Advances in Computer, Signals and Systems (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2026.100301
ISSN 2371-8838 Vol. 10 Num. 3

Spatial Evidence Consistency under Stain Perturbation in Whole Slide Classification

Jian Li, Zhaoguo Jiang*

University of Science and Technology Liaoning, Anshan, 114051, Liaoning, China
3761383069@qq.com

Keywords: Whole Slide Imaging; Multiple Instance Learning; Stain Perturbation; Attention Consistency; Histopathology

Abstract: A whole-slide classifier can assign nearly identical probabilities before and after a stain perturbation while changing which tissue regions it weights most heavily. This paper tests whether aligning the attention assigned to the same tissue coordinates improves classification beyond aligning predictions alone. A shared attention-based multiple instance learning model processes original and stain-perturbed tile bags with identical spatial support, and the objective combines supervised classification, prediction consistency and coordinate-aligned attention consistency. In three-seed leave-one-centre-out evaluation, adding the aligned objective raised worst-centre case-level AUROC from 75.8% to 78.2% with a paired gain of 2.4 percentage points, and macro AUROC from 82.0% to 83.9%. Attention Jensen–Shannon divergence fell from 0.182 to 0.121 and the intersection over union of the top 10% evidence sets rose from 0.41 to 0.58, while coordinate-shuffled and entropy-matched controls reached 81.6% and 82.1%. At a matched 10% tissue-area budget the deletion and retention advantages increased by 0.027 and 0.023, so spatial correspondence is a useful training signal for retaining tissue evidence under stain change.

1. Introduction

1.1. Research Background and Significance

A whole-slide classifier can return an almost unchanged probability before and after a stain perturbation while changing the tissue regions on which that probability depends. The slide probability describes the decision, whereas the attention distribution describes how the model pools its evidence. The distinction matters for hematoxylin and eosin (H&E) whole-slide images (WSIs), where preparation, staining and scanning give each laboratory a site-specific image signature that can shift model accuracy when a slide is transferred between centres[1]. In a slide with a small metastasis, a colour-perturbed view may emphasise the stroma instead of the lesion and still return a positive score.

Stain normalisation and colour augmentation reduce the dependence on colour appearance by estimating stain directions and concentrations in optical-density space, and prediction consistency couples the outputs of paired views[2]. In attention-based multiple instance learning (ABMIL) a slide score is computed only after several hundred instance features have been reduced to one

pooled representation, so distinct attention distributions can produce similar scores[3]. A same-coordinate perturbation makes the distinction measurable, because every tile in the first bag has exactly one counterpart in the second. We therefore ask whether attention alignment adds value to a model already trained with prediction consistency.

1.2. Research Status at Home and Abroad

Work on digital histology documents how site signatures arise from staining, preparation and acquisition, and how much accuracy is lost on transfer without correction. Colour normalisation estimates stain directions and concentrations per slide, whereas augmentation samples them during training; both act on appearance without stating which tissue the model should use. Attention consistency supplies that statement and has been enforced through attribution maps compared across paired inputs[4], but those comparisons involve different crops or domains, so the correspondence between attention entries is only approximate.

Evidence rankings are usually assessed by deleting or retaining the highest-ranked inputs. Two equally large regions of a WSI can differ in tissue content, colour and spatial structure, so an unmatched intervention confounds ranking with composition. Matching control regions within the same slide gives the intervention a common budget, and a coordinate-preserving perturbation then allows agreement in attention to be quantified directly.

1.3. Main Research Content and Structure

A shared ABMIL model processes original and perturbed tile bags with identical spatial support, and the objective combines supervised classification, prediction consistency and coordinate-aligned attention consistency. The primary comparison holds the encoder, classifier, training views and supervision fixed while adding attention alignment. Discrimination is measured at the weakest held-out centre, agreement of the selected tissue under stain perturbation, and the influence of the selected tissue under matched deletion and retention; two controls separate coordinate correspondence from attention entropy.

2. Related Work

2.1. Stain Variation and Whole-Slide Learning

Site signatures include stain appearance, specimen preparation and acquisition differences, and they produce systematic errors when a model is deployed away from its training centres. The common remedy separates colour from structure: stain directions and concentrations are estimated in optical-density space and recombined, as a normalisation step or as random augmentation[2]. ABMIL supplies the pooling mechanism used here, learning instance weights from slide-level labels that summarise which tiles support the decision[3].

