Controllable Latent Space Augmentation for Digital Pathology

Supplementary Material

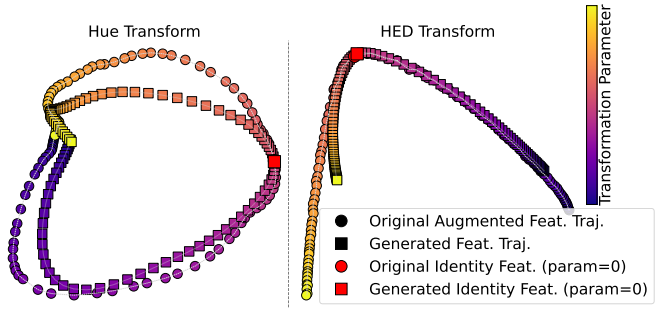

Figure S1. **Trajectories in augmentation space**: **Left** – PCA trajectories of Hue transform for CONCH. **Right** – PCA trajectories of HED transform for UNI. Circles represent the true augmented features extracted from the foundation model, while squares denote the features generated by our augmentation model.

A. Additional details

A.1. Dataset preprocessing

Slides are processed at $10\times$ (1 μm/pixel) and $20\times$ (0.5 μm/pixel) magnifications, with background regions removed and non-overlapping 256×256 pixel patches extracted from tissue regions using the CLAM toolbox [19]. Patch features are then computed using the UNI [4] and CONCH [20] foundation models and stored for downstream analysis.

A.2. Data splitting

Since our generator ρ is trained using TCGA data, and downstream tasks also rely on TCGA datasets, we carefully avoided data leakage as follows:

1. We split the dataset into two subsets: a training portion (70%), containing samples used to train the generator ρ, and a held-out portion (30%).
2. We generated five distinct training sets by bootstrapping (with replacement) from the 70% training subset:

$$\mathcal{D}_{\text{train}}^{(b)}, \quad b \in \{1, \ldots, 5\}.$$

3. The remaining 30% of held-out samples, which were never seen during generator training, were randomly partitioned into validation and test subsets five times (shuffle-split), creating:

$$\mathcal{D}_{\text{val}}^{(b)}, \quad \mathcal{D}_{\text{test}}^{(b)}, \quad b \in \{1, \ldots, 5\}.$$

This procedure resulted in five distinct dataset splits, each consisting of a training set $\mathcal{D}_{\text{train}}^{(b)}$, a validation set $\mathcal{D}_{\text{val}}^{(b)}$, and a test set $\mathcal{D}_{\text{test}}^{(b)}$, ensuring no sample overlap between the generator training and the validation/test sets used in downstream tasks.

A.3. MIL training

For both classification and survival tasks, models were trained for up to 200 epochs using the AdamW optimizer [18] (except for the TransMIL model, which uses Lookahead RAdam [36]) with a learning rate of 10^{-4}, weight decay of 10^{-5}, and gradient accumulation over 4 steps. Early stopping with a patience of 30 epochs was applied in both cases to prevent overfitting.

Classification was optimized using the cross-entropy loss, while survival relied on the negative log-likelihood (NLL) loss. The best classification model was selected based on validation balanced accuracy, and for survival, based on the validation concordance index (C-index), both after the early stopping criterion was met.

During MIL training, feature-level augmentation was applied with a probability of 75% across all augmentation strategies. For *HistAug* and *AugDiff*, augmentations were applied on-the-fly during training with a 75% probability. For *Patch Augmentation* (PAug), with the same probability, precomputed augmented features were used; otherwise, the original features were retained.

A.4. Transformations

In Table S1 we present details of the stochastic image transformations applied, along with their respective parameter ranges.

A.5. Augmented image retrieval

Figure 5 in the main paper presents the results of a study we conduct to get insights about the underlying transformations simulated by *HistAug* in the latent space. To this end, we sample patches and extract their embeddings with a foundation model, either *UNI* or *CONCH*. For each patch, we then predict the associated augmented embedding with *HistAug* for a given transform and associated hyperparameter (e.g. hue transform with hyperparameter -0.5). Finally, we apply all considered transforms with, for each, a set of hyperparameter values (e.g. $\{-0.5, -0.25, +0.25, +0.25\}$ for hue and contrast) , to the original patch (standard image-space augmentation), and extract associated embeddings with the same foundation model as used previously. This gives us a query embedding, i.e. feature vector predicted by HistAug for one specific transformation, and a pool of key embeddings (37 in total), i.e. embeddings of augmented patches with all known transformations. We then perform image retrieval to find the top-1 key embedding which is the

Table S1. **Stochastic Image Transformations and Parameter Ranges**

Transformation	Description	Sampling Range
Crop	Crop randomly from 4 corners or center	Crop side size : $\min(H, W)/2$
Dilation	Morphological dilation	Fixed 4×4 kernel
Erosion	Morphological erosion	Fixed 4×4 kernel
Blur	Gaussian blur	Fixed 15×15 kernel
Brightness	Brightness jitter	$b \sim U[0.5, 1.5]$
Contrast	Contrast jitter	$c \sim U[0.5, 1.5]$
Saturation	Saturation jitter	$s \sim U[0.5, 1.5]$
Hue	Hue jitter	$h \sim U[-0.5, 0.5]$
HED	Histology colour perturbation in HED space [10]. It separates hematoxylin, eosin, and DAB channels to enable fine-grained, stain-aware perturbations.	-
Flip	Horizontal *or* vertical flip	-
Rotate	Rigid rotation	Angle $\in \{90°, 180°, 270°\}$
Gamma	Power-law intensity transform	$\gamma \sim U[0.5, 1.5]$

closest (based on cosine similarity) to the query embedding. Figure 5 thus displays the original patches and for each of them, the top-1 retrieved key augmented patch when generating the query embedding for different transforms (hue, contrast, gaussian blur, erosion) and associated hyperparameters ($\{-0.5, -0.25, +0.25, +0.25\}$ for hue and contrast).

