

Controllable Latent Space Augmentation for Digital Pathology

Sofiène Boutaj¹[◊], Marin Scalbert^{2*}, Pierre Marza¹, Florent Couzinie-Devy³,
 Maria Vakalopoulou¹, Stergios Christodoulidis¹

¹MICS, CentraleSupélec – Université Paris-Saclay

²Bioptimus, Inc. ³VitaDX International

Abstract

Whole slide image (WSI) analysis in digital pathology presents unique challenges due to the gigapixel resolution of WSIs and the scarcity of dense supervision signals. While Multiple Instance Learning (MIL) is a natural fit for slide-level tasks, training robust models requires large and diverse datasets. Even though image augmentation techniques could be utilized to increase data variability and reduce overfitting, implementing them effectively is not a trivial task. Traditional patch-level augmentation is prohibitively expensive due to the large number of patches extracted from each WSI, and existing feature-level augmentation methods lack control over transformation semantics. We introduce HistAug, a fast and efficient generative model for controllable augmentations in the latent space for digital pathology. By conditioning on explicit patch-level transformations (e.g., hue, erosion), HistAug generates realistic augmented embeddings while preserving initial semantic information. Our method allows the processing of a large number of patches in a single forward pass efficiently, while at the same time consistently improving MIL model performance. Experiments across multiple slide-level tasks and diverse organs show that HistAug outperforms existing methods, particularly in low-data regimes. Ablation studies confirm the benefits of learned transformations over noise-based perturbations and highlight the importance of uniform WSI-wise augmentation. Code is available at <https://github.com/MICS-Lab/HistAug>.

1. Introduction

In recent years, several deep learning methods have been proposed for a growing number of tasks in histopathology, including sub-typing, time-to-event endpoints, gene marker prediction, and others [25]. Most of these approaches

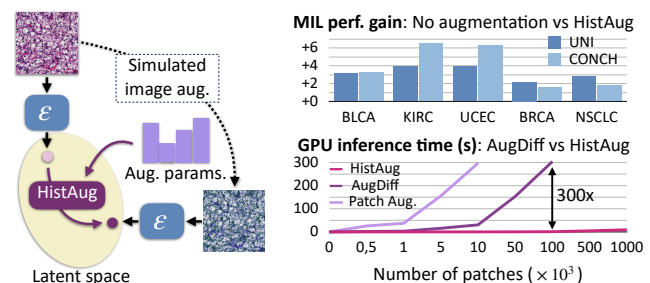


Figure 1. **HistAug**: We present a controllable and efficient method to augment WSI patches in the latent space of a vision encoder ($\mathcal{E}$) for histopathology MIL training. Such latent space augmentation leads to better downstream performance (gain in low-data shown here) as WSI datasets often lack data diversity. Our approach is faster and requires less memory than the state-of-the-art diffusion model AugDiff [6] and vanilla patch augmentation.

leverage digitized histopathology tissue slides referred to as whole slide images (WSIs). These high-resolution, gigapixel-sized images provide details of the tissue micro-environment, enabling precise and comprehensive analysis. To efficiently process WSIs, a standard multi-step workflow has emerged: detect the tissue on the slide, divide it into patches, extract features from each patch, and finally perform a slide-level aggregation of the features and obtain a prediction from the aggregated features. These steps are often executed sequentially offline and the final model is trained on the feature space. On these grounds, many foundation models have been trained with self-supervised learning [1, 4, 11, 20, 21, 29, 38] to serve as frozen patch-level feature extractors, removing the need for training task-specific vision encoders. Furthermore, since annotations are mainly available at the slide level and not for each patch individually, the aggregation models rely on Multiple Instance Learning (MIL) approaches [15, 19, 24].

A primary challenge for deep learning models in digital pathology is the limited number of annotated slides for each specific task. Although patch-level augmentation can

[◊]Corresponding email : sofiene.boutaj@centralesupelec.fr

^{*}Work done at CentraleSupélec and VitaDX.

improve robustness and generalization, performing it *online* for gigapixel WSIs is typically infeasible: it would require reading, transforming, and re-embedding on-the-fly tens or hundreds of thousands of patches per WSI. A common workaround is to pre-augment the patches offline and store multiple augmented versions of each WSI. However, this drastically increases both storage and preprocessing time, while offering limited diversity. Feature-level augmentation has thus emerged as a promising alternative, applying transformations directly to patch features in a way that simulates patch-level and pixel-level transformations (e.g. color jittering). The few existing solutions available are based on generative processes (e.g., diffusion-based [6] or GANs [34]) and suffer from two major limitations: (i) they lack direct control over transformations and (ii) they incur large memory and/or time overhead.

In this study, we propose a novel transformer-based model, named *HistAug*, performing latent-space augmentation that can be used in MIL (Figure 1). Given a patch embedding and a set of transformations together with their corresponding parameters (e.g. hue offsets), the model learns to predict the augmented patch embedding. The precise control over transformation parameters offers flexibility in selecting which transformations to apply and adjusting their intensity based on specific tasks. The lightweight nature of our model allows a fast processing of multiple patch embeddings in parallel, making it easy to use within MIL training.

Our approach builds on the key insight that foundation models *are not fully invariant* to image transformations [9, 30] and thus, that feature-space transforms can help during MIL training. This is validated on two state-of-the-art foundation models, the uni-modal vision encoder UNI [4] and the vision-language model CONCH [20]. Experiments on diverse histopathology datasets, with varying organs and tasks, show that *HistAug* has a positive impact on downstream performance. Additionally, our model is much faster and less memory-intensive than the current state-of-the-art diffusion-based model AugDiff [6] and naive patch augmentation. Since WSIs can also be processed at different magnification levels, we show that even if *HistAug* generator is trained at a given magnification, it generalizes well to others.

In summary, our main contributions are the following: (i) we introduce *HistAug*, a novel generative model, designed to perform fast and controllable augmentation in the latent space. The model leverages a transformer architecture with cross-attention mechanisms to predict how patch-level transformations affect features of non-augmented patches, (ii) the lightweight design of *HistAug* allows to process 1M patches in less than 10 seconds, making it easy to integrate into every MIL training setting, (iii) the parameters of the augmentations applied to the features can be fully controlled allowing task-specific transformations without re-

training the generator. At the same time, augmentations can be applied uniformly on a whole WSI, allowing, for example, consistent simulated color in the whole augmented bag.

A detailed validation of *HistAug* on several MIL settings is performed using two foundation models, diverse histopathology datasets, and multiple magnifications. The quality of the simulated image augmentations is validated both quantitatively and through visualizations.

