A Graph-Based Multi-Scale Approach with Knowledge Distillation for WSI Classification

Gianpaolo Bontempo, Federico Bolelli, *Member, IEEE*, Angelo Porrello, Simone Calderara, *Member, IEEE*, Elisa Ficarra

***Abstract*— The usage of Multi Instance Learning (MIL) for classifying Whole Slide Images (WSIs) has recently increased. Due to their gigapixel size, the pixel-level annotation of such data is extremely expensive and time-consuming, practically unfeasible. For this reason, multiple automatic approaches have been raised in the last years to support clinical practice and diagnosis. Unfortunately, most state-of-the-art proposals apply attention mechanisms without considering the spatial instance correlation and usually work on a single-scale resolution. To leverage the full potential of pyramidal structured WSI, we propose a graph-based multi-scale MIL approach, DAS-MIL. Our model comprises three modules: *i)* a self-supervised feature extractor, *ii)* a graph-based architecture that precedes the MIL mechanism and aims at creating a more contextualized representation of the WSI structure by considering the mutual (spatial) instance correlation both inter and intra-scale. Finally, *iii)* a (self) distillation loss between resolutions is introduced to compensate for their informative gap and significantly improve the final prediction. The effectiveness of the proposed framework is demonstrated on two well-known datasets, where we outperform SOTA on WSI classification, gaining a $+2.7\%$ AUC and $+3.7\%$ accuracy on the popular Camelyon16 benchmark.**

***Index Terms*— Whole Slide Images (WSIs), Multiple Instance Learning (MIL), (Self) Knowledge Distillation, Weakly Supervised Learning**

I. Introduction

ANALYZING histological tissues is crucial for the diagnosis and treatment planning of multiple human body lesions and diseases [1], [2]. Fortunately, modern microscopes allow to scan of conventional glass slides into digital Whole Slide Images (WSIs), keeping the information in a multi-resolutions pyramidal structure (Fig. 1). This representation allows to store a large amount of information, preserving the ability to quickly analyze the tissue from different perspectives, *i.e.*, from higher scales to observe the cellular morphology and the different cellular compartments, and from lower ones to identify distinct tissues, such as tumor, stroma, and immune system cells, and their location.

The authors are with the Dipartimento di Ingegneria "Enzo Ferrari," Università degli Studi di Modena e Reggio Emilia, Italy. E-mail: {*name.surname*}@unimore.it

This project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No. 965193 and from the Department of Engineering "Enzo Ferrari" of the University of Modena through the FARD-2022 (Fondo di Ateneo per la Ricerca 2022).

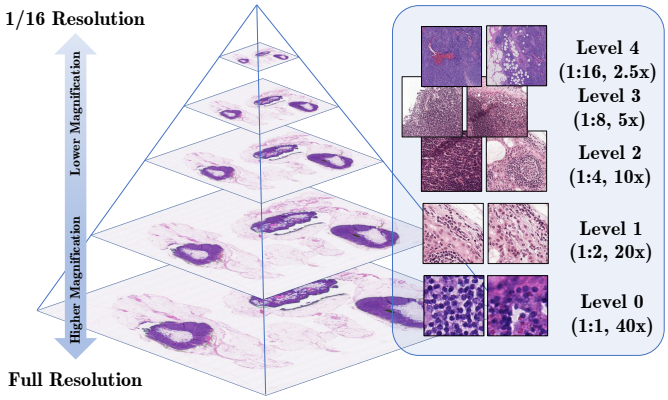

Fig. 1: This figure depicts the pyramidal structure of WSIs, highlighting the information available at different scales which ranges from structure level at lower resolutions to cellular interaction at higher resolutions (*i.e.* lower levels).

However, due to their gigapixel size, the manual analysis of WSIs requires specific tools [3] and is usually expensive and time-consuming for pathologists. For these reasons, several methodologies have been recently proposed to support clinicians with fast, accurate, and automatic analysis [4]–[6].

Feeding modern neural networks with the entire gigapixel image is not feasible, as it requires significant computational resources and time. Therefore, researchers have focused on developing algorithms that can analyze WSIs efficiently by cropping them into smaller patches. Additionally, since pixel-level annotation is time-consuming, annotation is generally provided at WSI level. As a consequence, an effective solution to analyze the WSIs is the use of weakly supervised paradigms such as Multiple Instance Learning (MIL) [7], [8], whose application is becoming increasingly popular in WSI learning-based analysis [9]–[13]. In MIL, each WSI is considered a "bag" composed of many patches called instances. By restricting the rule to a binary classification setting (*e.g.*, tumor/not tumor), if at least one instance (patch) is positive, the image (bag) is marked as positive [14].

Recent proposals integrate multiple resolutions extracted from the WSI in the MIL learning representation mechanism employing a mere features concatenation [10], arguing its effectiveness w.r.t. single-resolution algorithms. Other very recent approaches [9], [15], [16] resort to more complex hierarchical structures by adapting Vision Transformers (ViT) [17] and Graph Neural Networks (GNN) [18]. However, we argue

that such methods do not take advantage of the full potential of the WSI pyramidal structure. As an example, the flat concatenation of features extracted at different resolutions [10] does not consider the substantial difference in the informative content they provide. On the other hand, in [15] many efforts have been spent to model intra-scale structural connection through a cluster-based pooling layer while overlooking inter-scale patterns.

A proficient learning approach should instead consider the differences in the information content available at different scales, including global structures and local cellular regions. Our proposed approach aims to leverage this heterogeneity by preserving the scale diversity of the input data. This paper proposes a pyramidal Graph Neural Network combined with (Self) Knowledge Distillation to leverage the multi-resolution structure of WSI in a MIL classification setting. In this context, the graph message passing brings information from higher to lower scales. More specifically, the proposed framework employs a 2-tiers GNN to provide a more *contextualized representation*. The first tier processes each input scale separately, preserving the scale diversity in terms of information content and acting as an adapter. The second tier allows for full communication between scales, making the information flow between all the considered resolutions and capturing their relationships.

The *contextualized* features generated by our GNN module are transmitted to individual (one for each scale) attention-based MIL to derive bag labels. The (Self) Knowledge Distillation mechanism is introduced to encourage agreement across the predictions delivered at different resolutions. At the same time, individual scale features are learned in isolation to preserve the diversity in terms of information content. By transferring knowledge across scales, connected by the GNN, the model self-improves as information flows during training.