B. Trajectories in augmentation space

Figure S1 presents additional latent trajectory visualizations.

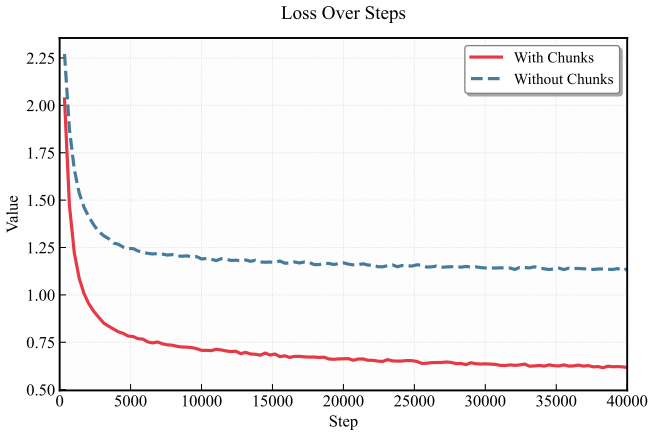

Figure S2. **Impact of chunking input features**: Chunking input features achieves lower training loss, reducing reconstruction error and leading to a faster training.

C. Impact of chunking input features

Training a d-dimensional (e.g., 1024) transformer directly on $\mathbf{z} \in \mathbb{R}^{1024}$ without chunking results in slower convergence and larger reconstruction errors. Indeed, as shown in Figure S2, chunking $\mathbf{z}$ into smaller segments and letting the transformer learn cross-chunk interactions achieves lower training MSE loss. We report the chunking details, including the original embedding dimension, the number of segments, and the resulting input dimension to the transformer network in Table S2.

Table S2. **Chunking for UNI and CONCH Feature Extractors**

Feature Extractor	Orig. Embed. Dimension	Num. of Segments	Transformer Input Dim
UNI	1024	8	128
CONCH	512	4	128

D. Evaluation on the BRACS Dataset for Breast Carcinoma Subtyping

To assess the generalizability of our method beyond TCGA datasets, we report results on the BRACS dataset [2], which focuses on multiclass classification for breast carcinoma subtyping. Specifically, we trained five MIL models (Abmil [15], CLAM variants [19], DSMIL [16], and TransMIL [24]) on five different seeds each. For each MIL model, we computed the mean performance on the official test set by averaging the results across the five runs (with different seeds). The final reported values in Table S3 represent the mean and standard deviation across these five MIL mean performances. Results show improvements of up to 3% in both F1-score and balanced accuracy (Bacc), highlighting the effectiveness and generalizability of our approach beyond TCGA and on a multiclass setting.

Table S3. **Additional results on BRACS dataset using 100% of the training data**. The results are based on the official test set, aggregated across mean performance of five MIL models.

		AUC	Bacc	F1 score
UNI	Base	$79.1_{\pm 2.6}$	$38.5_{\pm 4.8}$	$36.0_{\pm 4.7}$
	Ours(WSI)	$79.7_{\pm 2.4}$	$39.7_{\pm 5.0}$	$37.5_{\pm 4.9}$
	Ours(Inst)	$\mathbf{80.3}_{\pm 2.5}$	$\mathbf{40.0}_{\pm 5.2}$	$\mathbf{37.6}_{\pm 5.3}$
CONCH	Base	$82.3_{\pm 2.3}$	$44.2_{\pm 3.5}$	$41.9_{\pm 4.4}$
	Ours(WSI)	$\mathbf{83.5}_{\pm 2.3}$	$\mathbf{47.4}_{\pm 2.8}$	$\mathbf{44.6}_{\pm 3.2}$
	Ours(Inst)	$82.3_{\pm 2.2}$	$44.8_{\pm 3.0}$	$42.9_{\pm 3.6}$

E. AugDiff Performance Scaling

Despite its computational inefficiency, we trained AugDiff using the *100% training* setup on the BLCA dataset with the CONCH extractor—this being the only setting feasible to run within a reasonable timeframe. Notably, this experiment required approximately 500 GPU-hours for AugDiff, compared to only 5 GPU-hours for HistAug.

As shown in Table S4, HistAug continues to outperform AugDiff even when using the full training set. These results are consistent with the observations under the *10% training* setting reported in Table 2 of the main paper, further validating the robustness and efficiency of our augmentation strategy.

Table S4. Performance comparison on the BLCA dataset (CONCH extractor) using **100% training data**. Mean (± standard deviation) across 5 folds is reported. AugDiff is significantly more computationally expensive but still underperforms compared to our HistAug variants.

