MHD-Net: Memory-Aware Hetero-Modal Distillation Network for Thymic Epithelial Tumor Typing With Missing Pathology Modality

Huaqi Zhang, Jie Liu, Weifan Liu, Huang Chen, Zekuan Yu, Yixuan Yuan, *Senior Member, IEEE*, Pengyu Wang, and Jing Qin, *Senior Member, IEEE*

***Abstract*—Fusing multi-modal radiology and pathology data with complementary information can improve the accuracy of tumor typing. However, collecting pathology data is difficult since it is high-cost and sometimes only obtainable after the surgery, which limits the application of multi-modal methods in diagnosis. To address this problem, we propose comprehensively learning multi-modal radiology-pathology data in training, and only using uni-modal radiology data in testing. Concretely, a *Memory-aware Hetero-modal Distillation Network* (MHD-Net) is proposed, which can distill well-learned multi-modal knowledge with the assistance of memory from the teacher to the student. In the teacher, to tackle the challenge in hetero-modal feature fusion, we propose a novel *spatial-differentiated hetero-modal fusion module* (SHFM) that models spatial-specific tumor information correlations across modalities. As only radiology data is accessible to the student, we store pathology features in the proposed *contrast-boosted typing memory module* (CTMM) that achieves type-wise memory updating and stage-wise contrastive memory boosting to ensure the effectiveness and generalization of memory items. In the student, to improve the cross-modal distillation, we propose a *multi-stage memory-aware distillation* (MMD) scheme that reads memory-aware pathology features from CTMM to remedy missing modal-specific information. Furthermore, we construct a Radiology-Pathology Thymic Epithelial Tumor (RPTET) dataset containing paired CT and WSI images with annotations. Experiments on the RPTET and CPTAC-LUAD datasets demonstrate that MHD-Net significantly improves tumor typing and outperforms existing multi-modal methods on missing modality situations.**

***Index Terms*—Knowledge distillation, missing modality, memory network, thymic epithelial tumor typing.**

I. Introduction

THYMIC Epithelial Tumor (TET) [1] is a common primary anterior mediastinal tumor, which mainly occurs in adults between 30–50 years old. According to the definition by the World Health Organization (WHO), TETs can be divided into eight types in terms of thymomas (A, AB, B1, B1+B2, B2, B2+B3, B3) and thymic carcinomas (TC) [2]. Considering that different types of TETs show distinguishability in tumor sizes and the morphology of thymic epithelial cells, doctors typically develop type-specific treatment and prognosis schemes for patients. The TET diagnosis contains two main steps as shown in Fig. 1: 1) Before the surgery, doctors usually make a Masaoka-Koga staging [3] according to radiology data (computed tomography, CT) to obtain preliminary information; 2) After the surgery, doctors will perform a quantitative assessment based on pathology data (whole-slide image, WSI) for obtaining accurate typing results.

Hence, jointly exploiting radiology and pathology data is a promising approach for accurate TET diagnosis. Many recent works [4], [5], [6], [7], [8], [9], [10] also recommend adopting multi-modal learning due to the significant complementarity between multi-modal data, and show that multi-modal methods generally perform better than uni-modal methods. However, radiology and pathology data have distinct clinical importances and acquisition difficulties, so the patient-level data collected from real medical scenes usually have the missing modality. For a TET patient, we can easily obtain radiology data before surgery as the radiology examination is non-invasive and low-cost. In

Manuscript received 2 July 2023; revised 11 February 2024; accepted 8 March 2024. Date of publication 18 March 2024; date of current version 7 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 81571836, and in part by the Theme-based Research Scheme of Hong Kong Research under Grant Council T45-401/22-N. *(Corresponding authors: Jie Liu; Pengyu Wang; Jing Qin.)*

Huaqi Zhang is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100091, China, and also with the School of Nursing, The Hong Kong Polytechnic University, Hong Kong SAR 999077, China.

Jie Liu is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100091, China (e-mail: 20112040@bjtu.edu.cn).

Weifan Liu is with the College of Science, Beijing Forestry University, Beijing 100091, China.

Huang Chen is with the Department of Pathology, China-Japan Friendship Hospital, Beijing 100029, China.

Zekuan Yu is with the Academy of Engineering and Technology, Fudan University, Shanghai 200433, China.

Yixuan Yuan and Pengyu Wang are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR 999077, China (e-mail: y10180292@mail.ecust.edu.cn).

Jing Qin is with the School of Nursing, The Hong Kong Polytechnic University, Hong Kong SAR 999077, China (e-mail: harry.qin@polyu.edu.hk).

Digital Object Identifier 10.1109/JBHI.2024.3376462

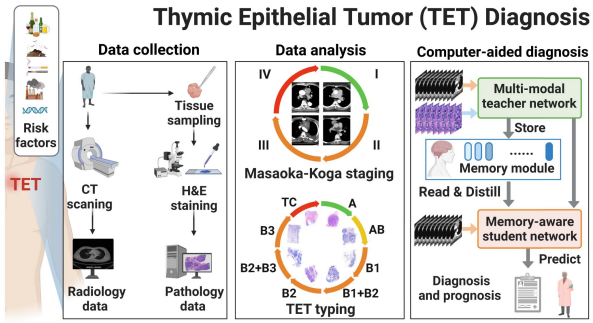


Fig. 1. Workflow of thymic epithelial tumor (TET) diagnosis using the proposed memory-aware hetero-modal distillation network (MHD-Net), which solves a clinical issue that multi-modal radiology-pathology data is available in training, and only uni-modal radiology data is accessible in test. (created with biorender.com).

contrast, pathology data provides more accurate information for TET diagnosis. Nevertheless, pathology data is high-cost and time-consuming, and we can only obtain credible it after surgery since TET is a typical mixed tumor [3]. Therefore, doctors expect to achieve an effective preoperative diagnosis using only radiology data, thereby providing reliable guidance for surgery or conservative treatment. From the method perspective, we can summarize the above clinical requirement as multi-modal radiology-pathology data can be collected and applied to train a model but only uni-modal radiology data is available for testing. To meet this requirement, it is desirable to develop a model, which can effectively learn radiology-pathology knowledge to ensure a high-performance radiology data typing (missing pathology modality).

Recently, many works have been proposed to tackle the missing modality problem. But for our task, shared information extraction methods [11], [12], [13], [14] may extract insufficient modal-invariant features from hetero-modal data, and modal generation methods [15], [16], [17], [18] cannot directly produce hetero-modal data with distinct colors, appearances, and structures. By comparison, knowledge distillation methods [19], [20], [21] can leverage the teacher and student with distinct architectures to learn and transfer different knowledge. Existing studies [22], [23], [24] also show that the knowledge distillation from multi-modal data to uni-modal data can be well implemented. Motivated by these, we propose constructing a multi-modal teacher to effectively learn complementary radiology-pathology knowledge, as well as a uni-modal student to achieve accurate radiology data typing by bridging the teacher and student. Although multi-modal knowledge distillation methods [22], [23], [24], [25], [26] have certain achievements, they still face the following problems in our task: 1) How to learn effective multi-modal knowledge [24], [25] in the teacher from radiology-pathology data is the first challenge. Actually, there is a large gap in tumor information distributions of radiology and pathology data. Directly fusing these hetero-modal features usually leads to suboptimal teacher performance. 2) How to transfer modal-specific information from the teacher to student [22], [23], [26] is the second challenge. As the student can only access radiology data, while the multi-modal teacher and uni-modal student features have distinct distributions. Some pathology-specific information is difficult to directly distill into the student, resulting in suboptimal student performance.

