ScoreNet: Learning Non-Uniform Attention and Augmentation for Transformer-Based Histopathological Image Classification

Thomas Stegmüller[1] Behzad Bozorgtabar[1,2,3] Antoine Spahr[1] Jean-Philippe Thiran[1,2,3]
[1]EPFL, Switzerland [2]CHUV, Switzerland [3]CIBM, Switzerland
{firstname.lastname}@epfl.ch

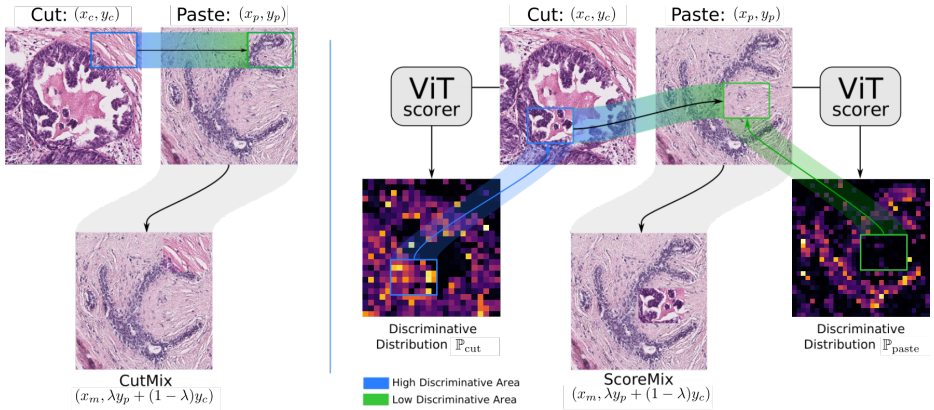

Figure 1: CutMix (**left**) randomly mixes samples, yielding label misallocation, while our ScoreMix (**right**) creates a coherent artificial training pair (x_m, y_m) by pasting a region of high semantic content from the *cut* image, x_c, at a non-discriminative region of the *paste* image, x_p, and obtains the new label via a convex combination of the *cut* and *paste* labels.

Abstract

Progress in digital pathology is hindered by high-resolution images and the prohibitive cost of exhaustive localized annotations. The commonly used paradigm to categorize pathology images is patch-based processing, which often incorporates multiple instance learning (MIL) *to aggregate local patch-level representations yielding image-level prediction. Nonetheless, diagnostically relevant regions may only take a small fraction of the whole tissue, and current* MIL*-based approaches often process images uniformly, discarding the inter-patches interactions. To alleviate these issues, we propose* ScoreNet*, a new efficient transformer that exploits a differentiable recommendation stage to extract discriminative image regions and dedicate computational resources accordingly. The proposed transformer leverages the local and global attention of a few dynamically recommended high-resolution regions at an efficient computational cost. We further introduce a novel mixing data-augmentation, namely* ScoreMix*, by leveraging the image's semantic distribution to guide the data mixing and produce coherent sample-label pairs.* ScoreMix *is embarrassingly simple and mitigates the pitfalls of previous augmentations, which assume a uniform semantic distribution and risk mislabeling the samples. Thorough experiments and ablation studies on three breast cancer histology datasets of Haematoxylin & Eosin* (H&E) *have validated the superiority of our approach over prior arts, including transformer-based models on tumour regions-of-interest* (TRoIs) *classification.* ScoreNet *equipped with proposed* ScoreMix *augmentation demonstrates better generalization capabilities and achieves new state-of-the-art (SOTA) results with only 50% of the data compared to other mixing augmentation variants. Finally,* ScoreNet *yields high efficacy and outperforms SOTA efficient transformers, namely TransPath [37] and SwinTransformer [20], with throughput around* $3\times$ *and* $4\times$ *higher than the aforementioned architectures, respectively. Our code is publicly available*[1].

[1]https://github.com/stegmuel/ScoreNet

1. Introduction

Due to the increasing availability of digital slide scanners enabling pathologists to capture high-resolution whole slide images (WSI), computational pathology is becoming a ripe ground for deep learning and recently witnessed a lot of advances. Nonetheless, the diagnosis from H&E stained WSIs remains challenging. The difficulty of the task is a consequence of two inherent properties of histopathology image datasets: *i)* the huge size for images and *ii)* the cost of exhaustive localized annotations, making the usage of most deep learning models computationally infeasible. Patch-based processing approaches [31, 23, 13] have become a *de facto* practice for high dimensional pathology images that aggregate individual patch representation/classification predictions by, e.g., a convolutional neural network (CNN) for image-level prediction. Nonetheless, patch-based methods increase the requirement of patch-level labeling and further regions of interest (RoI) detection as diagnostic-related tissue sections might only take a small fraction of the whole tissue, leading to considerable uninformative patches. Prior CNN methods [14, 18] have adopted multiple instance learning (MIL) [22] to address the above issues, which incorporates an attention-based aggregation operator to identify tissue sub-regions of high diagnostic value automatically. Nonetheless, these MIL methods embed all the patches independently and discard the inter-patches correlation or only incorporate it at a later stage.

Recently, self-supervised learning (SSL) methods [18, 17, 32, 7] aimed to construct semantically meaningful visual representations via pretext tasks for histopathological images. Despite their notable success using CNN backbones in improving classification performances, CNN's receptive field often restricts the learning of global context features. In another line of research, to compensate for the lack of diverse and large datasets, mixing augmentation techniques [36, 39, 40] have been developed to further enhance the performance of these models. While there have been substantial performance gains on natural image datasets, we argue that such data augmentations may not be helpful for histopathological images, as they risk creating locally ambiguous images or mislabelled samples. Furthermore, contrary to CNNs, vision transformer (ViT) models [10, 35] can capture long-range visual dependencies due to their flexible receptive fields via self-attention mechanisms. More recently, self-supervised ViTs method [37, 19] combined the advantages of ViT and SSL to efficiently learn visual representations from less curated pre-training data. Despite their usefulness, there is relatively little research on the impact of data augmentation design, efficiency and robustness of ViT for histopathological image classification. For example, can we train an efficient transformer by selecting only informative regions of high diagnostic value (RoIs) from high-resolution images? What data augmentation strategies can improve the transformer's representation learning for TRoIs classification? This paper addresses these questions by uncovering insights about key aspects of data augmentation and exploits the self-attention maps to identify the most relevant regions for the end task and train an efficient transformer.