Method	Base	AugDiff	Ours (Inst)	Ours (WSI)
Abmil	55.4 (±3.0)	**63.9** (±4.0)	60.4 (±5.0)	62.9 (±3.0)
ClamMb	58.7 (±4.0)	63.4 (±4.0)	**64.4** (±3.0)	62.6 (±4.0)
ClamSb	59.8 (±4.0)	64.2 (±3.0)	**64.9** (±4.0)	62.6 (±2.0)
DSMIL	55.6 (±5.0)	61.4 (±4.0)	**64.0** (±4.0)	61.9 (±3.0)
TransMIL	60.4 (±4.0)	60.0 (±5.0)	**63.6** (±3.0)	62.7 (±2.0)
Mean	58.0	62.6	**63.5**	62.5

F. Detailed result tables

Tables S5 - S10 are detailed counterparts of tables presented in the main paper, showing more fine-grained results.

Table S5. Mean cosine similarity (%) between true augmented features and generated augmented features over 10 000 patches at 10X, alongside feature extractor invariance (%). We also report the 95% bootstrap confidence intervals (CI). The † symbole denotes out of training distribution.

		Feature Reconstruction		Feature Extractor Invariance	
		Cosine Sim (%)	**95% CI**	**Cosine Sim (%)**	**95% CI**
UNI	BLCA	81.0	[80.9, 81.2]	11.9	[11.6, 12.1]
	BRCA	81.6	[81.4, 81.7]	11.9	[11.7, 12.1]
	LUSC	81.1	[80.9, 81.3]	12.6	[12.4, 12.8]
	LUAD †	80.3	[80.2, 80.5]	15.0	[14.8, 15.3]
	UCEC †	80.5	[80.4, 80.7]	9.5	[9.3, 9.7]
	KIRC †	80.5	[80.4, 80.7]	13.0	[12.8, 13.2]
CONCH	BLCA	90.3	[90.2, 90.4]	20.3	[19.9, 20.7]
	BRCA	90.4	[90.3, 90.5]	24.3	[24.0, 24.7]
	LUSC	90.4	[90.3, 90.5]	19.3	[18.9, 19.6]
	LUAD †	90.2	[90.1, 90.3]	24.9	[24.5, 25.3]
	UCEC †	90.4	[90.3, 90.5]	19.7	[19.3, 20.0]
	KIRC †	89.9	[89.8, 90.0]	23.1	[22.7, 23.5]

Table S6. Mean Cosine similarity (%) between true augmented features and generated augmented features over 10 000 patches at 20X, alongside feature extractor invariance (%). We also report the 95% bootstrap confidence intervals (CI). The † symbole denotes out of training distribution.

		Feature Reconstruction		Feature Extractor Invariance	
		Cosine Sim (%)	**95% CI**	**Cosine Sim (%)**	**95% CI**
UNI	BLCA	74.9	[74.7, 75.1]	11.6	[11.6, 11.8]
	BRCA	75.8	[75.6, 76.0]	11.7	[11.5, 11.9]
	LUSC	74.9	[74.7, 75.1]	11.9	[11.6, 12.0]
	LUAD †	74.6	[74.4, 74.8]	11.3	[11.1, 11.5]
	UCEC †	73.4	[73.2, 73.5]	12.9	[12.7, 13.2]
	KIRC †	72.1	[71.9, 72.3]	11.0	[10.8, 11.1]
CONCH	BLCA	88.1	[88.0, 88.3]	27.7	[27.4, 28.1]
	BRCA	88.6	[88.5, 88.7]	32.7	[32.4, 33.1]
	LUSC	87.0	[86.9, 87.1]	27.7	[27.4, 28.1]
	LUAD †	87.6	[87.5, 87.7]	23.7	[23.3, 24.0]
	UCEC †	88.5	[88.4, 88.6]	29.6	[29.2, 29.9]
	KIRC †	87.5	[87.4, 87.6]	29.8	[29.5, 30.2]

F.1. Generator evaluation and and Feature Extractor Invariance

Tables S5 and S6 present the mean cosine similarity between true and generated augmented features at 10X and 20X magnifications, alongside feature extractor invariance. The tables are the same as in the main paper, the main difference is that we include 95% bootstrap confidence intervals.

F.2. MIL evaluation at $10\times$ magnification

Performance at 10% Training Data — We compare several MIL architectures (Abmil [15], CLAM variants [19], DSMIL [16], and TransMIL [24]) with different augmen-

tation methods at 10% training data. Results in Table S7 clearly indicate that instance-wise (Inst) and WSI-wise augmentation methods (WSI) consistently outperform the baseline, the diffusion-based augmentation (AugDiff), the noise-based augmentation (Noise) and the offline patch augmentation (PAug) across nearly all cancer types. Specifically, on the CONCH embeddings, our augmentation methods achieves substantial improvements in survival prediction (C-Index) and classification tasks (AUC) compared to baselines. For example, in BLCA, KIRC, and UCEC cancers, WSI-wise augmentation resulted in notable improvements of 3–6 points in C-index compared to baseline models. The impact of noise perturbations at $10\times$ magnification is also evaluated in these tables. Our learned augmentation strategies (Inst and WSI) consistently outperform the noise perturbation baseline across both 10% and 100% training data setups.

Performance at 100% Training Data — Using 100% of the available training data (Table S8), our augmentation methods enhance MIL performance in most cases over baselines across multiple cancer types. The improvements are more pronounced on the UNI embeddings, where instance-wise augmentation results in a mean improvement of approximately 4.5 points in C-index for survival prediction tasks (e.g., BLCA improved from 54.5 to 60.3). For classification tasks (BRCA and NSCLC), WSI-wise augmentation methods maintain or slightly improve upon strong baseline performance.