2. Related Work

Foundation models in histopathology — are trained with self-supervised learning on large patch-level datasets to later be used as general feature extractors. They can either be uni-modal vision encoders [1, 4, 11, 21, 29, 38] or multi-modal vision-language models [7, 13, 14, 20, 23, 31, 37]. UNI [4] is a state-of-the-art VisionTransformer (ViT) model [8, 28] trained with a DINOv2 objective [3, 22] on a large dataset composed of over 100M patches extracted from more than 100,000 WSIs representing 20 tissue types. CONCH [20] is a vision-language model trained on over 1M image-caption pairs with a contrastive objective [33]. Its vision encoder is Transformer [28]-based and can be leveraged along with the text encoder to perform many downstream tasks such as image captioning or text-to-image retrieval. Since our method is backbone-agnostic, in this work, we show its positive impact on the two different embedding spaces of UNI and CONCH as they are considered strong candidates to process WSIs.

Multiple Instance Learning (MIL) — is an efficient approach to train aggregator functions predicting information of interest from a bag of features. In histopathology in particular, such embeddings are extracted from WSI patches with foundation models as the ones presented above. MIL approaches involve simple attention mechanisms [15, 19], or even self-attention mechanisms through Transformer layers [24]. This MIL setting is convenient for training a model to efficiently aggregate patch-level features, but processing such a high number of embeddings simultaneously makes any image-based augmentation technique hard to consider due to compute and memory constraints.

We believe that image augmentations are too important to be ignored when considering the small scale of many WSI datasets, and thus, finding approaches to allow for more efficient training is important. To this end, we propose a modular and efficient method to augment patch features in the context of MIL training.

Latent augmentations for MIL — have been proposed to mitigate the limitations expressed previously. By augmenting features directly in latent space to simulate real image augmentations, the benefits of data augmentations can be maintained while controlling computational costs. Feature-level augmentation methods can be divided into three fami-

lies: (i) feature mixing, (ii) feature generation, and (iii) on-line representation sampling. Techniques in the first category rely on MixUp [35] to interpolate features between instances [5], instance prototypes [17, 32] or WSIs [12], and thus cannot perform standard augmentations such as geometric transforms, color jittering or H&E-tailored transforms, known to be useful in pathology [27]. Generative approaches are mainly based on GANs [34], or diffusion models [6]. Diffusion-based methods, such as AugDiff [6] which is considered as the current state of the art, introduce controlled noise into feature embeddings and iteratively refine them. Their effectiveness at inference depends critically on the number of diffusion steps: too few results in minimal augmentation, while too much alters features. Additionally, diffusion models are computationally expensive in both time and memory, particularly in MIL settings where entire bags of features must be augmented, significantly limiting scalability. Both GANs and diffusion models lack explicit control over specific transformations, making them less effective for MIL tasks that require precise augmentation strategies. Recent self-supervised learning (SSL) approaches, such as SSRDL [26], fall in the third category. SSRDL introduces an online representation sampling strategy to enhance MIL feature diversity. However, it requires training a dedicated patch encoder, making it incompatible with foundation models like UNI or CONCH, which offer strong generalization across pathology datasets.

To summarize, existing feature augmentation methods either lack transformation controllability, produce unrealistic augmentations, or require specialized feature extractors. Our work introduces a transformer-based generative model for feature augmentation that explicitly conditions augmentations on transformations while leveraging foundation models for feature extraction. Unlike previous methods, our approach enables consistent bag-wise augmentation while efficiently scaling to large datasets, providing a robust solution for MIL in histopathology.

3. Method

Performing latent augmentations is particularly convenient in the context of MIL training where standard image-level augmentations can be time and memory-consuming. We propose to train a generator to transform latent features from a frozen foundation encoder into an augmented version. It is thus critical to learn to simulate standard image augmentations in latent space without distorting the information initially present in the image embedding. An overview of the method is presented in Figure 2.

3.1. Problem Setup

Let us consider a set of N available image augmentations $\{T_i\}_{i=1}^N$, and for each T_i , a space of available parameter values A_{T_i} defining how the augmentation should be applied

(e.g. the space of hue shifts when considering a hue transform). For transformations that only have a finite number of parameter values (e.g. flipping), A_{T_i} will be a set of one-hot vectors.

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ be a patch from a WSI, and let $\mathcal{E}$ be a frozen feature extractor (i.e., foundation model), parametrized by weight $\theta_{\mathcal{E}}$, producing an embedding $\mathbf{z} = \mathcal{E}(\mathbf{x}; \theta_{\mathcal{E}}) \in \mathbb{R}^d$. We denote the function applying K image transforms sequentially τ and define it as,

$$\tau(\mathbf{x}; (T_k, \alpha_k)_{k=1}^K) = T_K(\cdots T_2(T_1(\mathbf{x}; \alpha_1); \alpha_2) \cdots; \alpha_K), \quad (1)$$

where $(T_k, \alpha_k)_{k=1}^K$ is a sequence of transformations together with their associated parameter values, i.e. $T_k \sim \{T_i\}_{i=1}^N$ is a single transformation (e.g. change in hue) with parameter $\alpha_k \sim A_{T_k}$ (e.g. hue shift value).

Our goal is to train a generator ρ parametrized by weights θ_{ρ} to approximate the embedding of a transformed patch produced by the encoder $\mathcal{E}$:

$$\rho(\mathbf{z}, (T_k, \alpha_k)_{k=1}^K; \theta_{\rho}) \approx \mathcal{E}(\tau(\mathbf{x}; (T_k, \alpha_k)_{k=1}^K; \theta_{\mathcal{E}})). \quad (2)$$

While simulating real image transformations in feature space, the generator should maintain as much of the initial encoded information as possible. As we will present later in this section, this will be enforced by training ρ to reconstruct the identity, i.e. the input latent feature, if all parameters for the different transforms in the sequence have a specific value, i.e. if $(\alpha_k)_{k=1}^K = (\alpha_{\text{id},k})_{k=1}^K$, where $\alpha_{\text{id},k}$ is the parameter value for transformation T_k leading to identity transform to be applied (i.e. the transform is not applied). We will then want the following,

$$\rho(\mathbf{z}, (T_k, \alpha_{\text{id},k})_{k=1}^K; \theta_{\rho}) \approx \mathbf{z}. \quad (3)$$

3.2. Generator Architecture

Chunked Inputs — Since $\mathbf{z}$ is high-dimensional, we split it into C segments,

$$\mathbf{z} \mapsto (\mathbf{z}_i)_{i=1}^C, \quad \mathbf{z}_i \in \mathbb{R}^{\frac{d}{C}}. \quad (4)$$

Each chunk $\mathbf{z}_i$ is treated as a separate token in a transformer-based architecture. An ablation study on the interest of chunking is presented in Figure S2 of the Supplementary Material. We encode the order of each token with sinusoidal positional encoding (PE) [28].