Contributions. Our contributions are as follows:

1. We propose ScoreNet, a new efficient transformer-based architecture for histopathological image classification. It combines a fine-grained local attention mechanism with a coarse-grained global attention module to extract cell- and tissue-level features. Benefiting from a differentiable recommendation module, the proposed architecture only processes the most discriminative regions of the high-resolution image, making it significantly more efficient than competitive transformer architectures without compromising accuracy;

2. A novel mixing data-augmentation, namely ScoreMix for histopathological images is presented. ScoreMix works in synergy with our architecture, as they build upon the same observation: the different regions of the images are not equally relevant for a given task. Using the learned self-attention w.r.t. the `[CLS]` token, we determine the distribution of the semantic regions in images during training to ensure sampling of informed cutting and pasting locations (see Fig. 1);

3. We empirically show consistent improvements of ScoreNet over SOTA methods for TRoIs classification on the BRACS dataset, and similarly for ScoreNet's generalization capability on the CAMELYON16 and BACH datasets. The interpretability of ScoreNet behaviour is also investigated. Finally, we demonstrate ScoreNet throughput improvements over existing efficient transformers, making it an ideal candidate for applications on WSIs. **Our code and models will be publicly available upon acceptance.**

2. Related work

TRoIs Classification. Conventionally, deep convolutional neural networks [31, 30, 23, 13, 38] process pathology images in a patch-wise manner using a MIL formulation [22] and aggregate patch-level features extracted by CNNs. Nonetheless, current MIL methods discard the inter-patches interaction or only integrate it at the very end of the pipeline. Similarly, the computational resources dedicated to a specific region are independent of its pertinence for the task. Current methods rely on attention-based

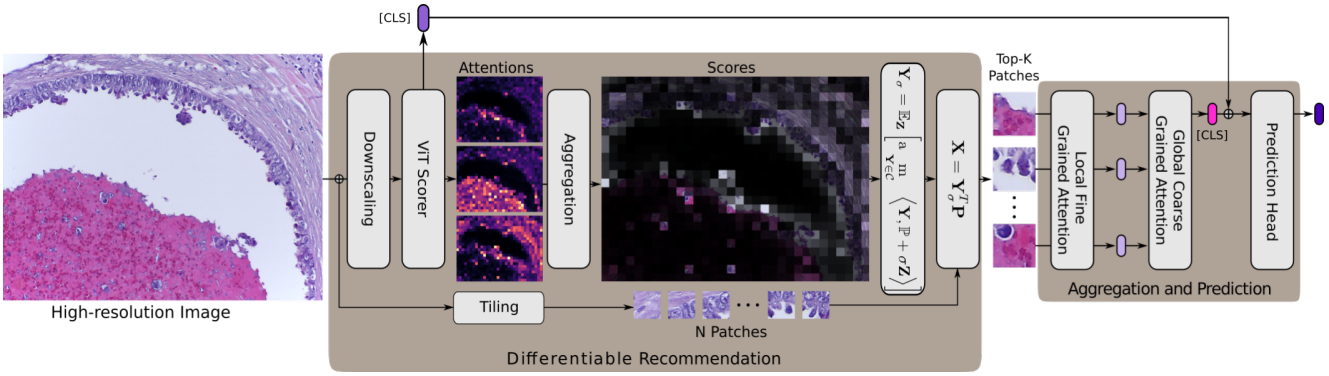

Figure 2: An overview of the proposed ScoreNet. The recommendation stage provides tissue-level features, and **differentiably selects** the most discriminative high-resolution patches. The aggregation stage independently extracts cell-level features and embeds the patches via a *local fine-grained attention* mechanism and endows them with contextual information with the *global coarse-grained attention* mechanism.

MIL techniques [14, 18, 15, 5, 28] to account for the non-uniform relevance of patches. On the contrary, the integration of contextual cues remains almost untouched, as all the aforementioned methods rely on a pipeline where the patch embedding and patch contextualization tasks are disconnected w.r.t. the gradient flow. For example, [15] processes representative patches extracted by an external tool [16]. Thus, their patch extraction is fixed and not data-driven as ours. Alternatively, [33] resort to using a multiple field-of-views/resolutions strategy to endow local patches with contextual information. In another line of research, graph neural network (GNN)-based methods [41, 27] have been proposed to capture global contextual information. These approaches build a graph model that operates on the cell-level structure or combines the cell-level and tissue-level context. However, graph generation can be cumbersome and computationally intensive, prohibiting its use in real-time applications. Recently, SSL methods [18, 17, 32] have demonstrated their capabilities to improve classification for histopathological images. Most of these methods harness pretext tasks, e.g., contrastive pre-training, to learn semantically meaningful features. Nonetheless, the CNN backbone used in these approaches inevitably abandons learning of global context features. The transformer-based architectures [37, 19] can be an alternative solution for processing images as a de-structured patch sequence and capturing their global dependencies. More recently, hybrid-based vision transformer models [6, 29, 37] have been used in digital pathology, either based on MIL framework [29] or SSL pre-training [37] on unlabeled histopathological images. Nevertheless, these methods process the whole image uniformly and do not allow dynamic extraction of the region of interest.

Mixing Data-Augmentation Methods. Recently, mixing data augmentations strategies [36, 39, 39] have been proposed to enhance the generalization capabilities of deep network classifiers. These improvements are further exacerbated when the augmentations model the interactions between the classes [39]. These methods create a new augmented sample by cutting an image region from one image and pasting it on another image, while a convex combination of their labels gives the ground-truth label of the new sample. Despite the strong performances of the existing methods, none of them is genuinely satisfying as they either create samples that exhibit atypical local features as in MixUp [40] or produce potentially mislabeled samples as in CutMix [39]. CutMix approach has been improved by [5] via re-weighting the mixing factor w.r.t. the sum of the attention map values in the randomly sampled image region, which is still at risk of producing mislabelled samples. In addition, recent CutMix based augmentation methods [36, 34] bear additional disadvantages. For example, Attentive CutMix [36] requires an auxiliary pre-trained model to select the most salient patches from the *cut* image and disregards the location of the informative regions in the *paste* image. SaliencyMix [34] assumes that discriminative parts in an image are highly correlated with the saliency map, which is typically not the case for histopathological images.