F.3. MIL evaluation at $20\times$ magnification

Tables S9 and S10 present the MIL evaluation results at 20X magnification, where we use our generator trained at 10X to generate augmented tiles. The trends observed at 10X remain consistent with WSI augmentations continuing to provide improvements. At 10% training data (Table S9), wsi-wise augmentation improves both survival prediction (C-Index) and classification (AUC), particularly for survival prediction. At 100% training data (Table S10), survival prediction still benefits from augmentation strategies. The results further demonstrate the generalizability of our method across magnifications.

G. Code and Data Availability

The source code of our project will be made publicly available at https://github.com/MICS-Lab/HistAug.

All TCGA datasets used in this study can be accessed via the Genomic Data Commons (GDC) portal at https://portal.gdc.cancer.gov.

The BRACS dataset [2] is publicly available at https://www.bracs.icar.cnr.it.

The script for slide pre-processing and patch extraction is available at https://github.com/mahmoodlab/CLAM.

The code and pre-trained weights for the UNI [4] and CONCH [20] models can be found on Hugging Face at https://huggingface.co/MahmoodLab/UNI and https://huggingface.co/MahmoodLab/CONCH, respectively.

Table S7. **MIL evaluation with limited data**: Comparison at 10% training data for 10X magnification. Survival (C-Index) on BLCA, KIRC, UCEC; classification (AUC) on BRCA, NSCLC. UNI (top), CONCH (bottom), **AugD**=AugDiff, **Noise**=feature-wise Gaussian noise, **PAug**=patch-wise augmentation, **Inst**=instance-wise augmentation (ours), **WSI**=wsi-wise augmentation (ours). Values are %. Means (± standard deviations) are reported over five splits.