To address the above problems, we propose a memory-aware hetero-modal distillation network (MHD-Net), which learns multi-modal knowledge from the complementary radiology-pathology data, and transfers the well-learned knowledge from the teacher to student, thereby achieving an accurate TET typing by only using radiology data. MHD-Net consists of three key components: 1) In the multi-modal teacher network, we propose a novel spatial-differentiated hetero-modal fusion module (SHFM) to combine heterogeneous radiology and pathology features. Concretely, SHFM is the first attempt to adaptively model the correlations between radiology and pathology tumor information via spatial-specific 3D involutions for effective hetero-modal feature fusion. 2) To mitigate the impact of missing modal-specific information, between the teacher and student, we propose a contrast-boosted typing memory module (CTMM) to store type-wise pathology features from the teacher, which guides the student's training for improving uni-modal typing performance. We also design a contrastive memory boosting (ConMB) loss to improve intra-stage memory effectiveness and inter-stage memory complementarity. (3) In the memory-aware student network, we develop a multi-stage memory-aware distillation (MMD) scheme to read memory-aware pathology features from CTMM and then guide the hetero-modal feature fusion in the student. This scheme successfully bridges the gap between multi-modal teacher and uni-modal student features, resulting in better student performance. In addition, we construct a radiology-pathology thymic epithelial tumor (RPTET) dataset, which collects the paired CT and WSI data from 126 TET patients, and provides accurate labels according to experts' annotations. The main contributions are:

- We propose MHD-Net, which is the first attempt to address a clinically relevant issue that improves TET typing solely using the preoperative radiology data by distilling the well-learned radiology-pathology knowledge.
- We propose SHFM to obtain reliable radiology-pathology fusion features for the teacher, which correlates cross-modal tumor information on each spatial position to ensure high-quality hetero-modal feature fusion.
- We propose CTMM to remedy missing modal-specific information for the student, which stores and boosts type-wise pathology features as memories to ensure efficacious cross-modal knowledge distillation.
- We publish the RPTET dataset, which makes up for the lack of publicly available benchmarks. Experiments show that MHD-Net improves TET typing significantly and performs favorably against state-of-the-art methods.

II. Related Works

A. Learning With Incomplete Multi-Modal Data

Modality-incomplete data is common in multi-modal learning and can significantly affect the model performance. Hence, many solutions [27], [28], [29], [30], [31], [32], [33] have been developed for multi-modal learning with missing modalities.

In general, they can be divided into three categories. The first category of methods directly integrates all available modalities by designing reliable multi-modal fusion methods. For example, mmformer [27] adopts hybrid modality-specific encoders to extract each modality feature, then designs a modality-correlated encoder to establish the correlations across modalities. TFusion [28] learns the relations between existing modalities using Transformer, and estimates modal-specific weight maps for producing effective multi-modal fusion features. However, such multi-modal fusion methods are sensitive to serious missing modality situations, where more than one missing modality may significantly degrade the model's performance.

In comparison, shared representation extraction methods [12], [13], [14], [34] explicitly disentangle multi-modal features in the latent space to extract modal-invariance information for improving model robustness. Among them, Yang et al. [12] propose a dual disentanglement network to transfer modality-specific features to tumor-specific representations for brain tumor segmentation with missing modalities. Zhou et al. [13] construct a latent correlation model to combine individual modality representations for obtaining the shared multi-source representations. Ouyang et al. [14] propose a margin loss to model the relations between modality representations and gain modality-invariant information by fusing disentangled modality representations. Although these methods show tolerance performance, they ignore the influence of modal-specific information. In other words, model performance will be reduced dramatically if an important modality is missing.

The third category is modality generation methods [15], [16], [17], [18], which leverage available modalities to generate missing modalities for completing multi-modal data in training and testing. Pan et al. [17] build a Generative Adversarial Network (GAN) with a feature-consistency constraint to generate missing modalities, and present a spatial cosine module to ensure the disease-image specificity in data. MouseGAN++ [18] introduces a disentanglement mechanism into GAN to transfer multi-modal data as the attribute and content features. Nevertheless, the above methods are unsuitable for handling heterogeneous multi-modal data due to distribution differences, such as the radiology-pathology data in our task.

B. Knowledge Distillation

Knowledge distillation methods [20], [21], [23], [24] are also applicable to missing modalities, which can learn complete knowledge using the multi-modal teacher, and then transfer useful knowledge to the uni-modal student. KD-Net [20] introduces a knowledge distillation strategy, where the well-trained multi-modal teacher network can guide the uni-modal student network for brain tumor segmentation. Xing et al. [23] propose a discrepancy and gradient-guided distillation network, which can selectively absorb reliable multi-modal knowledge from the teacher via the discrepancy-induced contrastive distillation and gradient-guided knowledge refinement schemes in the student. Li et al. [24] construct a dense adaptive grouping distillation network, which decouples multi-modal data into modality-shared and modality-specific information in the teacher, and performs hierarchical and grouping distillation for enhancing the uni-modal student. Overall, knowledge distillation has been shown to be effective in methods that use multi-modal data available in training, and only access uni-modal data in testing. However, the heterogeneity between multi-modal data usually leads to modal-specific knowledge being insufficiently distilled from the teacher to the student. In the proposed MHD-Net, we further build CTMM between the teacher and student to store generic missing modality features for guiding cross-modal distillation. With this design, missing modal-specific information can be supplemented to improve uni-modal classification results.

C. Memory Network

Memory networks [35], [36], [37], [38] are widely studied in various computer vision tasks, which aim to store universal features as parameters to improve model generalization and stability. For example, Alonso et al. [35] establish a contrastive learning module to maintain the class-wise memory bank, which can produce high-quality representations for the same-class labeled and unlabeled samples. Liu et al. [36] develop a memory-guided semantic learning network that first aligns multi-modal features via graph convolutions, and then stores cross-modal shared semantic information to enhance video understanding. Lee et al. [37] propose to store the long-term motion context using memory alignment learning. By recalling well-stored memories, it can achieve better prediction from inputs with limited dynamics. Motivated by them, we not only construct CTMM, but also propose a ConMB loss to constrain CTMM to store high-quality complementary information in multiple stages, which enables the student to absorb multi-modal knowledge more completely with the assistance of memories.

III. Method

As shown in Fig. 2, MHD-Net contains three main components: 1) Multi-modal teacher network with spatial-differentiated hetero-modal fusion modules (SHFMs) for fusing radiology and pathology features and learning multi-modal knowledge; 2) Contrast-boosted typing memory module (CTMM) for storing type-wise pathology features and remedying missing modal-specific information; 3) Memory-aware student network with a multi-stage memory-aware distillation (MMD) scheme for TET typing using only radiology data.

We adopt 3D ResNet-34 and ResNet-18 [39] as the backbones of the teacher and student, respectively. 3D ResNet-34 and ResNet-18 follow the hierarchical architecture consisting of four stages, and have [3, 4, 6, 3] and [2, 2, 2, 2] ResBlocks in four stages. Concretely, each ResBlock contains two 3D convolutions and a shortcut connection, and each stage will down-sample its input feature to produce the low-scale feature. The 1-st stage produces the feature $f^1 \in \mathbb{R}^{12\times64\times56\times56}$, the 2-nd stage produces the feature $f^2 \in \mathbb{R}^{6\times128\times28\times28}$, the 3-rd stage produces the feature $f^3 \in \mathbb{R}^{3\times256\times14\times14}$, and the 4-th stage produces the feature $f^4 \in \mathbb{R}^{2\times512\times7\times7}$. Firstly, the teacher learns multi-modal knowledge by sufficiently exploiting complementary radiology and pathology data. It consists of three branches: radiology branch B_R, pathology branch B_P, and fusion branch B_F. The

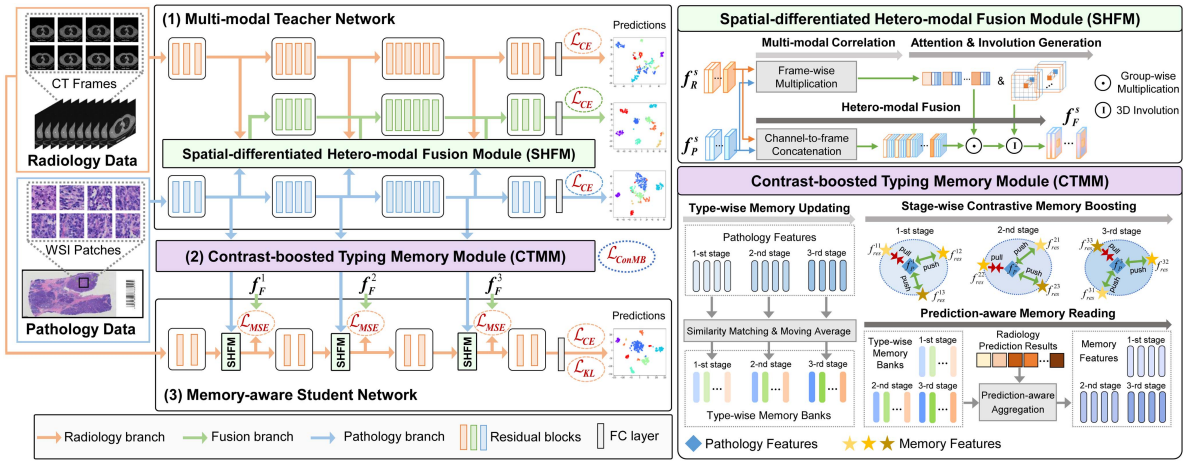


Fig. 2. Overview of the proposed memory-aware hetero-modal distillation network (MHD-Net) for TET typing, which consists of the multi-modal teacher network with spatial-differentiated hetero-modal fusion modules (SHFMs), contrast-boosted typing memory modules (CTMMs) and the memory-aware student network.