3. Methods

Model Overview. An overview of the proposed training pipeline for H&E stained histology TRoIs' representation learning is illustrated in Fig. 2. Histopathological image classification requires capturing cellular and tissue-level microenvironments and learning their respective interactions. Motivated by the above, we propose an efficient transformer, ScoreNet that captures the cell-level structure and tissue-level context at the most appropriate resolutions. Provided sufficient contextual information, we

postulate and empirically verify that a tissue's identification can be achieved by only attending to its sub-region in a high-resolution image. As a consequence, ScoreNet encompasses two stages. The former (*differentiable recommendation*) provides contextual information and selects the most informative high-resolution regions. The latter (*aggregation and prediction*) processes the recommended regions and the global information to identify the tissue and model their interactions simultaneously.

More precisely, the recommendation stage is implemented by a ViT and takes as input a downscaled image to produce a semantic distribution over the high-resolution patches. Then, the most discriminative high-resolution patches for the end task are **differentiably extracted**. These selected patches (tokens) are then fed to a second ViT implementing the *local fine-grained attention* module, which identifies the tissues represented in each patch. Subsequently, the **embedded patches attend to one another via a transformer encoder** (*global coarse grained attention*). This step concurrently refines the tissues' representations and model their interactions. As a final step, the concatenation of the `[CLS]` tokens from the recommendation's stage and that of the *global coarse-grained attention*'s encoder produces the image's representation. Not only does ScoreNet's workflow allows for a significantly increased throughput compared to SOTA methods (see Table 4), it further enables the independent pre-training and validation of its constituent parts.

3.1. ScoreNet

Semantic Regions Recommendation. Current MIL-based approaches [14, 18] based on patch-level features aggregation often process histopathological images uniformly and discard the inter-patches interactions. To alleviate these issues, we exploit **a differentiable recommendation stage to extract discriminative image regions relevant to the classification**. More specifically, we leverage the self-attention map of a ViT as a distribution of the semantic content. Towards that end, the high-resolution image is first downscaled by a factor s and subsequently fed to the recommendation's stage ViT. The resulting self-attention map captures the contribution of each patch to the overall representation. Let's assume a ViT, that processes a low-resolution image $x_l \in \mathbb{R}^{C\times h\times w}$ of spatial resolution $h \times w$ and encompassing N patches of dimension $P_l \times P_l$. The attended patches (tokens) of the $(L-1)$ layer are conveniently represented as a matrix $\mathbf{Z} \in \mathbb{R}^{(N+1)\times d}$, where d is the embedding dimension of the model, and the extra index is due to the `[CLS]` token. Up to the last MLP and for a single attention head, the representation of the complete image is given by:

$$y_{[\text{CLS}]} = \underbrace{\text{softmax}(a_1^T)}_{1\times(N+1)} \underbrace{\mathbf{Z}\mathbf{W}_{\text{val}}}_{(N+1)\times d} \tag{1}$$

where $\mathbf{W}_{\text{val}} \in \mathbb{R}^{d\times d}$ is the value matrix, and a_1^T is the first row of the self-attention matrix $\mathbf{A}$:

$$\mathbf{A} = \mathbf{Z}\mathbf{W}_{\text{qry}} \left(\mathbf{Z}\mathbf{W}_{\text{key}}\right)^T \tag{2}$$

where $\mathbf{W}_{\text{qry}}$ and $\mathbf{W}_{\text{key}}$ are the query and key matrices, respectively. The first row of the self-attention matrix captures the contribution of each token to the overall representation (Eq. 1). This is in line with the discriminative capacity of the `[CLS]` token that patches having the highest contribution are the ones situated in the highest semantic regions of the images. The distribution of the semantic content over the patches is therefore defined as:

$$\mathbb{P}_{\text{patch}} = \text{Softmax}(\tilde{a}_1^T) \in \mathbb{R}^N \tag{3}$$

where $\tilde{a}_1$ stands for a_1 without the first entry, namely the one corresponding to the `[CLS]` token. Since ViTs typically encompasses multiple heads, we propose to add an extra learnable parameter, which weights the relative contributions of each head to the end task; after aggregation of the multiples self-attention maps, the formulation is identical to that of Eq. 3.

Concurrently with acquiring the above defined semantic distribution, the high-resolution image, $x_h \in \mathbb{R}^{C\times H\times W}$, is tiled in a regular grid of large patches ($P_h \times P_h$), stored in a tensor $\mathbf{P} \in \mathbb{R}^{N\times C\times P_h\times P_h}$. At inference time, a convenient way to select the K most semantically relevant high-resolution regions is to encode the *top-K* indices as one-hot vectors: $\mathbf{Y} \in \mathbb{R}^{N\times K}$, and to extract the corresponding K patches, $\mathbf{X} \in \mathbb{R}^{K\times C\times P_h\times P_h}$ via:

$$\mathbf{X} = \mathbf{Y}^T\mathbf{P} \tag{4}$$

At training time, since the above formulation is not differentiable, we propose to adopt the differentiable approach of [8]. Following the perturbed optimizers scheme, the *top-K* operation is bootstrapped by applying a Gaussian noise, $\sigma\mathbf{N} \in \mathbb{R}^{N\times K}$, to the semantic distribution. The noisy indicators, $\mathbf{Y}_\sigma$, are subsequently computed as:

$$\mathbf{Y}_\sigma = \mathbb{E}_{\mathbf{N}}\left[\arg\max_{\mathbf{Y}\in\mathcal{C}} \left\langle \mathbf{Y}, \tilde{\mathbb{P}} + \sigma\mathbf{N}\right\rangle\right] \tag{5}$$

where σ is the standard deviation of the noise, $\tilde{\mathbb{P}} \in \mathbb{R}^{N\times K}$ is obtained by broadcasting $\mathbb{P}_{\text{patch}}$ to match the dimension of $\mathbf{Y}$, and $\mathcal{C}$ is a restriction of the domain ensuring the equivalence between solving Eq. 5 and the *top-K* operation [8]. The extraction of the high-resolution regions follows the procedure described in Eq. 4. Similarly, the Jacobian of the indicators w.r.t. the semantic distribution, $\mathbb{P}_{\text{patch}}$ can be computed as:

$$J_{\tilde{\mathbb{P}}}\mathbf{Y}_\sigma = \mathbb{E}_{\mathbf{N}}\left[\arg\max_{\mathbf{Y}\in\mathcal{C}} \left\langle \mathbf{Y}, \tilde{\mathbb{P}} + \sigma\mathbf{N}\right\rangle \mathbf{N}^T/\sigma\right] \tag{6}$$