		Survival (C-Index)																		**Classification (AUC)**											
		BLCA						KIRC						UCEC						BRCA						NSCLC					
	Model	**Base**	**AugD**	**Noise**	**PAug**	**Inst**	**WSI**	**Base**	**AugD**	**Noise**	**PAug**	**Inst**	**WSI**	**Base**	**AugD**	**Noise**	**PAug**	**Inst**	**WSI**	**Base**	**AugD**	**Noise**	**PAug**	**Inst**	**WSI**	**Base**	**AugD**	**Noise**	**PAug**	**Inst**	**WSI**
UNI	Abmil	46.1 (±5.0)	53.1 (±7.0)	46.0 (±5.0)	48.5 (±5.0)	48.8 (±4.0)	48.3 (±7.0)	56.2 (±5.0)	61.5 (±8.0)	56.1 (±6.0)	58.4 (±4.0)	59.1 (±7.0)	59.9 (±6.0)	55.3 (±5.0)	60.4 (±10.0)	54.9 (±5.0)	54.7 (±6.0)	60.0 (±11.0)	61.8 (±12.0)	85.7 (±2.0)	87.3 (±2.0)	85.8 (±2.0)	87.0 (±2.0)	89.4 (±2.0)	89.3 (±2.0)	89.3 (±4.0)	89.1 (±3.0)	89.4 (±4.0)	91.7 (±3.0)	91.2 (±3.0)	92.4 (±4.0)
	ClamMb	49.2 (±6.0)	48.4 (±7.0)	49.4 (±6.0)	47.9 (±5.0)	50.8 (±7.0)	49.6 (±8.0)	59.0 (±9.0)	67.2 (±3.0)	58.9 (±9.0)	57.3 (±6.0)	64.5 (±5.0)	64.3 (±5.0)	57.5 (±6.0)	62.6 (±8.0)	57.5 (±6.0)	59.5 (±9.0)	62.1 (±7.0)	60.6 (±12.0)	87.0 (±2.0)	87.2 (±2.0)	87.0 (±2.0)	89.5 (±2.0)	89.4 (±2.0)	88.9 (±2.0)	90.9 (±4.0)	87.4 (±5.0)	91.7 (±3.0)	92.7 (±4.0)	92.6 (±3.0)	93.7 (±2.0)
	ClamSb	45.8 (±7.0)	51.7 (±7.0)	45.6 (±7.0)	48.5 (±6.0)	51.8 (±7.0)	52.9 (±5.0)	55.5 (±9.0)	63.3 (±6.0)	56.0 (±9.0)	57.6 (±3.0)	58.3 (±5.0)	61.3 (±6.0)	54.1 (±4.0)	53.8 (±6.0)	53.8 (±4.0)	55.8 (±7.0)	51.0 (±9.0)	59.5 (±6.0)	86.4 (±2.0)	86.4 (±1.0)	86.2 (±2.0)	89.5 (±1.0)	88.2 (±2.0)	89.0 (±1.0)	91.3 (±4.0)	90.1 (±5.0)	91.0 (±4.0)	92.8 (±3.0)	92.5 (±3.0)	93.8 (±3.0)
	DSMIL	46.2 (±3.0)	46.7 (±4.0)	46.3 (±4.0)	46.3 (±6.0)	50.3 (±5.0)	50.8 (±3.0)	63.3 (±8.0)	67.5 (±3.0)	61.4 (±8.0)	63.4 (±8.0)	64.5 (±4.0)	66.5 (±4.0)	63.0 (±2.0)	67.2 (±4.0)	62.7 (±4.0)	66.8 (±5.0)	64.2 (±7.0)	64.8 (±6.0)	85.2 (±4.0)	73.2 (±2.0)	87.3 (±2.0)	87.0 (±3.0)	86.9 (±3.0)	87.0 (±3.0)	81.2 (±3.0)	79.3 (±4.0)	87.0 (±4.0)	84.3 (±3.0)	87.2 (±3.0)	85.6 (±2.0)
	TransMIL	50.3 (±2.0)	49.5 (±1.0)	50.4 (±2.0)	50.7 (±3.0)	50.9 (±2.0)	51.3 (±2.0)	58.7 (±10.0)	54.5 (±12.0)	58.2 (±10.0)	63.7 (±5.0)	59.9 (±12.0)	60.3 (±11.0)	66.5 (±9.0)	65.6 (±3.0)	66.1 (±4.0)	67.8 (±6.0)	69.0 (±6.0)	69.5 (±5.0)	86.4 (±2.0)	86.6 (±2.0)	86.6 (±2.0)	87.8 (±2.0)	87.0 (±3.0)	87.5 (±2.0)	85.3 (±5.0)	88.0 (±4.0)	86.0 (±5.0)	83.1 (±5.0)	84.9 (±5.0)	86.7 (±4.0)
