Pattern Recognition 179 (2026) 113642

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Attention-guided knowledge distillation based deep multiple instance learning for gigapixel whole slide image analysis

Weiheng Fu [a,1], Yike Zhou [b,1], Meilian Xu [c], Tianfu Wen [d], Xiaoshuang Shi [a,*], Xiaofeng Zhu [a]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[b] Glasgow College, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[c] School of Electronic Information and Artificial Intelligence, Leshan Normal University, China
[d] Division of Liver Surgery, Department of General Surgery, Sichuan University, West China Hospital Sichuan, Chengdu, China

ARTICLE INFO

Keywords:
Whole slide images
Deep multiple instance learning
Knowledge distillation
Attention-semantic consistent

ABSTRACT

Recently, a **M**ulti-scale **R**epresentation **A**ttention based deep multiple instance learning **N**etwork (**MRAN**) has been proposed to directly extract patch-level features from gigapixel whole slide images, and achieved promising performance on multiple popular datasets. However, it still faces three major limitations: (i) the negative impact of noisy bag-level labels; (ii) semantic inconsistencies between bag-level attention weights and bag-level predictions; and (iii) neglecting the relations of cell images and patch images. To overcome these issues, we propose a novel **A**ttention-**G**uided **K**nowledge **D**istillation based deep multiple instance learning framework, namely **AGKD**. Specifically, it employs knowledge distillation in MRAN to generate soft targets for bag-level images to reduce the impact of noisy labels. Additionally, we propose a novel attention-semantic consistent loss to align attention weights with bag-level predictions, to reduce the semantic inconsistencies. Moreover, we design cell- and patch-relation modules to explore the relationship among cell-level images and patch-level ones, respectively. Extensive experiments on multiple whole slide image datasets demonstrate the superior performance of the proposed framework over MRAN and other recent state-of-the-art methods, with better model interpretability. *All source codes are available at:* https://github.com/zhouyike1313/AGKD.

1. Introduction

Automated analysis of histopathological images plays a crucial role in clinical cancer diagnosis, prognosis assessment, and prediction of treatment outcomes. The advent of digital whole-slide images (WSIs) has greatly advanced the use of deep learning techniques in these image analysis [1]. Unfortunately, WSIs have a gigapixel size (ranging from 1 million to 10 billion pixels), making it impractical to directly input them into deep learning models. To overcome this obstacle, WSIs are typically segmented into non-overlapping small patches to facilitate more efficient processing and analysis. However, this segmentation has a significant challenge in obtaining detailed manual annotations for an enormous amount of patch images, which is costly and time-consuming and restricts the feasibility of traditional supervised learning methods. In light of these challenges, multiple instance learning (MIL) has been widely applied with deep learning for WSI analysis. In MIL-based methods [2–5], each WSI is often regarded as a bag, with the patches containing them as instances. Within the MIL paradigm, in a positive bag, there exists at least one positive instance, while for a negative bag, all instances should be negative.

Existing deep MIL-based methods for WSI analysis view each WSI as a bag and crop it into a set of patches as instances, and then utilize the bag label to train the model [6,7]. These methods have several popular strategies. One of them is to first employ a pre-trained model on ImageNet [8] to obtain feature representations of patches. However, the intrinsic distribution between different data types is often different, and thus this strategy cannot obtain the optimal feature representations of the patches. Another popular strategy for extracting features of WSIs is to employ unsupervised methods, e.g., contrastive learning [9,10], prototypes [11,12], clustering [13]. Unfortunately, these methods often fail to extract optimal patch-level feature representations [14], and even lose some significant semantic information [15]. Compared to previous MIL-based methods, MRAN [14] utilizes a two-branch deep neural network and a multi-scale representation attention mechanism to directly extract patch-level features from gigapixel WSIs, and obtains state-of-the-art performance on some popular datasets. However, MRAN still

* Corresponding author.
E-mail address: xsshi2013@gmail.com (X. Shi).
[1] Co-first authors: They have equal contribution.

https://doi.org/10.1016/j.patcog.2026.113642
Received 8 April 2025; Received in revised form 3 March 2026; Accepted 27 March 2026
Available online 30 March 2026

encounters three limitations as follows: (i) It employs some incorrect bag-level images for model training, since one branch of MRAN utilizes the same labels for bags as their corresponding WSIs; (ii) Although MRAN adopts bag-level attention to reduce the effect of wrong labels at the bag level, its performance is still restricted by the semantic inconsistency of attention weights [16], that is, a significant semantic bag often has a very small attention weight and vice versa; (iii) It does not take into account the relations among different cells and patches, respectively.

To overcome aforementioned limitations of MRAN, we first propose a novel deep framework, Knowledge Distillation based Dual-Branch Network (KDDB) (first introduced in our previous conference paper [17]), so as to reduce the negative impact of noisy bag-level labels by using knowledge distillation and designing an attention-semantic consistent loss, where KDDB utilizes the first and second branches as the student and teacher models, respectively, and proposes a novel attention-semantic consistent loss to further reduce the inconsistency between the bag-level attention weight and the bag-level image-class prediction probability, thereby further mitigating the effect of wrong labels at the bag level. However, KDDB does not consider the relations among cell-level images and patch-level ones, thereby degrading the teacher model performance. To address this issue, we extend KDDB to a novel **A**ttention-**G**uided **K**nowledge **D**istillation based deep MIL framework, namely AGKD, whose structure pipeline is shown in Fig. 1. Specifically, while KDDB using the bag representations generated by the first branch as the input of the second branch, AGKD utilizes the cell-level features extracted by the first branch as the input of the second branch. Then, AGKD designs cell- and patch-relation modules, which leverage Vision Transformer (ViT) blocks [18] to explicitly model the relations of cell-level images and patch-level ones, respectively, and adopt attention mechanisms with sparse connections for improved feature aggregation, so as to boost the performance of the teacher model.

In summary, our main contributions are listed as follows:

- We propose a novel attention-guided knowledge distillation framework based on MRAN for WSI analysis.
- We design an attention-semantic consistent loss to reduce the inconsistency between the attention weight and the class prediction probability of bag-level images, so as to reduce the effect of bag-level wrong labels.
- We further introduce the cell- and patch-relation modules to boost the performance of the teacher model.
- Extensive experimental results on three WSI datasets demonstrate the superior classification and interpretation performance of the proposed framework over recent state-of-the-art methods. We also conduct sensitivity experiments on backbone networks (ViT/CNN) to verify the generality and transferability of the proposed distillation strategy.

The remainder of this paper is organized as follows. Section 2 briefly reviews some popular related methods; Section 3 introduces the preliminaries; Section 4 presents the proposed framework. Section 5 shows and analyzes the experimental results; Finally, Section 6 concludes this paper and points out the future work.

2. Related work

In this section, we briefly review some popular deep multiple instance learning methods and knowledge distillation based methods.

2.1. Deep multiple instance learning

Deep multiple instance learning: which utilizes deep learning models to automatically learn representations of instances within a bag. In the MIL framework, each WSI is considered as a bag consisting of multiple instances. A bag is labeled as positive if it contains at least one positive instance, and negative if all instances are negative. In contrast to standard supervised learning, where each instance is individually labeled, MIL models learn from the bag-level labels and aim to infer the instance-level information indirectly. Due to the gigapixel image size of WSIs, one popular strategy for WSI analysis is to utilize pre-trained models to extract patch-level features. For example, ICMIL [19] generates patch embeddings using a fixed ResNet50 model pre-trained on ImageNet [8], and then iteratively couples the patch feature embedding process with the bag-level classification. However, this reliance on pre-trained models might not provide the optimal features for WSI analysis, because the feature representations learned from general datasets like ImageNet might not align well with the specific characteristics of histopathological images.

To address this limitation, the second strategy employs unsupervised learning to learn domain-specific features, e.g., cluster-based and prototype-based approaches. Cluster-based methods aim to group instances within a bag into clusters based on their similarities, and then use these clusters to help classify the bag. DGMIL [13] proposes a cluster-conditioned feature distribution modeling method and a pseudo label-based iterative feature space refinement strategy to separate positive/negative instances. By contrast, prototype-based methods aim to learn representative prototypes which are used to guide the classification process by comparing the similarity of instances to the prototypes. For instance, ProtoMIL [11] measures patch similarity using prototypes to provide a fine-grained interpretation of the model's predictions, while ProDiv [12] leverage prototypes to guide the division of bags into consistent pseudo-bags, which are used to augment the training data and improve model generalization.

While prototype-based methods rely on learning explicit representative prototypes to guide instance and bag classification, Sconf-MIL [20] utilizes the similarity confidence between unlabeled bag pairs, which represents the probability of two bags sharing the same label, to train a bag-level classifier. It adheres to empirical risk minimization, derives a generalization error bound for convergence, and uses risk correction to mitigate overfitting, offering a novel MIL framework for low-label-cost scenarios.

Although these methods can achieve impressive performance on various types of WSIs, they share a significant limitation: they often fail to interpret the significance of individual instances within the bag. Since MIL models only have access to bag-level labels and lack instance-level supervision, they struggle to identify which specific instances (patches) contribute the most to the bag-level prediction.

Attention-based deep multiple instance learning: To further improve the bag classification performance and interpret significant instances, attention-based deep MIL methods have been proposed by embedding the attention mechanism into the neural network to automatically mine significant features or instances. For example, the attention-based MIL method [6] proposes a learnable attention-based MIL pooling operator, which computes bag-level feature representations as weighted averages of instance features, where weights reflect the predicted contribution of individual patches. Subsequent developments have extended this method along multiple dimensions while addressing histological specificities. Recent advances focus on capturing multi-scale contextual information in WSIs. MSF-WSI [21] integrates a Context-Target Fusion Module (CTFM) and Dense SimSiam Learning (DSL) to effectively learn complementary multi-resolution features during self-supervised pre-training, thereby enhancing semantic segmentation performance in histopathology images. HAG-MIL [7] exemplifies this trend by developing an Integrated Attention Transformer that dynamically identifies discriminative regions across different magnification levels, contrasting with earlier single-scale approaches. Additionally, spatial relationships between neighboring tissue structures have also been incorporated through novel attention formulations. CAMIL [22] introduces a neighbor-constrained attention module that explicitly models spatial dependencies between adjacent tiles using contrastive learning, addressing the limitation of conventional attention mechanisms that process patches in isolation. Moreover, the integration of hierarchical feature

learning represents another evolutionary direction. HA-CSL [23] integrates a hierarchical attention (HA) module to extract significant features at slice-, region-, and channel-level, and a consistency learning (CSL) module to align attention weights between 2D and 3D branches, thereby enhancing both classification performance and model interpretability for 3D MRI image analysis. TransMIL [24] employs self-attention to jointly model morphological characteristics and spatial relationships among patches, creating a unified representation that captures both cellular patterns and tissue architecture. Building upon this, MRAN [14] proposes a multi-scale attention network that simultaneously processes cellular, patch, bag, and whole-slide features. However, three critical limitations emerge in MRAN: first, its performance degrades with noisy bag-level labels due to direct label propagation; second, it overlooks potential inconsistencies between attention weights of bags and their predictions; third, it fails to explicitly explore interdependencies between cellular and patch-level features.