parallel B_R and B_P take paired CT and WSI data r and p as inputs for extracting radiology and pathology features f_R and f_P, respectively. Between B_R and B_P, the fusion branch B_F is constructed to learn multi-modal knowledge, where we propose SHFM to combine features f_R and f_P for producing hetero-modal fusion feature f_F. The teacher adopts three cross-entropy (CE) losses ($\mathcal{L}_{CE}$) to separately optimize B_R, B_P and B_F for improving classification performance. Secondly, CTMM stores each stage pathology feature f_P extracted from the teacher as type-wise memory items, which can remedy missing modal-specific information in the student. Under the constraint of the proposed ConMB loss ($\mathcal{L}_{ConMB}$), the intra-stage memory effectiveness and inter-stage memory complementarity are improved. Thirdly, in the memory-aware student network, a MMD scheme is presented to transfer well-learned multi-modal knowledge from the teacher. We first read the pathology memory feature $f_{mem'}$ from CTMM according to the radiology predicted results, and then fuse $f_{mem'}$ and the radiology feature $f_{R'}$ extracted from the student as the memory-aware hetero-modal fusion feature $f_{F'}$ via SHFM. In cross-modal knowledge distillation, we adopt the mean square error (MSE) losses ($\mathcal{L}_{MSE}$) to align f_F and $f_{F'}$, and the Kullback-Leibler (KL) loss ($\mathcal{L}_{KL}$) to align the teacher and student logits. The student is also optimized by the CE loss for better performance.

A. Multi-Modal Teacher Network With SHFM

In MHD-Net, the quality of hetero-modal fusion features determines the teacher's performance. Therefore, we propose SHFM to fuse radiology and pathology features by comprehensively considering their differences and correlations. Fig. 2 displays the multi-modal teacher network with SHFMs. To be specific, the teacher's inputs are given as radiology CT images $\{r_n, y_n\}_{n=1}^{N}$ and pathology WSI images $\{p_n, y_n\}_{n=1}^{N}$, where $r_n \in \mathbb{R}^{T\times 1\times H\times W}$ and $p_n \in \mathbb{R}^{T\times 3\times H\times W}$, T is the number of frames, H and W are the height and width of images respectively, and each pair of r_n and p_n share the same ground truth y_n. In the s-th stage, given the radiology, pathology, and fusion

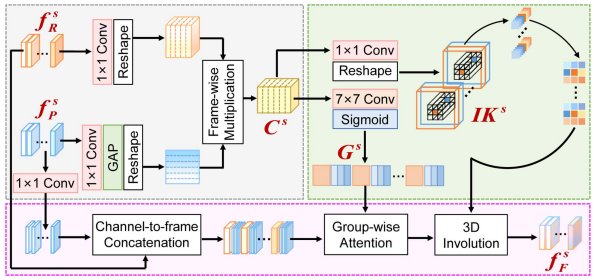


Fig. 3. Structure of spatial-differentiated hetero-modal fusion module (SHFM). **Gray part** produces the multi-modal correlation embedding C^s via a frame-wise multiplication; **Green part** produces the group-wise spatial attention map G^s and 3D involution kernels Ik^s from C^s; **Pink part** produces the hetero-modal fusion feature f_F^s using G^s and Ik^s.

branches B_R^s, B_P^s and B_F^s, the extracted radiology and pathology features are denoted as $f_R^s = B_R^s(f_R^{s-1})$ and $f_P^s = B_P^s(f_P^{s-1})$, where SHFM is used to fuse them, and the hetero-modal fusion feature is given by $f_F^s = SHFM^s(f_R^s, f_P^s) \in \mathbb{R}^{t_s\times c_s\times h_s\times w_s}$, which is the input of B_F^s.

SHFM: From the imaging and clinical perspectives, the tumor in CT is mainly concentrated in a local region across continuous frames. The clustered cancer cells in WSI also reflect tumor information, which shows the potential relations with CT tumor regions. On this basis, we propose a spatial-differentiated hereto-modal fusion module to adaptively model the correlations between different-position tumor information on CT and WSI, thus achieving a high-quality hetero-modal feature fusion. As shown in the gray part of Fig. 3, given the model-specific features f_R^s and f_P^s as inputs of SHFM, where the radiology feature f_R^s first passes through a 1×1 3D convolution, and then we reshape it as $\widehat{f}_R^s \in \mathbb{R}^{t_s\times c_s\times h_s w_s}$. Simultaneously, the pathology feature f_P^s is processed by another 1×1 3D convolution. Considering that cancer cell clusters in WSI are usually larger, then we adopt a global average pooling (GAP) layer for feature down-sampling and reshape the down-sampled feature as $\widehat{f}_P^s \in \mathbb{R}^{c_s\times \kappa^2 t_s}$. Next, we utilize a frame-wise multiplication [40] to model the correlations between each voxel of the radiology feature f_R^s and whole

local regions of the pathology feature f_P^s. For the t-th frame, the expression is given as:

$$C^s(t) = \widehat{f}_R^s(t) \otimes \widehat{f}_P^s, \tag{1}$$

where $\otimes$ is the matrix multiplication, t is the index of frames.

Then, we generate the multi-modal correlation embedding $C^s \in \mathbb{R}^{t_s \times \kappa^2 t_s \times h_s w_s}$. Motivated by the involution [41] that has opposite characteristics with convolution in terms of the spatial-specific and channel-agnostic, we generate the 3D involution kernels by using multi-model correlations, thereby exploring the spatial relations between each CT frame and all WSI patches (frames) for hetero-modal feature fusion. As shown in the green part of Fig. 3, given the multi-modal correlation embedding C^s, we first reshape it as $C^s \in \mathbb{R}^{t_s \times \kappa^2 t_s \times h_s \times w_s}$, and then project it to a group of 3D involution kernels $Ik^s \in \mathbb{R}^{t_s \times \kappa^2 \times h_s \times w_s}$ through a 1×1 3D convolution. Specifically, Ik^s includes a total of $t_s h_s w_s$ involution kernels with $\kappa \times \kappa$ size, where each kernel can aggregate the corresponding local neighborhoods across channels on a frame. The pink part of Fig. 3 illustrates the fusion process of hetero-modal features. We first adopt a 1×1 3D convolution to compress the channels of f_P^s as c_s/r, where r is the reduction rate. Next, we merge the channels and frames of f_P^s to obtain $\tilde{f}_P^s \in \mathbb{R}^{(t_s c_s/r) \times h_s \times w_s}$, and then concatenate $\tilde{f}_P^s$ on each frame of f_R^s in channel dimension to obtain the initial fusion feature $\tilde{f}_F^s \in \mathbb{R}^{t_s \times (c_s + c_s/r) \times h_s \times w_s}$. In SHFM, there are two schemes for hetero-modal feature fusion. We first employ a 7×7 3D convolution and a Sigmoid layer to transfer C^s to the group-wise spatial attention $G^s \in \mathbb{R}^{t_s \times (t_s+1) \times h_s \times w_s}$. For each frame of $\tilde{f}_F^s$, the 1-st channel of G^s enhances its whole CT components, while the 2-nd to the $(t_s{+}1)$-th channels of G^s enhance its components corresponding to each WSI frame:

$$\bar{f}_F^s(t) = [\tilde{f}_{F\{1\}}^s(t) \cdot G_{\{1\}}^s(t), \ldots, \tilde{f}_{F\{1+t_s\}}^s(t) \cdot G_{\{1+t_s\}}^s(t)], \tag{2}$$

where $\cdot$ is element-wise multiplication, $\bar{f}_{F\{b\}}^s(t)$ and $G_{\{b\}}^s(t)$ are the b-th group of the t-th frame of $\tilde{f}_F^s$ and G^s, respectively.

