DAS-MIL: Distilling Across Scales for MIL Classification of Histological WSIs

Gianpaolo Bontempo[1,2], Angelo Porrello[1], Federico Bolelli[1], Simone Calderara[1], and Elisa Ficarra[1]

[1] University of Modena and Reggio Emilia, Italy
{*name.surname*}@unimore.it
[2] University of Pisa, Italy
{*name.surname*}@phd.unipi.it

Abstract. The adoption of Multi-Instance Learning (MIL) for classifying Whole-Slide Images (WSIs) has increased in recent years. Indeed, pixel-level annotation of gigapixel WSI is mostly unfeasible and time-consuming in practice. For this reason, MIL approaches have been profitably integrated with the most recent deep-learning solutions for WSI classification to support clinical practice and diagnosis. Nevertheless, the majority of such approaches overlook the multi-scale nature of the WSIs; the few existing hierarchical MIL proposals simply flatten the multi-scale representations by concatenation or summation of features vectors, neglecting the spatial structure of the WSI. Our work aims to unleash the full potential of pyramidal structured WSI; to do so, we propose a graph-based multi-scale MIL approach, termed DAS-MIL, that exploits message passing to let information flows across multiple scales. By means of a knowledge distillation schema, the alignment between the latent space representation at different resolutions is encouraged while preserving the diversity in the informative content. The effectiveness of the proposed framework is demonstrated on two well-known datasets, where we outperform SOTA on WSI classification, gaining a +1.9% AUC and +3.3% accuracy on the popular Camelyon16 benchmark. The source code is available at https://github.com/aimagelab/mil4wsi.

Keywords: Whole-slide Images · Multi-instance Learning · Knowledge Distillation

1 Introduction

Modern microscopes allow the digitalization of conventional glass slides into gigapixel Whole-Slide Images (WSIs) [18], facilitating their preservation and retrieval, but also introducing multiple challenges. On the one hand, annotating WSIs requires strong medical expertise, is expensive, time-consuming, and labels are usually provided at the slide or patient level. On the other hand, feeding modern neural networks with the entire gigapixel image is not a feasible approach, forcing to crop data into small patches and use them for training. This process

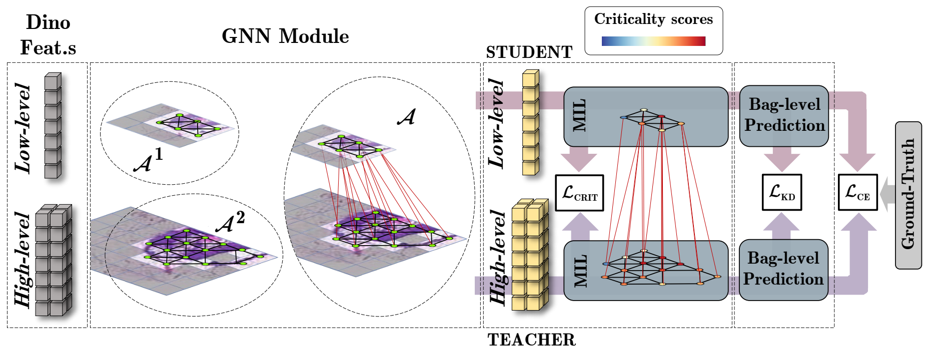

Fig. 1. Overview of our proposed framework, DAS-MIL. The features extracted at different scales are connected (8-connectivity) by means of different graphs. The nodes of both graphs are later fused into a third one, respecting the rule "part of". The contextualized features are then passed to distinct attention-based MIL modules that extract bag labels. Furthermore, a knowledge distillation mechanism encourages the agreement between the predictions delivered by different scales.

is usually performed considering a single resolution/scale among those provided by the WSI image.

Recently, Multi-Instance Learning (MIL) emerged to cope with these limitations. MIL approaches consider the image slide as a bag composed of many patches, called instances; afterwards, to provide a classification score for the entire bag, they weigh the instances through attention mechanisms and aggregate them into a single representation. It is noted that these approaches are intrinsically flat and disregard the pyramidal information provided by the WSI [15], which have been proven to be more effective than single-resolution [4,13,15,19]. However, to the best of our knowledge, none of the existing proposals leverage the full potential of the WSI pyramidal structure. Indeed, the flat concatenation of features [19] extracted at different resolutions does not consider the substantial difference in the informative content they provide. A proficient learning approach should instead consider the heterogeneity between global structures and local cellular regions, thus allowing the information to flow effectively across the image scales.