Our prior work KDDB [17] addresses the limitations of MRAN by proposing a dual-branch knowledge distillation framework, which integrates knowledge distillation and an attention-semantic consistent loss, effectively mitigating the impact of noisy bag-level labels. However, KDDB does not consider the relationships between cell-level and patch-level instances. To address this, AGKD extends KDDB by introducing cell- and patch-relation modules composed of ViT blocks, which explore the relationships between these instances. AGKD further enhances these modules with sparse attention mechanisms, improving feature aggregation and enabling more effective capture of semantic dependencies within and between cell- and patch-level images.

For clarity, we summarize the characteristics of different deep MIL methods for WSI analysis in Table 1. As we can see, AGKD uniquely addresses both noise robustness and relation modeling that limit prior methods. By integrating attention-guided knowledge distillation with cell- and patch-relation modules, AGKD effectively mitigates label noise while explicitly modeling inter-dependencies of cell- and patch-level representations, resulting in improved accuracy and interpretability.

2.2. Knowledge distillation

Offline distillation: which employs a one-way knowledge transfer and a two-phase training procedure, and focuses on compressing pre-trained teacher models into lightweight student ones. For instance, EKD [25] proposes an ensemble-based knowledge distillation framework that compresses multiple pre-trained teacher models into student model through parallel branch learning, thereby enhancing classification accuracy and generalization in scenarios with limited training data. While EKD focuses on leveraging ensemble diversity to improve classification performance, PKT [26] transfers knowledge by matching the probability distributions of teacher and student feature spaces, and extends knowledge transfer to representation learning and enables cross-modal transfer.

Online distillation: which updates the teacher model and the student model simultaneously, and the whole knowledge distillation framework is end-to-end trainable. For instance, DML [27] proposes a collaborative framework where multiple student networks learn together, enabling them to co-evolve without requiring a pre-trained teacher by utilizing mutual teaching and learning throughout the training process.To overcome the limitation of naive group-based learning like DML, where all peers end up with similar behaviors, OKDDip [28] introduces auxiliary peers and group leaders to form a diverse set of peer models.

Self-distillation: which exploits internal knowledge within single networks by eliminating separate teacher models. For example, SAD [29] establishes layer-wise attention propagation, distilling upper-layer attention maps to guide lower layers. Additionally, distillation-based training [30] encourages the early exit layer to mimic the output of later exit layer during the training in a multi-exit framework. Similarly, SD [31] also exploits intermediate network states for self-supervised learning, but operates across training epochs rather than within a single forward pass, which transfers the knowledge in the earlier epochs of the network.

Similar to online distillation and self-distillation strategies, our method updates student and teacher models in an end-to-end trainable framework. However, differing from them using the same networks for teacher and student models, our method consists of two branches with different network structures, where the first branch is used as the student model to output bag-level feature representations, and the second branch is used as the teacher one by using bag-level feature representations as the input and outputting the soft targets of bag-level images for the student model.

3. Preliminaries

3.1. Multiple instance learning

Given N WSIs $\{\mathbf{X}_n\}_{n=1}^{N}$, where $\mathbf{X}_n$ denotes the nth WSI, $Y_n \in \{0,1\}$ is the corresponding label of $\mathbf{X}_n$. Each slide $\mathbf{X}_n$ can be divided into a set of small patches $\{\mathbf{X}_{ni}\}_{i=1}^{N_b}$ without overlapping, and N_b is the number of patches in $\mathbf{X}_n$. In the MIL formulation, the slide $\mathbf{X}_n$ is viewed as a bag and Y_n is the bag label. $\mathbf{X}_{ni}$ is viewed as the ith instance in $\mathbf{X}_n$, and $y_{n,i}$ denotes the instance label of $\mathbf{X}_{ni}$. The bag label Y_n is represented as:

$$Y_n = \begin{cases} 0, & \text{if } \sum_i y_{n,i} = 0, \\ 1, & \text{else}, \end{cases} \tag{1}$$

which means that all instances within negative bags are assigned negative labels, while positive bags contain at least one instance with a positive label. Additionally, in the MIL scheme, only the bag labels are available in the training set, while the labels of individual instances are unknown.

3.2. WSI preprocessing

Given a dataset consisting of N WSIs $\{\mathbf{X}_n\}_{n=1}^{N}$, and the WSI $\mathbf{X}_n$ has a corresponding label $Y_n \in \{0,1\}$, where $\mathbf{X}_n \in \mathbb{R}^{C\times H_w\times W_w}$ denotes the nth WSI, C, H_w and W_w represent the number of channels, height and width of each WSI (40x magnification), respectively. In our work, we first resize the size of WSIs to $\frac{H_w}{2}\times\frac{W_w}{2}$ in order to reduce the computational and memory costs of model training. Next, a global binary thresholding algorithm [32] in OpenCV is used to localize the tissue regions in each WSI, which is then divided into a set of non-overlapping bag-level images $\{\mathbf{X}_{ni}\}_{i=1}^{N_b}$, where $\mathbf{X}_{ni} \in \mathbb{R}^{C\times H_b\times W_b}$ denotes the ith bag in $\mathbf{X}_n$, H_b and W_b represent the height and width of the bag, respectively, $N_b = \frac{H_w W_w}{H_b W_b}$ is the number of bags in $\mathbf{X}_n$. Similarly, the bag-level images would be further divided into small patches $\{\mathbf{X}_{nij}\}_{j=1}^{N_p}$, where $\mathbf{X}_{nij} \in \mathbb{R}^{C\times H_p\times W_p}$ denotes the jth patch of $\mathbf{X}_{ni}$, H_p and W_p represent the height and width of each patch, respectively, and $N_p = \frac{H_b W_b}{H_p W_p}$ is the number of patches in $\mathbf{X}_{ni}$. Moreover, in order to extract crucial cell-level information, the patch-level image is considered as consisting of a bunch of cell-level images $\{\mathbf{X}_{nijk}\}_{k=1}^{N_c}$, where $\mathbf{X}_{nijk} \in \mathbb{R}^{C\times H_c\times W_c}$ denotes the kth cell within $\mathbf{X}_{nij}$, H_c and W_c represent the height and width of each cell-level image, respectively, and $N_c = \frac{H_p W_p}{H_c W_c}$ is the number of cells in $\mathbf{X}_{nij}$. In this paper, same as MRAN [14], we set $H_b = W_b = 1024$, $H_p = W_p = 128$ and $H_c = W_c = 32$.

3.3. Multi-scale representation attention network

In this subsection, we briefly introduce the most related work MRAN. As shown in Fig. 1 (the blue dashed box), the MRAN framework begins with a preprocessing stage (a), where each WSI is divided into a set of bags, and each bag is further divided into multiple patches. These patch-level images are then fed into the first branch (b) of the network. This branch employs convolutional layers from ResNet18 [33] to extract feature maps for each patch (with each feature map corresponding to a set

Table 1
The characteristics of deep MIL methods for WSI analysis.

Method	Noisy Label Handling	Relation Modeling	Key Innovations
ICMIL [19]	△ (*No explicit label noise handling; refines patch features via iterative coupling of bag classifier & embedder*)	× (*Implicit patch weight correlations; no spatial/hierarchical modeling*)	•Fixed ResNet50 features extractor •Iterative coupling of bag classifier and patch embedder
DGMIL [13]	△ (*Indirectly filters noise via K-means clustering + extreme instance pseudo-label selection; iteratively refines feature space*)	△ (*Captures instance-cluster correlations; no hierarchical modeling*)	•Cluster-based pseudo-label generation •Iterative feature space refinement
ProtoMIL [11]	△ (*Filters noisy instances via prototype similarity; cluster/separation loss strengthens prototype-instance consistency*)	△ (*Implicit instance correlations via prototypes; no spatial modeling*)	•Prototype-guided instance/bag classification •Similarity-based fine-grained model interpretation
ProDiv [12]	△ (Reduces pseudo-bag noise via prototype-guided division; retains parent bag phenotype)	△ (Instance-cluster correlations via prototype similarity; no hierarchical modeling)	•Prototype generation (mean/attention-based) •Consistent pseudo-bag division
Sconf-MIL [20]	△ (No explicit label noise handling; mitigates overfitting via risk correction)	△ (Bag-bag correlations via similarity confidence; no instance/hierarchical modeling)	•First MIL with similarity-confidence bags •Unbiased empirical risk estimator
ABMIL [6]	△ (No explicit label noise handling; filters trivial instances via attention weights)	△ (Instance-level correlations via attention-weighted average; no spatial/hierarchical modeling)	•Attention-based MIL pooling (gated/non-gated) •End-to-end neural network training
MSF-WSI [21]	× (*Resists noise via CTFM masking + DSL low-level feature learning*)	✓ (*Models cross-resolution (context-target) dependencies via CTFM; DSL links multi-stage features*)	•CTFM for multi-resolution fusion •DSL for multi-stage feature alignment
HAG-MIL [7]	× (*Focus on multi-magnification ROI detection*)	✓ (*Models patch spatial correlations via Integrated Attention Transformer*)	•Integrated Attention Transformer •Dynamic region identification
CAMIL [22]	× (*Focus on spatial dependency*)	✓ (*Models patch spatial correlations via neighbor-constrained attention + Nystromformer*)	•Neighbor-constrained attention for spatial dependencies •Contrastive learning for patch-level representation
HA-CSL [23]	× (*Focus on MRI multi-scale attention*)	✓ (*Models MRI regional correlations via hierarchical attention; aligns 2D-3D attention weights*)	•Hierarchical attention •Consistency learning for 2D-3D branch alignment
Trans-MIL [24]	× (*Enhances feature discriminability via self-attention*)	✓ (*Models patch morphological/spatial correlations via Nystrom self-attention + PPEG*)	•Transformer-based self-attention for instance modeling •PPEG for spatial relations
MuRCL [10]	△ (*Indirectly filters noisy features via RL-selected discriminative sets + set-mixup*)	✓ (*Implicit instance semantic correlations via contrastive learning*)	•Reinforcement learning for feature set selection •Contrastive learning to maximize set agreement
IBMIL [16]	✓ (*Indirectly reduces label confusion via causal deconfounding(backdoor adjustment)*)	× (*No cell/patch relations; focus on causal intervention for bag-level prediction*)	•Causal intervention •Deconfounded predictions
DTFD-MIL [2]	△ (*Mitigates pseudo-bag noise via Grad-CAM + double-tier distillation*)	× (*Only aggregates pseudo-bag features; no spatial relations*)	•Grad-CAM-based instance probability calculation •Double-tier MIL
MRAN [14]	✓ (*Filters trivial instances via 3-level attention*)	✓ (*Models cell-patch-bag hierarchy via $L_{2,1}$-norm attention*)	•Multi-scale representation attention •End-to-end patch feature extraction
KDDB [17]	✓ (*Reduces bag noise via teacher soft labels + attention-semantic loss*)	△ (*Inherits MRAN's cell-patch-bag hierarchical attention; enhances bag-level feature correlations via distillation; no cell-patch relation modeling*)	•Dual-branch knowledge distillation •Attention-semantic consistent loss

Table 1
Continue.