To achieve the hetero-modal feature fusion, we use spatial-specific 3D involution kernels Ik^s to process $\bar{f}_F^s$, where $\kappa \times \kappa$ neighborhoods of each pixel are aggregated by the corresponding involution kernel. The 3D involution procedure is defined as the following multiplication and addition operations:

$$f_F^s(t,c,h,w) = \sum_{u,v \in \Delta\mathcal{K}} \bar{f}_F^s(t,c,h{+}u,w{+}v) \cdot Ik^s(t,u{+}\lfloor \kappa/2 \rfloor, v{+}\lfloor \kappa/2 \rfloor, h, w), \tag{3}$$

where $\lfloor \cdot \rfloor$ indicates the rounding down operation, and $\Delta\mathcal{K}$ is the offset set.

Through the above schemes, we obtain each stage hetero-modal fusion feature f_F^s. Compared with existing multi-model fusion methods, the proposed SHFM comprehensively considers the tumor-related properties of radiology and pathology data, thus explicitly modeling the spatial correlations between potential tumor regions on CT and WSI for enhancing hetero-model fusion features. Moreover, to ensure that the radiology, pathology, and fusion branches B_R^s, B_P^s, and B_F^s can simultaneously extract effective semantic information, we adopt three CE losses to optimize the teacher jointly as:

$$\mathcal{L}_{CE} = \mathrm{CE}\langle z_R, y\rangle + \mathrm{CE}\langle z_P, y\rangle + \mathrm{CE}\langle z_F, y\rangle, \tag{4}$$

where $\mathrm{CE}\langle\cdot\rangle$ is the CE loss, z_R, z_P and z_F denote the predict results of the radiology, pathology, and fusion branches.

B. Contrast-Boosted Typing Memory Module (CTMM)

Considering the complementary modal-specific information between radiology and pathology data, only extracting modal-invariant features is suboptimal for addressing the missing hetero-modality problem. Hence, we propose CTMM between the teacher and student to store pathology features from the teacher, and then guide the hetero-modal feature fusion in the student. The proposed CTMM consists of three phases: 1) Type-wise memory updating, which records type-wise generic pathology information from the teacher; 2) Stage-wise contrastive memory boosting, which improves the effectiveness and distinctiveness of different-stage memory items; 3) Prediction-aware memory reading, which uses the teacher's predicted results as priors to achieve reliable memory reading.

1) Type-Wise Memory Updating: To better retain tumor-related semantic information, we first construct type-wise memory banks $\{\chi_s\}_{s=1}^3$ that store multi-stage pathology features extracted by the pathology branch of the multi-modal teacher network according to data labels, where $\chi_s \in \mathbb{R}^{C \times M \times K}$ denotes the s-th stage memory bank, M and C are the number and dimension of memory items, K is the number of tumor types. Given a pathology feature $f_P^{s,k} \in \mathbb{R}^{t_s \times c_s \times h_s \times w_s}$ from the s-th stage pathology branch, we regard each pixel on $f_P^{s,k}$ as a query and reshape this feature as $f_P^{s,k} \in \mathbb{R}^{c_s \times t_s h_s w_s} = \{q_1^{\mathrm{T}}, q_2^{\mathrm{T}}, \ldots, q_J^{\mathrm{T}}\}$, where the query $q_j \in \mathbb{R}^{c_s} (j = 1, 2, \ldots, J)$, $J = t_s h_s w_s$ and k is the data label. To improve the inter-type semantic information differences between memory items, we only adopt the k-th type pathology features to update the k-th type memory bank $\chi_s(k)$. Specifically, we first compute the cosine similarity matrix $S \in \mathbb{R}^{M \times J}$ between $\chi_s(k)$ and $f_P^{s,k}$ to quantify their correlations. The cosine similarity for the i-th memory item $e_{i,k}$ and the j-th query q_j is computed as:

$$S_{i,j} = \frac{e_{i,k} \cdot q_j}{||e_{i,k}||||q_j||}. \tag{5}$$

After that, we define $r_{(j)} = \arg\max_i S_{i,j}$ to match the most similar memory items $e_{r_{(j)},k}$ for the query q_j. Following the moving average scheme, the memory item e_i is updated by its most relevant queries according to the above computed $r_{(j)}$:

$$e_{i,k} \leftarrow \lambda e_{i,k} + (1-\lambda)\frac{\sum_{j=1}^J \mathbb{1}(r_{(j)} = i) q_j}{\sum_{j=1}^J \mathbb{1}(r_{(j)} = i)}, \tag{6}$$

where $\lambda \in [0, 1]$ represents the decay rate of the moving average, and $\mathbb{1}(\cdot)$ is the indicator function with 0 or 1 value.

2) Stage-Wise Contrastive Memory Boosting: Considering that memories determine the reconstruction quality of the missing modality, we not only need to improve the reliability and generalization of each stage memory items, but also ensure the complementarity between different stage memory items. Inspired by self-supervised contrastive learning [42],

[43], [44], [45], we propose the stage-wise contrastive memory boosting. This phase uses the 1-st, 2-nd and 3-rd stage pathology features $f_P^1 \in \mathbb{R}^{12\times64\times56\times56}$, $f_P^2 \in \mathbb{R}^{6\times128\times28\times28}$ and $f_P^3 \in \mathbb{R}^{3\times256\times14\times14}$ from the teacher, and memory banks $\chi_1 \in \mathbb{R}^{768\times64}$, $\chi_2 \in \mathbb{R}^{768\times128}$ and $\chi_3 \in \mathbb{R}^{196\times256}$ to produce different reconstruction features. For convenience, we denote the s-th stage pathology feature as f_P^s and memory bank as χ_s, $s \in [1, 2, 3]$. Taking the s-th stage as an example, given the pathology feature f_P^s to query the current-stage (s-th) memory bank χ_s, the current-stage reconstruction feature f_{res}^{ss} is generated. Here, we define features f_P^s and f_{res}^{ss} as the intra-stage features, i.e., the positive pair. We also use the pathology feature f_P^s to query the other-stage (s'-th, $s' \neq s$) memory bank $\chi_{s'}$ to generate the other-stage reconstruction feature $f_{res}^{ss'}$. Next, we define features f_P^s and $f_{res}^{ss'}$ as the inter-stage features, which denote the negative pair. By simultaneously pulling the positive pairs and pushing the negative pairs in the feature space, we can effectively enhance different stage memory banks. Referring to well-established memory networks [37], we adopt soft attention to aggregate the memory bank for producing $f_{rec}^{ss'}$. Following the (5), we recompute the cosine similarity matrix $\widehat{S} \in \mathbb{R}^{M\times J\times K}$ between the updated memory bank $\chi_{s'}$ and the pathology feature f_P^s. Then, we use a softmax layer to transfer the cosine similarity matrix $\hat{S}$ into the attention matrix $A \in \mathbb{R}^{M\times J\times K}$:

$$A_{i,j,k} = \frac{\exp(\hat{S}_{i,j,k})}{\sum_{c=1}^{K}\sum_{m=1}^{M}\exp(\hat{S}_{m,j,c})}, \tag{7}$$