	Mean	47.5	49.9	47.5	48.4	50.5	**50.6**	58.5	**62.8**	58.1	60.1	61.3	62.5	59.3	61.9	59.0	60.9	61.3	**63.2**	86.1	84.1	86.6	88.2	88.2	**88.3**	87.6	86.8	89.0	88.9	89.7	**90.4**
CONCH	Abmil	49.7 (±5.0)	52.2 (±5.0)	51.6 (±3.0)	51.0 (±4.0)	52.3 (±7.0)	52.2 (±8.0)	60.4 (±5.0)	65.8 (±3.0)	61.3 (±6.0)	61.6 (±11.0)	65.3 (±9.0)	67.6 (±8.0)	56.1 (±5.0)	61.3 (±6.0)	55.6 (±6.0)	64.2 (±6.0)	63.2 (±9.0)	65.8 (±2.0)	91.6 (±2.0)	91.9 (±1.0)	90.2 (±2.0)	88.2 (±4.0)	90.9 (±1.0)	91.7 (±2.0)	95.2 (±1.0)	93.8 (±2.0)	93.0 (±2.0)	94.2 (±1.0)	94.5 (±2.0)	95.0 (±2.0)
	ClamMb	53.1 (±6.0)	52.3 (±7.0)	53.2 (±6.0)	55.1 (±3.0)	56.7 (±5.0)	57.6 (±4.0)	63.0 (±7.0)	67.3 (±3.0)	62.9 (±7.0)	67.7 (±2.0)	70.2 (±1.0)	70.8 (±3.0)	54.7 (±11.0)	60.6 (±7.0)	54.8 (±10.0)	64.7 (±6.0)	64.1 (±8.0)	65.0 (±4.0)	90.3 (±1.0)	92.5 (±2.0)	90.3 (±1.0)	90.8 (±3.0)	91.2 (±1.0)	91.4 (±1.0)	93.0 (±1.0)	93.8 (±2.0)	92.8 (±1.0)	93.3 (±2.0)	94.9 (±1.0)	95.7 (±1.0)
	ClamSb	48.2 (±4.0)	50.7 (±7.0)	48.6 (±4.0)	53.6 (±5.0)	53.8 (±5.0)	49.7 (±6.0)	65.6 (±5.0)	66.0 (±3.0)	66.2 (±5.0)	65.4 (±9.0)	70.5 (±4.0)	72.0 (±6.0)	58.5 (±10.0)	62.3 (±7.0)	58.6 (±10.0)	65.3 (±5.0)	64.4 (±11.0)	63.3 (±5.0)	90.2 (±1.0)	92.3 (±1.0)	90.6 (±1.0)	88.1 (±3.0)	90.6 (±1.0)	91.4 (±2.0)	92.5 (±3.0)	94.6 (±2.0)	92.4 (±2.0)	94.1 (±2.0)	94.5 (±2.0)	95.6 (±1.0)
	DSMIL	49.5 (±5.0)	55.0 (±5.0)	51.7 (±5.0)	57.2 (±7.0)	55.9 (±7.0)	57.0 (±7.0)	67.2 (±5.0)	69.7 (±1.0)	65.9 (±6.0)	71.5 (±3.0)	72.0 (±3.0)	72.7 (±3.0)	59.6 (±9.0)	63.1 (±9.0)	61.7 (±8.0)	64.1 (±6.0)	62.0 (±8.0)	64.8 (±5.0)	82.3 (±3.0)	82.5 (±4.0)	89.7 (±2.0)	82.5 (±3.0)	88.1 (±3.0)	87.4 (±2.0)	91.3 (±2.0)	93.7 (±2.0)	93.1 (±2.0)	92.6 (±1.0)	93.0 (±2.0)	93.8 (±2.0)
	TransMIL	53.7 (±4.0)	55.0 (±6.0)	53.5 (±4.0)	53.9 (±4.0)	53.9 (±5.0)	53.8 (±5.0)	59.3 (±9.0)	60.8 (±8.0)	59.5 (±10.0)	60.5 (±8.0)	64.1 (±6.0)	64.7 (±5.0)	64.0 (±7.0)	62.0 (±9.0)	63.3 (±8.0)	64.2 (±9.0)	64.6 (±7.0)	65.5 (±5.0)	91.4 (±1.0)	91.2 (±1.0)	91.4 (±1.0)	90.9 (±2.0)	91.2 (±2.0)	92.1 (±1.0)	92.0 (±2.0)	93.0 (±2.0)	92.3 (±2.0)	92.0 (±2.0)	92.7 (±1.0)	93.0 (±2.0)
	Mean	50.8	53.0	51.7	54.2	**54.5**	54.1	63.1	65.9	63.2	65.3	68.4	**69.6**	58.6	61.9	58.8	64.5	63.7	**64.9**	89.2	90.1	90.4	88.1	90.4	**90.8**	92.8	93.8	92.7	93.2	93.9	**94.6**