Our main contributions can be resumed as follows:

- We propose the use of GNNs to naturally exploit the heterogeneous information contained in WSI images, employing the message passing for sharing the informative content provided at different scales;
- We devise a learning strategy based on (Self) Knowledge Distillation, that promotes the agreement across the predictions delivered by different scales. This approach allows valuable secondary information, which can only be inferred from a specific resolution, to be transferred to the other resolutions. Since connected through a GNN, the agreement between scales enables for self-improvement of the teacher module;
- An exhaustive set of experiments on two publicly available datasets, Camelyon-16[1] [19] and TCGA-Lung[2], validate and confirm the effectiveness of our proposal for analyzing a variety of WSI and supporting clinical decision;
- The source-code is available on GitHub[3] to ensure experiment reproducibility and future comparisons.

[1] https://camelyon16.grand-challenge.org/

[2] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[3] https://github.com/aimagelab/mil4wsi

II. Related Work

A. MIL for Disease Detection in WSI

Existing MIL approaches for WSI classification can be clustered based on two different aspects: the number of resolutions employed and the aggregation mechanism used to provide the final prediction. With the former, we can distinguish single-scale algorithms [10], [12], [20], where the tiling process is done at a unique resolution, and multi-scale approaches [10], [21], [22].

Regarding the aggregation mechanism, a further distinction between instance-level predicting algorithms [23], [24] and those that deal with bag-level [10]–[12], [20]–[22] predictions can be highlighted. In the first case, the patch probability is employed to produce the final result (*e.g.*, mean or max pooling), while bag-level approaches aggregate instances into bag representation and feed a classifier with it. The main difference between recent proposals concerns the attention mechanisms employed for aggregating instance-level information.

Single Scale. Regarding the single resolution, the classical AB-MIL [11] is based on a side-branch network to calculate the attention scores. In [20] a similar attention mechanism is employed as support for a double-tier feature distillation approach, which distills features from pseudo-bags to the original slide. Those features are selected by relevance as MaxMin [20] or by aggregation as AFS [20]. Another approach proposed in the literature, DS-MIL [10], is to apply non-local attention aggregation measuring the distance with the most relevant patch. In 2021, Lu *et al.* [25] proposed an algorithm based on a clustering loss applied on single or multiple branches (CLAM-SB and CLAM-MB), a variant of the classic AB-MIL. In Trans-MIL [12], a typical transformer architecture is used. The issue related to all the aforementioned solutions is that they miss to consider the mutual instance correlation. In [9], the authors leverage the well-performing Dino [26] feature extractor, proving its effectiveness also in this scenario. Beyond the classical attention mechanism, there are also algorithms based on Recurrent Neural Networks (RNNs) [22] and GNNs [27], [28] able to solve the same task. Also, these proposals miss to consider multiple resolutions and ignore the pyramidal structure of the WSIs.

Multi Scale. Recently, different authors focused on multi-resolution approaches. DSMIL-LC [10] concatenates representations obtained at different resolutions (*e.g.*, a low instance representation is concatenated with all the ones obtained at a higher resolution). MS-RNNMIL [22], instead, fed an RNN with instances extracted at different scales. In [29], a self-supervised hierarchical transformer is applied at each scale. In MS-DA-MIL [21], multi-scale features are included in the same attention algorithm. In H^2MIL [15] the multi-resolution is exploited by a GNN architecture and the information is aggregated through a specifically designed iterative pooling layer based on patches' location. The proposed aggregation uses the WSI pyramidal structure, but forcing a local attention that may affect performances. Previously, Chen *et al.* [30]

proposed the PTree-Net that selects patches at different resolutions based on the thumbnail level attention map. Those patches are then connected in a tree structure, later investigated through a relevance-enhanced GNN. However, this approach misses considering the hierarchical structure of patches and their heterogeneity [15].

B. Knowledge Distillation

Distilling knowledge from a more extensive network (teacher) to a smaller one (student) has been widely investigated in recent years [20], [31]–[34]. Typically, the loss evaluates the mimic capabilities of the student observing the teacher. Recent self-supervised representation learning approaches have also used this idea. In [26], [35], and [36], knowledge distillation is used to realize an agreement between networks that receive as input variants of the original image. In literature, knowledge distillation has been applied to different fields ranging from model compression [37] to WSI analysis [20], [31]. In [20], distillation is used to transfer the knowledge between MIL tiers applied on different subsampled bags. In [38], a self-distillation term is used in a regression task where the student model is trained on soft labels provided by the teacher. Moreover, [39] includes the weighted-ground truth targets in the loss term to better guide the student. In [40], the authors analyze the importance of teacher diversity and they provide a series of label smoothing methods to directly increase predictive diversity. Differently, our model applies (self) knowledge distillation between WSI scale resolutions. This way, improving the worst scale benefits the best one since they are connected through the GNN layer.

III. Method

Our proposal aims to enhance the information flow through WSI resolutions. In this respect, while existing works [10], [12], [25] take into account the interactions between scales by mostly leveraging trivial operations (such as concatenation of related feature representations), we instead provide a novel technique that builds upon: *i)* a GNN module that propagates patches' representation taking into account the natural multi-resolution structure of WSI; *ii)* a regulation term based on (self) Knowledge Distillation. This regulation term guides the most effective resolution to train the others more effectively self-improving at the same time. In the following section, we will discuss the proposed architecture in detail.

A. Architecture

Our approach can be decomposed into three main stages:

- **Feature Extraction**. Given the whole slide image, we re-sample it at various resolutions; for the i-th scale, we divide the resulting image into regular patches $x^i_1, x^i_2, \ldots, x^i_N$, which are then fed into a self-supervised feature extractor $f(\cdot;\theta_i)$. This way, we obtain multiple grids of latent representations;
- **Context Enrichment**. The representations are then re-arranged as nodes of a multi-Graph Neural Network (GNN), which spreads the information between resolutions through a message passing mechanism;
- **Multiple Instance Learning**. An auxiliary module weights the contribution of each instance to the representation of the whole slide used to perform classification.

Feature Extraction. Recent studies have shown the effectiveness of self-supervised learning techniques in capturing patch-level characteristics. For instance, the authors of [10] resorted to SimCLR [41], an approach that aims to align the representations of positive pairs and keep the representations of negative pairs distant. This approach suffers from a slow convergence rate and has a huge memory footprint. To address the shortcomings of self-supervised learning, our work advocates for Dino [26], which has recently shown promising results in the field of image understanding. In particular, it does not reckon on forging negative pairs during optimization, but rather, the training objective focuses solely on aligning the representations of positive pairs. In more detail, the latter is computed by two distinct networks, playing the roles of teacher and student. This way, the training is faster and requires lower computational resources, thus obtaining excellent representations in a few days. Similarly to [36], the risk of collapsing solutions is avoided through slower teacher network updates obtained through the Exponential Moving Average (EMA).

