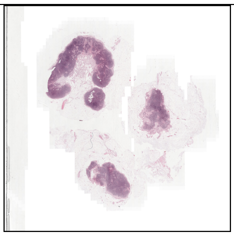
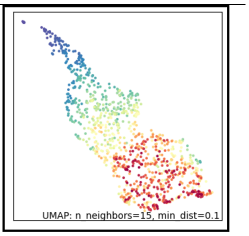
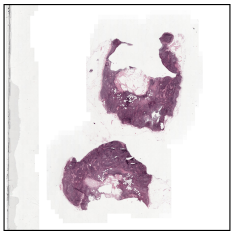
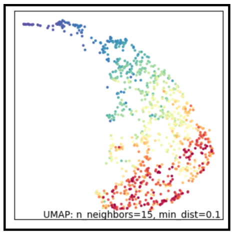
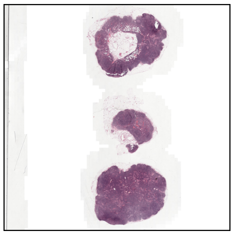
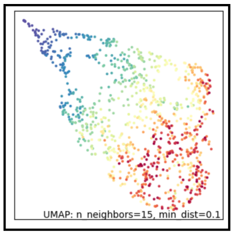
Table 5. Interpretability study on tumor Camelyon16 slide images

Slide Name	Slide Image	Attention Score
tumor_002.tif		
tumor_005.tif		
tumor_012.tif		

11. Supplementary materials

11.1. Loss Function

If we have prototype instances denoted as $P = \{p_1, p_2, \ldots, p_K\}$, we can feed them into the Transformer-MIL and multilayer perceptron (MLP) to predict the label at the end of our Transformer block by:

$$Y_{pred} = MLP(p_1, p_2, \ldots, p_K). \tag{12}$$

Then, we apply a Binary Cross-Entropy (BCE) for calculating the loss function as:

$$\mathcal{L} = BCE(Y_{pred}, Y), \tag{13}$$

where $\mathcal{L}$ calculates the error between the true slide-label and the predicted slide-label.

In our approach, motivated by the need to diversify the minority contexts, we modify the loss function as follows:

$$\mathcal{L}_{total} = w_1 \cdot \mathcal{L}(Y_{pred}, Y) + w_2 \cdot \mathcal{L}(\tilde{Y}_{\text{pred}}, Y), \tag{14}$$

where the first term denotes the loss resulting from the predicted label without Attention-Guided Prototype Mixing, while the second term represents the loss resulting from the predicted label under the Attention-Guided Prototype Mixing condition.

11.2. Interpretation

In Table 5, we present the interpretation of our proposed method. To conduct the interpretation, we used 1024 prototype instances (P) on slide images from the Camelyon16 dataset. The set P was then projected onto a 2D space using the uniform manifold approximation and projection (UMAP) technique [25], which transformed the 512-dimensional representation of each instance into a 2D space. For visualization, the attention score obtained from the self-attention module was normalized using the softmax function. The resulting 2D plot shows the most attentive instances as red points and the least attentive instances as blue points.

11.3. Extended Study on Comparison with SOTA WSI-Augmentation in Section 6.1

Both our method and ReMix [37] aim at diversifying the contexts of normal and tumor slides. However, ReMix sometimes removes the original contexts through interpolate mixing. As presented in Table 6, it is observed that our method achieves improvements of up to 1.02% in AUC and 1.05% in PR AUC that slightly outperforms ReMix.

Table 6. **Extended Comparison with SOTA WSI-Augmentation**.

	AUC			PR AUC		
	Vanilla	ReMix [37]	Ours	Vanilla	ReMix [37]	Ours
TransMIL [28]						
+ BCE	93.00	$93.65_{\uparrow 0.65}$	$93.77_{\uparrow 0.77}$	92.50	$92.70_{\uparrow 0.20}$	$92.95_{\uparrow 0.45}$
+ Bal-BCE [36]	92.75	$93.70_{\uparrow 0.75}$	$93.77_{\uparrow 0.82}$	93.00	$93.10_{\uparrow 0.10}$	$93.45_{\uparrow 0.45}$
GTP [41]						
+ BCE	92.75	$93.70_{\uparrow 0.95}$	$93.77_{\uparrow 1.02}$	93.00	$93.10_{\uparrow 0.10}$	$93.35_{\uparrow 0.45}$
+ Bal-BCE [36]	93.77	$94.00_{\uparrow 0.23}$	$94.20_{\uparrow 0.43}$	92.95	$93.80_{\uparrow 0.85}$	$94.05_{\uparrow 1.05}$

Table 7. **Comparison on SOTA aggregator MILs at WSI-level** in (%).