Method	Noisy Label Handling	Relation Modeling	Key Innovations
AGKD (Ours)	✓ (*Reduces bag noise via teacher soft labels + attention-semantic loss; filters trivial instances via 3-level attention*)	✓ (*Explicitly models cell-cell/patch-patch correlations via ViT-Blocks*)	•Knowledge distillation soft targets •Attention-semantic consistent loss •**ViT-based cell/patch relations** •Multi-scale attention

Note: ✓ = Explicitly addressed; × = Not explicitly addressed; △ = Partially/indirectly handled.

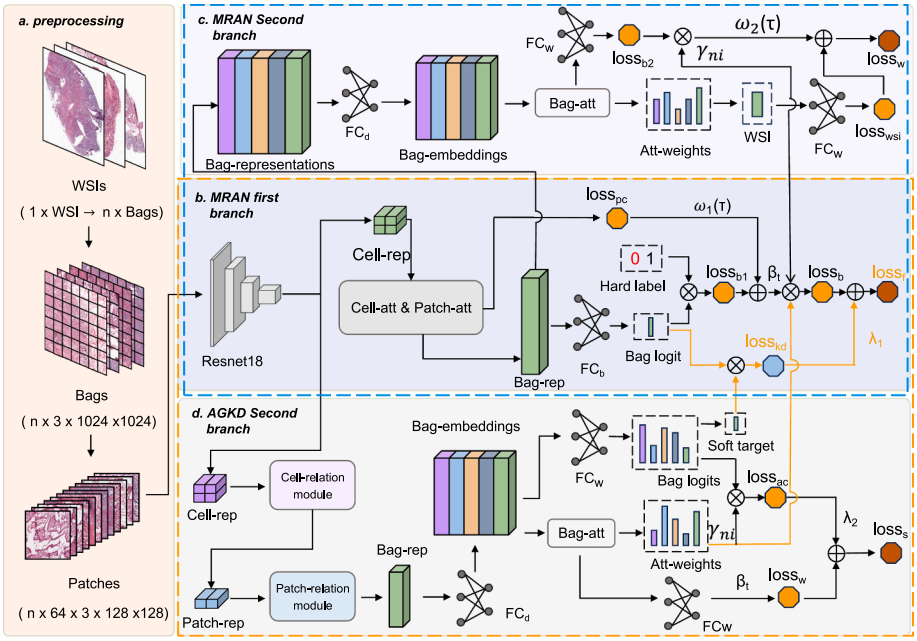


Fig. 1. The structure pass of MRAN and the proposed AGKD, where the blue and yellow dashed boxes contain the structures of MRAN and AGKD, respectively. Note that except the $loss_{kd}$ and λ_1 that are only used for AGKD, the other modules in the MRAN first branch are shared by MRAN and AGKD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of cell-level features) and a fully-connected layer FC_b. These features pass through the cell-level and patch-level attention modules to obtain $loss_{pc}$ and the bag-level image feature representations, which are used with hard labels of the bag-level images to attain $loss_{b_1}$ for bag prediction, and then obtain $loss_b$ (which is made up of $loss_{pc}$ and $loss_{b_1}$) to train the first branch. Additionally, the bag-level image features are also used as the input of MRAN second branch (c), which consists of two fully-connected layers (FC_d and FC_w), employs bag-level attention to learn bag-level attention weights (e.g., γ_{ni}), and utilizes $loss_w$ (which is composed by the loss $loss_{b_2}$ for bag prediction and the loss $loss_{wsi}$ for WSI prediction) to train the second branch.

In WSI analysis, cells can often provide key information for classification. Additionally, because directly dividing each WSI into small patches will cause a large number of patches with wrong labels, MRAN divides each WSI into bag-level images and then further subdivides each bag-level image into patches, so as to reduce the number of wrong labels. Moreover, to mine the significant cell- and patch-level information and reduce the effect of the noisy bag labels, MRAN proposes a multi-scale representation attention as follows:

Multi-scale Representation Attention: Specifically, it uses cell-level attention to calculate weights for cell features within each patch-level image, so as to remove noise and mine significant cell-level information. Patch-level attention is then used to identify and retain significant patches within each bag-level image. Finally, bag-level attention in the MRAN second branch is to determine crucial bags within each WSI. All three attentions are on the basis of the $\ell_{2,1}$-norm based attention [4]. For clarity, we show the formulations of cell-, patch- and bag-level attention in the following.

First, after obtaining the cell-level feature representations $\left\{\mathbf{Z}^c_{nijk}\right\}_{k=1}^{N_c}$ of the patch-level image $\mathbf{X}_{nij}$ from the last convolutional layer of ResNet18, they are fed into the cell-level attention:

$$\mathbf{P}^c_{nijk} = f_b\left(\mathbf{Z}^c_{nijk}\right), \tag{2a}$$

$$\kappa_{nijk} = \frac{\sqrt{\sum_{r=1}^{R}\left(\mathbf{P}^c_{nijkr}\right)^2}}{\sum_{t=1}^{N_c}\sqrt{\sum_{r=1}^{R}\left(\mathbf{P}^c_{nijtr}\right)^2}}, \tag{2b}$$

$$\kappa_{nijk} = \frac{\max\left(\kappa_{nijk} - \frac{\lambda}{N_c}, 0\right)}{\sum_{t=1}^{N_c}\max\left(\kappa_{nijt} - \frac{\lambda}{N_c}, 0\right)}, \tag{2c}$$

$$\mathbf{Z}^p_{nij} = \sum_{k=1}^{N_c}\kappa_{nijk}\mathbf{Z}^c_{nijk}, \tag{2d}$$

where $f_b(\cdot)$ is the fully connected layer FC_b, $\mathbf{P}^c_{nijk}$ and κ_{nijk} denote the logit vector and attention weight of the cell-level image $\mathbf{X}_{nijk}$, respectively, R is the number of classes, and $\mathbf{Z}^p_{nij}$ is the feature representation of the patch-level image $\mathbf{X}_{nij}$. Additionally, $\frac{\lambda}{N_c}$ is a threshold to remove trivial cell-level images from each patch.

Based on the cell-level attention, it can obtain the patch-level feature representations $\mathbf{Z}^p_{nij}$, which are fed into the patch-level attention:

$$\mathbf{P}^p_{nij} = f_b\left(\mathbf{Z}^p_{nij}\right), \tag{3a}$$

$$\xi_{nij} = \frac{\sqrt{\sum_{r=1}^{R}\left(\mathbf{P}^p_{nijr}\right)^2}}{\sum_{t=1}^{N_p}\sqrt{\sum_{r=1}^{R}\left(\mathbf{P}^p_{nitr}\right)^2}}, \tag{3b}$$

$$\xi_{nij} = \frac{\max\left(\xi_{nij} - \frac{\lambda}{N_p}, 0\right)}{\sum_{t=1}^{N_p}\max\left(\xi_{nit} - \frac{\lambda}{N_p}, 0\right)}, \tag{3c}$$

$$\mathbf{Z}^b_{ni} = \sum_{j=1}^{N_p}\xi_{nij}\mathbf{Z}^p_{nij}, \tag{3d}$$

where $\mathbf{P}^p_{nij}$ and ξ_{nij} denote the logit vector and attention weight of the patch-level image $\mathbf{X}_{nij}$, respectively, and $\mathbf{Z}^b_{ni}$ is the feature representation of the bag-level image $\mathbf{X}_{ni}$, $\frac{\lambda}{N_p}$ is a threshold to remove trivial patch-level images in each bag.

After obtaining the bag-level feature representations $\mathbf{Z}^b_{ni}$, they are fed into the bag-level attention in the second branch. The formation of the bag-level attention is:

$$\mathbf{P}^{b_2}_{ni} = f_c\left(\mathbf{Z}^b_{ni}\right), \tag{4a}$$

$$\gamma_{ni} = \frac{\sqrt{\sum_{r=1}^{R}\left(\mathbf{P}^{b_2}_{nir}\right)^2}}{\sum_{t=1}^{N_b}\sqrt{\sum_{r=1}^{R}\left(\mathbf{P}^{b_2}_{nir}\right)^2}}, \tag{4b}$$

$$\gamma_{ni} = \frac{\max\left(\gamma_{ni} - \frac{\lambda}{N_b}, 0\right)}{\sum_{t=1}^{N_b}\max\left(\gamma_{ni} - \frac{\lambda}{N_b}, 0\right)}, \tag{4c}$$

$$\mathbf{Z}^w_n = \sum_{i=1}^{N_b}\gamma_{ni}\mathbf{Z}^b_{ni}, \tag{4d}$$

where $f_c(\cdot)$ denotes two fully connected layer FC_d and FC_w, $\mathbf{P}^{b_2}_{ni}$ and γ_{ni} are the class vector and the attention weight of the bag-level image $\mathbf{X}_{ni}$ in the second branch, respectively, $\frac{\lambda}{N_b}$ is a threshold to remove trivial bags in each WSI, and $\mathbf{Z}^w_n$ is the feature representation of the nth WSI $\mathbf{X}_n$. After obtaining $\mathbf{Z}^w_n$, the class vector $\mathbf{P}^w_n$ can be obtained by $\mathbf{P}^w_n = f_w(\mathbf{Z}^w_n)$, where $f_w(\cdot)$ represents the fully connected layer FC_w.

Loss functions of MRAN: MRAN consists of two branches to update the model parameters. For the first branch, MRAN first connects the cell-level attention and patch-level attention with the loss function for cell-level and patch-level image classification, respectively, it is:

$$loss_{pc} = -\sum_{j=1}^{N_p}\xi_{nij}\left(loss_p + \omega_1(\tau)\sum_{k=1}^{N_c}\kappa_{nijk}\log\left(s\left(\mathbf{P}^c_{nijk}\right)[t]\right)\right), \tag{5}$$

where $loss_p = \log\left(s\left(\mathbf{P}^p_{nij}\right)[t]\right)$ is used for the classification of patch-level images, $\sum_{k=1}^{M_c}\kappa_{nijk}\log\left(s\left(\mathbf{P}^c_{nijk}\right)[t]\right)$ is used for cell-level image classification, $\mathbf{X}_n$ belongs to the tth class so that all patch-level images $\left\{\mathbf{X}^p_{nij}\right\}_{k=1}^{N_p}$ and cell-level images $\left\{\mathbf{X}^c_{nijk}\right\}_{k=1}^{N_c}$ are assumed to be in the tth class, $s(\cdot)$ denotes the softmax function, $\omega_1(\tau)$ is an unsupervised function to balance the patch-level and cell-level image classification, and τ is the number of current training epochs.