To profit from the multi-resolution structure of WSI, we propose a pyramidal Graph Neural Network (GNN) framework combined with (self) Knowledge Distillation (KD), called DAS-MIL (Distilling Across Scales). A visual representation of the proposed approach is depicted in Fig. 1. Distinct GNNs provide contextualized features, which are fed to distinct attention-based MIL modules that compute bag-level predictions. Through knowledge distillation, we encourage agreement across the predictions delivered at different resolutions, while individual scale features are learned in isolation to preserve the diversity in terms of information content. By transferring knowledge across scales, we observe that the classifier self-improves as information flows during training. Our proposal has proven its effectiveness on two well-known histological datasets, Camelyon16 and TCGA lung cancer, obtaining state-of-the-art results on WSI classification.

2 Related Work

MIL approaches for WSI classification. We herein summarize the most recent approaches; we refer the reader to [11,26] for a comprehensive overview.

Single-Scale. A classical approach is represented by AB-MIL [16], which employs a side-branch network to calculate the attention scores. In [28], a similar attention mechanism is employed to support a double-tier feature distillation approach, which distills features from pseudo-bags to the original slide. Differently, DS-MIL [19] applies non-local attention aggregation by considering the distance with the most relevant patch. The authors of [20] and [25] propose variations of AB-MIL, which introduce clustering losses and transformers, respectively. In addition, SETMIL [31] makes use of spatial-encoding transformer layers to update the representation. The authors of [7] leverage DINO [5] as feature extractor, highlighting its effectiveness for medical image analysis. Beyond classical attention mechanisms, there are also algorithms based on Recurrent Neural Networks (RNN) [4], and Graphs Neural Networks (GNN) [32].

Multi-Scale. Recently, different authors focused on multi-resolution approaches. DSMIL-LC [19] merges representations from different resolutions, *i.e.*, low instance representations are concatenated with the ones obtained at a higher resolution. MS-RNNMIL [4], instead, fed an RNN with instances extracted at different scales. In [6], a self-supervised hierarchical transformer is applied at each scale. In MS-DA-MIL [13], multi-scale features are included in the same attention algorithm. [10] and [15] exploit multi-resolution through GNN architectures.

Knowledge Distillation. Distilling knowledge from a more extensive network (*teacher*) to a smaller one (*student*) has been widely investigated in recent years [21,24] and applied to different fields, ranging from model compression [3] to WSI analysis [17]. Typically, a tailored learning objective encourages the student to mimic the behaviour of its teacher. Recently, self-supervised representation learning approaches have also employed such a schema: as an example, [5,9] exploit KD to obtain an agreement between networks fed with different views of the same image. In [28], KD is used to transfer the knowledge between MIL tiers applied on different subsamples bags. Taking inspiration from [23] and [30], our model applies (self) knowledge distillation between WSI scale resolutions.

3 Method

Our approach aims to promote the information flow through the different employed resolutions. While existing works [19,20,25] take into account inter-scales interactions 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 based on message passing, which propagates patches' representation according to the natural structure of multi-resolutions WSI; *ii)* a regulation term based on (self) knowledge distillation, which pins the most effective resolution to further guide the training of the other one(s). In the following, we are delving into the details of our architecture.

Feature Extraction Our work exploits DINO, the self-supervised learning approach proposed in [5], to provide a relevant representation of each patch. Differently from other proposals [19,20,28], it focuses solely on aligning positive pairs during optimization (and hence avoids negative pairs), which has shown to require a lower memory footprint during training. We hence devise an initial stage with multiple self-supervised feature extractors $f(\cdot;\theta_1),\ldots,f_M(\cdot;\theta_M)$, one dedicated to each resolution: this way, we expect to promote feature diversity across scales. After training, we freeze the weights of these networks and use them as patch-level feature extractors. Although we focus only on two resolutions at time (*i.e.*, $M=2$) the approach can be extended to more scales.

Architecture. The representations yield by DINO provide a detailed description of the local patterns in each patch; however, they retain poor knowledge of the surrounding context. To grasp a global guess about the entire slide, we allow patches to exchange local information. We achieve it through a Pyramidal Graph Neural Network (PGNN) in which each node represents an individual WSI patch seen at different scales. Each node is connected to its neighbors (8-connectivity) in the euclidean space and between scales following the relation "part of"[3]. To perform message passing, we adopt Graph ATtention layers (GAT) [27].