Then, we utilize A to aggregate the updated memory bank $\chi_{s'}$ to generate the reconstruction feature $f_{rec}^{ss'}$:

$$f_{rec,j}^{ss'} = \sum_{k=1}^{K}\sum_{i=1}^{M} A_{i,j,k}e_{i,k}, \tag{8}$$

when $s = s'$, the intra-stage reconstruction feature is obtained. In contrast, we obtain the inter-stage reconstruction features. Concretely, given f_P^1, f_P^2 and f_P^3 from the 1-st, 2-nd and 3-th stage pathology branches respectively, we use them to query the 1-st, 2-nd and 3-th stage memory banks χ_1, χ_2 and χ_3 to obtain the intra-stage reconstruction features f_{rec}^{11}, f_{rec}^{22} and f_{rec}^{33}, and the inter-stage reconstruction features f_{rec}^{12}, f_{rec}^{13}, f_{rec}^{21}, f_{rec}^{23}, f_{rec}^{31} and f_{rec}^{32}. Next, we propose the ConMB loss, which aims to improve the similarity between intra-stage features f_P^s and f_{res}^{ss}, as well as to increase the difference between inter-stage features f_P^s and $f_{res}^{ss'}$. Under the constraint of this loss, each stage memory bank can recover higher-quality reconstruction features, and different stage memory banks can record more discriminative pathology information. In each stage, the MSE of the positive pair is defined as the numerator, and the MSE of the negative pair is defined as the denominator. By minimizing the proposed loss, the difference between positive samples is reduced while the difference between negative samples is amplified, thereby implementing contrastive learning-based memory boosting. Finally, the total ConMB loss is defined as:

$$\mathcal{L}_{ConMB} = \frac{\eta_1||f_{rec}^{11} - f_P^1||}{||f_{rec}^{12} - f_P^1|| + ||f_{rec}^{13} - f_P^1||} + \frac{\eta_2||f_{rec}^{22} - f_P^2||}{||f_{rec}^{21} - f_P^2|| + ||f_{rec}^{23} - f_P^2||} + \frac{\eta_3||f_{rec}^{33} - f_P^3||}{||f_{rec}^{31} - f_P^3|| + ||f_{rec}^{32} - f_P^3||}, \tag{9}$$

where η_s is the weight of the s-th $\mathcal{L}_{ConMB}$, which is defined as the softmax-normalized cosine similarity between the intra-stage reconstruction and pathology features. To be specific, $\eta_s = \sigma(1 - \frac{f_{rec}^{ss} \cdot f_P^s}{||f_{rec}^{ss}||||f_P^s||})$, where $\sigma(\cdot)$ is the softmax layer. A lower cosine similarity value indicates the low-quality reconstruction feature f_{rec}^{ss}, which shows that the current stage memory bank fails to store effective information. Hence, this stage should be enhanced in $\mathcal{L}_{ConMB}$. The ConMB and CE losses optimize the multi-modal teacher network jointly as:

$$\mathcal{L}_{Teacher} = \mathcal{L}_{CE} + \alpha_{ConMB}\mathcal{L}_{ConMB}. \tag{10}$$

3) Prediction-Aware Memory Reading: As discussed above, both the memory updating and boosting schemes are applied in teacher training. In the student, to extract missing modal-specific information from well-stored memory banks for hetero-modal feature fusion, we design a prediction-aware memory reading scheme, which can produce memory-aware pathology features according to the radiology prediction results. Specifically, in student training, given a CT image r_n as the input, we first feed it into the radiology branch of the teacher to obtain the uni-modal prediction result $z_{R'} = \sigma(l_{R'})$, where $l_{R'}$ is the logit output by the radiology branch. Considering that $z_{R'}$ contains relatively accurate semantic information reflecting the possible class of the input. Then, we employ it as a prior to read the information corresponding to the input class from well-stored memory banks. Here, each element in $z_{R'}$ is regarded as a typing confidence, and then each stage memory bank χ_s is aggregated by these confidences to generate the memory-aware pathology feature f_{mem}^s, where the s-th stage is expressed as:

$$f_{mem}^s = \text{reshape}\left(\sum_{k=1}^{K} z_{R'}(k) \cdot \chi_s(k)\right), \tag{11}$$

where the reshape operation changes the size of f_{mem}^s as $\mathbb{R}^{t_s\times c_s\times h_s\times w_s}$. Next, the radiology feature extracted by the s-th stage student is denoted as $f_{R'}^s$, and we adopt the proposed SHFM to combine $f_{R'}^s$ with the corresponding stage f_{mem}^s to produce the memory-aware hetero-modal fusion feature $f_{F'}^s$.

C. Memory-Aware Student Network

Since most missing modal-specific information is difficult to distill directly from the teacher to the student due to the difference between multi-modal and uni-modal features, we propose a multi-stage memory-aware distillation scheme in the student, which first reconstructs the pathology memory features from CTMM according to the radiology prediction results, and then produces memory-aware hetero-modal fusion features under the teacher's supervision.

As shown in Fig. 2, the memory-aware student network adopts a four-stage 3D ResNet-18 as the backbone, and takes radiology data $\{r_n, y_n\}_{n=1}^{N}$ as inputs. To distill the well-learned

multi-modal knowledge, we generate the memory-aware hetero-modal fusion features in the student, and then align them with the corresponding stage hetero-modal fusion features of the teacher. Specifically, we first extract the s-th stage radiology feature $f^s_{R'}$ and adopt the prediction-aware memory reading scheme to produce the pathology memory feature f^s_{mem}. Then, we feed $f^s_{R'}$ and f^s_{mem} into the proposed SHFM to adaptively produce the memory-aware hetero-modal fusion feature $f^s_{F'} = M^s_{SHF}(f^s_{R'}, f^s_{mem})$. After that, to ensure the effectiveness of the memory-aware hetero-modal fusion feature $f^s_{F'}$, we adopt the MSE loss to align $f^s_{F'}$ and f^s_F in the feature space, where a 1×1 3D convolution is used to compress the dimension of $f^s_{F'}$, and the s-th stage alignment process is defined as:

$$\mathcal{L}_{MSE} = \sum_{s=1}^{3} \| f^s_F - con_s(f^s_{F'}) \|_2, \tag{12}$$

where $con_s(\cdot)$ indicates the k-th stage 1×1 3D convolution.

We also introduce the KL loss to narrow the semantic gap between the logits output by the teacher and student as:

$$\mathcal{L}_{KL} = -\sum \sigma(l_{Tea}) \log \left(\frac{\sigma(l_{Stu})}{\sigma(l_{Tea})} \right), \tag{13}$$

where l_{Tea} is the teacher logits, and l_{Stu} is the student logits.

Furthermore, we adopt the CE loss with the above MSE and KL losses to optimize the memory-aware student network jointly. The total loss function is written as:

$$\mathcal{L}_{Student} = \mathcal{L}_{CE} + \alpha_{MSE}\mathcal{L}_{MSE} + \alpha_{KL}\mathcal{L}_{KL}, \tag{14}$$

where α_{MSE} and α_{KL} are parameters of the MSE and KL losses to balance their contributions. In summary, compared with previous knowledge distillation methods [20], [23], [24], the proposed MMD scheme avoids the loss of modal-specific information in the student by reading well-stored memory banks and transferring well-learned teacher knowledge, which improves model performance in scenes with missing modality.

IV. Experiments

A. Datasets

1) The Proposed RPTET Dataset: In this work, we construct a multi-modal benchmark called the radiology-pathology thymic epithelial tumor (RPTET) dataset. To the best of our knowledge, this is the first public radiology-pathology dataset with TET typing annotations. This dataset collects 126 TET patients' clinical data, where each patient has paired 3D CT (radiology) and 2D WSI (pathology) data. To produce radiology data, all patients undergo the CT examination before surgery using multiple cross-sectional spiral CT scanners, including the 320-row multi-detector CT (Toshiba Aquilion TM ONE, Tokyo, Japan) and 256-row multi-detector CT (GE revolution, Boston, USA). To obtain pathology data, the patients' thymic epithelial tumors resected after surgery are first fixed in 10% buffered formalin. Next, each whole tumor is divided into multiple pathology slides for Haematoxylin & Eosin (H&E) staining, and the slice scanning image system SQS-600P (Shengqiang Technol. Co. Ltd, Shenzhen, China) is adopted to produce WSIs with the spatial resolution 0.09 μm/pixel. Finally, doctors will select 2-4 WSIs near the largest cross section of the tumor and closely related to the clinical diagnosis for each patient as the initial pathology data for this study. For each patient, its CT contains multiple continuous frames, but the beginning- and ending-phase frames are irrelevant to tumors. Therefore, the initial 20% and final 30% frames are discarded since they do not contain any tumor-related information. We uniformly divide each CT into numerous CT sequences, where each sequence contains 24 frames with 224×224 resolution. Considering that not every CT sequence covers the tumor, we invite experts to annotate CT sequences with tumors into eight types: A, AB, B1, B1+B2, B2, B2+B3, B3, and TC, and to assign another label representing 'other' to CT sequences without tumors. In addition, as WSIs have some cell-sparse and all-white regions, we abandon such regions and divide each WSI into uniform WSI sequences with 24 patches, where each patch has the 224×224 resolution. In the training, we use a total of 9 labels, including 8 typing labels and 1 other label. In the test, we find that the proposed MHD-Net can well identify the tumor and non-tumor CT sequences as it achieves a classification accuracy of over 99% on other-label data. Accordingly, we only display the test performance on 8 typing labels for convenience. Specifically, a total of 850 CT (645 tumor CT and 205 non-tumor CT) and 895 WSI sequences are obtained, where the data from 100 patients (516 tumor CT, 205 non-tumor CT, and 716 WSI sequences) and 26 patients (129 tumor CT sequences) are defined as the training and test sets, respectively. The corresponding clinical attributes are counted in Table I, and we publish the RPTET dataset and the source code at https://github.com/wangpengyu0829.