Additionally, the loss for bag-level image classification is:

$$\mathrm{loss}_{b_1} = -\log\left(s\left(\mathbf{P}^{b_1}_{ni}\right)[t]\right), \tag{6}$$

where $\mathbf{P}^{b_1}_{ni} = f_b(\mathbf{Z}^b_{ni})$ is the class vector of $\mathbf{X}_{ni}$ in the first branch.

Then, connecting $loss_{pc}$ with loss_{b_1}, it obtains the loss of the first branch in MRAN as follows:

$$\mathrm{loss}_b = \frac{1}{|S|}\sum_{\mathbf{X}_{ni}\in S}\gamma_{ni}\beta_t(\mathrm{loss}_{b_1} + \omega_1(\tau)loss_{pc}), \tag{7}$$

where $|S|$ denotes the number of bag-level images in each batch, $\beta_t = \frac{\frac{N}{N_t}}{\sum_{r=1}^{R}\frac{N}{N_r}}$ represents the weight of the tth class, $N = \sum_{r=1}^{R}N_r$ is the total number of WSIs, N_r is the number of WSIs belonging to the rth class, and $\omega_1(\tau)$ is also used to balance the bag-level and patch-level image classification.

Similarly, for the second branch, MRAN connects the bag-level attention with the loss for the classification of bag-level images to obtain $loss_{b_2}$, it is:

$$\mathrm{loss}_{b_2} = -\sum_{i=1}^{N_b}\gamma_{ni}\log\left(s\left(\mathbf{P}^{b_2}_{ni}\right)[t]\right), \tag{8}$$

Then, combining $loss_{b_2}$ with the loss $loss_{wsi} = -\log\left(s\left(\mathbf{P}^w_n\right)[t]\right)$, which is used for WSI classification, we can obtain the loss for the second branch as follows:

$$\mathrm{loss}_w = \frac{1}{|\mathcal{B}|}\sum_{\mathbf{X}_{ni}\in\mathcal{B}}\beta_t\left(loss_{wsi} + \omega_2(\tau)loss_{b_2}\right), \tag{9}$$

where $|\mathcal{B}|$ denotes the number of WSIs in each batch, $\mathbf{P}^{b_2}_{ni}$ is the class vector of $\mathbf{X}_{ni}$ in the second branch, and $\omega_2(\tau)$ is an unsupervised weight function to balance the WSI and bag-level image classification.

4. Methodology

4.1. Overall of the proposed framework

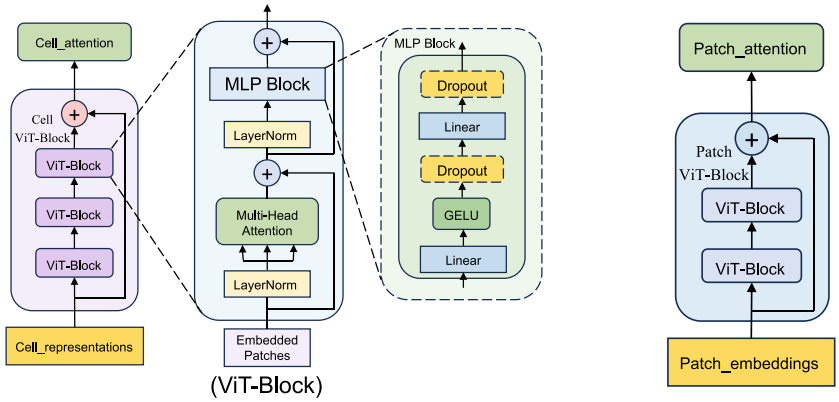
In this subsection, we first briefly introduce the proposed framework and its differences compared to MRAN. As shown in Fig. 1 (the orange dashed box), the proposed AGKD employs MRAN as the backbone, and thus it also consists of a preprocessing stage and two branches. AGKD utilizes the same preprocessing stage as MRAN. Besides incorporating $loss_{kd}$ and λ_1 for knowledge distillation, the first branch of AGKD is identical to that in MRAN.

Moreover, its second branch consists of cell- and patch-relation modules, two fully connected layers, and one bag-level attention. Specifically, AGKD has three major differences to MRAN as follows: (i) Different from MRAN feeding bag-level feature representations into its second branch and without considering the relations among cell- and patch-level images, respectively, AGKD feeds the cell-level features extracted by ResNet18 into its second branch, which utilizes cell-relation and patch-relation modules to explore the relationship among cell- and patch-level images, respectively; (ii) Unlike MRAN directly using hard labels from the corresponding WSIs as the targets for bag-level images to train the first branch, the proposed AGKD first utilizes its second branch as the teacher model to learn soft targets for bag-level images, and then employs the loss $loss_{kd}$ to train the first branch, which is viewed as a student model, so as to largely reduce the negative effect of wrong labels; (iii) the proposed AGKD introduces a novel attention-semantic consistent loss (loss_{ac}) to reduce the semantic inconsistency of bag-level attention, in order to further reduce the negative effect of wrong labels.

For clarity, we introduce the cell- and patch-relation modules in Section 4.2, show the knowledge distillation method in Section 4.3, present the consistent loss and the objective function in Sections 4.4 and 4.5, respectively.

4.2. Cell- and patch-relation modules

For clarity, Fig. 2 shows the structure of cell- and patch-relation modules, whose basic unit is ViT-Block [18]. As we can see, each ViT-Block is composed of three main components: a Multi-Head Attention mechanism, Layer Normalization, and an MLP block. The MLP block itself



(a) Cell-relation module (b) Patch-relation module

Fig. 2. The structure of the cell- and patch-relation modules.

includes linear layers, GELU activation, and dropout layers to prevent overfitting. The whole process of ViT-Block is formulated as:

$$\mathrm{ViT}^{(m)}(\mathbf{O}^{(m)}) = \mathrm{LN}(\mathbf{O}^{(m)} + \mathrm{FFN}(\mathrm{LN}(\mathbf{O}^{(m)} + \mathrm{MH}(\mathbf{O}^{(m)})))), \tag{10}$$

where LN and MH denote LayerNorm and the Multi-Head attention mechanism, respectively, $\mathrm{ViT}^{(m)}$ represents the operations within the mth ViT-Block, $\mathbf{O}^{(m)}$ is the input of the mth ViT-Block, FFN represents the feed forward network.

The cell-relation module consists of three ViT-Blocks to obtain cell-level features, which are then aggregated using cell-level attention to assign weights to different cell representations to emphasize the most relevant features. The whole process of the cell-relation module is formulated as:

$$\kappa^{t}_{nijk} = \mathrm{CA}\left(\mathrm{ViT}^{(3)}\left(\mathrm{ViT}^{(2)}\left(\mathrm{ViT}^{(1)}\left(\mathbf{Z}^{c}_{nijk}\right)\right)\right)\right), \tag{11a}$$

$$\mathbf{Z}^{pt}_{nij} = \sum_{k=1}^{N_c} \kappa^{t}_{nijk}\mathbf{Z}^{c}_{nijk}, \tag{11b}$$

where CA represents the cell-level attention that follows Eqs. (2b) and (2c) to calculate the attention weights of cell-level images, κ^{t}_{nijk} denotes the attention weight of the cell-level image $\mathbf{X}_{nijk}$ in the second branch, $\mathbf{Z}^{pt}_{nij}$ denotes the feature representation of patch-level image $\mathbf{X}_{nij}$ in the second branch.

Similar to the cell-relation module, patch-relation module includes two ViT-blocks to obtain patch-level features, which are then aggregated using patch-level attention to calculate the weights of different patches. It can be represented by:

$$\xi^{t}_{nij} = \mathrm{PA}\left(\mathrm{ViT}^{(2)}\left(\mathrm{ViT}^{(1)}\left(\mathbf{Z}^{pt}_{nij}\right)\right)\right), \tag{12a}$$

$$\mathbf{Z}^{bt}_{ni} = \sum_{j=1}^{N_p} \xi^{t}_{nij}\mathbf{Z}^{pt}_{nij}, \tag{12b}$$

where PA denotes patch-level attention that follows Eq. (3b) and (3c) to calculate the attention weights of patch-level images, ξ^{t}_{nij} represents the attention weight of the patch-level image $\mathbf{X}_{nij}$ in the second branch, $\mathbf{Z}^{bt}_{ni}$ denotes the representation of bag-level image $\mathbf{X}_{ni}$ in the second branch.

4.3. Knowledge distillation

The process of knowledge distillation involves training a teacher model and subsequently transferring its knowledge to a student model. In the proposed framework, the first and second branches are utilized as the student and teacher models, respectively.

After obtaining bag-level image representations through the cell- and patch-relation modules, they are fed into a fully connected layer to attain bag-level feature representations, so as to obtain compact multi-scale information from the input data. To facilitate training and supervision, the bag-level feature representations are used to compute bag-level logits, which are subsequently transformed into soft targets as follows:

$$\mathbf{P}^{t}_{ni} = f_c\left(\mathbf{Z}^{bt}_{ni}\right), \tag{13a}$$

$$\mathrm{s}\left(\mathbf{P}^{t}_{ni}, T\right) = \frac{\exp\left(\mathbf{P}^{t}_{ni}/T\right)}{\sum_{r=1}^{R}\exp\left(\mathbf{P}^{t}_{nir}/T\right)}, \tag{13b}$$

where $s\left(\mathbf{P}^{t}_{ni}, T\right) \in \mathbb{R}^R$ denotes the soft targets of the bag-level image $\mathbf{X}_{ni}$, $\mathbf{P}^{t}_{ni} \in \mathbb{R}^R$ is the logit vector of bag-level image $\mathbf{X}_{ni}$ in the teacher model, R is the number of classes and T is the distillation temperature.

Additionally, the student model is initially trained using hard targets, which correspond to the labels of their respective WSIs. It then leverages soft targets to mitigate the negative impact of incorrect bag-level labels. As stated in [34], soft targets contain the informative dark knowledge from the teacher model. To leverage the soft targets, we adopt the following loss function:

$$\mathrm{loss}_{kd} = \frac{1}{|S|}\sum_{\mathbf{X}_{ni}\in S} s\left(\mathbf{P}^{t}_{ni}, T\right)\log\frac{s\left(\mathbf{P}^{t}_{ni}, T\right)}{s\left(\mathbf{P}^{s}_{ni}, T\right)}, \tag{14}$$

where $\mathbf{P}^{s}_{ni}$ denotes the logit vector of $\mathbf{X}_{ni}$ obtained by the student model.