The authors of Dino introduced additional enhancements such as the exploitation of the Vision Transformers [42] in the design of the backbone networks. However, we refer the reader to the original paper for a deep understanding of Dino. Concerning our work, we devise an initial stage where multiple feature extractors $f(\cdot;\theta_1), \ldots, f_M(\cdot;\theta_M)$ are trained, each of which ends up being an expert of one of the M zoom scales of interest. On top of that, we freeze the weights of these networks and use them as patch-level feature extractors during the next step.

In our analysis, we focus only on two resolutions at the time (*e.g.* the $10\times$ and $20\times$ magnitudes) so that $M = 2$, but the approach can be extended to consider even more scales. Therefore, the overall input $\mathcal{X}$ for the subsequent layers can be summarized as follows:

$$\mathcal{X} = \Big[\mathcal{X}^1, \mathcal{X}^2\Big] = \Big[\underbrace{[f(x^1_1;\theta_1), \ldots, f_1(x^1_{n_1};\theta_1)]}_{\text{Low-level representations}}, \underbrace{[f_2(x^2_1;\theta_2), \ldots, f_2(x^2_{n_2};\theta_2)]}_{\text{High-level representations}}\Big].$$

Enabling Message Passing Between Scales Through Graph Neural Networks (GNNs). Although the representations generated by Dino offer a detailed portrait of local patterns in individual patches, they need to gain knowledge about the surrounding context. For this reason, a Graph Neural Network module is introduced to enable information exchange among local patches. In general terms, such a module takes as input multi-scale patch-level representations $\mathcal{X}$ and produces the resulting $\mathcal{Y}$ through a cascade of two neural sub-networks GNN_1 and GNN_2:

$$\begin{aligned}\mathcal{Y} &= \mathrm{GNN}(\mathcal{X}; \mathcal{A}^1, \mathcal{A}^2, \mathcal{A}^{1\cup 2}, \theta_{\mathrm{GNN}}) \\ &= \mathrm{GNN}_1(\mathcal{X}; \mathcal{A}^1, \mathcal{A}^2) \circ \mathrm{GNN}_2(\mathcal{X}; \mathcal{A}^{1\cup 2})\end{aligned}$$

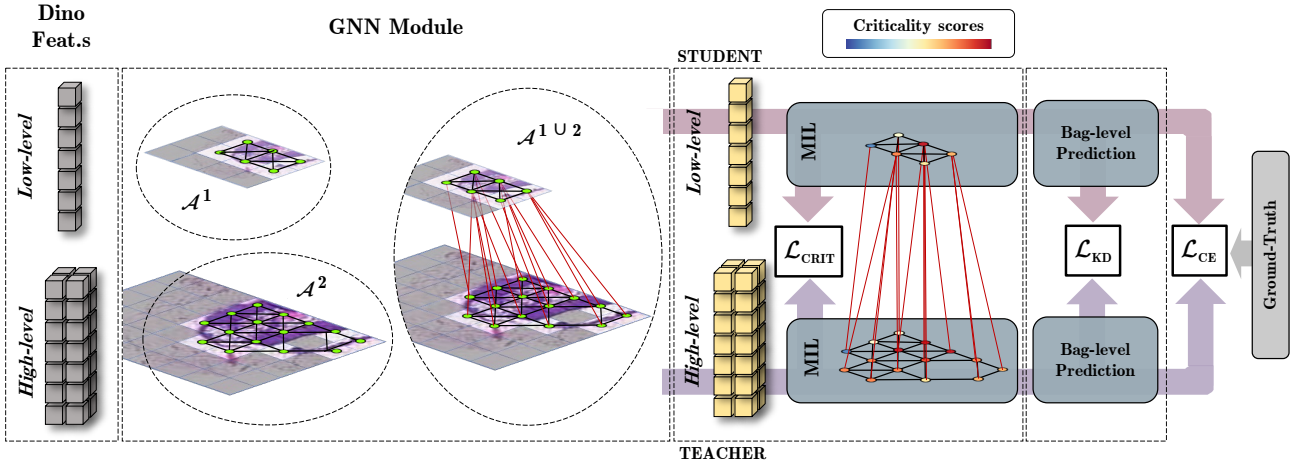

Fig. 2: Overview of the proposed DAS-MIL framework. The features extracted at different scales are connected considering both the image (as depicted in this Figure) and similarity space, by means of different graphs. The nodes of both graphs ($\mathcal{A}^1$ and $\mathcal{A}^2$) are later fused into a third one ($\mathcal{A}^{1\cup2}$), respecting the relation "part of". The contextualized features generated by the GNN module are passed to distinct attention-based MIL blocks (one for each scale) that extract bag labels. (Self-) Knowledge Distillation mechanism encourages agreement across the predictions delivered at different scales, both at bag ($\mathcal{L}_{\text{KD}}$) and instance level ($\mathcal{L}_{\text{CRIT}}$).

where $\mathcal{A}^1, \mathcal{A}^2$ and $\mathcal{A}^{1\cup2}$ identify the adjacency matrices used within this block and later described in this paragraph.

The former module, GNN_1, focuses only on *intra-scale* relations and no interactions are hence allowed between nodes from different sub-graphs. The computation of the resulting hidden activations $\mathcal{H}$ can be summarized as follows:

$$\begin{aligned}\mathcal{H} &= \Big[\mathcal{H}^1, \mathcal{H}^2\Big] \\ &= \Big[\mathcal{X}^1 + \text{GAT}(\mathcal{X}^1; \mathcal{A}^1; \theta_1), \mathcal{X}^2 + \text{GAT}(\mathcal{X}^2; \mathcal{A}^2; \theta_2)\Big],\end{aligned}$$

where GAT identifies the graph attention layer proposed in [43]. As can be understood, the two sub-graphs are processed through two independent graph layers. To explain such a design choice, we recall that the input embedding spaces $\mathcal{X}^1$ and $\mathcal{X}^2$ originate from different feature extractors. This way, individual scale features are learned in isolation to preserve the diversity in terms of information content. However, a direct shared vertex-wise transformation would clash with that input distribution shift[4]. For this reason, we ask the two independent modules to act as adapters, projecting the representations into a shared embedding space. The second module, GNN_2, allows for inter-scale relations. In particular, data are no longer considered as two distinct sub-graphs but rather as a unique and complex structure in which nodes from different scales can interact. This way, we aim to propagate neighborhood information across low and high resolutions. Briefly:

$$\mathcal{Y} = \Big[\mathcal{Y}^1, \mathcal{Y}^2\Big] = \mathcal{H} + \text{GAT}(\mathcal{H}; \mathcal{A}^{1\cup2}). \tag{1}$$

We obtain the adjacency matrix $\mathcal{A}^{1\cup2}$ by preserving the edges already in $\mathcal{A}^1$ and $\mathcal{A}^2$ and adding those connections between a parent WSI patch (lying in the low resolution, we call this relation "part of") and its *children*, *i.e.* the higher-scale patches it contains.