Transformation Embeddings & Cross-Attention — Suppose we have a sequence of K transformations $(T_k)_{k=1}^K$, each with parameters α_k . For each T_k , we encode α_k into a parameter vector $\mathbf{p}_k \in \mathbb{R}^{\frac{d}{C}}$ with a linear projection layer φ_{T_k} parametrized with weights $\theta_{\varphi}^{T_k}$. We have one projection layer per type of transformation:

$$\mathbf{p}_k = \varphi_{T_k}(\alpha_k; \theta_{\varphi}^{T_k}) = \theta_{\varphi}^{T_k} \cdot \alpha_k \in \mathbb{R}^{\frac{d}{C}}, \quad (5)$$

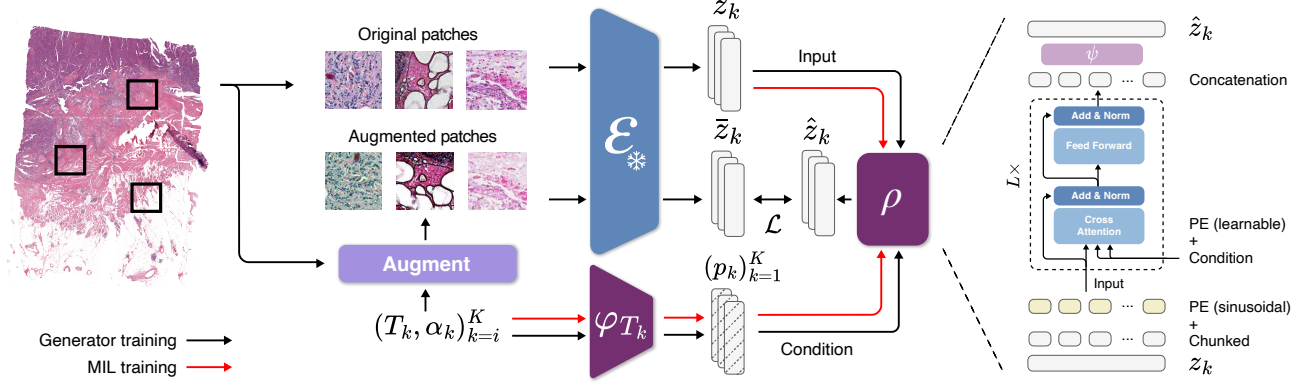


Figure 2. **HistAug Overview:** A frozen foundation model $\mathcal{E}$ (e.g., UNI) is used to encode both the original and augmented tissue patches to the latent spaces z_k and $\tilde{z}_k$ respectively. The augmentation hyperparameters together with the original features are passed to a conditional transformer architecture that simulates the augmentation in the latent space producing $\hat{z}_k$. The whole pipeline is trained with a two-term loss function that combines a reconstruction and an identity loss. During MIL training, ρ and the φ_{T_k} are frozen, and all operations are performed in the latent space.

where α_k is a one-hot vector for transformations with a discrete number of parameters, and $\theta_{\varphi}^{T_k} \subset \theta_{\rho}$. To explicitly capture the order of transformations, we also add learnable positional embeddings to the $(p_k)_{k=1}^K$.

The generator is composed of a sequence of L transformer blocks $(\phi_j)_{j=1}^L$. Each ϕ_j block is parametrized by weights θ_{ϕ}^j and performs cross-attention from the chunked tokens (queries) to the transformation tokens (keys/values), followed by a residual (skip) connection, layer normalization, and a feed-forward MLP with another skip connection,

$$(\tilde{z}_i^{j+1})_{i=1}^C = \phi_j((\tilde{z}_1^j)_{i=1}^C, (p_k)_{k=1}^K; \theta_{\phi}^j), \quad (6)$$

where $\tilde{z}_i^j$ is the representation token at block j ($\tilde{z}_i^1 = z_i$), and $\theta_{\phi}^j \subset \theta_{\rho}$. Such cross-attention mechanism allows sharing information between image features and augmentation parameter embeddings. After such L blocks, we concatenate the updated chunked tokens back into $\mathbb{R}^d$, then apply an MLP head ψ to obtain the final augmented feature as,

$$\hat{z} = \psi([\tilde{z}_1^L, \dots, \tilde{z}_C^L]; \theta_{\psi}), \quad (7)$$

where $[\cdot]$ is the concatenation operator, and $\theta_{\psi} \subset \theta_{\rho}$.

3.3. Objective Function

The generator is trained to reconstruct the features produced by the frozen encoder $\mathcal{E}$ when given an augmented image, effectively mimicking the transformations we apply to the input. Additionally, it is trained to retain as much of the original feature information as possible, minimizing any distortion of the original latent representation. This is implemented as the combination of a **reconstruction loss** and

an **identity loss**:

$$\mathcal{L} = \underbrace{\|\rho(z, (T_k, \alpha_k)_{k=1}^K; \theta_{\rho}) - \mathcal{E}(\tau(x; (T_k, \alpha_k)_{k=1}^K; \theta_{\mathcal{E}}))\|_2^2}_{\text{Reconstruction}} + \lambda_{\text{id}} \underbrace{\|\rho(z, (T_k, \alpha_{\text{id}, k})_{k=1}^K; \theta_{\rho}) - z\|_2^2}_{\text{Identity}}. \quad (8)$$

The first term enforces that when the patch x is transformed by $T_1, \dots, T_K$ (with parameters $\alpha_1, \dots, \alpha_K$), the output from the generator aligns with the embedding from $\mathcal{E}$. The second term ensures that when no transformation is applied, the generator recovers the original embedding z .

3.4. Integration into MIL Training

Once ρ is trained, we can augment any patch embedding in feature space, without recomputing image features. This allows us to efficiently augment data when training a model to aggregate patch features with MIL. Let $\{z_m\}_{m=1}^M$ be a bag of M embeddings extracted from a WSI. We distinguish two ways to augment patch features at training time:

Instance-wise — Each patch is augmented with a different sequence of K transformations sampled at random. For a given patch z_m , we thus have a specific sequence of transformations $(T_k^m)_{k=1}^K$ and parameters $(\alpha_k^m)_{k=1}^K$, yielding $\hat{z}_m = \rho(z_m, (T_k^m, \alpha_k^m)_{k=1}^K; \theta_{\rho})$.