	Highly Imbalanced Camelyon16		
	Acc	AUC	PR AUC
ABMIL [18] + **Our Attention-Guided Prototype Mixing**	86.04	87.00	86.59
DSMIL [21] + **Our Attention-Guided Prototype Mixing**	88.37	89.20	89.00
TransMIL [28] + **Our Attention-Guided Prototype Mixing**	89.14	92.00	**92.60**
GTP [41] + **Our Attention-Guided Prototype Mixing**	**89.92**	**93.00**	92.50

11.4. The Reason for Choosing Transformer-MIL

To further validate our design choice, we analyze the effectiveness of different multiple instance learning (MIL) techniques for WSI classification, including ABMIL [18], DSMIL [21], TransMIL [28], and GTP [41]. Since these aggregators have different architectures, we modify them to produce the attention scores for our use. Specifically, we incorporate the following specific equation with each MIL technique:

- **ABMIL [18]**: The attention score in ABMIL is obtained by taking a weighted average of our prototype instances $P = \{p_1, p_2, \ldots, p_K\}$. The attention score in ABMIL is calculated as:

$$\mathbf{A} = \sum_{k=0}^{N-1} \alpha_k p_k, \tag{15}$$

  where α_k is computed using the softmax function:

$$\alpha_k = \frac{\exp\left(\mathbf{w}^T \tanh\left(\mathbf{V} p_k^T\right)\right)}{\sum_{j=0}^{N-1} \exp\left(\mathbf{w}^T \tanh\left(\mathbf{V} p_j^T\right)\right)}, \tag{16}$$

  where the parameters $\mathbf{w}$ and $\mathbf{V}$ are learned from a neural network.

- **DSMIL [21]**: In DSMIL, the attention score is computed by measuring the distance between our prototype instances P and critical instances. The critical instances are derived from max-pooling , where the highest scored instances are selected. The attention score in DSMIL is calculated as:

$$U(p_i, p_m) = \frac{\exp(\langle p_i, p_m \rangle)}{\sum_{k=0}^{N-1} \exp(\langle p_k, p_m \rangle)}, \tag{17}$$

  where U represents the distance measurement between each prototype instance (p_i) and the critical instance (p_m) and the inner product between two vectors is denoted by $\langle\rangle$.

- **TransMIL [28]**. TransMIL designed Pyramid Position Encoding Generator (PPEG) to acquire spatial information. Therefore, we project our prototype instances P onto PPEG and subsequently feed them into the Transformer module.

- **GTP [41]**. GTP designed Graph Convolution Network (GCN) to learn the inter-relationship between instances. Therefore, we project our prototype instances P onto GCN and subsequently feed them into the Transformer module.

We compared these MIL aggregators to motivate our research. We conducted evaluations on 64 prototype instances extracted from the highly imbalanced Camelyon16 dataset, as shown in Table 7. Our primary objective is to design an MIL aggregator that incorporates pairwise correlations between prototype instances during the training process. This modification provides a significant advantage in the feature diversification process carried out during training. By contrast, the existing MIL methods, such as ABMIL, do not consider the correlation between instances, while DSMIL focuses on a single critical instance and overlooks the correlation with other instances.

11.5. Math Symbols

Table 8 shows the mathematical notations of our proposed method.

Table 8. Mathematical Notations

Symbol	Meaning/Definition
W	Slide image
X	Bag of patches
H	Bag of instance embeddings
f_θ	Feature extractor transforming X to H
P	Prototype instances
P^m/P^M	Prototypes from minority/majority class
$\tilde{P}$	Mixed protoypes
d	Length of instance embedding
n	Number of instances in a single WSI (bag)
K	Number of prototype instances
x_{class}	Learnable classification token
$\mathbf{x}$	Embedded sequence as input to the Transformer
$\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$	Projected queries, keys, and values
$\mathbf{A}$	Attention score of Transformer-MIL module
λ	Combination ratio of the two prototypes