4.4. Consistent loss

In order to reduce the gap between the attention weight of bag-level image and its prediction class probability, we introduce a novel attention-semantic consistent loss as follows:

$$\mathrm{loss}_{ac} = \frac{1}{|\mathcal{B}|}\sum_{\mathbf{X}_{n}\in\mathcal{B}}\frac{1}{|S|}\sum_{\mathbf{Z}_{ni}\in S}\sum_{i=1}^{N_b}\left|\left|\gamma_{ni} - \frac{s(\mathbf{Z}^{bt}_{ni})[t]}{\sum_{i=1}^{N_b} s(\mathbf{Z}^{bt}_{ni})[t]}\right|\right|^2_F, \tag{15}$$

where N_b denotes the number of bag-level images in the WSI $\mathbf{X}_n$. Eq. (15) aims to obtain that the bag-level image $\mathbf{X}_{ni}$ with a larger γ_{ni} should also have a larger prediction class probability among all bag-level images in its corresponding WSI.

4.5. Objective function

Based on the MRAN first branch loss function Eq. (7), in order to reduce the effect of bag-level wrong labels, we add the knowledge distillation loss in the first branch. Thus, the loss for the first branch becomes:

$$\mathrm{loss}_f = loss_b + \lambda_1\,\mathrm{loss}_{kd}, \tag{16}$$

where λ_1 is a constant to adjust the weight of $loss_{kd}$.

Similarly, in the second branch (teacher model), based on the MRAN second branch loss Eq. (9), then, we combine $loss_w$ with $loss_{ac}$ to obtain the loss for the second branch:

$$\text{loss}_s = \text{loss}_w + \lambda_2 \, \text{loss}_{ac}, \tag{17}$$

where λ_2 is a constant to adjust the weight of $loss_{ac}$.

Because the first branch and the second branch are alternatively trained and nested, based on Eqs. (16) and (17), the objective function of our framework is:

$$loss = v loss_f + (1 - v) loss_s, \tag{18}$$

where $v \in \{0, 1\}$ determines which branch's parameters should be updated.

For clarity, we present the detailed training procedure in Algorithm 1.

Algorithm 1: AGKD.

Input: Training WSIs $\{\mathbf{X}_n\}_{i=1}^{N}$, labels $\{\mathbf{y}_n\}_{n=1}^{N}$, weight function $\omega(\tau)$, network $f_\theta(\cdot)$, augmentation function $g(\cdot)$, distillation temperature T, hyperparameters λ_1, λ_2, class weight β_t, training epochs Γ.

Output: Model parameters $\{\theta^l\}_{l=1}^{L+2}$.

Preprocess $\mathbf{X}_n$ to obtain $\mathbf{X}_{ni}$;
for $\tau \in [1, \Gamma]$ **do**
  for *Stage* $\in [\zeta, \eta]$ **do**
    for *minibatch* $\in \mathcal{S}$ **do**
      Divide $\mathbf{X}_{ni}$ into $\mathbf{X}_{nij}$;
      $\mathbf{Z}^c_{nijk} \leftarrow f_\theta\left(g(\mathbf{X}_{nij})\right)$;
      $\mathbf{Z}^p_{nij}, \kappa_{nijk} \leftarrow att\left(\mathbf{Z}^c_{nijk}\right)$ // cell-attention;
      $\mathbf{Z}^b_{ni}, \xi_{nij} \leftarrow att\left(\mathbf{Z}^p_{nij}\right)$ // patch-attention;
      if *Stage* $= \zeta$ **then**
        $loss_f \leftarrow \text{loss}_b$ // Warm up stage;
      else
        $loss_f \leftarrow \text{loss}_b + \lambda_1 \, \text{loss}_{kd}$ // Eq. (16);
      end
      Back-propagate $loss_f$ to update θ^l $(1 \le l \le L)$ in student model;
    end
    for *minibatch* $\in \mathcal{B}$ **do**
      $\mathbf{Z}^{pt}_{nij}, \kappa^t_{nijk} \leftarrow CA\left(ViT^3\left(\mathbf{Z}^c_{nijk}\right)\right)$ // Cell-relation module;
      $\mathbf{Z}^{bt}_{ni}, \xi^t_{nij} \leftarrow PA\left(ViT^2\left(\mathbf{Z}^{pt}_{nij}\right)\right)$ // Patch-relation module;
      $\mathbf{Z}_n, \gamma_{ni} \leftarrow att\left(\mathbf{Z}^{tv}_{ni}\right)$ // bag-attention;
      $\text{s}\left(\mathbf{P}^t_{ni}, T\right) = \frac{\exp\left(\mathbf{P}^t_{ni}/T\right)}{\sum_{r=1}^{R} \exp\left(\mathbf{P}^t_{nir}/T\right)}$ // soft targets. Eq. (13);
      $loss_s \leftarrow \text{loss}_w + \lambda_2 \, \text{loss}_{ac}$ // Eq. (17);
      Back-propagate $loss_s$ to update θ^l $(L \le l \le L+2)$ in teacher model;
    end
  end
end

5. Experiments

5.1. Datasets

STAD (Stomach Adenocarcinoma) [35]: which is collected from The Cancer Genome Atlas (TCGA) and contains 755 slides from 443 subjects, with 632 positive slides and 123 negative ones. We randomly split the subjects into 60%, 20%, and 20% for training, validation, and testing, respectively.

C-16 (CAMELYON16) [36]: which consists of 399 WSIs of lymph nodes, with 160 positive slides and 239 negative ones. We randomly select 60%, 20%, 20% WSIs to construct training, validation and testing sets, respectively.

HCC (Hepatocellular carcinoma): which is a private dataset collected by West China Hospital, Sichuan University and has 326 WSIs from 326 subjects, with 156 positive slides and 170 negative ones, and each one has the average size of 102,952 × 90,369 pixels. If the patient has an early recurrence of liver cancer within 2 years after liver cancer surgery, the corresponding slide is marked as positive; otherwise, the slide is denoted as negative. We randomly utilize 80%, 10%, and 10% WSIs for model training, validation, and testing, respectively.

5.2. Comparison methods

CLAM_SB [37]: which first utilizes a pre-trained model of ResNet-18 on ImageNet [8] to extract features from patches, and then based on the attention scores obtained from Gated-Attention [6], employs a SVM-based loss to classify patches and single-branch attention to classify WSIs, respectively.

CLAM_MB [37]: which also adopts the same pre-trained model as CLAM-SB for feature extraction, but adopts the SVM-based loss and multi-branch attention to aggregate patch-level features for WSI classification.

DSMIL [9]: which first utilizes self-supervised contrast learning to extract features from patch-level images, and then models the relations of the instances using a trainable distance measurement and classifies WSIs by MIL.

TransMIL [24]: which adopts a pre-trained model of ResNet-18 on ImageNet to extract features from patch-level images, and then classifies WSIs by using the Transformer-based attention MIL method.

DTFD-MIL [2]: which uses the pre-trained model of ResNet-18 on ImageNet to extract features from patch-level images, and then employs the gradient-based calculation of instance probabilities and the attention-based MIL to classify WSIs.

MuRCL [10]: which utilizes reinforcement learning to select discriminative feature sets from patches and trains the model via contrastive learning to maximize agreement between these sets, followed by fine-tuning for WSI classification.

IBMIL [16]: which employs causal intervention via backdoor adjustment to eliminate bias from bag contextual priors, improving MIL robustness by learning deconfounded bag-level predictions.

MRAN [14]: which proposes an end-to-end deep MIL framework with a multi-scale representation attention mechanism, so as to directly extract patch-level features from WSIs and simultaneously mine the significant information at different scales for WSI classification.

5.3. Implementation details

We implement the proposed method with the PyTorch framework on a server with 8 NVIDIA GeForce 3090 (each one has 24 GB memory). We adopt the Adam [38] optimizer with default parameters, and then totally run 20 epochs to alternatively train the first and the second branches. For the first branch, we adjust the learning rate by using the OneCycleLR scheduler [39], with an initialized learning rate of 0.0001. For the second branch, we initialize the learning rate as 0.0001 and then adjust the learning rate to 0.00005 after 5 epochs. Additionally, we set the batch size as 32 for the first branch, and adopt the batch size as 1 for the second branch.

We adopt the same parameter settings as MRAN [14] and empirically set $T = 3$, $\lambda_1 = 5$, and $\lambda_2 = 5$ for our framework. Additionally, for fairness, we employ the same patch size and downsampling rate as the proposed framework for comparison methods, but we adopt their default augmentation to obtain the best performance of themselves.

Table 2
Classification results(%) on three dataset. Best and second-best are bold and underlined.