For what concerns $\mathcal{A}^1$ and $\mathcal{A}^2$, we propose to use two types of distances for the grid calculation:

- **Chessboard distance.** In this case, a patch neighborhood is defined by the 8-connectivity, *i.e.*, its surrounding 8 patches in *image space*.
- **Similarity distance.** The similarity distance is calculated as $(\mathcal{X}^i) \cdot (\mathcal{X}^i)^T$, and each node is linked to its K closest neighbors. This way, semantically similar nodes are connected in *feature space*.

A visual representation of the proposed multi-layer graph structure is provided in Fig. 2.

Bag-Level Representations via Critical Instance Selection. Following existing and well-established works [10], [15], [20], we compute the overall representation for the whole slide through a technique based on Multiple Instance Learning. Our model is inspired by DSMIL [10] and employs a self-attention mechanism that provides bag-level feature vectors, one for each of the two WSI scales involved in the pipeline. Such a module is based on self-attention [33] and builds upon the prior individuation of the **critical patches** $\mathcal{Y}_{\text{CRIT}} = [y^1_{\text{CRIT}}, y^2_{\text{CRIT}}]$, selected in a winner-takes-it-all fashion:

$$\begin{aligned} y^1_{\text{CRIT}} &= \mathcal{Y}^1_* \\ \text{s.t. } {}_* &\equiv \underset{i \le n_1}{\arg\max}\ z^1(\mathcal{Y}^1_i) \end{aligned}$$

where $z^1(\cdot) = \mathbf{W}^1_{\text{CRIT}}\, \text{g}(\mathcal{Y}^1_i; \mathbf{W}^2_{\text{CRIT}})$ is a scoring projection network (with weights matrices $\mathbf{W}^1_{\text{CRIT}}$ and $\mathbf{W}^2_{\text{CRIT}}$) that assigns a single scalar weight to each patch. It is noted that the same equation holds for the selection of the critical instance y^2_{CRIT} of the high-level resolution. Once these instances have been individuated, the bag-level representations y^1_{BAG} and

[4]The presence of such a distribution shift has been verified by applying a simple classifier to recognize the origin scales from the embedding.

y^2_{BAG} are computed as:

$$y^1_{\text{BAG}} = \mathbf{W}_{\text{CLS}} \sum_{i}^{n_1} \underbrace{U(\mathcal{Y}^1_{i=1}, \mathcal{Y}^1_{\text{CRIT}})}_{\substack{\text{Attention scores w.r.t.} \\ \text{the critical patch.}}} * \underbrace{\text{V}(\mathcal{Y}^1_i; \mathbf{W}^1_{\text{V}})}_{\text{Patch-level value.}}. \quad (2)$$

For an in-depth description of how self-attention weights and values have been calculated, we refer the reader to [44].

B. Aligning Scales with (Self) Knowledge Distillation

We have hence obtained two distinct sets of predictions for the two resolutions: namely, a bag-level score (*e.g.*, a tumor is either present or not) and a patch-level one (*e.g.*, which instances contribute the most to the target class). However, as these learned metrics are inferred from different WSI magnitudes, a disagreement may emerge: indeed, as previously observed in this Section, the higher resolutions generally yield better classification performance w.r.t. lower ones. In this work, we exploit such a disparity to introduce two additional optimization objectives, which exploit the predictions of the high scale as a teaching signal for the low one. Further than improving the results of the low scale, the aim is to propagate the benefits of such an improvement to the shared message-passing module and back to the high resolution.

To achieve this goal, we propose extending the transfer by promoting consistency between scales. As the poorer scale is additionally encouraged to *chase* the richer one, the message-passing module placed in the middle turns out to be crucial in discovering valuable patterns and relations, which could help not only the lower scale but also, the higher one. In formal terms, we propose extending the training objective with a twofold term. The former asks the predictions from the two scales to be close as much as possible via (Self) Knowledge Distillation (KD) [45]:

$$\mathcal{L}_{\text{KD}} = \tau^2 \ \text{KL}(\text{softmax}(\frac{y^1_{\text{BAG}}}{\tau}) \parallel \text{softmax}(\frac{y^2_{\text{BAG}}}{\tau})) \quad (3)$$

where KL identifies the Kullback–Leibler divergence and τ is a temperature that lets secondary information emerge from the teaching signal. Under the light of Bayesian statistical inference, the probability distribution given by the higher scale represents the prior distribution we enforce while fitting the lower one.

The second aligning term regards the scores computed through Eq. (3). More in detail, it encourages the two resolutions to assign criticality scores in a **consistent** manner: intuitively, if a low-resolution instance has been considered critical, then the average score attributed to its children instances should be likewise high. Such desiderata are carried out by minimizing the Euclidean distance between the low-resolution criticality grid map $\text{z}^1(\mathcal{Y}^1)$ and its sub-sampled counterpart computed by the high-resolution branch:

$$\mathcal{L}_{\text{CRIT}} = \|\text{z}^1(\mathcal{Y}^1) - \text{GraphPooling}(\text{z}^2(\mathcal{Y}^2))\|^2_2 \quad (4)$$

In the equation above, GraphPooling identifies a pooling layer applied over the higher scale: to do so, it considers the relation "part of" between scales.

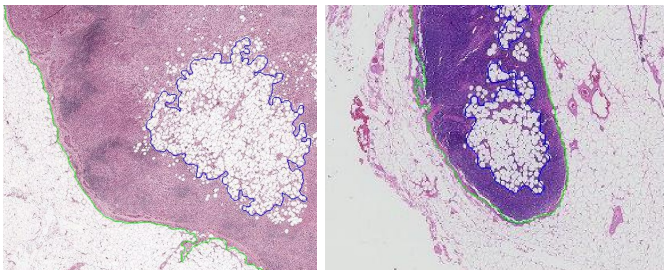

Fig. 3: Examples of WSI pre-processing. Green contours represent the considered tissue, the blue ones are holes the algorithm will discard. In this examples, the holes are mainly composed by fat tissue.