In general terms, such a module takes as input multi-scale patch-level representations $\mathcal{X}=[\mathcal{X}_1\|\mathcal{X}_2]$, where $\mathcal{X}_1\in\mathbb{R}^{N_1\times F}$ and $\mathcal{X}_2\in\mathbb{R}^{N_2\times F}$ are respectively the representations of the lower and higher scale. The input undergoes two graph layers: while the former treats the two scales as independent subgraphs $\mathcal{A}_1\in\mathbb{R}^{N_1\times N_1}$ and $\mathcal{A}_2\in\mathbb{R}^{N_2\times N_2}$, the latter process them jointly by considering the entire graph $\mathcal{A}$ (see Fig. 1, left). In formal terms:

$$\begin{aligned}\mathcal{H} &= \mathrm{PGNN}(\mathcal{X};\mathcal{A},\mathcal{A}_1,\mathcal{A}_2,\theta_{\mathrm{PGNN}})\\ &= \mathrm{GAT}([\mathrm{GAT}(\mathcal{X}_1;\mathcal{A}_1,\theta_1)\|\mathrm{GAT}(\mathcal{X}_2;\mathcal{A}_2,\theta_2)];\mathcal{A},\theta_3),\end{aligned}$$

where $\mathcal{H}\equiv[\mathcal{H}_1\|\mathcal{H}_2]$ stands for the output of the PGNN obtained by concatenating the two scales. These new contextualized patch representations are then fed to the attention-based MIL module proposed in [19], which produces bag-level scores $y_1^{\mathrm{BAG}},\ y_2^{\mathrm{BAG}}\in\mathbb{R}^{1\times C}$ where C equals the number of classes. Notably, such a module provides additional importance scores $\mathrm{z}_1\in\mathbb{R}^{N_1}$ and $\mathrm{z}_2\in\mathbb{R}^{N_2}$, which quantifies the importance of each original patch to the overall prediction.

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

[3] The relation "part of" connects a parent WSI patch (lying in the lower resolution) with its children, *i.e.*, the higher-scale patches it contains.

metrics are inferred from different WSI zooms, a disagreement may emerge: indeed, we have observed (see Tab. 4) that the higher resolutions generally yield better classification performance. In this work, we exploit such a disparity to introduce two additional optimization objectives, which pin the predictions out of the higher scale as teaching signal for the lower one. Further than improving the results of the lowest scale only, we expect its benefits to propagate also to the shared message-passing module, and so to the higher resolution.

Formally, the first term seeks to align bag predictions from the two scales through (self) knowledge distillation [14,29]:

$$\mathcal{L}_{\mathrm{KD}} = \tau^2 \ \mathrm{KL}(\mathrm{softmax}(\frac{y_1^{\mathrm{BAG}}}{\tau}) \parallel \mathrm{softmax}(\frac{y_2^{\mathrm{BAG}}}{\tau})), \tag{1}$$

where KL stands for the Kullback–Leibler divergence and τ is a temperature that lets secondary information emerge from the teaching signal.

The second aligning term regards the instance scores. It encourages the two resolutions to assign criticality scores in a *consistent* manner: intuitively, if a low-resolution patch has been considered critical, then the average score attributed to its children patches should be likewise high. We encourage such a constraint by minimizing the Euclidean distance between the low-resolution criticality grid map z_1 and its subsampled counterpart computed by the high-resolution branch:

$$\mathcal{L}_{\mathrm{CRIT}} = \|\mathrm{z}_1 - \mathrm{GraphPooling}(\mathrm{z}_2)\|_2^2. \tag{2}$$

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 and then averages the child nodes, hence allowing the comparison at the instance level.

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 $\mathbf{y}$ and carried out through the Cross-Entropy loss $\mathcal{L}_{\mathrm{CE}}(\cdot;\mathbf{y})$; the other one based on knowledge distillation:

$$\min_{\theta}\ (1-\lambda)\mathcal{L}_{\mathrm{CE}}(y_2^{\mathrm{BAG}}) + \mathcal{L}_{\mathrm{CE}}(y_1^{\mathrm{BAG}}) + \lambda\mathcal{L}_{\mathrm{KD}} + \beta\mathcal{L}_{\mathrm{CRIT}}, \tag{3}$$

where λ is a hyperparameter weighting the tradeoff between the teaching signals provided by labels and the higher resolution, while β balances the contributions of the consistency regularization introduced in Eq. (2).

4 Experiments

WSIs Pre-processing. We remove background patches through an approach similar to the one presented in the CLAM framework [20]: after an initial segmentation process based on Otsu [22] and Connected Component Analysis [2], non-overlapped patches within the foreground regions are considered.