2.2. Consistency and Tissue Evidence

Prediction agreement across augmented views is the standard consistency signal. Attention agreement constrains where the model looks and has been enforced through attribution maps compared across paired inputs[4]. The difficulty is alignment: two crops or domains do not share a coordinate system, so the comparison must be made between aggregated maps. The paired views used here remove that difficulty, and the comparator already enforces prediction agreement.

Deletion and retention probe a ranking by removing or preserving the inputs it selects.

Comparing selected regions with tissue-matched subsets of the same slide gives the intervention a common area budget and tests whether highly ranked tissue matters more than matched tissue; a ranking that is reproducible across stain views should also dominate the slide score.

3. Methods

3.1. Paired Views of the Same Tissue

For each slide a tissue mask is defined and non-overlapping tiles are extracted at a fixed imaging scale, with tile coordinates stored in the reference frame of the original slide. The second view modifies H&E concentrations in optical-density space and reconstructs the RGB tile: the optical density is written as a stain basis and non-negative concentrations, and the perturbation acts on the concentrations while the basis and the coordinates are preserved[2]. Both views therefore share the tissue mask, tile indices and tile ordering, as shown in Figure 1, and every group uses the same perturbation recipe, set on source training data only.

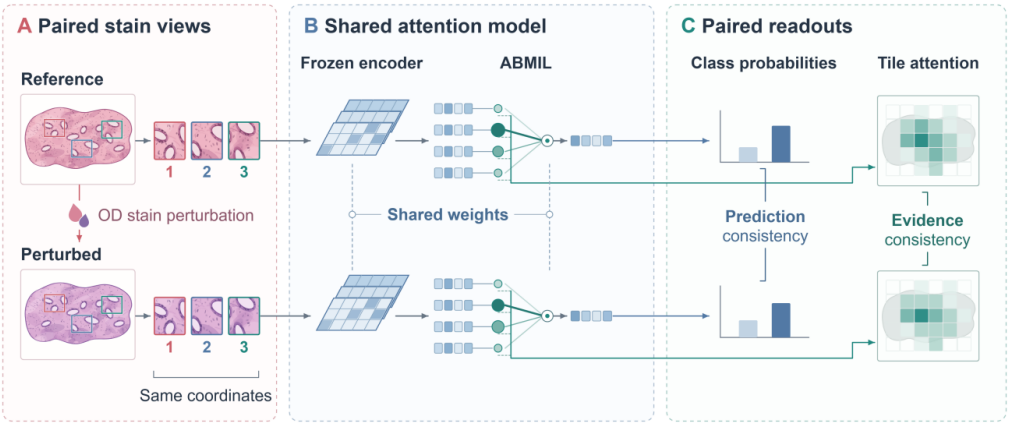

Figure 1 Same-coordinate stain views, shared ABMIL pooling and paired evidence alignment. Tissue tiles, class probabilities and attention maps are schematic.

3.2. Shared Attention Pooling and Training Objective

The same frozen encoder maps each tile to an embedding. For each view the shared ABMIL head converts embeddings into instance features and attention logits, and a softmax produces an attention distribution with one weight per coordinate. The slide representation is the attention-weighted sum of the tile features, and a shared classifier maps it to the two-class probability vector; the paired view yields a second probability vector and attention distribution under the same parameters. Because both distributions share the same support, element i of each refers to the same tissue location, so the loss can separate a change in the amount of attention from a change in where it is placed.

Let y be the slide label. The supervised losses of the two views are averaged, a prediction term compares the two probability vectors, and an evidence term compares the two attention distributions in the original coordinate order; both use the Jensen–Shannon divergence, and two coefficients weight their contributions. At inference a single original view provides the slide score and its attention ranking, so the path remains one encoder and one pooling pass.

3.3. Baselines and Mechanism Controls

B0 is single-view ABMIL and B1 adds the paired view while supervising both predictions; B2

adds prediction consistency and is the main comparator. E1 couples attention across the paired views while retaining both supervised losses, and E2 combines prediction and coordinate-aligned attention consistency. N1 uses the E2 objective with the second attention distribution permuted, preserving the marginal distribution of weights but breaking their correspondence across locations. N2 adds a position-independent entropy penalty to B2, tuned to match the attention entropy of E2.