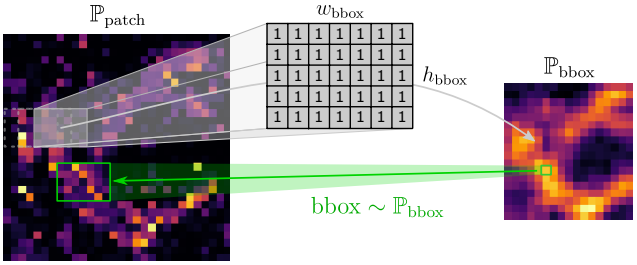

Figure 3: The bounding box selection scheme for ScoreMix. The score distribution for each bounding box ($\mathbb{P}_{\text{bbox}}$) is obtained by convolving the patch distribution map $\mathbb{P}_{\text{patch}}$ with a kernel of **1**s of the bounding box dimensions $(h_{\text{bbox}}, w_{\text{bbox}})$. The bbox is then sampled from $\mathbb{P}_{\text{bbox}}$, which we often refer to as $\mathbb{P}_{\text{cut}}$ or $\mathbb{P}_{\text{paste}}$.

Computational Complexity. Vision transformers heavily rely on the attention mechanism to learn a high-level representation from low-level regions. The underlying assumption is that the different sub-regions of the image are not equally important for the overall representation. Despite this key observation, the computation cost dedicated to a sub-region is independent of its contribution to the high-level representation, which is inefficient. Our ScoreNet attention mechanism overcomes this drawback by learning to attribute more resources to regions of high interest. For a high-resolution input image $x_h \in \mathbb{R}^{C\times H\times W}$, the asymptotical time and memory cost is $\mathcal{O}\left(\left(\frac{H}{s\cdot P_l}\cdot\frac{W}{s\cdot P_l}\right)^2\right)$, when the recommendation stage uses inputs downscaled by a factor s and processes them with a patch size of P_l. The derivation of this cost, including that of the recommendation stage, which is independent of the input size, can be found in the **Supplementary Material**.

3.2. ScoreMix

We propose a new mixing data augmentation for histopathological images by learning the **distribution of the semantic image regions** using the learned self-attention for `[CLS]` token of the ViT without requiring architectural changes or additional loss. More formally, let $x_c, x_p \in \mathbb{R}^{C\times H\times W}$ be the *cut* and *paste* images respectively and let y_c and y_p be their corresponding labels. We aim to mix the *cut* and *paste* samples to generate a new training example (x_m, y_m). To do so, we first compute the semantic distributions using the current parameters of the model and the input samples; namely, we compute $\mathbb{P}_{\text{cut}}(x_c, \theta)$ and $\mathbb{P}_{\text{paste}}(x_p, \theta)$. Given these distributions and a randomly defined bounding box size, we sample the cutting and pasting locations from the *cut* and *paste* distributions, respectively:

$$\begin{aligned} M_c &\sim \frac{1}{Z_c}\cdot \mathbb{P}_{\text{cut}}(x_c, \theta, \lambda) \\ M_p &\sim \frac{1}{Z_p}\cdot (1 - \mathbb{P}_{\text{paste}}(x_p, \theta, \lambda)) \end{aligned} \tag{7}$$

where Z_c and Z_p are normalization constants, and $1-\lambda \sim \mathcal{U}([0,1])$ defines the strength of the mixing, i.e. the size of the bounding box. The locations of the cutting and pasting regions are encoded as binary masks, i.e., $M_c, M_p \in \{0,1\}^{H\times W}$, where a value of 1 encodes for a patch in the cutting/pasting region. Under the above formalism, the mixing operation can be defined as:

$$\begin{aligned} x_m &= (\mathbf{1} - M_p)\odot x_p \\ M_p \otimes x_m &\leftarrow M_c \otimes x_c \\ y_m &= \lambda y_p + (1-\lambda) y_c \end{aligned} \tag{8}$$

where $\mathbf{1}$ is a mask of ones, $\odot$ denotes the element-wise multiplication, and $\otimes$ indicates an indexing w.r.t. a mask.

Computing the Semantic Distributions. Computing the semantic distributions of the *paste* and *cut* images is an essential part of the pipeline as it allows for a data-driven selection of the cutting/pasting sites, thereby avoiding the pitfalls of random selection. When the size of the bounding box matches that of a single patch, the distribution can be directly deduced from the self-attention map, as described in Sec. 3. As a consequence, and when the bounding box's size matches that of a single patch, the semantic distribution can be directly obtained from $\mathbb{P}_{\text{patch}}$ (see Eq. 3). In practice, we would typically use bounding boxes encompassing more than a single patch. In that case, the distribution of the semantic content at the bounding box resolution can be obtained by a local aggregation of the above distribution:

$$\mathbb{P}_{\text{bbox}}(i) = \frac{1}{Z_{\text{bbox}}}\sum\nolimits_{j\in\mathcal{N}(i)} \mathbb{P}_{\text{patch}}(j) \tag{9}$$

where Z_{bbox} is a normalization constant and $\mathcal{N}(i)$ returns the indices of the patches situated in the bounding box whose top left corner is the patch i. In practice, this can be efficiently implemented by first unflattening the patch distribution $\mathbb{P}_{\text{patch}}$, and convolving it with a kernel of ones and of the same dimension as the desired bounding box (see Fig. 3).

4. Experiments

Datasets. The primary dataset used in our experiments is the BReAst Carcinoma Sub-typing (**BRACS**) [27]. BRACS consists of 4391 RoIs acquired from 325 H&E stained breast carcinoma WSI (at 0.25 μm/pixel) with varying dimensions and appearances. Each RoI is annotated with one of the seven classes: Normal, Benign, Usual Ductal Hyperplasia (UDH), Atypical Ductal Hyperplasia (ADH), Flat Epithelial Atypia (FEA), Ductal Carcinoma In Situ (DCIS), and Invasive. Our experiments follow the same data splitting scheme as [27] for training, validation, and test set at the WSI level to avoid test leakage. In addition, we use publicly available BreAst Cancer Histology **(BACH)** dataset [1]

to show ScoreNet generalization capabilities. It contains 400 training and 100 test images from four different breast cancer types: Normal, Benign, In Situ, and Invasive. All images have a fixed size of 1536×2048 pixels and a pixel scale of 0.42×0.42 μm. To assess the interpretability of ScoreNet, we further evaluate our model on the **CAME-LYON16** dataset [3] for binary tumour classification. We extract a class-balanced dataset of 1920×1920 pixels from high-resolution WSIs.