Metrics	CLAM_SB	CLAM_MB	DSMIL	TransMIL	DTFD-MIL	MuRCL	IBMIL	MRAN	AGKD
Dataset	STAD								
ACC	$92.1_{\pm 1.2}$	$93.1_{\pm 1.0}$	$68.7_{\pm 11.8}$	$92.7_{\pm 3.0}$	$91.6_{\pm 2.8}$	$95.2_{\pm 1.7}$	$74.9_{\pm 7.5}$	$\underline{96.2}_{\pm 2.5}$	$\mathbf{98.7}_{\pm 1.0}$
AUC	$96.1_{\pm 1.7}$	$97.2_{\pm 31.1}$	$71.8_{\pm 9.4}$	$94.1_{\pm 3.8}$	$94.9_{\pm 2.2}$	$95.7_{\pm 3.5}$	$75.7_{\pm 6.1}$	$\underline{98.1}_{\pm 1.7}$	$\mathbf{99.6}_{\pm 0.8}$
F_1	$95.3_{\pm 0.7}$	$95.9_{\pm 0.5}$	$75.5_{\pm 6.8}$	$81.5_{\pm 9.2}$	$94.9_{\pm 2.2}$	$97.2_{\pm 1.0}$	$83.2_{\pm 6.5}$	$\underline{97.4}_{\pm 1.7}$	$\mathbf{99.2}_{\pm 0.6}$
SE	$95.4_{\pm 3.4}$	$95.5_{\pm 2.7}$	$67.1_{\pm 14.4}$	$77.4_{\pm 10.6}$	$92.3_{\pm 3.4}$	$\underline{98.1}_{\pm 1.0}$	$76.8_{\pm 10.4}$	$97.4_{\pm 1.8}$	$\mathbf{98.6}_{\pm 1.5}$
SP	$73.2_{\pm 16.1}$	$80.1_{\pm 5.7}$	$78.7_{\pm 13.1}$	$\underline{90.9}_{\pm 6.7}$	$87.4_{\pm 6.3}$	$78.5_{\pm 7.3}$	$67.9_{\pm 14.3}$	$90.2_{\pm 8.8}$	$\mathbf{98.9}_{\pm 2.2}$
Metrics	CLAM_SB	CLAM_MB	DSMIL	TransMIL	DTFD-MIL	MuRCL	IBMIL	MRAN	AGKD
Dataset	C-16								
ACC	$76.7_{\pm 3.4}$	$75.5_{\pm 2.7}$	$51.5_{\pm 8.1}$	$70.3_{\pm 2.4}$	$59.8_{\pm 2.7}$	$72.0_{\pm 5.5}$	$67.3_{\pm 5.1}$	$\underline{77.5}_{\pm 2.2}$	$\mathbf{81.5}_{\pm 1.7}$
AUC	$77.8_{\pm 3.5}$	$77.3_{\pm 3.9}$	$53.4_{\pm 8.1}$	$67.4_{\pm 2.1}$	$61.6_{\pm 4.5}$	$76.9_{\pm 5.2}$	$63.3_{\pm 6.2}$	$\underline{79.6}_{\pm 4.1}$	$\mathbf{87.2}_{\pm 1.8}$
F_1	$65.0_{\pm 3.9}$	$61.7_{\pm 2.1}$	$57.6_{\pm 1.8}$	$63.9_{\pm 2.9}$	$56.0_{\pm 4.8}$	$63.2_{\pm 7.0}$	$56.9_{\pm 4.4}$	$\underline{65.9}_{\pm 1.2}$	$\mathbf{73.4}_{\pm 2.7}$
SE	$58.5_{\pm 5.3}$	$51.4_{\pm 4.7}$	$\mathbf{82.2}_{\pm 18.9}$	$64.4_{\pm 2.1}$	$\underline{68.4}_{\pm 16.8}$	$60.3_{\pm 5.1}$	$55.8_{\pm 3.2}$	$56.5_{\pm 2.9}$	$67.3_{\pm 10.7}$
SP	$88.4_{\pm 9.0}$	$90.6_{\pm 5.5}$	$29.9_{\pm 21.3}$	$70.1_{\pm 3.9}$	$54.5_{\pm 13.8}$	$78.3_{\pm 8.0}$	$74.5_{\pm 10.0}$	$\underline{90.6}_{\pm 3.2}$	$\mathbf{90.8}_{\pm 7.0}$
Metrics	CLAM_SB	CLAM_MB	DSMIL	TransMIL	DTFD-MIL	MuRCL	IBMIL	MRAN	AGKD
Dataset	HCC								
ACC	$50.6_{\pm 5.7}$	$52.9_{\pm 10.4}$	$\underline{61.8}_{\pm 4.2}$	$50.0_{\pm 9.8}$	$59.4_{\pm 2.5}$	$57.7_{\pm 6.1}$	$59.4_{\pm 5.3}$	$60.0_{\pm 1.4}$	$\mathbf{66.5}_{\pm 2.4}$
AUC	$49.7_{\pm 10.4}$	$56.3_{\pm 13.3}$	$53.9_{\pm 9.7}$	$60.9_{\pm 10.2}$	$56.9_{\pm 5.8}$	$62.7_{\pm 8.8}$	$53.3_{\pm 10.9}$	$\underline{64.0}_{\pm 7.1}$	$\mathbf{66.6}_{\pm 4.3}$
F_1	$52.1_{\pm 9.7}$	$58.2_{\pm 11.3}$	$\underline{59.8}_{\pm 11.6}$	$43.9_{\pm 12.2}$	$51.8_{\pm 14.2}$	$59.6_{\pm 3.5}$	$52.8_{\pm 18.4}$	$59.2_{\pm 7.3}$	$\mathbf{66.9}_{\pm 8.7}$
SE	$56.2_{\pm 9.7}$	$68.2_{\pm 10.3}$	$67.8_{\pm 31.4}$	$52.9_{\pm 7.4}$	$46.6_{\pm 23.3}$	$\mathbf{72.4}_{\pm 17.9}$	$51.5_{\pm 22.5}$	$61.8_{\pm 10.7}$	$\underline{67.8}_{\pm 12.5}$
SP	$43.6_{\pm 19.1}$	$39.3_{\pm 12.1}$	$57.0_{\pm 35.1}$	$49.9_{\pm 18.8}$	$\mathbf{71.6}_{\pm 24.8}$	$48.9_{\pm 10.1}$	$\underline{69.2}_{\pm 18.8}$	$58.4_{\pm 15.1}$	$61.8_{\pm 15.4}$

To assess the proposed framework and comparison methods, we employ five popular metrics: accuracy (ACC), Area Under the Curve (AUC), F_1 score, Sensitivity (SE) and Specificity (SP). Additionally, we randomly split each dataset five times and report their average results obtained by each method.

5.4. Classification results

Table 2 shows the classification results of the proposed AGKD and eight comparison methods on three datasets. As we can see, AGKD outperforms existing methods on all the three datasets: STAD, C-16, and HCC. Specifically, on the STAD dataset, the gain of AGKD is 2.5%, 1.5% and 1.8% over the best competitors in terms of the three major metrics ACC, AUC and F_1-score, respectively. Similarly, on the C-16 dataset, AGKD achieves 4.0%, 7.6% and 7.5% higher ACC, AUC and F_1-score than the best competitors, respectively. Meanwhile, on the HCC dataset, AGKD achieves improvements of 4.7%, 2.6%, and 7.1% over the best competitor in terms of ACC, AUC, and F_1-score, respectively.

5.5. Ablation study

In this subsection, we present the results of our ablation experiments, designed to evaluate the three key components of our framework: knowledge distillation, attention-semantic consistent loss, and the relation modules. Table 3 shows the classification results of different components in our method on the three datasets. Specifically, Baseline refers to the MRAN method, Baseline + kd incorporates knowledge distillation into Baseline, 'Baseline + kd + ac' adds both knowledge distillation and attention-semantic consistent loss, which represents the KDDB method [17], and 'Baseline + kd + ac + pr' further includes the patch-relation module. Finally, AGKD is the proposed method that consists of knowledge distillation, attention-semantic consistent loss, and the cell- and patch-relation modules.

As we can see from Table 3, the two components, knowledge distillation and attention-semantic consistent loss, can boost the classification performance on all the three datasets in terms of the three major metrics ACC, AUC and F_1-score. The patch-relation module can only improve the model classification performance on STAD and C-16 in terms of ACC, AUC and F_1-score. However, by combining the cell-relation module with the patch-relation one, they can further boost the performance

Table 3
Classification results(%) of ablation study on the three datasets. Best and second-best are bold and underlined.

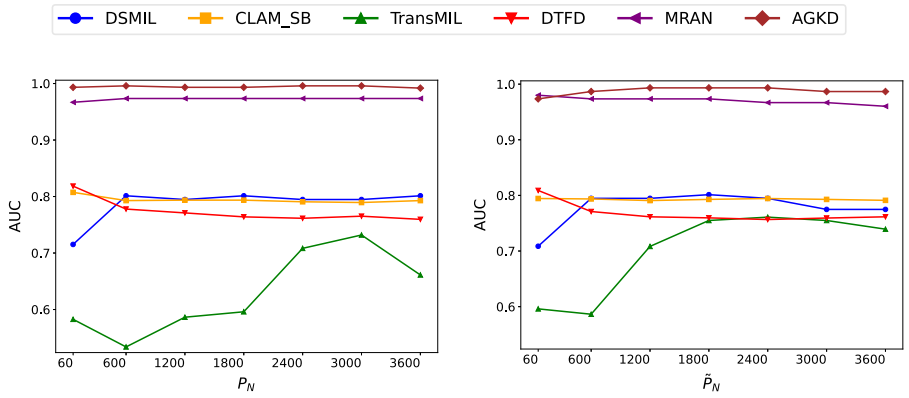
Dataset	Method	ACC	AUC	F_1	SE	SP
STAD	Baseline	$96.2_{\pm 2.5}$	$98.1_{\pm 1.7}$	$97.4_{\pm 1.7}$	$97.4_{\pm 1.8}$	$90.2_{\pm 8.8}$
	Baseline + kd	$97.7_{\pm 0.9}$	$99.1_{\pm 0.5}$	$98.6_{\pm 0.5}$	$\mathbf{98.7}_{\pm 1.0}$	$92.2_{\pm 4.0}$
	Baseline + kd + ac	$97.8_{\pm 1.0}$	$99.3_{\pm 0.6}$	$98.7_{\pm 0.5}$	$\underline{98.6}_{\pm 0.9}$	$93.2_{\pm 5.2}$
	Baseline + kd + ac + pr	$\underline{98.0}_{\pm 1.1}$	$\underline{99.3}_{\pm 0.5}$	$\underline{98.9}_{\pm 0.6}$	$98.5_{\pm 0.7}$	$\underline{95.0}_{\pm 6.2}$
	AGKD	$\mathbf{98.7}_{\pm 1.0}$	$\mathbf{99.6}_{\pm 0.8}$	$\mathbf{99.2}_{\pm 0.6}$	$98.6_{\pm 0.9}$	$\mathbf{98.9}_{\pm 2.2}$
Dataset	Method	ACC	AUC	F_1	SE	SP
C-16	Baseline	$77.5_{\pm 2.2}$	$79.6_{\pm 4.1}$	$65.9_{\pm 1.2}$	$56.5_{\pm 4.4}$	$\underline{90.6}_{\pm 3.2}$
	Baseline + kd	$78.5_{\pm 2.4}$	$81.3_{\pm 3.5}$	$68.0_{\pm 7.2}$	$61.6_{\pm 13.4}$	$89.0_{\pm 6.6}$
	Baseline + kd + ac	$79.5_{\pm 2.5}$	$83.6_{\pm 2.2}$	$71.6_{\pm 3.7}$	$\underline{67.8}_{\pm 6.9}$	$87.3_{\pm 6.1}$
	Baseline + kd + ac + pr	$\underline{80.5}_{\pm 1.5}$	$\underline{86.2}_{\pm 2.8}$	$\underline{73.3}_{\pm 3.1}$	$\mathbf{68.4}_{\pm 9.6}$	$88.3_{\pm 7.1}$
	AGKD	$\mathbf{81.5}_{\pm 1.7}$	$\mathbf{87.2}_{\pm 1.8}$	$\mathbf{73.4}_{\pm 2.7}$	$67.3_{\pm 10.7}$	$\mathbf{90.8}_{\pm 7.5}$
Dataset	Method	ACC	AUC	F_1	SE	SP
HCC	Baseline	$60.0_{\pm 1.4}$	$64.0_{\pm 7.1}$	$59.2_{\pm 7.3}$	$61.8_{\pm 10.7}$	$58.4_{\pm 15.1}$
	Baseline + kd	$62.4_{\pm 7.8}$	$63.2_{\pm 7.9}$	$61.2_{\pm 11.3}$	$59.9_{\pm 14.3}$	$\mathbf{64.4}_{\pm 12.1}$
	Baseline + kd + ac	$64.1_{\pm 6.0}$	$\underline{66.0}_{\pm 7.2}$	$\underline{65.9}_{\pm 9.6}$	$\mathbf{73.0}_{\pm 10.1}$	$55.1_{\pm 4.0}$
	Baseline + kd + ac + pr	$\underline{65.3}_{\pm 6.6}$	$61.4_{\pm 8.0}$	$62.5_{\pm 9.6}$	$\underline{68.7}_{\pm 15.5}$	$59.4_{\pm 20.9}$
	AGKD	$\mathbf{66.5}_{\pm 2.4}$	$\mathbf{66.6}_{\pm 4.3}$	$\mathbf{66.9}_{\pm 8.7}$	$67.8_{\pm 12.5}$	$\underline{61.8}_{\pm 15.4}$

of 'Baseline + kd + ac' on all the three datasets in terms of ACC, AUC and F_1-score. Therefore, experimental results demonstrate the effectiveness of the three key components in our framework.