Optimization. We use Adam as optimizer, with a learning rate of 2×10^{-4} and a cosine annealing scheduler (10^{-5} decay w/o warm restart). We set $\tau = 1.5$, $\beta =$

Table 1. 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 [19,28].

	Method	Camelyon16		TCGA Lung	
		Accuracy	**AUC**	**Accuracy**	**AUC**
Single Scale	Mean-pooling †	0.723	0.672	0.823	0.905
	Max-pooling †	0.893	0.899	0.851	0.909
	MILRNN [4]	0.806	0.806	0.862	0.911
	ABMIL [16]	0.845	0.865	0.900	0.949
	CLAM-SB [20]	0.865	0.885	0.875	0.944
	CLAM-MB [20]	0.850	0.894	0.878	0.949
	Trans-MIL † [25]	0.883	0.942	0.881	0.948
	DTFD (AFS) [28]	0.908	0.946	0.891	0.951
	DTFD (MaxMinS) [28]	0.899	0.941	0.894	0.961
	DSMIL † [19]	0.915	0.952	0.888	0.951
Multi Scale	MS-DA-MIL [13]	0.876	0.887	0.900	0.955
	MS-MILRNN [4]	0.814	0.837	0.891	0.921
	HIPT † [6]	0.890	0.951	0.890	0.950
	DSMIL-LC † [19]	0.909	0.955	0.913	0.964
	H^2-MIL † [15]	0.859	0.912	0.823	0.917
	DAS-MIL (ours)	**0.945**	**0.973**	**0.925**	**0.965**

1, and $\lambda = 1$. The DINO feature extractor has been trained with two RTX5000 GPUs: differently, all subsequent experiments have been performed with a single RTX2080 GPU using Pytorch-Geometric [12]. To asses the performance of our approach, we adhere to the protocol of [19,28] and use the accuracy and AUC metrics. Moreover, the classifier on the higher scale has been used to make the final overall prediction. Regarding the KD loss, we apply the temperature term to both student and teacher outputs for numerical stability.

Camelyon16. [1] We adhere to the official training/test sets. To produce the fairest comparison with the single-scale state-of-the-art solution, the 270 remaining WSIs are split into training and validation in the proportion 9:1.

TCGA Lung Dataset. It is 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 aim is to classify LUAD *vs* LUSC; we follow the split proposed by DSMIL [19].

4.1 Comparison with the State-of-the-art

Tab. 1 compares our DAS-MIL approach with the state-of-the-art, including both single- and multi-scale architectures. As can be observed: *i)* the joint exploitation of multiple resolutions is generally more efficient; *ii)* our DAS-MIL yields robust and compelling results, especially on Camelyon16, where it provides 0.945 of accuracy and 0.973 AUC (*i.e.*, an improvement of +3.3% accuracy and +1.9%

Table 2. Impact (AUC, Camelyon16) of Eq. 3 hyperparameters.

λ	20×	10×	β	20×	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 3. Impact (Camelyon16) of KD temperature (Eq. 1), $\alpha = \beta = 1.0$.

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

AUC with respect to the SOTA). Finally, we remark that most of the methods in the literature resort to different feature extractors; however, the next subsections prove the consistency of DAS-MIL benefits across various backbones.

4.2 Model Analysis

On the Impact of Knowledge Distillation. To assess its merits, we conducted several experiments varying the values of the corresponding balancing coefficients (see Tab. 2). As can be observed, lowering their values (even reaching $\lambda = 0$, *i.e.*, no distillation is performed) negatively affects the performance. 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. 3 on the bidirectional benefits of knowledge distillation in our multi-scale architecture.

We have also performed an assessment on the temperature τ, which controls the smoothing factor applied to teacher's predictions (Tab. 3). We found that the lowest the temperature, the better the results, suggesting that the teacher scale is naturally not overconfident about its predictions, but rather well-calibrated.

Table 4. Comparison between scales. The target column indicates the features passed to the two MIL layers: the "‖" symbol indicates that they have been previously concatenated.

Input Scale	MIL Target(s)	Accuracy	AUC
10×	10×	0.818	0.816
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

Single-Scale *vs* Multi-Scale. Tab. 4 demonstrates the contribution of hierarchical representations. For single-scale experiments, the model is fed only with patches extracted at a single reference scale. For what concerns multi-scale results, representations can be combined in different ways. Overall, the best results are obtained with 10× and 20× input resolutions; the table also highlights that 5× magnitude is less effective and presents a worst discriminative capability. We ascribe it to the specimen-level pixel size relevant for cancer diagnosis task; different datasets/tasks may benefit from different scale combinations.

Table 5. 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×).