Overall Objective. To sum up, the overall optimization problem is formulated as a mixture of two objectives: the one requiring higher conditional likelihood w.r.t. ground truth labels $\mathbf{y}$ (*e.g.* a tumor is either present or not) and carried out through the Cross Entropy loss $\mathcal{L}_{\text{CE}}(\cdot; \mathbf{y})$; the other one based on (Self) Knowledge Distillation:

$$\min_{\theta} \ (1-\lambda)\mathcal{L}_{\text{CE}}(y^2_{\text{BAG}}) + \mathcal{L}_{\text{CE}}(y^1_{\text{BAG}}) + \lambda\mathcal{L}_{\text{KD}} + \beta\mathcal{L}_{\text{CRIT}}, \quad (5)$$

where λ is a hyperparameter weighting the trade-off between the teaching signals provided by labels and the higher resolution, while β balances the contributions of the consistency regularization introduced in Eq. (4).

IV. Experiments

Datasets. The proposed framework is evaluated on two different benchmarking datasets: the Camelyon16 [19] and a Lung dataset from The Cancer Genome Atlas (TCGA) project [10], which are widely employed for the evaluation of state-of-the-art proposals for WSI analysis [10], [20], [25].

More specifically, the *Camelyon16 dataset* —which derives from the homonym challenge that took place at the International Symposium on Biomedical Imaging (ISBI) in 2016— is designed for the automatic detection of metastases in Hematoxylin and Eosin (H & E) stained WSIs of lymph node sections [19]. It contains 398 WSIs, 128 of which are part of the official test set. Images have been acquired with two different slide scanners, named RUMC and UMCU, with $20\times$ and $40\times$ objective lenses, respectively, and comparable specimen-level pixel sizes, *i.e.* 0.243 μm × 0.243 μm for the RUMC and 0.226 μm × 0.226 μm for the UMCU. In this respect, the knowledge about the specimen-level pixels sizes allows to combine data acquired with different scanners and resolutions: we exploit this metadata to extract different scales from the original whole-slide images. To evaluate the performance of our proposal, we have adhered to the official training/test sets.

The second dataset —the *TCGA Lung dataset*— is publicly available on the GDC Data Transfer Portal and comprises two subsets of cancer: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC), counting 541 and 513 WSIs respectively. The task tackled in this case is the classification of LUAD vs LUSC. To split the dataset into

Camelyon16			TCGA Lung		
5×	10×	20×	5×	10×	20×
427	1 528	5 771	743	2 773	10 744

TABLE I: Mean number of patches per WSI obtained by our pre-processing algorithm at different magnification levels for both the considered datasets.

training and test, we have followed what has been devised by DSMIL [10], also removing ten corrupted slides from the dataset, according to [10].

Pre-processing. Before feeding our framework, WSIs are pre-processed to extract and filter background patches: to this aim, we adopt the CLAM framework [25]. In particular, after an initial segmentation process based on Otsu [46], non-overlapped patches within the foreground regions are considered. The original WSI is memory-loaded with a down-sampled resolution, usually 32×, and converted into the HSV color space. The saturation channel of the image is filtered with median blur to smooth edges and then thresholded to produce a binary foreground map of tissue regions. Additional morphological operators [47] and connected components labeling [48] are employed to fill gaps and holes within the map. Eventually, the foreground objects are filtered based on their area to remove spurious noise blobs. Examples of the filtering process are depicted in Fig. 3. 256×256 patches are finally extracted within the segmented contours, ensuring no overlap between patches. The process is repeated multiple times for each slide at different target resolutions (5×, 10×, and 20× in this study). When changing the resolution, the number of patches can vary significantly (from a hundred to hundreds of thousands). Moreover, given that the process is independently performed at different scales, the number of patches at different levels does not necessarily respect a fixed ratio (*e.g.*, 4:1 between 20× and 10× magnification levels). To provide a summary, Tab. I reports the average number of resulting patches per WSI at different magnification levels.

Metrics. The performances of our model have been measured through the AUC (Area Under the Curve) and the accuracy. The AUC measures the area under the ROC (Receiver Operator Characteristics) curve by varying the probability threshold. The accuracy is evaluated by selecting the best threshold suggested by the ROC.

Additionally, as proposed for the Camelyon16, the detection/localization performance is measured by means of the Free Response Operating Characteristic (FROC) curves. FROC differs from the ROC analysis since it substitutes the false positive rate on the x-axis with the average number of false positives per image. Specifically, a detection is considered a true positive, if the location of the detected region is within the ground truth lesion. If there are multiple findings for a single ground truth region, they are counted as a single true positive. Moreover, all the detections not within a specific distance from the ground truth annotations are considered false positives.

Implementation Details. We leveraged the Pytorch-Geometric codebase [49] to build the graph structure proposed in this paper. To train our models, we used the Adam [50] optimizer with a learning rate of 2×10^{-4} and a cosine annealing scheduler with a 1×10^{-5} decay without warm restart [51]. The Dino [26] feature extractor has been trained with two NVIDIA RTX5000 GPUs and all subsequent experiments have been performed on a SLURM server [52] with a single NVIDIA RTX2080 GPU.

		Camelyon16		TCGA Lung	
	Method	**Accuracy**	**AUC**	**Accuracy**	**AUC**
Single Scale	Mean-pooling	0.798	0.762	0.886	0.937
	Max-pooling	0.830	0.864	0.809	0.901
	MILRNN [22]	0.806	0.806	0.862	0.911
	ABMIL [11]	0.845	0.865	0.900	0.949
	CLAM-SB [25]	0.865	0.885	0.875	0.944
	CLAM-MB [25]	0.850	0.894	0.878	0.949
	Trans-MIL [12]	0.868	0.937	0.883	0.949
	DTFD (AFS) [20]	0.908	0.946	0.891	0.951
	DTFD (MaxMinS) [20]	0.899	0.941	0.894	0.961
	DSMIL [10]	0.868	0.894	0.919	0.963
Multi Scale	MS-DA-MIL [21]	0.876	0.887	0.900	0.955
	MS-MILRNN [22]	0.814	0.837	0.891	0.921
	HIPT † [29]	0.898	0.951	0.890	0.950
	DSMIL-LC [10]	0.899	0.917	**0.929**	0.958
	H^2-MIL † [15]	0.859	0.912	0.823	0.917
	DAS-MIL (ours)	**0.945**	**0.973**	0.925	**0.965**

TABLE II: Comparison with state-of-the-art solutions. Results marked with "†" have been calculated on our premises as the original papers lack the specific settings; all the other numbers are taken from [10] and [20]. DAS-MIL performance are obtained with $T = 1.5$, $\beta = \lambda = 1$.

Input Scale	MIL Target(s)	Accuracy	AUC
5×	5×	0.818	0.816
10×	10×	0.859	0.891
20×	20×	0.891	0.931
5×, 20×	5×, 20×	0.891	0.938
5×, 20×	5×, [5× ‖ 20×]	0.898	0.941
10×, 20×	10×, 20×	**0.945**	**0.973**
10×, 20×	10×, [10× ‖ 20×]	0.922	0.953

TABLE III: Comparison between scales. The target column indicates the features passed to the two MIL layers: the "‖" symbol indicates that they have been previously concatenated.