Experimental Setup. We base ScoreNet' ViTs, namely the one used by the recommendation stage and by the local fine-grained attention mechanism on a modified ViT-tiny architecture (see **Supplementary Material**) and follow the self-supervised pre-training scheme of [4] for both of the aforementioned ViTs. Noteworthy that an end-to-end pre-training of ScoreNet is also feasible. After pre-training, the ScoreNet is optimized using the SGD optimizer (momentum=0.9) with a learning rate chosen with the linear scaling rule [11] ($lr = 10^{-2} \cdot \text{batchsize}/256 = 3.125 \cdot 10^{-4}$) annealed with a cosine schedule until 10^{-6}. ScoreNet is finetuned for 15 epochs with a batch-size of 8. We empirically determine the top $K = 20$ regions, and a downscaling factor $s = 8$ by a hyperparameter sweep (cf. ablation experiment in the **Supplementary Material**). All experiments are implemented in PyTorch 1.9 [25] using a single GeForce RTX3070 GPU.

4.1. TRoIs Classification Results and Discussion

In Table 1, we compare the TRoIs classification performance of ScoreNet on the BRACS dataset against the state-of-the-arts, including MIL-based [23, 29, 21], GNN-based, e.g., [27], and self-supervised transformer-based [37] approaches. The first MIL-based baseline [23] aggregates independent patch representations from the penultimate layer of a ResNet-50 [12] pre-trained on ImageNet [9]. The patch model is further finetuned on 128×128 patches at different magnification, e.g., $10\times$, $20\times$ or $40\times$. The latter operate either on multi- or single-scale images to benefit from varying levels of context and resolution. Similarly, we report the performances of the recent MIL-based methods, TransMIL [29], and CLAM [21] using the original implementations and setup. Both methods are tested with different magnifications (see Table 1). Additionally, the single-head (-SB) and multi-head (-MB) variants of CLAM are used with the small (-S) and big (-B) versions of the models (see CLAM's implementation). We further use various GNN-based baselines, particularly HACT-Net [27], the current SOTA approach for TRoIs classification on the BRACS. Finally, we report the performance of the recent self-supervised transformer approach, TransPath [37], which is a hybrid transformer/convolution-based architecture. ScoreNet reaches a new state-of-the-art weighted F1-score of 64.4% on the BRACS TRoIs classification task outperforming the second-best method, HACT-Net, by a margin of 2.9% (Table 1). The results are reported for two variants of ScoreNet, namely ScoreNet$/4/1$ and ScoreNet$/4/3$, which use the four last `[CLS]` tokens of the scorer and the last or the three last `[CLS]` tokens from the coarse attention mechanism (aggregation stage). ScoreNet$/4/3$ variant puts more emphasis on the features available at ($40\times$), whereas ScoreNet$/4/1$ is more biased towards the global representation available at ($5\times$) (with a downscaling factor $s = 8$). One can observe that both model variants significantly outperform the existing baseline in terms of weighted F1-scores and for almost every class, but DCIS. A potential explanation for this behaviour, could be that the relevant features for the classification of DCIS tissues are mostly texture-based, which favors CNN-based architectures that are more sensitive to texture than transformer-based models [24]. As it happens, the baselines outperforming ScoreNet for this class, all rely on a CNN features extractor. More interestingly, the architectural differences directly translate to differences in the classification results. ScoreNet$/4/3$ is more suitable for classes where the **discriminative features are at the cell level than** ScoreNet$/4/1$**, which is more suited when the tissue organization is the discriminative criterion**. Nonetheless, both of these architectures indeed benefit from the information available at each scale. This observation is well supported by the classification results obtained when a linear layer is trained independently on the scorer's `[CLS]` tokens (Lin. scorer's `[CLS]` in Table 1) or using only the `[CLS]` tokens from the aggregation stage (Lin. encoder's `[CLS]` in Table 1). Despite the difference in results between the two model variants, it is clear that they both perform worse when separated, which indicates that the representations of both stages are complementary. In brief, ScoreNet allows for an easily tuning to meet prior inductive biases on the ideal scale for a given task.

ScoreMix & Data-Regime Sensitivity. We also show that ScoreNet equipped with the proposed ScoreMix augmentation achieves superior TRoIs classification performances compared to CutMix [39] and SaliencyMix [34] augmentations for different data regimes, e.g., low-regime with only 10% of the data. **Our proposed ScoreMix outperforms SOTA methods with only 50% of the data** and is on-par or better than most baselines with only 20% of the data (Table 2). We argue that these improvements are primarily due to the generation of more coherent sample-label pairs under the guidance of the learned semantic distribution. This alleviates randomly cutting and pasting non-discriminative patches, as is the case with CutMix. Our results further support that image saliency used in the SaliencyMix is not correlated with discriminative regions.

Generalization Capabilities. To gauge the generaliza-

Table 1: Comparison with the prior art for TRoIs classification using weighted and class-wise F1-scores averaged over three independent runs on the BRACS dataset. The best results are in **bold**. ScoreNet/x/y refers to an instance of ScoreNet using the recommendation module's last x [CLS] tokens and the last y tokens from the global coarse-grained attention.