TABLE I
Clinical Attribute Statistics of the RPTET Dataset

Clinical Attributes		Number of Patients
Tumor Size (cm)	6.59 ± 2.15	–
Age Range (years)	17-81 (Avg. 48)	–
Sex	Male	66 (52.38%)
	Female	60 (47.62%)
Clinical Symptom	Myasthenia Gravis	31 (24.60%)
	Mediastinal Mass	65 (51.59%)
	Space-occupying Lesions	14 (11.11%)
	Chest Mass	16 (12.70%)
WHO Typing	A	13 (10.32%)
	AB	28 (22.22%)
	B1	15 (11.90%)
	B1+B2	18 (14.29%)
	B2	19 (15.08%)
	B2+B3	10 (7.94%)
	B3	17 (13.49%)
	TC	6 (4.76%)

2) The CPTAC-LUAD Dataset: Considering that existing public datasets contain un-processed radiology-pathology data, we also collect and compile the data from the Clinical Proteomic Tumor Analysis Consortium Lung Adenocarcinoma (CPTAC-LUAD) dataset [46] to further evaluate the proposed MHD-Net. It contains lung radiology-pathology data from 50 patients, which can be classified into grades G1, G2 and G3. Through pre-processing, 517 CT and 671 WSI sequences with 24 frames are obtained. Among them, we randomly choose 80% data as the training set (413 CT and 536 WSIs sequences), and 20% data as the test set (104 CT sequences).

TABLE II
DEFINITIONS OF MACC, PRE, REC, F1 AND QWK

Metrics	Definitions
Precision	$\frac{1}{K}\sum_{k=1}^{K}(\frac{\mathrm{TP}_k}{\mathrm{TP}_k+\mathrm{FP}_k})$
Recall	$\frac{1}{K}\sum_{k=1}^{K}(\frac{\mathrm{TP}_k}{\mathrm{TP}_k+\mathrm{FN}_k})$
F-measure	$\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$
mAcc	$\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$
QWK	$1-(\sum_{i=1}^{K}\sum_{j=1}^{K}D_{ij}O_{ij})/(\sum_{i=1}^{K}\sum_{j=1}^{K}E_{ij}O_{ij})$

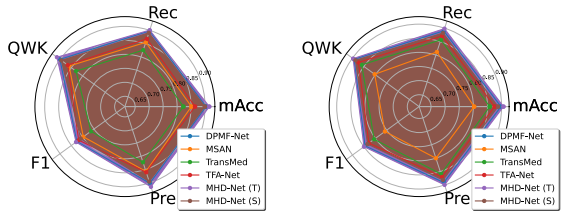


Fig. 4. Comparisons with multi-modal methods on the RPTET (left) and CPTAC-LUAD (right) datasets, where T and S denote the teacher and student networks respectively.

B. Implementation Details

Our framework is implemented by PyTorch 1.13.0, and all experiments are conducted on a server with Intel (R) Core (TM) i9-10850 K CPU and NVIDIA GeForce RTX 3090 GPU. Specifically, MHD-Net is optimized by a Adam optimizer with momentums $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the multi-modal teacher network, there are 100 epochs in network training with batch size 8 and the initial learning rate 0.001. Moreover, the memory-aware student network is trained in 100 epochs with batch size 8 and the initial learning rate 0.001.

C. Comparisons With Methods on the Missing Modality

We first compare the proposed MHD-Net with state-of-the-art missing modality methods, including knowledge distillation based SCKD [47], DAGD-Net [24] and DGMKD-Net [23], and modality dropout based AW3M [30], ModDrop++ [29] and mmFormer [27]. All comparison methods learn multi-modal radiology-pathology knowledge, and perform the prediction using uni-modal radiology data. To quantify their performance, we introduce five well-established evaluation metrics, namely the mean accuracy (mAcc), recall (Rec), precision (Pre), F-measure (F1), and quadratic weighted kappa (QWK) [48]. In this work, mAcc and QWK are suitable for evaluating the multi-classification task. As shown in Table II, mAcc is defined as the number of correct predictions divided by the total number of predictions, and QWK quantifies the agreement between the predicted results and ground truths. In the QWK's formula, K is the total number of classes, O is the confusion matrix and O_{ij} denotes the number of samples with ground truth i but predicted as j; D is the weight matrix and $D_{ij} = (i-j)^2/(K-1)^2$; E is the expected confusion matrix and E_{ij} is obtained rely on the histograms of the predicted label and ground truth distributions. By comparison, Pre, Rec, and F1 cannot directly measure the multi-classification performance. Therefore, we first compute the TP_k (true positive), FN_k (false negative) and FP_k (false positive) for each class to obtain the class-level $\{\mathrm{Pre}_k\}_{k=1}^{K}$, $\{\mathrm{Rec}_k\}_{k=1}^{K}$, $\{\mathrm{F1}_k\}_{k=1}^{K}$, and then adopt a mean operation to produce the multi-classification $\mathrm{Pre} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{Pre}_k$, $\mathrm{Rec} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{Rec}_k$ and $\mathrm{F1} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{F1}_k$. The quantitative results on the RPTET dataset are displayed in Table III, which shows that the proposed MHD-Net surpasses existing methods since it obtains the best performance of 89.89% mAcc, 89.5% Rec, 89.43% Pre, 89.46% F1, and 80.83% QWK, respectively. For knowledge distillation methods, SCKD [47] can customize an adaptive distillation approach for the student, which quantifies the teacher-student mismatch via the gradient similarity, and then selectively distills the privileged knowledge. DGMKD-Net [24] proposes a discrepancy-guided contrastive learning scheme, and constructs a gradient-guided knowledge refinement module for adaptively distilling knowledge from the multi-modal teacher to the uni-modal student. DAGD-Net [23] proposes an interactive gated-based grouping module to disentangle modal-share and -specific features for mitigating hetero-modal gaps, and then performs a multi-stage adaptive strategy for progressively distilling multi-modal knowledge. It can be seen that DGMKD-Net gains the performance improvement of 1.15% mAcc over DAGD-Net, while SCKD falls behind DGMKD-Net by 1.58% mAcc. By comparison, modality dropout methods exhibit relatively poor performance. The main reason is that dropping the pathology modality leads to only one modality being accessible for these methods. AW3M [30] adopts a self-supervised consistency loss to learn modal-specific and -invariant features during training, and designs a light-weight feature recovery block for handling missing modality situations. ModDrop++ [29] follows the modality dropout training strategy, it proposes a dynamic head that can respond to different missing modality situations, and develops a co-training strategy to align fully- and missing modality representations. mmFormer [27] leverages intra- and inter-modality Transformers to construct and align modal-invariant semantic features, and further introduces an auxiliary regularizer for improving generalization on missing modality situations. On the RPTET dataset, mmFormer achieves more competitive performance than AW3M and ModDrop++ in terms of 7.52% and 6.18% mAcc improvements. In contrast, the proposed MHD-Net stores pathology information in CTMM during teacher training, which ensures that the student can selectively absorb valuable missing modal-specific information from memories to improve cross-modal distillation. Finally, quantitative results on the CPTAC-LUAD dataset further confirm the effectiveness of MHD-Net, which obtains the excellent performance of 88.73% mAcc, 88.11% Rec, 88.69% Pre, 88.40% F1, and 83.37% QWK, respectively.