A. Comparison with the State-of-the-Art

Tab. II compares the results obtained by the proposed framework with the state-of-the-art architectures, considering both single-scale and multi-scale alternatives. Whenever available, the results published in the official papers are reported, otherwise, the experiments are replicated on our premises, using the code provided by the original authors. We compare our architecture with six different single resolution MIL mechanisms, *i.e.*, MILLRNN [22], ABMIL [11], CLAM [25], Trans-MIL [12], DTFT [20] and DS-MIL [10], as well as multi resolution proposals MS-DA-MIL [21], MS-MILRNN [22], DSMIL-LC [10], HIPT [29] and H^2-MIL [15].

Scales	Validation Acc.	Test Acc.
5×, 20×	0.922±0.011	0.89
10×, 20×	0.941±0.028	0.94

TABLE IV: 3-fold cross-validation for selecting the best scale combination on the tumor/not tumor task on Camelyon16 dataset.

Feature Extractor	Graph Mechanism	Camelyon16		TCGA Lung	
		Acc.	AUC	Acc.	AUC
SimCLR	✗	0.859	0.869	0.864	0.932
SimCLR	DAS-MIL	**0.906**	**0.928**	**0.883**	**0.9489**
SimCLR	H^2-MIL	0.836	0.857	0.826	0.916
Dino	✗	0.852	0.905	0.906	0.956
Dino	DAS-MIL	**0.891**	**0.938**	**0.925**	**0.965**
Dino	H^2-MIL	0.859	0.912	0.823	0.917

TABLE V: Comparison between DAS-MIL with and w/o (✗) the graph contextualization mechanism, and the most recent graph-based multi-scale approach H^2-MIL, when using different resolutions as input (5× and 20×).

As can be observed: *i)* the joint exploitation of multiple resolutions is generally more efficient; *ii)* DAS-MIL, the approach proposed in this work, yields robust and compelling results (especially on Camelyon16, where it provides 3.7% and 2.7% more than the SOTA accuracy and AUC, respectively). We acknowledge that methods reported in that table are based on different feature extractors, thus influencing the overall performance. The following subsection reports exhaustive ablation studies that shed light on the advantages of our technical contributions, disentangling the contribution of the feature extractor. Comparing C16 and TCGA datasets, there is a significant difference in the signal intensity. The former has roughly $< 10\%$ of tumor tissue, while the latter has $> 80\%$ of tumor regions per slide. In this sense, the contextualized representation provided by our DAS-MIL is much more effective in the first case.

B. Model Analysis

This section reports multiple ablation studies that detail the contribution of each component, highlighting their strengths and weaknesses.

Single-scale *vs* Multi-scale. Let's start by analyzing the contribution of different scales (see Tab. III). For single-scale experiments, we fed the model only with patches extracted at a single reference scale and the corresponding adjacency matrix. For what concerns multi-scale, instead, magnitudes can be combined in different ways[5]. The experiments revealed that the best results are obtained when the model is trained with 10× and 20× input resolutions. Tab. III also highlights that 5× magnitude is less effective and presents worse discriminative capabilities: we ascribe it to the type of samples in the WSIs (*e.g.* untreated tumor samples) and the specimen-level pixel size, underlying that different datasets and classification tasks may benefit from different scale combinations.

[5]Since the 20× resolution is experimentally proved to be the most effective one for tumor identification, all the considered combinations includes it.

λ	AUC 20×	AUC 10×	β	AUC 20×	AUC 10×
1.0	**0.973**	**0.974**	1.5	0.964	0.968
0.8	0.967	0.966	1.2	0.970	0.964
0.5	0.968	0.932	1.0	**0.973**	**0.974**
0.3	0.962	0.965	0.8	0.962	0.965
0.0	0.955	0.903	0.6	0.951	0.953

TABLE VI: Impact of Eq. 5 hyperparameters on Camelyon16.

τ	Accuracy		AUC	
	20×	10×	20×	10×
$\tau = 1$	0.883	0.962	0.906	0.957
$\tau = 1.3$	0.898	0.958	0.891	0.959
$\tau = 1.5$	**0.945**	**0.945**	**0.973**	**0.974**
$\tau = 2$	0.906	0.914	0.962	0.963
$\tau = 2.5$	0.922	0.914	0.951	0.952

TABLE VII: Impact of KD temperature (Eq. 3) on Camelyon16. Results are obtained with $\alpha = \beta = 1.0$.

At this point, the reader may wonder how to choose the scales to be employed without running all the possible combinations on the final test set. For this purpose, we highlight that such a selection may depend upon prior analyses of the domain and the nature of the task at hand, considering the peculiarities of the specific biological structure under analysis. In practice, it constitutes an *hyperparameter*; as such, common techniques based on Cross Validation (CV) can be employed to choose the proper resolutions: as reported in Tab. IV, the results on the validation set suggest and confirm our previous outcomes.

Interestingly, while [10] needs to concatenate the representations from different scales to attain satisfactory performance, such an operation has a negligible impact on our framework. In order to merge information from different resolutions, our approach already devises graph layers. While the concatenation assigns the same importance to the input representations, we argue that our mechanism can more effectively weight the contributions of patches since it can learn to route valuable features through the message-passing algorithm dynamically.

The Impact of Feature Extractors and Graph-Based Fusion. We refer the reader to Tab. V for an investigation of these aspects. In more detail, we consider both SimCRL and Dino, as well as the recently proposed graph mechanism H^2-MIL [15]: in doing so, we fix the input resolutions to 5× and 20×, according to our previous analysis. We can draw the following conclusions: *i)* when our DAS-MIL feature propagation layer is used, the selection of the optimal feature extractor (*i.e.* SimCLR *vs* Dino) has less impact on performance, as the message-passing can compensate for possible lacks in the initial representation; *ii)* DAS-MIL appears a better features propagator w.r.t. H^2-MIL. On this latter point, we conjecture that our approach has a characteristic that better captures the peculiarities of WSIs, *i.e.*, how the global representation of the slide is computed. Indeed, H^2-MIL exploits a global pooling layer that fulfills only the spatial structure of patches.

