIL. Through guidance and constraints derived from center information during training, the framework reduces reliance on center-related cues while preserving pathological discriminative information. Extensive experiments on multiple benchmark datasets demonstrate that CenPrime-MIL generally improves WSI classification performance across different base architectures, supporting its applicability to cross-center classification tasks. Conditional representation analysis further supports its ability to suppress center dependence while preserving task-relevant information. Future work will extend the evaluation to a broader range of independent external cohorts and pathological tasks to further assess the framework's generalization ability and the limits of its applicability.

5 AI Usage Report

AI was used to polish writing in this manuscript.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. doi: 10.48550/arXiv.1610.01644. URL https://arxiv.org/abs/1610.01644.

Peter Bandi, Oscar Geessink, Quirine Manson, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38:550–560, 2019. doi: 10.1109/TMI.2018.2867350. URL https://pubmed.ncbi.nlm.nih.gov/30716025/.

Kaustav Bera, Kurt A. Schalper, David L. Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16:703–715, 2019. doi: 10.1038/s41571-019-0252-y. URL https://pubmed.ncbi.nlm.nih.gov/31399699/.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25:1301–1309, 2019. doi: 10.1038/s41591-019-0508-1. URL https://pubmed.ncbi.nlm.nih.gov/31308507/.

Kaitao Chen, Shiliang Sun, and Jing Zhao. CaMIL: Causal multiple instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1120–1128, 2024. doi: 10.1609/aaai.v38i2.27873. URL https://ojs.aaai.org/index.php/AAAI/article/view/27873.

Xiaoyu Cui, Weixing Chen, and Jiandong Su. A multiscale frequency domain causal framework for enhanced pathological analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6xrDPHhwD3.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318:2199–2210, 2017. doi: 10.1001/jama.2017.14585. URL https://jamanetwork.com/journals/jama/fullarticle/2665774.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://www.jmlr.org/papers/v17/15-239.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://arxiv.org/abs/2004.07780.

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*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

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*, 12:4423, 2021. doi: 10.1038/s41467-021-24698-1. URL https://pubmed.ncbi.nlm.nih.gov/34285218/.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Dual-Stream_Multiple_Instance_Learning_Network_for_Whole_Slide_Image_Classification_CVPR_2021_paper.html.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19830–19839, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Lin_Interventional_Bag_Multi-Instance_Learning_on_Whole-Slide_Pathological_Images_CVPR_2023_paper.html.

Weiping Lin, Zhenfeng Zhuang, Lequan Yu, and Liansheng Wang. Boosting multiple instance learning models for whole slide image classification: A model-agnostic framework based on counterfactual inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3477–3485, 2024. doi: 10.1609/aaai.v38i4.28135. URL https://ojs.aaai.org/index.php/AAAI/article/view/28135.

Geert Litjens, Clara I. Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-van de Kaa, Peter Bult, Bram van Ginneken, and Jeroen van der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, 6:26286, 2016. doi: 10.1038/srep26286. URL https://www.nature.com/articles/srep26286.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w. URL https://arxiv.org/abs/2004.09666.

Tianhang Nan, Song Zheng, Siyuan Qiao, Hao Quan, Xin Gao, Jun Niu, Bin Zheng, Chunfang Guo, Yue Zhang, Xiaoqin Wang, Liping Zhao, Ze Wu, Yaoxing Guo, Xingyu Li, Mingchen Zou, Shuangdi Ning, Yue Zhao, Wei Qian, Hongduo Chen, Ruiqun Qi, Xinghua Gao, and Xiaoyu Cui. Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis. *Nature Communications*, 16:5493, 2025. doi: 10.1038/s41467-025-60307-1. URL https://www.nature.com/articles/s41467-025-60307-1.

Muhammad Khalid Khan Niazi, Anil V. Parwani, and Metin N. Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20:e253–e261, 2019. doi: 10.1016/S1470-2045(19)30154-8. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8711251/.

M. Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. StainGAN: Stain style transfer for digital histological images. In *IEEE 16th International Symposium on Biomedical Imaging*, pp. 953–956, 2019. doi: 10.1109/ISBI.2019.8759152. URL https://arxiv.org/abs/1804.01601.

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 *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

Yiran Song, Yikai Zhang, Shuang Zhou, Guojun Xiong, Xiaofeng Yang, Nian Wang, Fenglong Ma, Rui Zhang, and Mingquan Lin. MeCaMIL: Causality-aware multiple instance learning for fair and interpretable whole slide image diagnosis. *arXiv preprint arXiv:2511.11004*, 2025. doi: 10.48550/arXiv.2511.11004. URL https://arxiv.org/abs/2511.11004.

Yining Wang, Junjie Sun, Chenyue Wang, Mi Zhang, and Min Yang. Navigate beyond shortcuts: Debiased learning through the lens of neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12322–12331, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_Navigate_Beyond_Shortcuts_Debiased_Learning_Through_the_Lens_of_Neural_CVPR_2024_paper.html.

Han Yu, Xingxuan Zhang, Renzhe Xu, Jiashuo Liu, Yue He, and Peng Cui. Rethinking the evaluation protocol of domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21897–21908, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Yu_Rethinking_the_Evaluation_Protocol_of_Domain_Generalization_CVPR_2024_paper.pdf.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Computer Vision – ECCV 2016*, volume 9908, pp. 543–559, 2016. doi: 10.1007/978-3-319-46493-0_33. URL https://link.springer.com/chapter/10.1007/978-3-319-46493-0_33.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 7523–7532, 2019. URL https://proceedings.mlr.press/v97/zhao19a.html.