Bag-wise (WSI-wise) — The same sequence of transformations $(T_k)_{k=1}^K$ and parameters $(\alpha_k)_{k=1}^K$ is sampled for all patches in the bag. We thus have $\hat{z}_m = \rho(z_m, (T_k, \alpha_k)_{k=1}^K; \theta_{\rho})$.

While the *Instance-Wise* mode increases data diversity within a bag, an advantage of augmenting *WSI-Wise* is to preserve global consistency across the whole slide.

Finally, the augmented bag $\{\hat{z}_m\}_{m=1}^M$ can be fed into any

MIL model (e.g., ABMIL [15], CLAM [19], DSMIL [16], TransMIL [24]). Since the generator is lightweight, we can augment all patches from a WSI at training time in a single forward pass with minimal overhead.

4. Experiments and Implementation Details

4.1. Datasets

In this study, we use five TCGA datasets: BLCA (Bladder Urothelial Carcinoma), BRCA (Breast Invasive Carcinoma), NSCLC (Non-Small Cell Lung Carcinoma), UCEC (Uterine Corpus Endometrial Carcinoma) and KIRC (Kidney Renal Clear Cell Carcinoma). For training the generator, BLCA, BRCA, and LUSC (Lung Squamous Cell Carcinoma) from NSCLC are used: 60% of WSIs are sampled from each dataset, resulting in approximately 1200 training WSIs. Slides are processed at 10 $\times$ and 20 $\times$ magnification using the CLAM toolbox [19]. We evaluated our method in two different downstream tasks: cancer subtyping (BRCA, NSCLC) and survival analysis (BLCA, UCEC, KIRC).

4.2. Generator Training

We train two separate generators ρ_{UNI} and ρ_{CONCH} for the UNI [4] and CONCH [20] feature spaces, respectively. In our experiments, AugDiff [6] is trained on the same dataset.

Transformations — We apply a diverse set of transformations to image patches, extract their features using the respective feature extractor, and train the generator to reconstruct the augmented feature while being conditioned on both the original feature and transformation parameters. Each transformation is parameterized, and multiple transformations are applied sequentially in a randomized manner, ensuring a rich augmentation space. The generator learns to predict the feature embedding corresponding to the transformed patch while preserving the semantic structure of the input. The transformations we consider include: (i) **geometric transformations** such as rotations, horizontal and vertical flips, random cropping, as well as morphological operations such as dilation and erosion; (ii) **color transformations** that modify the brightness, contrast, hue, gamma, and saturation; (iii) **histology-specific transformation** specifically the HED transform [10], designed particularly for H&E-stained images. Transformation details are shown in Table S1 from the Supplementary Material.

4.3. MIL Training

We train MIL models using both UNI [4] and CONCH [20] foundation models as vision encoders. We generate five different data folds for each dataset and evaluate five MIL architectures: Abmil [15], CLAM [19], DSMIL [16], and TransMIL [24], each leveraging different aggregation mechanisms to predict slide-level outcomes. For each MIL model, we compare six augmentation settings: **Base**

Table 1. **Generator Evaluation:** Mean Cosine similarity between original and generated augmented features over 10k patches at 10X and 20X, alongside feature extractor invariance. The † symbol denotes out of training distribution.

		Feature Reconstruction		Feature Invariance	
		10 $\times$	20 $\times$ †	10 $\times$	20 $\times$ †
UNI	BLCA	81.0	74.9	11.9	11.6
	BRCA	81.6	75.8	11.9	11.7
	LUSC	81.1	74.9	12.6	11.9
	LUAD †	80.3	74.6	15.0	11.3
	UCEC †	80.5	73.4	9.5	12.9
	KIRC †	80.5	72.1	13.0	11.0
CONCH	BLCA	90.3	88.1	20.3	27.7
	BRCA	90.4	88.6	24.3	32.7
	LUSC	90.4	87.0	19.3	27.7
	LUAD †	90.2	87.6	24.9	23.7
	UCEC †	90.4	88.5	19.7	29.6
	KIRC †	89.9	87.5	23.1	29.8

Table 2. **Performance comparison across MIL models.** Mean and standard deviation of mean performance metrics aggregated across MIL models (Abmil, ClamMb, ClamSb, DSMIL, TransMIL). For BLCA, KIRC, and UCEC, C-index (%) is used, while the AUC (%) is used for the remaining datasets.

		BLCA	KIRC	UCEC	BRCA	NSCLC
UNI	10% Training					
	Base	47.5 $\pm$ 1.9	58.5 $\pm$ 2.7	59.3 $\pm$ 4.7	86.1 $\pm$ 0.6	87.6 $\pm$ 3.8
	AugD	49.9 $\pm$ 2.3	62.8 $\pm$ 4.7	61.9 $\pm$ 4.7	84.1 $\pm$ 5.5	86.8 $\pm$ 3.9
	PAug	48.4 $\pm$ 1.4	60.1 $\pm$ 2.9	60.9 $\pm$ 5.5	88.2 $\pm$ 1.1	88.9 $\pm$ 4.3
	Ours (Inst)	50.5 $\pm$ 1.0	61.3 $\pm$ 2.7	61.3 $\pm$ 5.9	88.2 $\pm$ 1.1	89.7 $\pm$ 3.1
	Ours (WSI)	50.6 $\pm$ 1.6	62.5 $\pm$ 2.5	63.2 $\pm$ 3.6	88.3 $\pm$ 0.9	90.4 $\pm$ 3.6
	100% Training					
	Base	54.5 $\pm$ 3.7	65.8 $\pm$ 2.4	63.8 $\pm$ 4.2	92.7 $\pm$ 0.9	97.7 $\pm$ 1.1
	PAug	56.7 $\pm$ 3.7	66.9 $\pm$ 1.5	64.7 $\pm$ 2.2	93.3 $\pm$ 0.3	93.5 $\pm$ 0.9
	Ours (Inst)	60.3 $\pm$ 2.3	67.6 $\pm$ 1.5	65.5 $\pm$ 3.5	92.4 $\pm$ 0.2	97.3 $\pm$ 0.8
	Ours (WSI)	59.5 $\pm$ 2.6	68.2 $\pm$ 1.1	64.8 $\pm$ 2.6	92.4 $\pm$ 0.3	97.7 $\pm$ 0.4
CONCH	10% Training					
	Base	50.8 $\pm$ 2.2	63.1 $\pm$ 3.0	58.6 $\pm$ 3.2	89.2 $\pm$ 3.5	92.8 $\pm$ 1.3
	AugD	53.0 $\pm$ 1.7	65.9 $\pm$ 2.9	61.9 $\pm$ 0.9	90.1 $\pm$ 3.8	93.8 $\pm$ 0.5
	PAug	54.2 $\pm$ 2.0	65.3 $\pm$ 4.0	64.5 $\pm$ 0.5	88.1 $\pm$ 3.0	93.2 $\pm$ 0.8
	Ours (Inst)	54.5 $\pm$ 1.6	68.4 $\pm$ 3.1	63.7 $\pm$ 1.0	90.4 $\pm$ 1.2	93.9 $\pm$ 0.9
	Ours (WSI)	54.1 $\pm$ 3.0	69.6 $\pm$ 3.0	64.9 $\pm$ 0.9	90.8 $\pm$ 1.7	94.6 $\pm$ 1.1
	100% Training					
	Base	58.0 $\pm$ 2.1	68.5 $\pm$ 1.5	60.1 $\pm$ 1.3	93.0 $\pm$ 1.7	97.9 $\pm$ 0.3
	PAug	61.3 $\pm$ 1.5	69.9 $\pm$ 1.3	64.8 $\pm$ 5.0	92.4 $\pm$ 0.5	94.6 $\pm$ 1.1
	Ours (Inst)	63.5 $\pm$ 1.6	70.1 $\pm$ 1.2	63.7 $\pm$ 2.9	93.5 $\pm$ 0.7	98.1 $\pm$ 0.3
	Ours (WSI)	62.5 $\pm$ 0.3	71.2 $\pm$ 1.7	65.2 $\pm$ 1.7	93.7 $\pm$ 0.5	97.8 $\pm$ 0.6