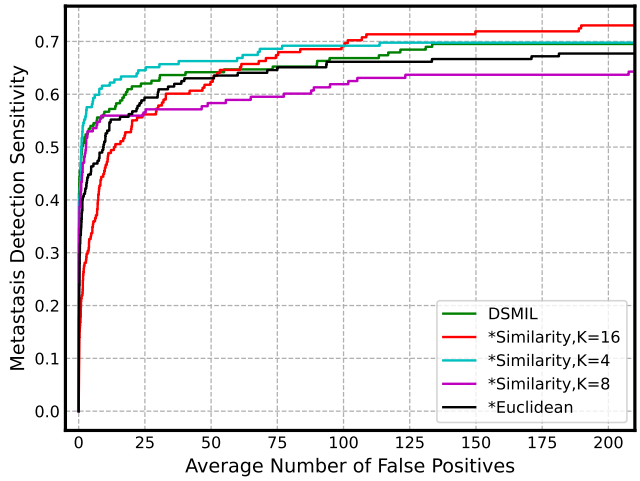

Fig. 4: FROC curves on Camelyon16. Attention scores are re-scaled from [min, max] to [0, 1] at slide level.

Consequently, if non-tumor patches surround a tumor patch, its contribution to the final prediction is likely to be outweighed by the IHPool module of H^2-MIL. Differently, our approach is not restricted in such a way, as it can dynamically route the information across the hierarchical structure (also based on the connections with the critical instance).

On the Role of (Self) Knowledge Distillation and Consistency Regularization. To assess the merits of the regularization objectives discussed earlier, we conducted several experiments varying the values of the corresponding balancing coefficients and present their results in Tab. VI. Lowering their values (even reaching $\lambda = 0$ *i.e.* no distillation is performed) negatively affects the performance. Notably, such a statement holds not only for the lower resolution (as one could expect), but also for the higher one, thus corroborating the claims we made in Sec. III-B on the bidirectional benefits of (Self) Knowledge Distillation in our multi-scale architecture.

We have also performed an assessment on the temperature τ, Tab. VII, which usually controls the smoothing factor applied to the teacher's predictions. We found out that the lower the temperature, the better the results, thus suggesting that the teacher scale is naturally not over-confident about its predictions but rather well-calibrated.

***Max vs Mean* Pooling with Different Graph Topologies.** The results presented in Tab. VIII indicate that (self-) knowledge distillation across different resolutions typically yield better results when mean pooling is used. Employing mean-pooling prevents knowledge transfer from misleading isolated samples, ensuring greater accuracy. However, when computing the 8-similarity graph in feature space, max pooling becomes more relevant. Not only it is more robust to overfitting (achieving 0.94% AUC at the last epoch), but it also outperforms the graph build on image space using the same pooling strategy. This is likely because using the feature space ensures the graph *homophily* property, dramatically reducing noise related to outlier instances.

Localization. In clinical scenarios, a discriminative model has to provide not only good predictive capabilities but also a reasonable explanation for its final prediction. The proposed architecture can provide such an explanation, highlighting the disease-positive patches extracted by the framework. An example is depicted in Fig. 5 where the model's output is compared with the ground-truth provided for a small subset of WSIs in the Camelyon dataset.

			Best Epoch		Last Epoch	
	Connectivity	**K**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
MAX	Image space	(8)	0.888	0.944	0.875	0.914
	Feature space	4	0.898	0.960	0.906	0.925
	Feature space	8	**0.930**	**0.966**	**0.906**	**0.942**
	Feature space	16	0.914	0.944	0.891	0.899
MEAN	Image space	(8)	**0.945**	0.973	0.890	**0.954**
	Feature space	4	0.914	0.959	**0.906**	0.933
	Feature space	8	0.922	0.962	0.875	0.912
	Feature space	16	0.930	**0.9755**	0.883	0.914

TABLE VIII: Exploring the effectiveness of *GraphPooling*, Eq. (4), implemented as *mean* or *max* pooling when using diverse graph topologies. Our analysis focuses on graphs build upon the image space considering the *chessboard* distance (8-connectivity) or K top nearest neighbors in feature space.

Moreover, the chart in Fig. 4 compares the localization properties (FROC curves) of DAS-MIL with different graph types and DSMIL. Notably, the use of a similarity graph can have a positive impact on the localization task. In particular, when the number of connected neighbors K is low (*e.g.*, K = 4), the model detects a high number of true positive instances at the cost of a few false positives. On the other hand, when K is high (*e.g.*, K = 16), although the number of false positives is higher, the model achieves higher sensitivity. We can stress that K implicitly regulates the smoothing operation performed on the attention scores by the graph.

V. Conclusion

This paper proposes a novel way to exploit multiple resolutions in the domain of histological WSI. We conceive a novel graph-based architecture that learns correlations between different WSI resolutions. Specifically, a GNN cascade is used to extract a context-aware and instance-level feature considering the spatial relationship between scales. This connection is further boosted during the training process by a distillation loss, asking for an agreement between scales.

On the one hand, an extensive set of experiments shows the effectiveness of the proposed distillation approach Tab. VI, Tab. VII, and of the graph mechanism employed Tab. V.

On the other hand, a few criticisms about the proposed architecture can be highlighted: *i)* while the criticality-based two-stage MIL approach is well suited for localization-related tasks (such as tumor detection), scenarios like tumor staging/survival prediction may require a deeper analysis across multiple critical regions [53]; *ii)* our DAS-MIL relies on a separate feature extractor for each target scale. While this is advantageous in terms of overall accuracy, it would require additional training stages for introducing new magnification(s) or adapting those already available for targeting a new task; *iii)* as currently devised, the patch-level feature extractors are trained in a self-supervised fashion and frozen in the

(a) WSI (b) Filtering (c) Ground-Truth (d) Attention Map

Fig. 5: Localization Example. From left to right we have (a) the original image, (b) the tissue segmented with pre-processing algorithm, (c) the positive-disease annotation provided by Camelyon16, and (d) the attention map extracted with DAS-MIL.

subsequent supervised stages of our pipeline. However, it could be beneficial to envision end-to-end training, which could promote representations more aligned with the task under consideration.

In addition to tackling the aforementioned issues, future works will explore proper and profitable ways to leverage on more than two scales. Such an analysis will focus not only on mere technicalities but also on the rationales behind using deeper pyramidal structures: which medical imaging tasks would effectively benefit from such an approach? Are all the scales equally crucial for all the tasks?