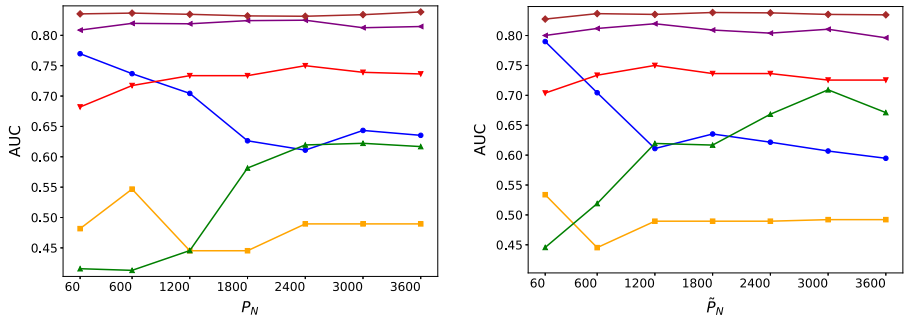
Across all three evaluation datasets, the proposed AGKD achieves state-of-the-art performance on the three core comprehensive metrics of ACC, AUC and F_1-score, which fully validates the effectiveness of each core module and the overall design of the proposed framework. It is observed that AGKD does not yield the optimal sensitivity (SE) on three datasets, nor the optimal specificity (SP) on the HCC dataset. This is not a reflection of insufficient module integration, but rather stems from the inherent SE-SP trade-off in binary classification and dataset-specific characteristics, which are key considerations for clinically applicable model design.

5.6. Interpretation study

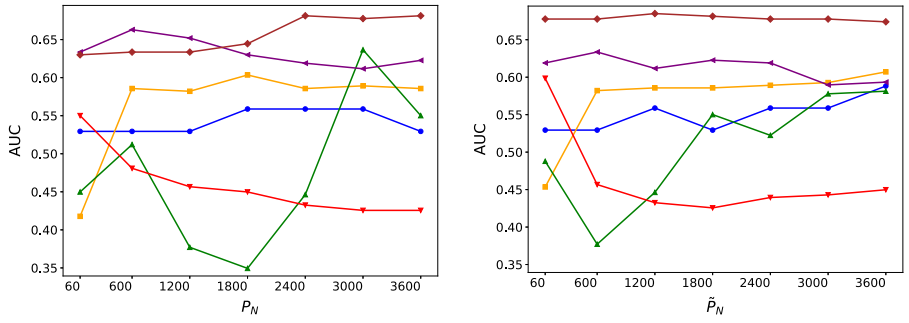
Apart from the classification experiments, we also conduct interpretation experiments to assess the model interpretability of the proposed



(a) STAD



(b) C-16



(c) HCC

Fig. 3. Classification results (AUC) of six methods at different numbers of selected patches on the three datasets. Both P_N and $\widetilde{P}_N$ are the total number of selected patches. In P_N, MRAN and AGKD select the top 30 patches for each bag, with the variable top selected numbers of bags in each WSI, while in $\widetilde{P}_N$, MRAN and AGKD select the top 60 bags for each WSI, with variable top selected numbers of patches in each bag.

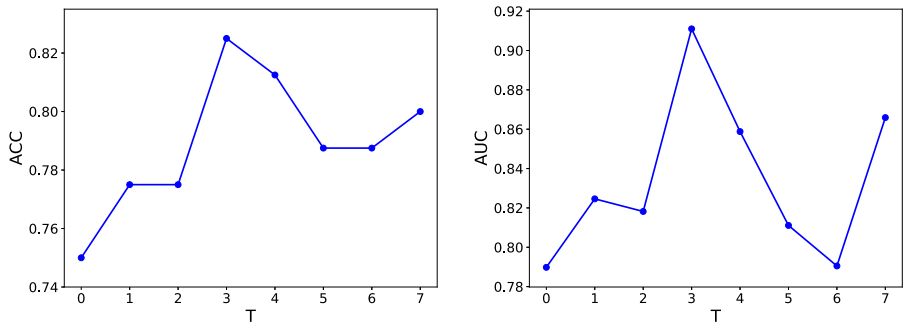
AGKD on the three datasets, by comparing it with five comparative interpretation methods: DSMIL, CLAM-SB, TransMIL, DTFD-MIL and MRAN. Similar to MRAN [14], we employ the popular metric AUC to assess their selected patch-level images. Specifically, we first utilize the selected patch-level images of each method from the original testing set to construct a new testing dataset for each method, and then assess their classification performance by using the new testing set. For fairness, all methods select the same number of patch-level images from each WSI. Fig. 3 shows the AUC of different methods on different numbers of selected patch-level images. As we can see, AGKD achieves the best AUC even using a very small number of selected patch-level images, which infers that our method has better model interpretability.

5.7. Parameters analysis

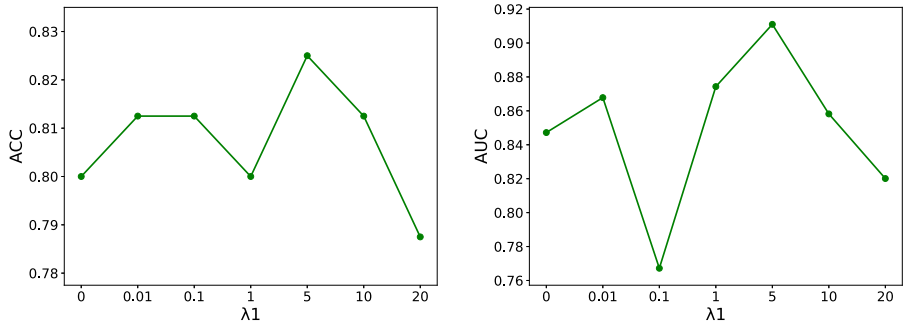
In this subsection, we explore the effect of the hyperparameter temperature (T), λ_1 and λ_2 on the performance of the proposed AGKD. Specifically, we conduct experiments with different values of T during [0,7], and λ_1 and λ_2 during [0,20] on the CAMELYON 16 dataset. As shown in Fig. 4, when $T = 3$, $\lambda_1 = 5$ and $\lambda_2 = 5$, AGKD can obtain the best performance. Similar observations can be found on the other two datasets.

5.8. Sensitivity test on backbone network

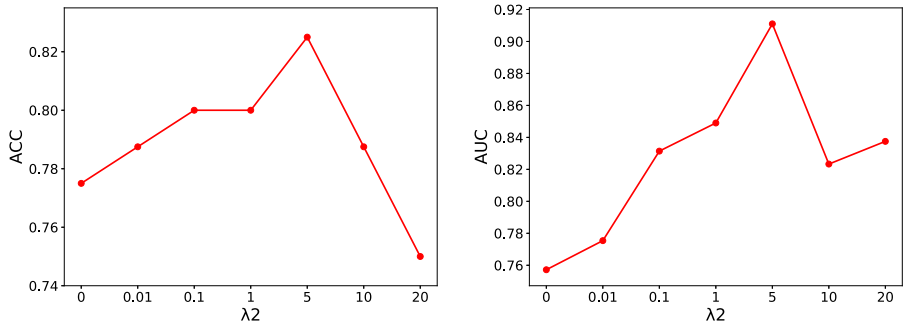
To verify the applicability of the proposed framework to CNN backbones, we conduct a sensitivity test on the CAMELYON16 dataset, replacing the ViT-Block structures in the cell-relation and patch-relation modules with a CNN module, while keeping all other training settings, hyperparameters and model components completely consistent with the original ViT-based AGKD. The classification results of the CNN-based variant are presented in Table 4 alongside the original ViT-based framework. The results show that the CNN-based AGKD still achieves a significant performance improvement (79.5% ACC and 83.8% AUC) over the



(a) Temperature



(b) λ_1



(c) λ_2

Fig. 4. Parameters analysis of the proposed AGKD on the CAMELYON16 dataset. (a) The accuracy and AUC of AGKD under different values of temperature with fixing $\lambda_1 = 5$ and $\lambda_2 = 5$; (b) The accuracy and AUC of AGKD under different values of λ_1 with fixing $T = 3$ and $\lambda_2 = 5$; (c) The accuracy and AUC of AGKD under different values of λ_2 with fixing $T = 3$ and $\lambda_1 = 5$.

Table 4
Classification results(%) of the backbone network sensitivity on the CAMELYON16 dataset. Best and second-best results are bold and underlined, respectively.

Dataset	Method	ACC	AUC	F_1	SE	SP
C-16	MRAN	$77.5_{\pm 2.2}$	$79.6_{\pm 4.1}$	$65.9_{\pm 1.2}$	$56.5_{\pm 4.4}$	$\underline{90.6}_{\pm 3.2}$
	CNN-based AGKD	$\underline{79.5}_{\pm 2.1}$	$\underline{83.8}_{\pm 3.7}$	$\underline{70.3}_{\pm 5.3}$	$\underline{66.4}_{\pm 10.2}$	$87.3_{\pm 5.2}$
	ViT-based AGKD	$\mathbf{81.5}_{\pm 1.7}$	$\mathbf{87.2}_{\pm 1.8}$	$\mathbf{73.4}_{\pm 2.7}$	$\mathbf{67.3}_{\pm 10.7}$	$\mathbf{90.8}_{\pm 7.5}$

MRAN baseline, which further confirms the backbone-agnostic nature of our core distillation strategy, whose effectiveness is not tied to ViT, but generalizes seamlessly to CNN architectures. The ViT-based AGKD achieves higher overall performance due to the global modeling capability of the self-attention mechanism, which is more suited to capturing long-range dependencies between discrete cell- and patch-level features.

5.9. Limitations of the proposed method

While the proposed AGKD demonstrates superior classification performance and interpretability for gigapixel WSI analysis across multiple datasets, it still has several limitations. The first limitation lies in the increased computational overhead introduced by the ViT-based relation modules. The average training time of AGKD per epoch on the C-16 dataset reaches 1.8 h on a single NVIDIA GeForce 3090 GPU, which is 50% longer than the baseline MRAN, and the peak GPU memory occupancy is also increased due to the multi-head self-attention mechanism in the ViT-Block structures. The second limitation is the dependence of the model performance on key hyperparameters. The parameter analysis results in Section 5.7 show that the distillation temperature T, and the λ_1 and λ_2 have a non-negligible impact on the final classification performance of the model. While we provide empirically optimal values for these hyperparameters that are applicable to most WSI analysis

tasks, fine-tuning is still required when migrating the framework to new datasets with different data distributions.