Table S8. **MIL evaluation with full data**: Results at 100% training data for 10X magnification. Survival (C-Index) on BLCA, KIRC, UCEC; classification (AUC) on BRCA, NSCLC. UNI (top), CONCH (bottom), **Noise**=feature-wise Gaussian noise, **PAug**=patch-wise augmentation, **Inst**=instance-wise augmentation (ours), **WSI**=wsi-wise augmentation (ours). Values are %. Means (± standard deviations) are reported over five splits.

		Survival (C-Index)																		**Classification (AUC)**											
		BLCA						KIRC						UCEC						BRCA						NSCLC					
	Model	**Base**	**Noise**	**PAug**	**Inst**	**WSI**		**Base**	**Noise**	**PAug**	**Inst**	**WSI**		**Base**	**Noise**	**PAug**	**Inst**	**WSI**		**Base**	**Noise**	**PAug**	**Inst**	**WSI**		**Base**	**Noise**	**PAug**	**Inst**	**WSI**	
UNI	Abmil	51.8 (±6.0)	54.1 (±5.0)	59.2 (±5.0)	59.1 (±4.0)	57.4 (±6.0)		67.1 (±3.0)	66.5 (±5.0)	68.5 (±3.0)	68.6 (±5.0)	69.1 (±4.0)		63.0 (±9.0)	63.5 (±10.0)	63.0 (±3.0)	61.5 (±12.0)	62.7 (±5.0)		93.0 (±2.0)	93.3 (±2.0)	93.8 (±2.0)	92.1 (±2.0)	92.1 (±2.0)		97.8 (±1.0)	98.0 (±1.0)	93.5 (±2.0)	97.9 (±1.0)	97.5 (±1.0)	
	ClamMb	58.0 (±7.0)	56.7 (±8.0)	55.2 (±4.0)	56.7 (±6.0)	56.7 (±4.0)		64.0 (±3.0)	64.9 (±4.0)	68.4 (±2.0)	67.3 (±3.0)	69.1 (±3.0)		65.5 (±5.0)	65.1 (±3.0)	62.2 (±4.0)	61.6 (±11.0)	62.2 (±5.0)		93.0 (±2.0)	93.1 (±2.0)	93.0 (±1.0)	92.5 (±2.0)	92.4 (±2.0)		98.6 (±1.0)	98.1 (±1.0)	94.1 (±1.0)	97.1 (±1.0)	97.5 (±2.0)	
	ClamSb	55.2 (±5.0)	52.5 (±5.0)	58.7 (±3.0)	62.8 (±2.0)	63.9 (±2.0)		69.9 (±5.0)	69.0 (±7.0)	65.3 (±3.0)	68.5 (±5.0)	67.8 (±4.0)		56.2 (±7.0)	57.4 (±3.0)	65.3 (±10.0)	69.4 (±6.0)	63.3 (±11.0)		93.5 (±2.0)	92.8 (±2.0)	93.2 (±2.0)	92.3 (±2.0)	92.5 (±3.0)		98.5 (±1.0)	98.3 (±1.0)	94.4 (±2.0)	97.6 (±1.0)	97.4 (±1.0)	
	DSMIL	48.8 (±5.0)	58.2 (±4.0)	50.1 (±6.0)	60.3 (±3.0)	58.5 (±4.0)		64.4 (±7.0)	66.1 (±4.0)	64.9 (±3.0)	69.0 (±4.0)	68.8 (±4.0)		65.3 (±7.0)	60.1 (±14.0)	64.6 (±3.0)	69.2 (±6.0)	68.9 (±6.0)		90.9 (±2.0)	92.5 (±2.0)	93.4 (±2.0)	92.7 (±1.0)	92.3 (±2.0)		98.1 (±0.0)	97.4 (±1.0)	91.9 (±1.0)	98.0 (±1.0)	98.4 (±1.0)	
	TransMIL	58.7 (±3.0)	58.3 (±2.0)	60.4 (±4.0)	62.5 (±3.0)	60.9 (±2.0)		63.5 (±3.0)	63.0 (±3.0)	67.4 (±3.0)	64.8 (±3.0)	66.3 (±3.0)		68.9 (±7.0)	68.3 (±9.0)	68.6 (±4.0)	65.6 (±5.0)	66.9 (±9.0)		92.9 (±2.0)	92.8 (±2.0)	93.3 (±2.0)	92.4 (±2.0)	92.9 (±3.0)		95.6 (±1.0)	95.9 (±1.0)	93.6 (±2.0)	95.8 (±1.0)	97.8 (±0.0)	
	Mean	54.5	56.0	56.7	**60.3**	59.5		65.8	65.9	66.9	67.6	**68.2**		63.8	62.9	64.7	**65.5**	64.8		92.7	92.9	**93.3**	92.4	92.4		**97.7**	97.5	93.5	97.3	**97.7**	
CONCH	Abmil	55.4 (±3.0)	56.6 (±3.0)	60.9 (±3.0)	60.4 (±5.0)	62.9 (±3.0)		67.6 (±4.0)	72.8 (±2.0)	69.6 (±3.0)	69.9 (±1.0)	71.6 (±2.0)		59.3 (±4.0)	59.6 (±6.0)	66.9 (±9.0)	65.3 (±12.0)	67.3 (±12.0)		93.8 (±1.0)	93.0 (±2.0)	92.7 (±2.0)	93.3 (±1.0)	94.0 (±2.0)		98.0 (±2.0)	97.3 (±1.0)	94.8 (±1.0)	98.4 (±1.0)	98.4 (±1.0)	
	ClamMb	58.7 (±4.0)	58.7 (±5.0)	62.7 (±4.0)	64.4 (±3.0)	62.6 (±4.0)		70.0 (±5.0)	69.9 (±4.0)	71.6 (±2.0)	71.9 (±3.0)	73.2 (±3.0)		62.1 (±7.0)	63.9 (±7.0)	71.8 (±10.0)	67.0 (±8.0)	67.0 (±4.0)		94.3 (±2.0)	94.2 (±2.0)	92.8 (±2.0)	93.5 (±2.0)	93.7 (±2.0)		98.4 (±1.0)	98.1 (±1.0)	95.3 (±1.0)	98.4 (±1.0)	98.5 (±1.0)	
	ClamSb	59.8 (±4.0)	58.6 (±3.0)	63.1 (±4.0)	64.9 (±4.0)	62.6 (±2.0)		66.6 (±7.0)	66.5 (±4.0)	71.2 (±5.0)	70.2 (±1.0)	70.9 (±2.0)		60.6 (±13.0)	59.0 (±11.0)	65.0 (±10.0)	62.9 (±9.0)	63.0 (±3.0)		94.1 (±2.0)	92.8 (±2.0)	92.6 (±2.0)	93.9 (±2.0)	94.2 (±2.0)		97.5 (±2.0)	95.9 (±1.0)	95.0 (±1.0)	97.7 (±1.0)	98.0 (±1.0)	
	DSMIL	55.6 (±5.0)	61.9 (±5.0)	59.0 (±7.0)	64.0 (±4.0)	61.9 (±3.0)		67.8 (±5.0)	69.0 (±7.0)	69.1 (±2.0)	70.6 (±4.0)	72.2 (±4.0)		58.1 (±8.0)	62.5 (±10.0)	63.9 (±6.0)	64.6 (±9.0)	64.8 (±5.0)		89.8 (±2.0)	93.9 (±2.0)	92.5 (±2.0)	94.4 (±2.0)	93.9 (±2.0)		98.2 (±1.0)	97.5 (±1.0)	95.4 (±1.0)	98.2 (±1.0)	97.5 (±2.0)	
	TransMIL	60.4 (±4.0)	60.2 (±4.0)	60.6 (±5.0)	63.6 (±3.0)	62.7 (±2.0)		70.7 (±4.0)	70.1 (±4.0)	68.0 (±5.0)	68.1 (±2.0)	68.1 (±4.0)		60.2 (±9.0)	60.0 (±9.0)	56.4 (±4.0)	58.6 (±13.0)	64.1 (±10.0)		93.0 (±2.0)	93.0 (±2.0)	91.3 (±1.0)	92.2 (±2.0)	92.7 (±2.0)		97.6 (±1.0)	96.2 (±1.0)	92.5 (±1.0)	97.8 (±1.0)	96.8 (±1.0)	
	Mean	58.0	59.2	61.3	**63.5**	62.5		68.5	69.7	69.9	70.1	**71.2**		60.1	61.0	64.8	63.7	**65.2**		93.0	93.4	92.4	93.5	**93.7**		97.9	97.0	94.6	**98.1**	97.8	