REFERENCES

[1] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, and B. Srinivasan, "A generalized deep learning framework for whole-slide image segmentation and analysis," *Scientific Reports*, vol. 11, pp. 1–14, 2021. 1

[2] S. Ortega, H. Fabelo, R. Camacho, M. De la Luz Plaza, G. M. Callicó, and R. Sarmiento, "Detecting brain tumor in pathological slides using hyperspectral imaging," *Biomedical Optics Express*, vol. 9, no. 2, pp. 818–831, 2018. 1

[3] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman *et al.*, "QuPath: Open source software for digital pathology image analysis," *Scientific reports*, vol. 7, no. 1, pp. 1–7, 2017. 1

[4] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep Learning for Whole Slide Image Analysis: An Overview," *Frontiers in medicine*, vol. 6, p. 264, 2019. 1

[5] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, vol. 67, p. 101813, 2021. 1

[6] F. Ponzio, G. Urgese, E. Ficarra, and S. Di Cataldo, "Dealing with Lack of Training Data for Convolutional Neural Networks: The Case of Digital Pathology," *Electronics*, vol. 8, no. 3, pp. 256–277, 2019. 1

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997. 1

[8] O. Maron and T. Lozano-Pérez, "A Framework for Multiple-Instance Learning," *Advances in Neural Information Processing Systems*, vol. 10, pp. 570–576, 1997. 1

[9] R. J. Chen and R. G. Krishnan, "Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology," *Learning Meaningful Representations of Life, NeurIPS 2021*, 2022. 1, 2

[10] B. Li, Y. Li, and K. 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*, 2021, pp. 14 318–14 328. 1, 2, 3, 4, 5, 6, 7

[11] M. Ilse, J. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, Jul 2018, pp. 2127–2136. 1, 2, 6

[12] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147, 2021. 1, 2, 3, 6

[13] G. Bontempo, N. Bartolini, M. Lovino, F. Bolelli, A. Virtanen, and E. Ficarra, "Enhancing PFI Prediction with GDS-MIL: A Graph-based Dual Stream MIL Approach," in *International Conference on Image Analysis and Processing*. Springer, 2023, pp. 550–562. 1

[14] E. Marconato, G. Bontempo, E. Ficarra, S. Calderara, A. Passerini, and S. Teso, "Neuro symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, 2023. 1

[15] W. Hou, L. Yu, C. Lin, H. Huang, R. Yu, J. Qin, and L. Wang, "H2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis," in *36th AAAI Conference on Artificial Intelligence*, 2022. 1, 2, 3, 4, 6, 7

[16] Y. Zhao, Z. Lin, K. Sun, Y. Zhang, J. Huang, L. Wang, and J. Yao, "SETMIL: Spatial Encoding Transformer-Based Multiple Instance Learning for Pathological Image Analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 66–76. 1

[17] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022. 1

[18] Y. Zheng, R. H. Gindra, E. J. Green, E. J. Burks, M. Betke, J. E. Beane, and V. B. Kolachalama, "A graph-transformer for whole slide image classification," *arXiv preprint arXiv:2205.09671*, 2022. 1

[19] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017. 2, 5

[20] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 802–18 812. 2, 3, 4, 5, 6

[21] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi, "Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3852–3861. 2, 6

[22] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019. 2, 6

[23] J. Feng and Z.-H. Zhou, "Deep MIML Network," in *Thirty-First AAAI conference on artificia intelligence*, 2017. 2

[24] P. O. Pinheiro and R. Collobert, "From Image-level to Pixel-level Labeling with Convolutional Networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1713–1721. 2

[25] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021. 2, 3, 5, 6

[26] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660. 2, 3, 6

[27] M. Tu, J. Huang, X. He, and B. Zhou, "Multiple instance learning with graph neural networks," *arXiv preprint arXiv:1906.04881*, 2019. 2

[28] Y. Zhao, F. Yang, Y. Fang, H. Liu, N. Zhou, J. Zhang, J. Sun, S. Yang, B. Menze, X. Fan *et al.*, "Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning with Deep Graph Convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4837–4846. 2

[29] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 144–16 155. 2, 6

[30] Z. Chen, J. Zhang, S. Che, J. Huang, X. Han, and Y. Yuan, "Diagnose Like A Pathologist: Weakly-Supervised Pathologist-Tree Network for Slide-Level Immunohistochemical Scoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2

[31] T. Ilyas, Z. I. Mannan, A. Khan, S. Azam, H. Kim, and F. De Boer, "TSFD-Net: Tissue specific feature distillation network for nuclei segmentation and classification," *Neural Networks*, vol. 151, pp. 1–15, 2022. 3

[32] L. Qi, J. Kuen, J. Gu, Z. Lin, Y. Wang, Y. Chen, Y. Li, and J. Jia, "Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 443–14 453. 3

[33] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3713–3722. 3, 4

[34] A. Monti, A. Porrello, S. Calderara, P. Coscia, L. Ballan, and R. Cucchiara, "How many Observations are Enough? Knowledge Distillation for Trajectory Forecasting," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6543–6552, 2022. 3

[35] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning," *Advances in Neural Information Processing ystems*, vol. 33, pp. 21 271–21 284, 2020. 3

[36] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758. 3

[37] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model Compression," in *KDD '06*, 2006. 3

[38] H. Mobahi, M. Farajtabar, and P. Bartlett, "Self-Distillation Amplifies Regularization in Hilbert Space," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3351–3361, 2020. 3

[39] K. Borup and L. N. Andersen, "Even your Teacher Needs Guidance: Ground-Truth Targets Dampen Regularization Imposed by Self-Distillation," in *Neural Information Processing Systems*, 2021. 3

[40] Z. Zhang and M. Sabuncu, "Self-Distillation as Instance-Specific Label Smoothing," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2184–2195, 2020. 3

[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. 3

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020. 3

[43] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018, accepted as poster. 4

[44] L. Zhang, C. Bao, and K. Ma, "Self-distillation: Towards efficient and compact neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4388–4403, 2021. 5

[45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015. 5

[46] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. 6

[47] F. Bolelli, S. Allegretti, and C. Grana, "One DAG to Rule Them All," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3647–3658, Jan 2021. 6

[48] S. Allegretti, F. Bolelli, M. Cancilla, F. Pollastri, L. Canalini, and C. Grana, "How does Connected Components Labeling with Decision Trees perform on GPUs?" in *Computer Analysis of Images and Patterns*. Springer, 2019, pp. 39–51. 6

[49] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 6

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015. 6

[51] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in *International Conference on Learning Representations*, 2017. 6

[52] A. B. Yoo, M. A. Jette, and M. Grondona, "SLURM: Simple Linux Utility for Resource Management," in *Job Scheduling Strategies for Parallel Processing*. Springer, 2003, pp. 44–60. 6

[53] G. Bontempo, L. Lumetti, A. Porrello, F. Bolelli, S. Calderara, and E. Ficarra, "Buffer-MIL: Robust Multi-instance Learning with a Buffer-based Approach," in *International Conference on Image Analysis and Processing*. Springer, 2023, pp. 1–12. 8