5.10. Discussion

Table 2 and Fig. 3 demonstrate that the proposed AGKD can achieve superior classification and interpretation performance over recent state-of-the-art methods. Additionally, Tables 3–4 illustrate the significance the designed modules. This suggests that the superior performance of the proposed framework is majorly caused by (i) the knowledge distillation to reduce the negative impact of bag-level wrong labels; (ii) the attention-semantic consistent loss to further reduce the negative effect of bag-level wrong labels and meanwhile boost the interpretation performance of the bag-level attention; (iii) the cell-/patch-relation module can mine the relations of cell- and patch-level images to further boost the teacher model's performance.

Although the proposed AGKD focuses on binary classification in this study, it can be easily extended to multi-class tasks by only modifying the output layer to accommodate multiple class labels. Additionally, in this paper the proposed framework is only applied to single-modal data (WSIs), however, it can also be extended to handle multi-modal data [40], such as genomic data, clinical records, and multi-modal imaging (e.g., CT and MRI). Specifically, a multi-input module can be used to process these diverse data types, our relation modules can be extended to explore the relations of multi-modals and the attention mechanism can be used to fuse their feature representations, and our knowledge distillation strategy can help handle noisy or missing data across modalities. Therefore, in the future, we will not only continue to improve the proposed framework on binary classification tasks, but also extend it to handle multi-class and multi-modal tasks, enhancing its applicability and performance in broader medical image understanding scenarios.

6. Conclusion

In this paper, we propose a novel deep MIL framework, AGKD, which integrates attention-guided knowledge distillation into MRAN to generate soft targets for bag-level images and mitigate the negative impact of noisy bag-level labels, and introduce the cell- and patch-relation modules to mine the relations of cell- and patch-level images, respectively, to boost the performance of the teacher model. Extensive experiments demonstrate the superior classification and interpretation performance of the proposed model over recent state-of-the-art methods, and the effectiveness of the proposed knowledge distillation strategy, attention-semantic consistent loss, and the cell-/patch-relation modules.

CRediT authorship contribution statement

Weiheng Fu: Writing – original draft, Validation; **Yike Zhou:** Methodology, Data curation, Conceptualization; **Meilian Xu:** Supervision; **Tianfu Wen:** Supervision, Data curation; **Xiaoshuang Shi:** Writing – review & editing, Supervision, Resources, Project administration; **Xiaofeng Zhu:** Supervision, Project administration.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276052), and Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2022YGRH014).

References

[1] M. Gadermayr, M. Tschuchnig, Multiple instance learning for digital pathology: a review of the state-of-the-art, limitations & future potential, Comput. Med. Imaging Graph. 112 (2024) 102337.
[2] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S.E. Coupland, Y. Zheng, DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 18802–18812.
[3] H. Jin, J. Shen, L. Cui, X. Shi, K. Li, X. Zhu, Dynamic graph based weakly supervised deep hashing for whole slide image classification and retrieval, Med. Image Anal. 101 (2025) 103468.
[4] X. Shi, F. Xing, K. Xu, P. Chen, Y. Liang, Z. Lu, Z. Guo, Loss-based attention for interpreting image-level prediction of convolutional neural networks, IEEE Trans. Image Process. 30 (2020) 1662–1675.
[5] T. Wang, Q. Dai, A patch distribution-based active learning method for multiple instance Alzheimer's disease diagnosis, Pattern Recognit. 150 (2024) 110341.
[6] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, 2018, pp. 2127–2136.
[7] C. Xiong, H. Chen, J.J.Y. Sung, I. King, Diagnose like a pathologist: transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2023, p. 9.
[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[9] B. Li, Y. Li, K.W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 14318–14328.
[10] Z. Zhu, L. Yu, W. Wu, R. Yu, D. Zhang, L. Wang, MuRCL: multi-instance reinforcement contrastive learning for whole slide image classification, IEEE Trans. Med. Imaging 42 (5) (2022) 1337–1348.
[11] D. Rymarczyk, A. Pardyl, J. Kraus, A. Kaczyńska, M. Skomorowski, B. Zieliński, ProtoMIL: multiple instance learning with prototypical parts for whole-slide image classification, in: Machine Learning and Knowledge Discovery in Databases, 2023, pp. 421–436.
[12] R. Yang, P. Liu, L. Ji, ProDiv: prototype-driven consistent pseudo-bag division for whole-slide image classification, Comput. Methods Programs Biomed. 249 (2024) 108161.
[13] L. Qu, X. Luo, S. Liu, M. Wang, Z. Song, DGMIL: distribution guided multiple instance learning for whole slide image classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022, pp. 24–34.
[14] H. Xiang, J. Shen, Q. Yan, M. Xu, X. Shi, X. Zhu, Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis, Med. Image Anal. 89 (2023) 102890.
[15] B. Shi, X. Liu, F. Zhang, MLCN: metric learning constrained network for whole slide image classification with bilinear gated attention mechanism, in: International Workshop on Computational Mathematics Modeling in Cancer Analysis, 2022, pp. 35–46.
[16] T. Lin, Z. Yu, H. Hu, Y. Xu, C.W. Chen, Interventional bag multi-instance learning on whole-slide pathological images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 19830–19839.
[17] W. Fu, M. Xu, J. Wu, X. Shi, K. Li, X. Zhu, Knowledge distillation based dual-branch network for whole slide image analysis, in: International Workshop on Machine Learning in Medical Imaging, 2024, pp. 392–401.
[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[19] H. Wang, L. Luo, F. Wang, R. Tong, Y. Chen, H. Hu, L. Lin, H. Chen, Iteratively coupled multiple instance learning from instance to bag classifier for whole slide image classification, in: International Conference on Medical Image Computing and Computer Assisted Intervention, 2023, pp. 467–476.
[20] X. Zhang, Y. Xu, X. Liu, Multiple instance learning from similarity-confidence bags, Pattern Recognit. 146 (2024) 109984.
[21] H. Wang, E. Ahn, J. Kim, A multi-resolution self-supervised learning framework for semantic segmentation in histopathology, Pattern Recognit. 155 (2024) 110621.
[22] O. Fourkioti, M. De Vries, C. Bakal, CAMIL: context-aware multiple instance learning for cancer detection and subtyping in whole slide images, in: International Conference on Learning Representations, 2024, pp. 1–16.
[23] S. Pang, Y. Chen, X. Shi, R. Wang, M. Dai, X. Zhu, B. Song, K. Li, Interpretable 2.5 D network by hierarchical attention and consistency learning for 3D MRI classification, Pattern Recognit. 164 (2025) 111539.
[24] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, y. zhang, TransMIL: transformer based correlated multiple instance learning for whole slide image classification, in: Advances in Neural Information Processing Systems, 34, 2021, pp. 2136–2147.

[25] U. Asif, J. Tang, S. Harrer, Ensemble knowledge distillation for learning improved and efficient networks, in: European Conference on Artificial Intelligence, IOS Press, 2020, pp. 953–960.
[26] N. Passalis, A. Tefas, Learning deep representations with probabilistic knowledge transfer, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 268–284.
[27] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4320–4328.
[28] D. Chen, J.-P. Mei, C. Wang, Y. Feng, C. Chen, Online knowledge distillation with diverse peers, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, 2020, pp. 3430–3437.
[29] Y. Hou, Z. Ma, C. Liu, C.C. Loy, Learning lightweight lane detection CNNs by self attention distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1013–1021.
[30] M. Phuong, C.H. Lampert, Distillation-based training for multi-exit architectures, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1355–1364.
[31] C. Yang, L. Xie, C. Su, A.L. Yuille, Snapshot distillation: teacher-student optimization in one generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2859–2868.
[32] C. Malathy, U. Sharma, C.M. Naidu, U.U. Pratheebha, A new approach for recognition of implant in knee by template matching, Indian J. Sci. Technol. 9 (37) (2016).
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[34] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: NIPS Deep Learning and Representation Learning Workshop, 2015, pp. 1–9.
[35] K. Tomczak, P. Czerwińska, M. Wiznerowicz, Review the cancer genome atlas (TCGA): an immeasurable source of knowledge, Contemporary Oncology/Współczesna Onkologia 2015 (1) (2015) 68–77.
[36] 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 318 (22) (2017) 2199–2210.
[37] M.Y. Lu, D.F.K. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570.
[38] D.P. Kingma, Adam: a method for stochastic optimization, arxiv preprint arXiv:1412.6980 (2014).
[39] L.N. Smith, N. Topin, Super-convergence: very fast training of neural networks using large learning rates, in: Artificial Intelligence and Machine Learning for Multi-domain Operations Applications, Vol. 11006, SPIE, 2019, pp. 369–386.
[40] C. Hong, J. Yu, J. Zhang, X. Jin, K.-H. Lee, Multimodal face-pose estimation with multitask manifold deep learning, IEEE Trans. Ind. Inf. 15 (7) (2019) 3952–3961.

Weiheng Fu received the BS degree in computer science and technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2022. He is currently working toward the MS degree in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China. His research interests include deep learning and medical image analysis.

Yike Zhou is currently working toward the BS degree in the Glasgow College at the University of Electronic Science and Technology of China. His research interests include large language model and medical image analysis.

Meilian Xu received her BE, MSc and PhD, all in computer science, from East China Normal University, Peking University and the Univerty of Manitoba, respectively. She is now a professor at the School of Electronic Information and Artificial Intelligence, Leshan Normal University. Her research interests include machine learning, nature inspired optimization algorithms, medical imaging, parallel algorithms design and performance improvement on different architectures.She is a member of the ACM society.

Tianfu Wen is a professor and serves as the director of division of liver surgery at West China Hospital of Sichuan University, and a member of the biomedical ethics committee of West China Hospital. He engages in clinical and basic research in the fields of liver cancer.

Xiaoshuang Shi is a faculty member in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China (UESTC). He obtained his PhD degree (2019) from University of Florida, Master degree (2013) from Tsinghua University, and Bachelor degree (2009) from Northwestern Polytechnical University. Before joining UESTC, he worked as a Postdoctoral fellow at the National Institutes of Health (NIH) (2020.01–2021.04), and as a research assistant at Tsinghua University (2013.09–2015.04). His major research interests include large-scale data retrieval, deep learning, medical image analysis.

Xiaofeng Zhu is a faculty member of University of Electronic Science and Technology of China, Chengdu, China. His current research interests include large-scale multimedia retrieval, feature selection, sparse learning, data preprocess, and medical image analysis.