Table S9. Comparison at 10% training data (20X magnification). **UNI** (top) and **CONCH** (bottom). Values are reported in percentage. Means (± standard deviations) are reported over five splits.

		Survival (C-Index)						Classification (AUC)			
		BLCA		KIRC		UCEC		BRCA		NSCLC	
	Model	Base	WSI	Base	WSI	Base	WSI	Base	WSI	Base	WSI
UNI	Abmil	45.8 (±6.0)	48.5 (±6.0)	55.3 (±3.0)	58.5 (±5.0)	54.6 (±4.0)	58.8 (±6.0)	86.4 (±3.0)	88.1 (±2.0)	90.9 (±4.0)	89.7 (±7.0)
	ClamMb	45.7 (±7.0)	49.1 (±9.0)	56.4 (±6.0)	61.4 (±6.0)	51.6 (±6.0)	55.5 (±8.0)	87.1 (±2.0)	88.5 (±1.0)	91.9 (±4.0)	92.8 (±4.0)
	ClamSb	46.3 (±6.0)	50.7 (±7.0)	56.6 (±5.0)	57.3 (±5.0)	52.8 (±5.0)	56.4 (±3.0)	87.0 (±2.0)	87.5 (±2.0)	92.1 (±4.0)	93.6 (±4.0)
	DSMIL	45.6 (±7.0)	50.2 (±8.0)	58.3 (±7.0)	63.6 (±2.0)	56.9 (±4.0)	60.0 (±3.0)	81.6 (±3.0)	85.2 (±2.0)	81.3 (±6.0)	84.1 (±4.0)
	TransMIL	51.5 (±7.0)	51.1 (±4.0)	61.5 (±4.0)	65.3 (±3.0)	65.2 (±5.0)	68.4 (±4.0)	85.5 (±1.0)	86.7 (±2.0)	83.5 (±7.0)	83.6 (±7.0)
	Mean	47.0	**49.9**	57.6	**61.2**	56.2	**59.8**	85.5	**87.2**	87.9	**88.8**
CONCH	Abmil	50.4 (±6.0)	52.1 (±6.0)	59.0 (±7.0)	65.7 (±9.0)	55.9 (±8.0)	65.4 (±10.0)	90.1 (±4.0)	90.7 (±3.0)	92.8 (±4.0)	95.0 (±3.0)
	ClamMb	51.7 (±7.0)	55.4 (±3.0)	64.9 (±6.0)	70.7 (±3.0)	60.2 (±10.0)	59.2 (±11.0)	90.2 (±3.0)	90.9 (±1.0)	92.9 (±4.0)	95.9 (±2.0)
	ClamSb	46.5 (±7.0)	53.9 (±6.0)	66.4 (±5.0)	70.2 (±6.0)	57.9 (±9.0)	62.5 (±11.0)	90.9 (±3.0)	90.5 (±2.0)	92.3 (±5.0)	95.4 (±3.0)
	DSMIL	50.2 (±5.0)	57.7 (±6.0)	69.8 (±3.0)	72.4 (±4.0)	60.5 (±9.0)	62.3 (±10.0)	83.0 (±3.0)	83.6 (±5.0)	92.4 (±3.0)	94.6 (±2.0)
	TransMIL	53.1 (±7.0)	55.4 (±8.0)	62.1 (±8.0)	64.8 (±6.0)	66.0 (±8.0)	66.2 (±8.0)	89.5 (±3.0)	90.4 (±1.0)	92.2 (±4.0)	94.5 (±2.0)
	Mean	50.4	**54.9**	64.4	**68.8**	60.1	**63.1**	88.7	**89.2**	92.5	**95.1**

Table S10. Comparison at 100% training data (20X magnification). **UNI** (top) and **CONCH** (bottom). Values are reported in percentage. Means (± standard deviations) are reported over five splits.

		Survival (C-Index)						Classification (AUC)			
		BLCA		KIRC		UCEC		BRCA		NSCLC	
	Model	Base	WSI	Base	WSI	Base	WSI	Base	WSI	Base	WSI
UNI	Abmil	47.2 (±3.0)	51.5 (±5.0)	67.0 (±6.0)	68.4 (±4.0)	60.1 (±5.0)	65.3 (±8.0)	93.7 (±2.0)	93.2 (±2.0)	96.6 (±2.0)	97.4 (±1.0)
	ClamMb	53.8 (±4.0)	54.6 (±4.0)	62.3 (±4.0)	64.1 (±6.0)	59.5 (±4.0)	63.2 (±9.0)	93.0 (±2.0)	93.2 (±1.0)	96.3 (±2.0)	96.6 (±1.0)
	ClamSb	51.9 (±6.0)	56.2 (±6.0)	66.9 (±4.0)	65.1 (±5.0)	57.9 (±6.0)	58.2 (±3.0)	92.2 (±2.0)	93.2 (±2.0)	96.9 (±2.0)	97.9 (±1.0)
	DSMIL	50.1 (±3.0)	56.7 (±2.0)	61.5 (±6.0)	68.8 (±3.0)	53.0 (±5.0)	65.3 (±7.0)	92.4 (±2.0)	92.2 (±2.0)	96.8 (±2.0)	97.0 (±1.0)
	TransMIL	52.6 (±3.0)	58.0 (±5.0)	59.1 (±4.0)	63.4 (±4.0)	66.0 (±5.0)	62.2 (±5.0)	92.9 (±2.0)	93.0 (±2.0)	94.6 (±2.0)	96.5 (±2.0)
	Mean	51.1	**55.4**	63.4	**66.0**	59.3	**62.8**	92.8	**93.0**	96.2	**97.1**
CONCH	Abmil	53.2 (±2.0)	63.2 (±5.0)	71.5 (±5.0)	69.9 (±1.0)	61.6 (±4.0)	70.0 (±8.0)	93.4 (±2.0)	93.5 (±2.0)	98.0 (±2.0)	98.7 (±1.0)
	ClamMb	59.2 (±6.0)	63.2 (±5.0)	69.1 (±2.0)	71.2 (±4.0)	62.6 (±8.0)	69.3 (±5.0)	93.4 (±1.0)	93.6 (±2.0)	98.7 (±1.0)	98.7 (±1.0)
	ClamSb	53.4 (±3.0)	65.9 (±6.0)	61.9 (±3.0)	70.2 (±5.0)	58.3 (±9.0)	68.4 (±10.0)	93.7 (±1.0)	93.7 (±1.0)	98.0 (±1.0)	98.6 (±1.0)
	DSMIL	57.7 (±4.0)	61.9 (±6.0)	68.9 (±6.0)	74.2 (±2.0)	61.7 (±7.0)	68.6 (±5.0)	91.7 (±1.0)	93.3 (±2.0)	98.3 (±1.0)	98.2 (±1.0)
	TransMIL	64.6 (±3.0)	66.3 (±3.0)	68.2 (±4.0)	68.9 (±3.0)	58.5 (±8.0)	59.4 (±11.0)	91.3 (±2.0)	91.9 (±2.0)	98.1 (±1.0)	96.9 (±3.0)
	Mean	55.9	**64.1**	67.9	**70.9**	60.5	**67.1**	92.7	**93.2**	**98.2**	**98.2**