4. Experimental Design

4.1. Cohort and Centre-Held-Out Evaluation

The evaluation follows the CAMELYON17 design for H&E lymph node WSIs, and the five centres are evaluated in leave-one-centre-out folds[5]. The official case group is the unit of splitting, evaluation and resampling, and both views of every slide stay within one partition; the case score is the maximum of its eligible slide scores. Source sets follow a 70/15/15 allocation by case group, the three seeds are 17, 23 and 41, and each consistency coefficient is chosen from 0.1, 1 and 10 under a common validation budget. The images and annotations are distributed through the public CAMELYON dataset release.

4.2. Discrimination, Stability and Matched Interventions

Case-level AUROC, F1 and accuracy are reported, with macro averages giving each held-out centre equal weight and the worst-centre AUROC taken as the minimum of the centre-specific values. Worst-centre AUROC is the primary endpoint because deployment is decided by the site where the model performs worst: a macro average can stay high while one centre fails, whereas the minimum is a conservative lower bound on unseen-centre performance that assumes nothing about the distribution of centre difficulty or case mix. The minimum is recomputed inside every bootstrap draw, because the identity of the weakest centre can change between resamples. The discrimination criterion for the E2 minus B2 contrast is a gain of at least 0.02 with a paired 95% confidence interval above zero; paired intervals use 2,000 bootstrap replicates that resample cases within centres and apply the same draws to both methods. Selected tissue agreement is measured with attention JSD and with the intersection over union of the top-10% evidence sets, while Shannon entropy describes the concentration of the weights. For the intervention analysis tiles are ranked by full-bag attention and the frozen model is re-evaluated on the modified bag, with twenty control sets per slide matched on tissue area, content, colour and brightness strata, connectivity and boundary distance.

5. Results

5.1. Discrimination at Held-Out Centres

E2 achieved the highest macro AUROC, F1 and accuracy of the seven arms, as shown in Table 1. Relative to B2, macro AUROC rose from 82.0% to 83.9%, F1 from 65.0% to 67.4% and accuracy from 66.9% to 69.1%. At the weakest centre the AUROC was 78.2% against 75.8%, a paired gain of 2.4 percentage points (95% CI 0.3 to 4.6), while the corresponding macro gain was 1.9 points (95% CI 0.4 to 3.4) and ranged from +1.4 to +2.6 points across the three seeds. Worst-centre performance is listed in Table 2. Figure 2 compares the B2 and E2 discrimination and reports
the paired E2 minus B2 effects with their paired 95% bootstrap intervals.

Table 1 Macro case-level performance across held-out centres. Mean ± SD across three seeds (%)

Method	AUROC	F1	Accuracy
Mean pooling	78.4 ± 4.1	60.2 ± 5.3	63.1 ± 4.6
B0 Single-view ABMIL	80.1 ± 3.6	62.5 ± 4.8	64.8 ± 4.2
B1 Paired cross-entropy	81.2 ± 3.2	64.1 ± 4.5	66.0 ± 4.0
B2 Prediction consistency	82.0 ± 3.0	65.0 ± 4.3	66.9 ± 3.8
E1 Aligned evidence only	81.9 ± 3.3	64.8 ± 4.6	66.7 ± 4.1
E2 Prediction and aligned evidence	83.9 ± 2.8	67.4 ± 4.0	69.1 ± 3.6
N1 Coordinate-shuffled evidence	81.6 ± 3.4	64.6 ± 4.7	66.5 ± 4.1
N2 Entropy-matched control	82.1 ± 3.1	65.2 ± 4.4	67.0 ± 3.9

Table 2 Worst-centre case-level performance. Mean ± SD across three seeds (%)

Method	AUROC	F1	Accuracy
Mean pooling	72.1 ± 5.3	54.3 ± 6.1	57.8 ± 5.7
B0 Single-view ABMIL	73.9 ± 4.8	56.0 ± 5.6	59.4 ± 5.2
B1 Paired cross-entropy	75.1 ± 4.4	57.6 ± 5.2	60.9 ± 4.8
B2 Prediction consistency	75.8 ± 4.2	58.4 ± 5.0	61.7 ± 4.6
N1 Coordinate-shuffled evidence	75.3 ± 4.5	57.9 ± 5.3	61.2 ± 4.9
E2 Prediction and aligned evidence	78.2 ± 3.9	61.2 ± 4.6	63.9 ± 4.2