D. Comparisons With Multi- and Uni-Modal Methods

This subsection first compares the proposed MHD-Net with current multi-modal methods, including MSAN [6], TFA-Net [48], TransMed [49], and DPMF-Net [50]. The quantitative results are displayed in Fig. 4. MSAN [6] combines two attentions for multi-modal learning, where multi-scale attention

TABLE III
QUANTITATIVE RESULTS (MACC, PRE, REC, F1 AND QWK) OF THE PROPOSED MHD-NET AND STATE-OF-THE-ART METHODS ON MISSING PATHOLOGY MODALITY ON THE RPTET AND CPTAC-LUAD DATASETS

Methods	RPTET dataset					CPTAC-LUAD dataset				
	mAcc	F1	Pre	QWK	Rec	mAcc	F1	Pre	QWK	Rec
AW3M [30]	77.71	76.29	76.23	71.02	76.35	76.60	74.13	74.05	70.57	74.22
ModDrop++ [29]	79.05	79.70	79.64	68.13	79.76	80.79	81.12	81.34	70.46	80.90
mmFormer [27]	85.23	84.86	84.98	77.14	84.75	83.91	82.51	82.62	76.43	82.40
SCKD [47]	82.48	80.52	80.92	76.14	80.14	83.98	82.56	82.13	77.43	82.99
DAGD-Net [24]	82.91	82.56	82.99	74.80	82.11	82.24	82.69	83.28	75.43	82.10
DGMKD-Net [23]	84.06	83.60	83.42	76.72	83.78	85.47	83.18	83.44	75.76	82.93
MHD-Net	**89.89**	**89.46**	**89.43**	**80.83**	**89.50**	**88.73**	**88.40**	**88.69**	**83.37**	**88.11**

The best results are represented as the boldface.

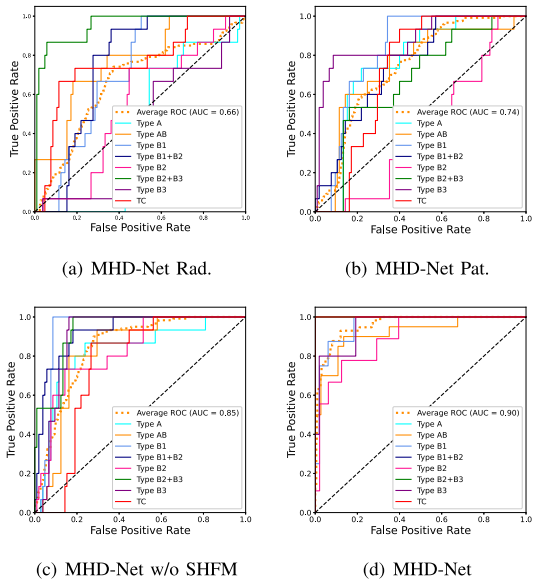
TABLE IV
QUANTITATIVE RESULTS (MACC, F1, QWK) OF THE PROPOSED MHD-NET AND UNI-MODAL METHODS ON THE RPTET AND CPTAC-LUAD DATASETS

Methods	RPTET dataset			CPTAC-LUAD dataset		
	mAcc	F1	QWK	mAcc	F1	QWK
SCM-Net [30]	73.37	72.02	66.02	72.43	71.07	65.54
PCCT [29]	76.82	74.19	68.31	74.85	73.20	65.02
CMMF-Net [27]	78.19	77.55	73.74	77.57	76.02	69.74
MHD-Net (student)	89.89	89.46	80.83	88.73	88.40	83.37
MHD-Net (teacher)	**91.50**	**91.42**	**82.92**	**90.67**	**89.83**	**84.97**

The best results are represented as the boldface.

is used to capture global-local features, and region-guided attention is presented to enhance foreground information. TFA-Net [48] builds a multi-scale backbone to extract modal-specific features, and then introduces a reverse cross-attention scheme to classify multi-modal data. By combining the advantages of CNN and Transformer to extract modal-invariant information, TransMed [49] achieves a simple but effective multi-modal classification framework. DPMF-Net [50] develops a switched attention module to enhance cross-modal semantic relatedness for multi-modal fusion, and achieves dual-polarization contrastive training for multi-modal classification. Among multi-modal methods, we observe that DPMF-Net obtains the optimal performance of 90.35% and 89.77% mAcc on the RPTET and CPTAC-LUAD datasets. Compared with them, the proposed MHD-Net (S: student) achieves a comparable performance of 89.89% and 88.73% mAcc since missing modal-specific information is well completed by memory features for multi-modal feature fusion.

We also compare the proposed MHD-Net with recent uni-modal methods, including SCM-Net [51], PCCT [52], and CMMF-Net [53]. For fair comparisons, we retrain these methods using only radiology data, and illustrate their performance in Table IV. Specifically, SCM-Net [51] is a light-weight CNN, which develops a spiking cortical model to capture the global and local lesion-related attention for medical image classification. PCCT [52] is implemented with the classical ResNet-50, it also designs a progressive class-center triplet loss to mitigate the impact of class-imbalanced data. CMMF-Net [53] is customized for CT data, this framework integrates the multi-scale feature fusion and distance-guided feature selection for effective classification. It can be seen that the proposed MHD-Net (student) performs favorably against the above methods in every metric. This demonstrates that distilling the multi-modal knowledge well-learned from the teacher to student can effectively improve the uni-modal typing performance. On the other hand, MHD-Net (teacher) gives more consideration to the heterogeneity and



(a) MHD-Net Rad. (b) MHD-Net Pat.

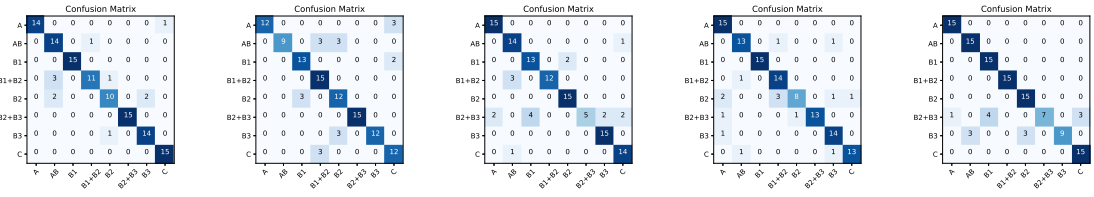
(c) MHD-Net w/o SHFM (d) MHD-Net

Fig. 5. Ablation study (ROC curves with AUC values) for the proposed SHFM on the RPTET dataset.

correlation between radiology and pathology data, and achieves an effective hetero-modal feature fusion by correlating spatial-specific tumor information across modalities via SHFM. As results, MHD-Net (teacher) achieves the performance of 91.50% and 90.67% on two datasets, which also improves the top-limit performance of the student.

E. Ablation Study

1) Effectiveness of Spatial-Differentiated Hetero-Modal Fusion Module: We first conduct the ablation study to evaluate the effectiveness of SHFM. We construct MHD-Net w/o SHFM by replacing the proposed SHFM with the most-common concatenation for hetero-modal feature fusion. To make more intuitive comparisons, we display the multi-modal teachers' performance using class-wise ROC curves in Fig. 5. Compared with the vanilla concatenation manner, SHFM considers the tumor information correlations between radiology and pathology features, and then models such correlations via spatial-specific 3D involutions for hetero-modal feature fusion. It can be observed that MHD-Net outperforms MHD-Net w/o SHFM with a clear superiority of 5% AUC on the RPTET dataset. In addition, we show the uni-modal classification performance by constructing models MHD-Net Rad. and MHD-Net Pat., where the former adopts a single branch to predict radiology data, and the latter uses a



(a) MHD-Net (b) MHD-Net w/o CTMM (c) w/o $\mathcal{L}_{MSE}$ (d) w/o $\mathcal{L}_{KL}$ (e) w/o $\mathcal{L}_{ConMB}$

Fig. 6. Ablation study (confusion matrices) for CTMM and loss functions on the RPTET dataset, where the horizontal axis represents ground truths, and the vertical axis represents prediction results.

single branch to predict pathology data. As reported in Fig. 5, MHD-Net Rad. and MHD-Net Pat. obtain the AUC of 66% and 74%, respectively. The above ablation study demonstrates that the hetero-modal feature fusion improves the teacher's performance, while the proposed SHFM further enables the teacher to learn reliable multi-modal knowledge from heterogeneous radiology and pathology features.