refers to models trained without augmentation. **AugDiff (AugD)** applies feature augmentation using the state-of-the-art AugDiff method [6]. **Patch augmentation (PAug)** applies transformations in image space offline, and stores the augmented features before MIL training. HistAug (Ours) includes both **instance-wise (Inst)** and **WSI-wise (WSI)**

feature augmentation, leveraging our trained generator. The last baseline performs **noise-based perturbations (Noise)**.

4.4. Evaluation Protocol

Generator Evaluation — We evaluate the performance of the generator using two key metrics. **Feature reconstruction** where we randomly sample 10,000 patches from WSIs that were not used during training and we compute the cosine similarity between the generated augmented feature $\hat{z}$ and the original augmented feature $\bar{z}$. A high cosine similarity indicates that the generator accurately replicates the effect of image-space transformations in feature space. **Feature extractor invariance** where we compute the cosine similarity between the original embedding $z = \mathcal{E}(x; \theta_\varepsilon)$ and the real augmented feature $\bar{z}$. Lower similarity values suggest that the feature extractor is sensitive to transformations in its latent space [9, 30].

Downstream Tasks — For each model, we conduct training across five different folds. Area Under the Curve (AUC) and C-Index are reported for subtyping and survival analysis respectively. Moreover, we perform experiments on two different data regimes namely (i) **10% data regime** at which the models are trained using only 10% of the available labeled training samples, while the testing sets remain the same and (ii) **100% data regime** where the models are trained on the full set of available training data.

5. Results

5.1. Generator Evaluation

Table 1 reports our performance in reconstructing augmented features at $10\times$ and $20\times$ magnification. Results demonstrate high cosine similarity between original and generated augmented features, consistently exceeding 80 for UNI and approximately 90 for CONCH. This is achieved despite the significant impact of the transformations on the latent space of feature extractors, i.e. cosine similarity drops to approximately 11 to 15 for UNI and around 20 for CONCH, highlighting the sensitivity of these foundation models to sequences of transformations. This implies that, despite the non-invariance of the UNI and CONCH feature spaces, our generator can successfully capture and reconstruct these altered representations.

Importantly, although our generator was trained using patches extracted at $10\times$ magnification, it can also reconstruct features from images at a higher magnification (i.e., $20\times$). As seen in Table 1, the cosine similarity at $20\times$ remains high (approximately 75 for UNI and 88 for CONCH), showcasing strong cross-scale generalization capabilities.

Furthermore, the effectiveness of the generator extends to external datasets on different organs. For instance, results on external lung (LUAD), kidney (KIRC), and endometrial (UCEC) datasets demonstrate similarly high cosine similar-

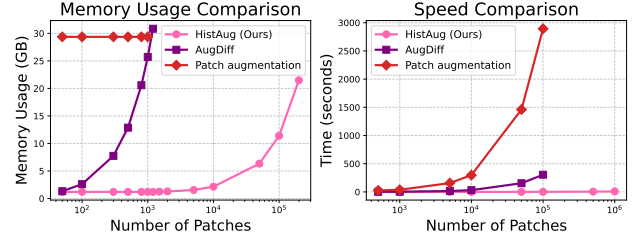


Figure 3. **Comparison in forward pass efficiency:** HistAug is significantly faster than AugDiff [6] and requires less GPU memory to augment a set of features in a single parallelized forward pass. The experiments were conducted on a V100 32GB GPU

ity (around 80 for UNI and 90 for CONCH at $10\times$), underscoring the robustness and generalizability of our generator to diverse histopathology datasets and tissue types unseen at training time. Tables S5 & S6 in the Supplementary Material provide cosine similarity bootstrap confidence intervals.

5.2. MIL Evaluation

Tables 2 and 3 report MIL performance averaged across MIL models and datasets. At 10% training data, instance-wise and WSI-wise augmentations with our method consistently outperform the baseline, AugDiff and PAug across nearly all cancer types and vision encoders. For instance, WSI-wise augmentation improves survival prediction (C-index) notably, from 59.3 to 63.2 for UNI and from 58.6 to 64.9 for CONCH in UCEC. Similar performance improvements are observed in classification tasks at 10% training data, with WSI-wise augmentation achieving gains such as an increase from 86.1 to 88.3 for UNI in BRCA and from 92.8 to 94.6 for CONCH in NSCLC. Moreover, HistAug matches or outperforms PAug as it introduces more diverse augmentations on-the-fly during training, unlike PAug which relies on precomputed and stored features.

At 100% training data, our method also yields improvements, particularly for survival prediction tasks e.g., instance-wise augmentation improves UNI-based BLCA from 54.5 to 59.5. For classification tasks, improvements are smaller, likely due to the already high baseline performance leaving limited room for further gains. AugDiff is not evaluated at 100% training data due to its high computational cost (at $10\times$ magnification it requires over 60 GPU hours per fold for BLCA dataset, see Section 5.3). An extensive experiment with AugDiff with 100% of the data can be found in the Supplementary Material in Table S4. Additional results on the BRACS dataset [2] can be found in the Supplementary Material in Table S3.