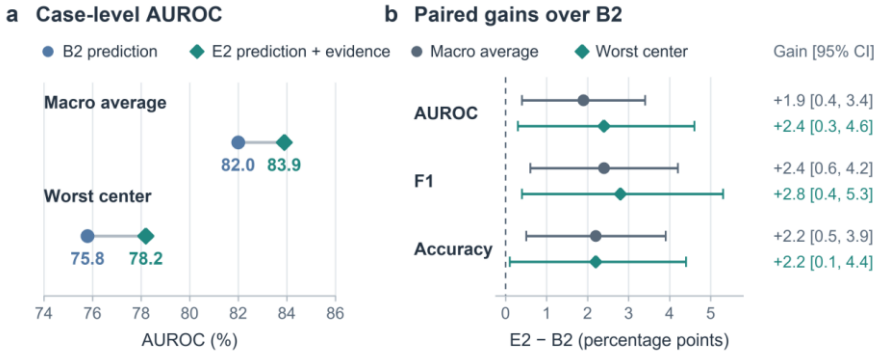

Figure 2 B2 and E2 discrimination and paired E2 minus B2 effects, with paired 95% bootstrap intervals

5.2. Stability of Tissue Evidence

Attention JSD fell from 0.182 with B2 to 0.121 with E2, and the intersection over union of the top-10% evidence sets rose from 0.41 to 0.58, so the highest-ranked tissue overlapped more strongly across the two stain appearances. Entropy fell from 3.42 to 2.87, as listed in Table 3.

Table 3 Evidence stability across original and stain-perturbed views

Method	Attention JSD	Attention entropy	Top-10% IoU
B0 Single-view ABMIL	0.214	3.71	0.35
B1 Paired cross-entropy	0.206	3.66	0.37
B2 Prediction consistency	0.182	3.42	0.41
N1 Coordinate-shuffled evidence	0.177	3.39	0.42
E2 Prediction and aligned evidence	0.121	2.87	0.58

5.3. Fidelity under Matched Interventions

At the 10% tissue-area budget the deletion advantage rose from 0.041 to 0.068 and the retention advantage from 0.038 to 0.061, and the deletion curve area increased from 0.612 to 0.674. Table 4 and Figure 3 show the fixed-budget contrasts; because each selected set is compared with twenty matched controls from the same slide, the contrast reflects the ranking rather than the amount of tissue removed.

Table 4 Matched deletion and retention at 10% tissue area and deletion curve area

Method	Deletion (10%)	Retention (10%)	Curve area
B0 Single-view ABMIL	0.028	0.025	0.581
B1 Paired cross-entropy	0.031	0.027	0.590
B2 Prediction consistency	0.041	0.038	0.612
N1 Coordinate-shuffled evidence	0.039	0.036	0.607
E2 Prediction and aligned evidence	0.068	0.061	0.674

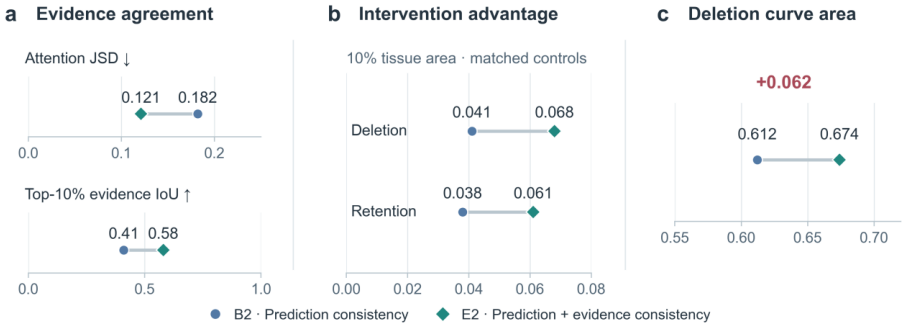

Figure 3 Evidence stability and matched tissue interventions with 10% of tissue area in the fixed-budget contrasts

5.4. Mechanism Controls and Computational Cost

The seven-arm comparison separates prediction agreement, attention agreement and the two control objectives, as listed in Table 1. E1 reached 81.9% macro AUROC and E2 83.9% when the objectives were combined, while the coordinate-shuffled and entropy-matched arms achieved 81.6% and 82.1%, close to the 82.0% of B2. All arms used 12.1 million parameters; E2 trained in 3.3 h against 3.1 h, and single-slide inference took 43 and 41 ms, because the evidence term acts only during training, as shown in Figure 4.