Feature Extractor	Graph Mechanism	Camelyon16 Acc.	Camelyon16 AUC	TCGA Lung Acc.	TCGA Lung 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

The Impact of the Feature Extractors and GNNs. Tab. 5 proposes an investigation of these aspects, which considers both SimCRL [8] 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×. We 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.

H^2-MIL exploits a global pooling layer (IHPool) that fulfils only the spatial structure of patches: as a consequence, 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).

5 Conclusion

We proposed a novel way to exploit multiple resolutions in the domain of histological WSI. We conceived a novel graph-based architecture that learns spatial correlation at different WSI resolutions. Specifically, a GNN cascade architecture is used to extract context-aware and instance-level features considering the spatial relationship between scales. During the training process, this connection is further amplified by a distillation loss, asking for an agreement between the lower and higher scales. Extensive experiments show the effectiveness of the proposed distillation approach.

6 Acknowledgement

This project has received funding from DECIDER, 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). We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

References

1. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., 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) 6
2. Bolelli, F., Allegretti, S., Grana, C.: One DAG to Rule Them All. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3647–3658 (2021) 5
3. Buciluă, C., Caruana, R., Niculescu-Mizil, A.: Model Compression. In: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 535–541 (2006) 3
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019) 2, 3, 6
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021) 3, 4
6. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16144–16155 (2022) 3, 6
7. Chen, R.J., Krishnan, R.G.: Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology. Learning Meaningful Representations of Life, NeurIPS (2022) 3
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020) 8
9. Chen, X., He, K.: Exploring Simple Siamese Representation Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15750–15758 (2021) 3
10. Chen, Z., Zhang, J., Che, S., Huang, J., Han, X., Yuan, Y.: Diagnose Like A Pathologist: Weakly-Supervised Pathologist-Tree Network for Slide-Level Immunohistochemical Scoring. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021) 3
11. Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep Learning for Whole Slide Image Analysis: An Overview. Frontiers in Medicine **6**, 264 (2019) 3
12. Fey, M., Lenssen, J.E.: Fast Graph Representation Learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds (2019) 6
13. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3852–3861 (2020) 2, 3, 6

14. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. In: NIPS Deep Learning and Representation Learning Workshop (2015) 5
15. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: H2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022) 2, 3, 6, 8
16. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: International Conference on Machine Learning. vol. 80, pp. 2127–2136. PMLR (2018) 3, 6
17. Ilyas, T., Mannan, Z.I., Khan, A., Azam, S., Kim, H., De Boer, F.: TSFD-Net: Tissue specific feature distillation network for nuclei segmentation and classification. Neural Networks **151**, 1–15 (2022) 3
18. Kumar, N., Gupta, R., Gupta, S.: Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. Journal of Digital Imaging **33**(4), 1034–1040 (2020) 1
19. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14318–14328 (2021) 2, 3, 4, 6
20. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021) 3, 4, 5, 6
21. Monti, A., Porrello, A., Calderara, S., Coscia, P., Ballan, L., Cucchiara, R.: How many Observations are Enough? Knowledge Distillation for Trajectory Forecasting. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6543–6552 (2022) 3
22. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979) 5
23. Porrello, A., Bergamini, L., Calderara, S.: Robust Re-Identification by Multiple Views Knowledge Distillation. In: Computer Vision – ECCV 2020. pp. 93–110. Springer (2020) 3
24. Qi, L., Kuen, J., Gu, J., Lin, Z., Wang, Y., Chen, Y., Li, Y., Jia, J.: Multi-Scale Aligned Distillation for Low-Resolution Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14443–14453 (2021) 3
25. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. Advances in Neural Information Processing Systems 34 (NeurIPS) **34**, 2136–2147 (2021) 3, 6
26. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis **67**, 101813 (2021) 3
27. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. International Conference on Learning Representations (2018), accepted as poster 4
28. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: 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). pp. 18802–18812 (2022) 3, 4, 6
29. Zhang, L., Bao, C., Ma, K.: Self-distillation: Towards efficient and compact neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4388–4403 (2021) 5

30. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3713–3722 (2019) 3
31. Zhao, Y., Lin, Z., Sun, K., Zhang, Y., Huang, J., Wang, L., Yao, J.: SETMIL: Spatial Encoding Transformer-Based Multiple Instance Learning for Pathological Image Analysis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 66–76. Springer (2022) 3
32. Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al.: Predicting Lymph Node Metastasis Using Histopathological Images Based on Multiple Instance Learning With Deep Graph Convolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4837–4846 (2020) 3