Finally, we compare our method with SSRDL [26], a state-of-the-art approach for Online Representation Sampling (ORS), on the TCGA-EGFR dataset (Table 4), as it is the only one with data splits provided in [26]. A crucial

Table 3. **Performance comparison across datasets.** Mean and standard deviation of mean performance metrics aggregated across tasks (BLCA, KIRC, UCEC for survival; BRCA, NSCLC for classification) are reported for each MIL model and feature augmentation strategy.

Augmentation	UNI					CONCH				
	Abmil	ClamMb	ClamSb	DSMIL	TransMIL	Abmil	ClamMb	ClamSb	DSMIL	TransMIL
Survival (C-Index, %) – 10% Training										
Base	52.5 $\pm$ 4.6	55.2 $\pm$ 4.3	51.8 $\pm$ 4.3	57.5 $\pm$ 8.0	58.5 $\pm$ 6.6	55.4 $\pm$ 4.4	56.9 $\pm$ 4.3	57.4 $\pm$ 7.1	58.8 $\pm$ 7.2	59.0 $\pm$ 4.2
AugD	58.3 $\pm$ 3.7	59.4 $\pm$ 8.0	56.3 $\pm$ 5.0	60.5 $\pm$ 9.7	56.5 $\pm$ 6.7	59.8 $\pm$ 5.7	60.1 $\pm$ 6.1	59.7 $\pm$ 6.5	62.6 $\pm$ 6.0	59.3 $\pm$ 3.1
PAug	53.9 $\pm$ 4.1	54.9 $\pm$ 5.0	54.0 $\pm$ 3.9	58.8 $\pm$ 9.0	60.7 $\pm$ 7.3	58.9 $\pm$ 5.7	62.5 $\pm$ 5.4	61.4 $\pm$ 5.5	64.3 $\pm$ 5.8	59.5 $\pm$ 4.3
Ours (Inst)	56.0 $\pm$ 5.1	59.1 $\pm$ 6.0	53.7 $\pm$ 3.3	59.7 $\pm$ 6.6	59.9 $\pm$ 7.4	60.3 $\pm$ 5.7	63.7 $\pm$ 5.5	62.9 $\pm$ 6.9	63.3 $\pm$ 6.6	60.9 $\pm$ 4.9
Ours (WSI)	56.7 $\pm$ 6.0	58.2 $\pm$ 6.2	57.9 $\pm$ 3.6	60.7 $\pm$ 7.0	60.4 $\pm$ 7.4	61.9 $\pm$ 6.9	64.5 $\pm$ 5.4	61.7 $\pm$ 9.2	64.8 $\pm$ 6.4	61.3 $\pm$ 5.3
Survival (C-Index, %) – 100% Training										
Base	60.6 $\pm$ 6.5	62.5 $\pm$ 3.2	60.4 $\pm$ 6.7	59.5 $\pm$ 7.6	63.7 $\pm$ 4.2	60.8 $\pm$ 5.1	63.6 $\pm$ 4.7	62.3 $\pm$ 3.0	60.5 $\pm$ 5.3	63.8 $\pm$ 4.9
PAug	63.6 $\pm$ 3.8	61.9 $\pm$ 5.4	63.1 $\pm$ 3.1	59.9 $\pm$ 6.9	65.5 $\pm$ 3.6	65.8 $\pm$ 3.6	68.7 $\pm$ 4.2	66.4 $\pm$ 3.5	64.0 $\pm$ 4.1	61.7 $\pm$ 4.8
Ours (Inst)	63.1 $\pm$ 4.0	61.9 $\pm$ 4.3	66.9 $\pm$ 2.9	66.2 $\pm$ 4.1	64.3 $\pm$ 1.3	65.2 $\pm$ 3.9	67.8 $\pm$ 3.1	66.0 $\pm$ 3.1	66.4 $\pm$ 3.0	63.4 $\pm$ 3.9
Ours (WSI)	63.1 $\pm$ 4.8	62.7 $\pm$ 5.1	65.0 $\pm$ 2.0	65.4 $\pm$ 4.9	64.7 $\pm$ 2.7	67.3 $\pm$ 3.6	67.6 $\pm$ 4.3	65.5 $\pm$ 3.8	66.3 $\pm$ 4.3	65.0 $\pm$ 2.3
Classification (AUC, %) – 10% Training										
Base	87.5 $\pm$ 1.8	89.0 $\pm$ 2.0	88.8 $\pm$ 2.4	83.2 $\pm$ 2.0	85.8 $\pm$ 0.6	93.4 $\pm$ 1.8	91.7 $\pm$ 1.4	91.3 $\pm$ 1.1	86.8 $\pm$ 4.5	91.7 $\pm$ 0.3
AugD	88.2 $\pm$ 0.9	87.3 $\pm$ 0.1	88.2 $\pm$ 1.9	76.2 $\pm$ 3.0	87.3 $\pm$ 0.7	92.8 $\pm$ 0.9	93.2 $\pm$ 0.6	93.5 $\pm$ 1.1	88.1 $\pm$ 5.6	92.1 $\pm$ 0.9
PAug	89.3 $\pm$ 2.4	91.1 $\pm$ 1.6	91.2 $\pm$ 1.7	85.7 $\pm$ 1.4	85.4 $\pm$ 2.4	91.2 $\pm$ 3.0	92.1 $\pm$ 1.3	91.1 $\pm$ 3.0	87.6 $\pm$ 5.1	91.5 $\pm$ 0.5
Ours (Inst)	90.3 $\pm$ 0.9	91.0 $\pm$ 1.6	90.3 $\pm$ 2.1	87.1 $\pm$ 0.1	85.9 $\pm$ 1.1	92.7 $\pm$ 1.8	93.0 $\pm$ 1.8	92.6 $\pm$ 1.9	90.5 $\pm$ 2.5	92.0 $\pm$ 0.8
Ours (WSI)	90.8 $\pm$ 1.6	91.3 $\pm$ 2.4	91.4 $\pm$ 2.4	86.3 $\pm$ 0.7	87.1 $\pm$ 0.4	93.3 $\pm$ 1.6	93.6 $\pm$ 2.1	93.5 $\pm$ 2.1	90.6 $\pm$ 3.2	92.6 $\pm$ 0.4
Classification (AUC, %) – 100% Training										
Base	95.4 $\pm$ 2.4	95.8 $\pm$ 2.8	96.0 $\pm$ 2.5	94.5 $\pm$ 3.6	94.2 $\pm$ 1.3	95.9 $\pm$ 2.1	96.3 $\pm$ 2.1	95.8 $\pm$ 1.7	94.0 $\pm$ 4.2	95.3 $\pm$ 2.3
PAug	93.7 $\pm$ 0.1	93.5 $\pm$ 0.5	93.8 $\pm$ 0.6	92.7 $\pm$ 0.8	93.5 $\pm$ 0.1	93.8 $\pm$ 1.0	94.1 $\pm$ 1.2	93.8 $\pm$ 1.2	93.9 $\pm$ 1.4	91.9 $\pm$ 0.6
Ours (Inst)	95.0 $\pm$ 2.9	94.8 $\pm$ 2.3	95.0 $\pm$ 2.6	95.3 $\pm$ 2.6	94.1 $\pm$ 1.7	95.9 $\pm$ 2.5	96.0 $\pm$ 2.5	95.8 $\pm$ 1.9	96.3 $\pm$ 1.9	95.0 $\pm$ 2.8
Ours (WSI)	94.8 $\pm$ 2.7	95.0 $\pm$ 2.5	94.9 $\pm$ 2.4	95.4 $\pm$ 3.0	95.3 $\pm$ 2.4	96.2 $\pm$ 2.2	96.1 $\pm$ 2.4	96.1 $\pm$ 1.9	95.7 $\pm$ 1.8	94.8 $\pm$ 2.0