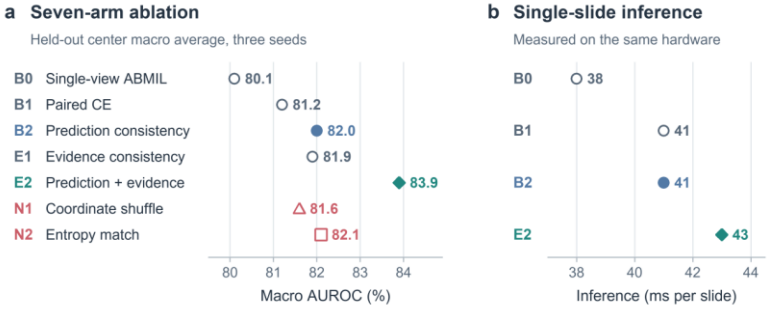

Figure 4 Macro AUROC for the seven training arms and measured single-slide inference time for four arms

6. Discussion

The spatial alignment objective raises discrimination at the lowest-performing held-out centre, and the tissue it ranks highest produces larger matched deletion and retention effects; on the paired stain test it also selects more overlapping regions, which connects agreement in tissue selection with the response of the classifier when that tissue is removed or retained.

The seven-arm comparison clarifies the role of each signal. B1 captures the effect of two supervised stain views, B2 adds agreement at the decision level, and E2 adds agreement at the tissue-coordinate level, where the views provide an exact correspondence. N1 preserves the distribution of weights but disrupts their spatial assignment, and N2 targets only their spread; both cluster around B2.

The objective also brings practical advantages.Because the supervision is obtained from the stain-perturbed view that colour augmentation already produces, no additional annotation or paired data are required, and the same perturbation recipe can be applied to any tile grid without extra acquisition cost.

These properties make the learned ranking easier to audit. The attention map of a slide is spatially reproducible under a change in stain appearance, so the tissue a pathologist is asked to review stays the same when the slide is rescanned or restained. The largest gains also fall at the weakest held-out centre, where cross-centre deployment is most at risk, which is precisely where a ranking that follows tissue rather than colour is most useful.

7. Conclusion

Coordinate-aligned attention consistency improved discrimination over a prediction-consistency baseline in multicentre whole-slide classification. It raised worst-centre AUROC by 2.4 percentage points, increased the spatial overlap of high-attention tissue across stain views, and produced stronger deletion and retention effects at a matched budget without changing the parameter count.

Acknowledgement

Liaoning Province College Students Innovation and Entrepreneurship Training Program.

References

[1] Howard F M ,Dolezal J ,Kochanny S , et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias[J].Nature Communications,2021,12:4423.DOI:10.1038/s41467-021-24698-1.

[2] Macenko M ,Niethammer M ,Marron J S , et al. A method for normalizing histology slides for quantitative analysis[C].IEEE International Symposium on Biomedical Imaging,2009:1107-1110.DOI:10.1109/ISBI.2009.5193250.

[3] Ilse M ,Tomczak J M ,Welling M .Attention-based deep multiple instance learning[C].Proceedings of the 35th International Conference on Machine Learning,2018,80:2127-2136.

[4] Mirzazadeh A ,Dubost F ,Pike M , et al. ATCON: attention consistency for vision models[C].IEEE/CVF Winter Conference on Applications of Computer Vision,2023:1880-1889.DOI:10.1109/WACV56688.2023.00192.

[5] Bandi P ,Geessink O ,Manson Q , et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge[J].IEEE Transactions on Medical Imaging,2019,38(2):550-560. DOI:10.1109/TMI.2018.2867350.