	Method	Normal	Benign	UDH	ADH	FEA	DCIS	Invasive	Weighted F1
MILs	Agg-Penultimate (10×) [30]	48.7 ± 1.7	44.3 ± 1.9	45.0 ± 5.0	24.0 ± 2.8	47.0 ± 4.3	53.3 ± 2.6	86.7 ± 2.6	50.8 ± 2.6
	Agg-Penultimate (20×) [30]	42.0 ± 2.2	42.3 ± 3.1	39.3 ± 2.0	22.7 ± 2.5	47.7 ± 1.2	50.3 ± 3.1	77.0 ± 1.4	46.8 ± 2.2
	Agg-Penultimate (40×) [30]	32.3 ± 4.6	39.0 ± 0.8	23.7 ± 1.7	18.0 ± 0.8	37.7 ± 2.9	47.3 ± 2.0	70.7 ± 0.5	39.4 ± 1.9
	Agg-Penultimate (10× + 20×) [30]	48.3 ± 2.0	45.7 ± 0.5	41.7 ± 5.0	32.3 ± 0.9	46.3 ± 1.4	59.3 ± 2.0	85.7 ± 1.9	52.3 ± 1.9
	Agg-Penultimate (10× + 20× + 40×) [30]	50.3 ± 0.9	44.3 ± 1.2	41.3 ± 2.5	31.7 ± 3.3	51.7 ± 3.1	57.3 ± 0.9	86.0 ± 1.4	52.8 ± 1.9
	CLAM-SB/S (10×) [21]	39.6 ± 4.6	45.5 ± 4.9	34.7 ± 2.0	30.4 ± 6.7	68.8 ± 1.9	64.3 ± 0.8	84.2 ± 2.6	53.9 ± 1.9
	CLAM-SB/S (20×) [21]	50.2 ± 3.2	45.5 ± 1.8	32.2 ± 1.6	25.5 ± 4.2	69.6 ± 1.0	60.8 ± 2.7	84.2 ± 1.6	54.0 ± 0.7
	CLAM-SB/S (40×) [21]	47.0 ± 5.2	38.8 ± 1.8	30.0 ± 7.7	29.4 ± 2.9	65.9 ± 1.2	52.2 ± 1.3	76.7 ± 1.6	49.9 ± 0.8
	CLAM-SB/B (10×) [21]	46.4 ± 6.0	42.4 ± 2.8	33.1 ± 1.0	29.3 ± 2.1	67.4 ± 1.4	63.0 ± 4.5	84.4 ± 2.1	53.7 ± 1.9
	CLAM-SB/B (20×) [21]	56.2 ± 1.2	42.3 ± 4.4	27.4 ± 2.4	30.1 ± 4.0	68.5 ± 2.1	60.9 ± 2.1	84.6 ± 1.2	54.3 ± 1.5
	CLAM-SB/B (40×) [21]	42.8 ± 1.1	43.3 ± 2.8	33.8 ± 0.7	29.6 ± 3.6	64.1 ± 2.6	52.0 ± 3.8	78.8 ± 2.2	50.5 ± 0.9
	CLAM-MB/S (10×) [21]	42.5 ± 3.3	43.4 ± 3.6	31.4 ± 3.2	32.1 ± 4.8	67.5 ± 2.2	59.7 ± 2.4	83.8 ± 2.0	52.9 ± 1.7
	CLAM-MB/S (20×) [21]	56.6 ± 0.8	47.4 ± 0.9	33.5 ± 5.2	17.0 ± 1.5	70.3 ± 1.1	56.9 ± 1.6	84.9 ± 1.2	53.8 ± 0.6
	CLAM-MB/S (40×) [21]	50.2 ± 7.7	39.3 ± 2.9	38.6 ± 2.4	26.5 ± 8.9	69.4 ± 2.6	54.1 ± 3.3	82.9 ± 2.5	52.9 ± 0.8
	CLAM-MB/B (10×) [21]	39.7 ± 1.6	41.0 ± 2.6	34.5 ± 1.0	29.8 ± 4.7	66.8 ± 1.5	63.4 ± 1.0	83.5 ± 0.4	52.7 ± 0.9
	CLAM-MB/B (20×) [21]	59.4 ± 2.0	47.7 ± 1.2	31.7 ± 0.7	20.1 ± 3.4	68.3 ± 0.4	59.9 ± 1.7	86.8 ± 0.6	54.8 ± 1.0
	CLAM-MB/B (40×) [21]	47.3 ± 3.2	39.5 ± 1.5	38.8 ± 4.5	30.2 ± 6.3	68.2 ± 1.9	59.2 ± 2.9	82.1 ± 2.7	53.5 ± 1.3
GNNs	CGC-Net [41]	30.8 ± 5.3	31.6 ± 4.7	17.3 ± 3.4	24.5 ± 5.2	59.0 ± 3.6	49.4 ± 3.4	75.3 ± 3.2	43.6 ± 0.5
	Patch-GNN (10×) [2]	52.5 ± 3.3	47.6 ± 2.2	23.7 ± 4.6	30.7 ± 1.8	60.7 ± 5.3	58.8 ± 1.1	81.6 ± 2.2	52.1 ± 0.6
	Patch-GNN (20×) [2]	43.9 ± 4.2	43.4 ± 3.2	19.5 ± 2.3	25.7 ± 2.9	55.6 ± 2.1	52.9 ± 1.8	79.2 ± 1.1	47.1 ± 0.7
	Patch-GNN (40×) [2]	41.7 ± 3.1	32.9 ± 1.0	25.1 ± 3.7	25.6 ± 2.0	49.5 ± 3.5	48.6 ± 4.2	71.6 ± 5.1	43.2 ± 0.6
	TG-GNN [26]	58.8 ± 6.8	40.9 ± 3.0	46.8 ± 1.9	40.0 ± 3.6	63.7 ± 10.5	53.8 ± 3.9	81.1 ± 3.3	55.9 ± 1.0
	CG-GNN [26]	63.6 ± 4.9	47.7 ± 2.9	39.4 ± 4.7	28.5 ± 4.3	72.1 ± 1.3	54.6 ± 2.2	82.2 ± 4.0	56.6 ± 1.3
	CONCAT-GNN	61.0 ± 4.5	43.1 ± 2.3	42.0 ± 4.7	26.1 ± 3.7	71.3 ± 2.1	60.8 ± 3.7	85.4 ± 2.7	57.0 ± 2.3
	HACT-Net [26]	61.6 ± 2.1	47.5 ± 2.9	43.6 ± 1.9	40.4 ± 2.5	74.2 ± 1.4	**66.4 ± 2.6**	88.4 ± 0.2	61.5 ± 0.9
Tansformers	TransPath [37]	58.5 ± 2.5	43.1 ± 1.8	34.9 ± 5.2	38.3 ± 6.0	66.9 ± 0.8	61.4 ± 1.2	85.0 ±1.4	56.7 ± 2.0
	TransMIL (10×) [29]	38.7 ± 5.4	44.0 ± 2.9	30.5 ± 4.1	31.0 ± 11.8	68.1 ± 2.6	61.8 ± 1.9	87.3 ± 2.6	53.2 ± 1.1
	TransMIL (20×) [29]	51.0 ± 0.1	44.5 ± 2.9	31.6 ± 2.1	31.4 ± 10.3	71.3 ± 4.8	63.0 ± 2.8	89.9 ± 1.6	56.2 ± 1.6
	TransMIL (40×) [29]	47.6 ± 9.8	42.9 ± 3.6	41.5 ± 5.3	38.4 ± 5.9	72.7 ± 2.6	62.7 ± 2.9	87.1 ± 3.9	57.5 ± 0.7
	Lin. encoder's [CLS]	52.7 ± 9.4	35.6 ± 3.4	34.5 ± 6.7	25.1 ± 3.6	53.5 ± 9.8	38.7 ± 2.8	63.3 ± 7.6	43.8 ± 3.4
	Lin. scorer's [CLS]	57.5 ± 4.2	48.8 ± 5.5	42.7 ± 3.5	42.7 ± 7.4	74.3 ± 5.2	60.5 ± 2.4	90.6 ± 0.2	60.9 ± 3.1
	ScoreNet/4/1	**64.6 ± 2.2**	52.6 ± 2.8	**48.4 ± 2.2**	**47.4 ± 2.4**	77.9 ± 0.7	59.3 ± 1.1	90.6 ± 1.5	64.1 ± 0.7
	ScoreNet/4/3	64.3 ± 1.5	**54.0 ± 2.2**	45.3 ± 3.4	46.7 ± 1.0	**78.1 ± 2.8**	62.9 ± 2.0	**91.0 ± 1.4**	**64.4 ± 0.9**