advantage of *HistAug* is that, unlike SSRDL, it does not require to train a patch encoder along with the augmentation model. Instead, we can leverage powerful, histopathology-specific foundation models such as UNI to train a strong baseline already outperforming SSRDL, which is then improved with our augmented features. Tables S7 & S8 in the Supplementary Material provide detailed results.

5.3. Speed and Memory Comparison

Since latent augmentation is performed at every training step in MIL, computational efficiency is a key consideration. For this reason, we compare the time and memory requirements of *HistAug*, AugDiff, and naive patch augmentation — in which transformations are applied in the pixel space and features are then extracted by a foundation model — for augmenting a bag of features on a single 32GB V100 GPU. As shown in Figure 3, *HistAug* is significantly faster than AugDiff while using much less GPU memory. *HistAug* can process up to 1M input patches in less than 10 seconds, while in the same time, AugDiff can only augment fewer than 5k patches. When considering a realistic number of patches as required in MIL training, i.e. 50k–100k patches, *HistAug* is approximately 300 $\times$ faster than AugDiff. Moreover, AugDiff reaches memory saturation on a 32GB GPU when augmenting just 1k patches in parallel, while *HistAug* can handle batches of up to 200k patches before reaching the same limit. Furthermore, patch augmentation scales

Table 4. **Comparison with SSRDL** on the TCGA-EGFR dataset. Performance is measured using ROC AUC.

Model	SSDRL	Baseline (UNI)	Ours (UNI)
TransMIL	79.7	86.5	87.9
CLAM	83.1	86.5	89.4

poorly, with high GPU memory usage ($\sim$ 32 GB) and long runtimes (3k s for 100k patches), making it impractical for MIL due to repeated foundation model forward passes. In addition to leading to better MIL performance, *HistAug* is thus more convenient to use at training time.

5.4. Underlying properties of latent augmentations

Cross-magnification Generalization — Our generator trained at 10 $\times$ magnification also improves MIL performance when applied at 20 $\times$ without additional training (Table 5). For example, WSI-wise augmentation significantly improves the C-index in UCEC from 60.5 to 67.1 (CONCH embeddings) in this out-of-distribution setting.

Superiority to Noise Perturbation — To verify that our method is not simply altering features without meaningful structure, we introduced a noise perturbation baseline which adds random noise to features. As shown in Table 6, our method (Inst and WSI) consistently outperforms it across both 10% and 100% training data settings. Noise pertur-

Table 5. **Gains at the $20\times$ OOD magnification** from augmented features highlight the generalizability properties of *HistAug*.

Model	Survival (C-Index)						Classification (AUC)			
	BLCA		KIRC		UCEC		BRCA		NSCLC	
	10%	100%	10%	100%	10%	100%	10%	100%	10%	100%
UNI	+2.9	+4.3	+3.6	+2.6	+3.6	+3.5	+1.7	+0.1	+0.8	+0.8
CONCH	+4.5	+8.2	+4.3	+2.8	+3.0	+6.6	+0.5	+0.5	+2.6	+0.0

Table 6. **Gain over Noise** for UNI and CONCH at 10% vs. 100% training at $10\times$ magnification for the two *HistAug* variants.

	10% UNI		10% CONCH		100% UNI		100% CONCH	
	Inst	WSI	Inst	WSI	Inst	WSI	Inst	WSI
BLCA	+3.0	+3.1	+2.8	+2.3	+4.3	+3.5	+4.3	+3.3
KIRC	+3.1	+4.3	+5.2	+6.4	+1.7	+2.3	+0.5	+1.5
UCEC	+2.3	+4.2	+4.9	+6.1	+2.6	+1.9	+2.7	+4.2
BRCA	+1.6	+1.7	-0.4	+0.3	-0.5	-0.5	+0.1	+0.3
NSCLC	+0.7	+1.4	+1.3	+2.0	-0.2	+0.2	+1.1	+0.8

bation does not significantly improve performance, showing that random perturbations alone are insufficient. This further underscores the effectiveness and necessity of our learned augmentation strategies. Tables S7 and S8 in the Supplementary Material present full results.

5.5. Visualizing the learned latent transformations

Trajectories in augmentation space — An important characteristic of *HistAug* is controllability. We visualize how augmented features evolve when navigating the space of parameter values. For a given patch and transformation, we sample a set of parameter values and, for each of them, both apply the original transformation followed by feature extraction with the encoder, and generate the augmented features with *HistAug*. We apply a 2-dim PCA on the whole set of original and generated augmented features and visualize the trajectory of latent codes. Trajectories in Figure 4 are close between original and generated features, highlighting the controllability of *HistAug*. Additional visualizations are shown in Figure S1 in the Supplementary Material.

Augmented image retrieval — is used to visualize the generated features from *HistAug*. We augment the features from a patch with *HistAug* given a transformation and associated parameter, and perform original image transformations on the same image patch for all considered augmentation types and parameter values. We then compute the cosine similarity between the generated features and the ones extracted from all the original image augmentations. As shown in Figure 5, for different patches and augmentations (hue, contrast, gaussian blue, erosion), top-1 retrieved images from generated features are correct, showing that

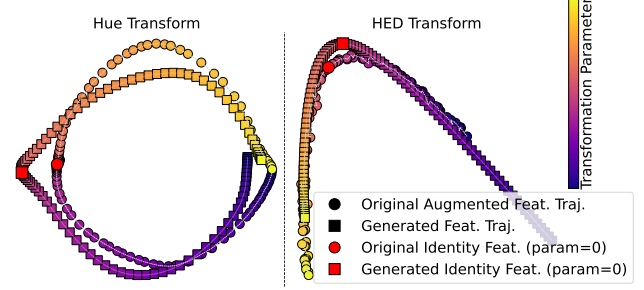


Figure 4. **Trajectories in augmentation space:** Left – PCA trajectories of Hue transform for UNI. Right – PCA trajectories of HED transform for CONCH. Circles represent the original augmented features extracted from the foundation model, while squares denote the features generated by our augmentation model.