Table 2: Comparison with SOTA Mixup-based augmentation methods [39, 34] and the standard random augmentation strategy using various fractions of the BRACS dataset and identical distribution for the bounding boxes' sizes.

Dataset	Random Aug.	CutMix [39]	SaliencyMix [34]	**ScoreMix**
BRACS 10%	52.9 ± 2.4	53.7 ± 2.9	53.5 ± 2.7	**55.9 ± 1.9**
BRACS 20%	57.6 ± 1.8	58.0 ± 1.4	57.8 ± 1.0	**58.7 ± 0.8**
BRACS 50%	60.4 ± 1.8	61.2 ± 2.5	59.8 ± 2.4	**62.3 ± 0.6**
BRACS 100%	62.7 ± 1.6	63.1 ± 1.1	62.8 ± 1.2	**64.0 ± 0.7**

tion capabilities of ScoreNet compared to other current SOTA methods, e.g., HACT-Net [26], we leverage two external evaluation datasets, namely CAMELYON16 and BACH. After training on the BRACS dataset, the weights of ScoreNet are frozen. To evaluate the quality of the learned features, we either train a linear classifier on top of the frozen features or apply a k-nearest-neighbor classifier (k = 1) without any finetuning. We perform stratified 5-fold cross-validation. For HACT-Net, we use the available pre-trained weights and follow the implementation of [27]. As HACT-Net sometimes fails to generate embeddings and to have a fair comparison, we only evaluate the samples for which HACT-Net could successfully produce embeddings (around 95% of the BACH and 80% of the CAMELYON16 datasets). Experimental results in Table 3 demonstrate the superiority of ScoreNet in learning generalizable features. It further demonstrates the robustness of ScoreNet to changes in magnification. Indeed, the model is pre-trained on BRACS (40×), while BACH's images were acquired at a magnification of 20×. Furthermore, the CAMELYON16 dataset contains WSIs collected from lymph nodes in the vicinity of the breast, while BRACS contains WSIs collected by mastectomy or biopsy (i.e., directly in the breast). The excellent knowledge transfer between the two datasets highlights the transferability of features learned by ScoreNet in various use cases.

Interpretability? To probe the internal behavior of ScoreNet, we finetune the model on CAMELYON16 images using image-level labels only. At test time, we scrutinize the learned semantic distributions of the tumour-positive images. The semantic distributions depicted in Fig. 4 seems to indicate that ScoreNet learns to identify the tumour area and interpret **cancer-related morphological information**, while never having been taught to do

Table 3: Generalization capabilities of **ScoreNet** compared to HACT-Net trained on BRACS and evaluated on the BACH's annotated images and 1000 images from CAMELYON16, respectively. The weighted F1-scores over a stratified 5-fold cross-validation fold is reported.

	BRACS → BACH		BRACS → CAMELYON16	
	Linear	k-NN	Linear	k-NN
TransPath [37]	61.8 ± 4.8	72.0 ± 2.9	58.1 ± 4.8	69.9 ± 2.5
TransMIL [29]	46.5 ± 10.2	74.0 ± 4.8	59.8 ± 3.0	60.8 ± 5.3
CLAM-SB/S [21]	53.3 ± 13.0	69.8 ± 4.5	56.7 ± 1.9	68.0 ± 3.5
CLAM-SB/B [21]	57.5 ± 3.6	75.3 ± 3.1	55.5 ± 4.1	68.0 ± 1.5
HACT-Net [26]	40.2 ± 2.8	32.8 ± 5.8	60.0 ± 4.6	61.0 ± 4.2
ScoreNet	$\mathbf{73.4 \pm 3.5}$	$\mathbf{76.9 \pm 6.1}$	$\mathbf{81.1 \pm 3.5}$	$\mathbf{77.0 \pm 4.6}$

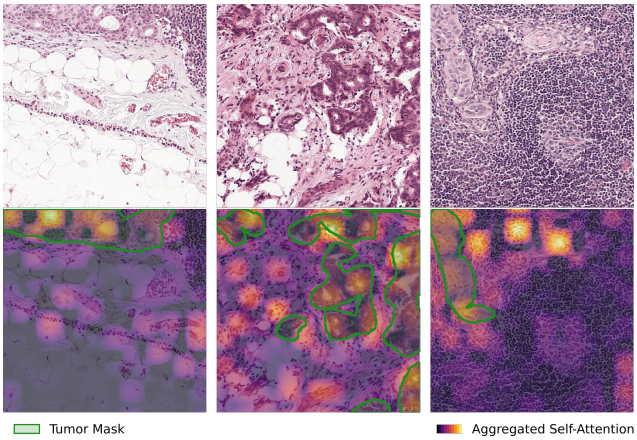

Figure 4: ScoreNet Interpretability. Visualization of the semantic distribution, overlaid with the tumour ground-truth mask on a few samples of the CAMELYON16 dataset. The semantic distributions are obtained from the recommendation stage, i.e., at low-resolution. ScoreNet is pre-trained on BRACS and finetuned on CAMELYON16.

so. Quantitatively, we observe that, on average, 74.6% of the 20 patches selected from positive images are tumour-positive. Furthermore, we report an average image-wise AuC of 73.6% when interpreting the probability of the recommendation stage to sample a patch as the probability of it being tumour-positive.

Ablation on Efficacy of ScoreNet. The critical aspect of ScoreNet is its improved efficiency compared to other transformer-based architectures. This improvement is due to the choice of a hierarchical architecture and the exploitation of redundancy in histology images. At inference time, we expect a gain in throughput compared to the vanilla ViT of the order of the squared downscaling factor, s, (see **Supplementary Material**), typically $s^2 = 64$, which is well reflected in practice, as shown in Table 4. Due to the self-supervised pre-training, ScoreNet does not require any stain normalization or pre-processing, unlike its competitor HACT-Net. Similarly, ScoreNet yields higher throughput than other SOTA efficient transformers architectures,

Table 4: Inference throughput comparison of ScoreNet, HACT-Net, and SOTA transformer-based architectures. All models were tested with the same image size and a single GeForce RTX 3070 GPU.

	Image size	Throughput (im./s)	Pre-processing
HACT-Net [26]	1536×2048	$4.95e\text{-}4 \pm 1.40e\text{-}3$	✓
Vanilla ViT [10]	1536×2048	3.8 ± 0.1	-
SwinTransformer [20]	1536×2048	76.8 ± 0.4	✗
TransPath [37]	1536×2048	97.6 ± 3.1	✗
ScoreNet	1536×2048	$\mathbf{335.0 \pm 7.9}$	✗

namely TransPath [37], and SwinTransformer [20], with throughput around $3\times$ and $4\times$ higher than these methods. The latter observation is interesting considering the linear asymptotic time and memory cost of the SwinTransformer, which is probably a consequence of the fact that SwinTransformers process a lot of uninformative high-resolution patches in the first layer(s).

Ablation on Shape Cues and Robustness. We investigate ScoreNet's ability to learn shape-related features. To do so, we study shape cues extracted by the recommendation model via the concatenated `[CLS]` tokens (see Fig. 2). Consequently, we implement shape removal by applying a random permutation of the downscaled image's tokens at test time. With this setup, a weighted F1-score of $59.8 \pm 0.8\%$ is reached, representing a significant drop in performance compared to $64.4 \pm 0.9\%$ without permutation. It demonstrates that *i)* the recommendation stage's concatenated `[CLS]` tokens contribute positively to the overall representation and *ii)* the latter is **not permutation invariant** and thus shape-dependent. In a second experiment, we show the whole recommendation stage is also shape-dependent. To that end, we repeat the same experiment, but the patches are extracted from the permuted images, reaching a weighted F1-score of $59.5 \pm 0.6\%$. We further observe that for a given image, the overlap of the selected patches with and without permutation is, on average, only 15.7%, which indicates that the semantic distribution learned by ScoreNet is shape-dependent.

5. Conclusion and Future Work

We have introduced ScoreNet, an efficient transformer-based architecture that dynamically recommends discriminative regions from large histopathological images, yielding rich generalizable representations at an efficient computational cost. In addition, we propose ScoreMix, a new attention-guided mixing augmentation that produces coherent sample-label pairs. We achieve new SOTA results on the BRACS dataset for TRoIs classification and demonstrate ScoreNet's superior throughput improvements compared to previous SOTA efficient transformer-based architectures.

References

[1] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[2] Bulut Aygüneş, Selim Aksoy, Ramazan Gökberk Cinbiş, Kemal Kösemehmetoğlu, Sevgen Önder, and Ayşegül Üner. Graph convolutional networks for region of interest classification in breast histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, page 113200K. International Society for Optics and Photonics, 2020.

[3] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[4] 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 International Conference on Computer Vision (ICCV)*, 2021.

[5] Jie-Neng Chen, Shuyang Sun, Ju He, Philip Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. *arXiv preprint arXiv:2111.09833*, 2021.

[6] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021.

[7] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[8] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2351–2360, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[11] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016.

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[15] Shivam Kalra, Mohammed Adnan, Sobhan Hemati, Taher Dehkharghanian, Shahryar Rahnamayan, and Hamid R Tizhoosh. Pay attention with focus: A novel learning scheme for classification of whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 350–359. Springer, 2021.

[16] Shivam Kalra, Hamid R Tizhoosh, Charles Choi, Sultaan Shah, Phedias Diamandis, Clinton JV Campbell, and Liron Pantanowitz. Yottixel–an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*, 65:101757, 2020.

[17] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 2021.

[18] 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.

[19] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[21] 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.

[22] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.

[23] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. From patch-level to roi-level deep feature representations for breast histopathology classification. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560H. International Society for Optics and Photonics, 2019.

[24] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[26] Pushpak Pati, Maria Frucci, and Maria Gabrani. Hactnet: A hierarchical cell-to-tissue graph neural network for histopathological image classification. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings*, volume 12443, page 208. Springer Nature, 2020.

[27] Pushpak Pati, Guillaume Jaume, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, Maurizio Di Bonito, Giuseppe De Pietro, Gerardo Botti, Jean-Philippe Thiran, Maria Frucci, Orcun Goksel, and Maria Gabrani. Hierarchical graph representations in digital pathology. *Medical Image Analysis*, 75:102264, 2022.

[28] Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zielinski. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1721–1730, 2021.

[29] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classication. *arXiv preprint arXiv:2106.00908*, 2021.

[30] Korsuk Sirinukunwattana, Nasullah Khalid Alham, Clare Verrill, and Jens Rittscher. Improving whole slide segmentation through visual context-a systematic study. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–200. Springer, 2018.

[31] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

[32] Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *arXiv preprint arXiv:2102.03897*, 2021.

[33] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12597–12606, 2019.

[34] AFM Shahab Uddin, Mst Sirazam Monira, Wheemyung Shin, TaeChoong Chung, and Sung-Ho Bae. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *International Conference on Learning Representations*, 2020.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[36] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3642–3646. IEEE, 2020.

[37] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021.

[38] Yan Xu, Zhipeng Jia, Yuqing Ai, Fang Zhang, Maode Lai, I Eric, and Chao Chang. Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 947–951. IEEE, 2015.

[39] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[40] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

[41] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.