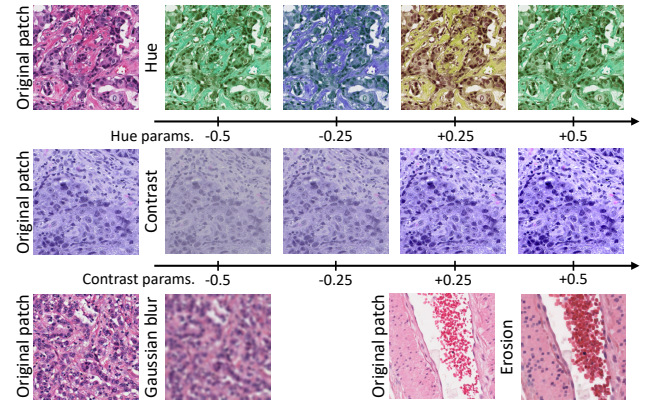


Figure 5. **Augmented image retrieval:** Visualization of the top-1 retrieved augmented images based on cosine similarity with *HistAug* augmented features. As all retrieved images are correct, they correspond to the expected transformed patches for the given augmentation and parameter.

HistAug properly simulates standard augmentations, even proper variations based on provided parameters.

6. Conclusion

Latent augmentation is promising in compute-demanding MIL training on small-scale datasets. We introduce *HistAug*, a lightweight generator performing controllable augmentations in latent space. We improve MIL performance across models and datasets while being faster and less memory intensive than state-of-the-art diffusion-based counterparts. Additional studies demonstrate the ability of *HistAug* to generalize to magnifications unseen during training, and its superiority against noise-based techniques. Qualitative studies highlight the controllability of our latent augmentations. Future work will study how our method can be extended to more types of augmentations, and how its controllability can further be improved.

Acknowledgments — This research was supported by the *French National Research Agency (ANR)* under project *ANR-23-CE45-0029*, and the *Health Data Hub (HDH)* as part of the second edition of the *France-Québec* call for projects *Intelligence Artificielle en santé*. It was carried out using HPC resources from *GENCI-IDRIS* (Grant 2024-AD011015593).

References

- [1] Maximilian Alber, Stephan Tietz, Jonas Dippel, Timo Milbich, Timothée Lesort, Panos Korfiatis, Moritz Krügener, Beatriz Perez Cancer, Neelay Shah, Alexander Möllers, et al. A novel pathology foundation model by mayo clinic, charit'e, and aignostics. *arXiv preprint arXiv:2501.05409*, 2025. 1, 2
- [2] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022. 6, 2, 4
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2
- [4] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 1, 2, 5, 4
- [5] Yuan-Chih Chen and Chun-Shien Lu. Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23936–23945, 2023. 3
- [6] Liuxi Dai, Yifeng Wang, Haoqian Wang, Yongbing Zhang, et al. Augdiff: Diffusion based feature augmentation for multiple instance learning in whole slide image. *IEEE Transactions on Artificial Intelligence*, 2024. 1, 2, 3, 5, 6
- [7] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahnong Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv*, 2024. 2
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [9] Matouš Elphick, Samra Turajlic, and Guang Yang. Are the latent representations of foundation models for pathology invariant to rotation?, 2024. 2, 6
- [10] Khrystyna Faryna, Jeroen van der Laak, and Geert Litjens. Tailoring automated data augmentation to h&e-stained histopathology. In *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, pages 168–178. PMLR, 2021. 5, 2
- [11] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024. 1, 2
- [12] Michael Gadermayr, Lukas Koller, Maximilian Tschuchnig, Lea Maria Stangassinger, Christina Kreutzer, Sebastien Couillard-Despres, Gertie Janneke Oostingh, and Anton Hittmair. Mixup-mil: A study on linear & multilinear interpolation-based data augmentation for whole slide image classification. *arXiv preprint arXiv:2311.03052*, 2023. 3
- [13] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 2023. 2
- [14] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *NeurIPS*, 2023. 2
- [15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2018. 1, 2, 5, 3
- [16] 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*, pages 14318–14328, 2021. 5, 2, 3
- [17] Pei Liu, Luping Ji, Xinyu Zhang, and Feng Ye. Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. *IEEE Transactions on Medical Imaging*, 43(5):1841–1852, 2024. 3
- [18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1
- [19] Ming Y Lu, Drew FK 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(6):555–570, 2021. 1, 2, 5, 3
- [20] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 2024. 1, 2, 5, 4
- [21] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology. *arXiv preprint arXiv:2406.05074*, 2024. 1, 2
- [22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

- [23] George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv*, 2024. 2
- [24] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 1, 2, 5, 3
- [25] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12): 930–949, 2023. 1
- [26] Kunming Tang, Zhiguo Jiang, Kun Wu, Jun Shi, Fengying Xie, Wei Wang, Haibo Wu, and Yushan Zheng. Self-supervised representation distribution learning for reliable data augmentation in histopathology wsi classification. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. 3, 6
- [27] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58: 101544, 2019. 3
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. 2, 3
- [29] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 2024. 1, 2
- [30] Georg Wölflein, Dyke Ferber, Asier R. Meneghetti, Omar S. M. El Nahhas, Daniel Truhn, Zunamys I. Carrero, David J. Harrison, Ognjen Arandjelović, and Jakob Nikolas Kather. Benchmarking pathology feature extractors for whole slide image classification, 2024. 2, 6
- [31] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 2025. 2
- [32] Jiawei Yang, Hanbo Chen, Yu Zhao, Fan Yang, Yao Zhang, Lei He, and Jianhua Yao. Remix: A general and efficient framework for multiple instance learning based whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2022. 3
- [33] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 2
- [34] Imaad Zaffar, Guillaume Jaume, Nasir Rajpoot, and Faisal Mahmood. Embedding space augmentation for weakly supervised learning in whole-slide images. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2023. 2, 3
- [35] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3
- [36] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back, 2019. 1
- [37] Xiao Zhou, Luoyi Sun, Dexuan He, Wenbin Guan, Ruifen Wang, Lifeng Wang, Xin Sun, Kun Sun, Ya Zhang, Yanfeng Wang, et al. A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. *arXiv*, 2024. 2
- [38] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024. 1, 2