2) Effectiveness of Contrast-Boosted Typing Memory Module: Then, we conduct the ablation study to verify the effectiveness of CTMM. As shown in Fig. 6, we ablate the ConMB loss in CTMM as well as the whole CTMM to demonstrate their contributions, respectively. The confusion matrices in Fig. 6 illustrate that MHD-Net w/o $\mathcal{L}_{ConMB}$ reduces the mAcc by 5.00% since the quality of memory items can be improved by the ConMB loss. Moreover, ablating the whole CTMM leads to a serious mAcc decline of 7.50% due to missing pathology information in the student. Compared to radiology data, pathology data provides more accurate diagnosis information in clinical applications. The above ablation study also indicates that storing pathology features in memories is a straightforward but effective strategy to address the problem of missing modal-specific information in cross-modal distillation, resulting in significant performance improvements.

3) Effectiveness of Multi-Stage Memory-Aware Distillation: Finally, we conduct the ablation study to confirm that the KL and MSE losses can effectively align the teacher and student features for improving model performance, as shown in Fig. 6. It can be seen that when ablating the KL loss, MHD-Net suffers non-ideal performance. We consider that due to some discriminative information not being distilled from the teacher to student. When ablating the MSE loss, MHD-Net has imperfect results on mixed types B1+B2 and B2+B3, which prove that the MSE loss enables type-wise memories to be transferred as reliable memory features, and ensures the student can selectively absorb beneficial multi-modal knowledge from the teacher. Overall, this ablation study shows that both the KL and MSE losses are crucial for MMD in the memory-aware student network. Employing a combination of the KL and MSE losses can avoid confusion between difficult-to-distinguish types.

V. Discussion

A. Analysis of Trade-Off Parameters

In the proposed MHD-Net, we study the impacts of loss functions by analyzing the trade-off parameters. For convenience, we set the trade-off parameter of the primary CE loss as 1 in both the teacher and student, and observe the changes in model

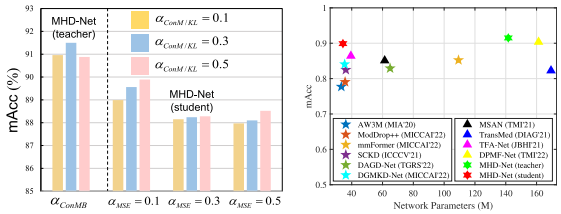


Fig. 7. Evaluations of trade-off parameters (left) and model complexity (right) on the RPTET dataset.

performance by varying other trade-off parameters within a range of 0-1 for the KL, MSE, and ConMB losses. As shown in Fig. 7 (left), in the teacher, when the trade-off parameter α_{ConMB} increases from 0.1 to 0.3 the model performance exhibits a positive growth trend in mAcc. However, with the continuous increments of α_{ConMB}, the model performance presents a sluggish downward trend. Moreover, on the RPTET dataset, the student achieves an optimal performance of 89.89% mAcc with $\alpha_{MSE} = 0.1$ for the MSE loss and $\alpha_{KL} = 0.5$ for the KL loss respectively, which further demonstrates that proper trade-off parameters can improve cross-modal distillation performance.

B. Evaluation for Model Complexity

To further display the relationship between model performance and complexity, we compare MHD-Net with state-of-the-art methods in terms of parameters. As can be seen from Fig. 7 (right), MHD-Net (S) built upon 3D ResNet-18 only has 34.30 M parameters, which are similar to other ResNet-18 based methods AW3M, SCKD, and DGMKD-Net, and more efficient than ResNet-34 and 50 based methods DPMF-Net and DAGD-Net. Since our teacher network consists of three 3D ResNet-34 branches, it has a moderate model complexity (141.65 M parameters) compared to the multi-modal comparison methods. In summary, MHD-Net not only shows the performance advantage in both the missing modality and multi-modal situations, but also maintains a good balance between model performance and complexity.

C. Analysis of Receptive Fields in SHFM

In SHFM, hetero-modal feature fusion is achieved by spatial-specific 3D involutions. Theoretically, different involution receptive fields have noticeable impacts on fusion results. To investigate a suitable receptive field, we conduct an experiment to analyze the impacts of various receptive fields on model performance. As shown in Table V, we employ 3 × 3, 5 ×

TABLE V
EVALUATIONS OF RECEPTIVE FIELDS (LEFT) IN SHFM AND MEMORY CAPACITY (RIGHT) IN CTMM ON THE RPTET DATASET

Receptive Field	mAcc	Param.	Memory Capacity	mAcc	Param.
3×3	88.27	3.01M	3136; 784; 196	88.40	6.32M
5×5	88.65	3.97M	784; 784; 196	89.89	2.71M
7×7	89.89	5.44M	784; 196; 196	86.14	1.81M
9×9	87.83	7.46M			

5, 7×7 and 9×9 involution kernels, respectively. It can be confirmed that despite the small-scale involution kernels (3×3 and 5×5) having lower parameters, they are insufficient to gain high-quality fusion results. In addition, the large-scale involution kernel (9×9) seems to fail to bring significant performance improvement, but increases parameters dramatically. In summary, the proposed MHD-Net achieves the best performance when using the 7×7 involution kernel, so it is more reasonable to set a larger receptive field in SHFM.

D. Analysis of Memory Capacity in CTMM

Next, we discuss the impact of memory capacity, i.e., the number of each stage memory items in CTMM on model performance. In general, excessive memory items tend to make a model remember specific features rather than generic information. In contrast, insufficient memory items may make it difficult for a model to remember enough effective information. As can be seen from Table V, we analyze three memory item settings in CTMM. In this work, since memory items are directly used to produce memory-aware pathology features, the number of each stage memory items should match the feature scale. An intuitive setting is that the number of each stage memory items is equal to the number of pixels, so we define the 1-st setting with 3136, 784 and 196 memory items, respectively. Considering that the nearest up-sampling can be applied to convert memory items into memory-aware pathology features to reduce the number of memory items. We define the 2-nd setting with 784, 784 and 196 memory items, where the 1-st stage memory items require up-sampling; and the 3-rd setting with 784, 196 and 196 memory items, where both the 1-st and 2-nd stage memory items require up-sampling. Table V shows that our MHD-Net achieves the best performance in the 2-nd setting, which confirms that using appropriate numbers of memory items helps to improve model performance and reduce model complexity.

VI. CONCLUSION

We propose a memory-aware hetero-modal distillation network, which can perform accurate TET typing using only radiology data by distilling the well-learned radiology-pathology knowledge. In the multi-modal teacher, the proposed SHFM improves the quality of hetero-modal fusion features by correlating radiology and pathology tumor information, which enables the teacher to learn more reliable multi-modal knowledge. Between the teacher and student, the proposed CTMM stores effective pathology features via type-wise memory updating and stage-wise contrastive memory boosting, which provide missing modal-specific information to improve cross-modal knowledge distillation. In the memory-aware student, the proposed MMD scheme well distills multi-modal knowledge from the teacher to student by introducing memory features, resulting in better uni-modal performance. We also construct the RPTET dataset, which contributes to studying radiology-pathology data typing. Experiments on the RPTET and CPTAC-LUAD datasets show that the proposed MHD-Net effectively addresses a new clinical problem, and outperforms state-of-the-art missing modality methods.

The proposed MHD-Net makes the first effort to the radiology data typing guided by multi-modal knowledge, but it still has several limitations: 1) MHD-Net needs to take the pre-processed CT data from TET patients as inputs. Although MHD-Net can provide accurate typing results after the doctor confirms a patient with TET, it fails to achieve the full-automatic TET diagnosis using raw CT data. In the future, we will collect more radiology-pathology data from non-TET patients to train MHD-Net, thereby giving it the ability to diagnose the presence of TET, and exploring how to use raw CT data for TET typing. 2) MHD-Net models the global spatial correlation between radiology and pathology tumor information for hetero-modal feature fusion. However, using spatial-specific 3D involutions to traverse all positions brings extra computation costs. Aiming at this issue, we will further explore the positional correlation between pathology slices of the tumor and tumor frames in CT to produce WSIs. This explicitly provides position priors for promoting hetero-modal feature fusion and reducing computational costs.

ETHICS STATEMENT

Ethics is approved by the Ethical Committee of China-Japan Friendship Hospital and Anhui University Of Science & Technology (2021001).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] M. Scorsetti et al., "Thymoma and thymic carcinomas," *Crit. Rev. Oncol. Hemat.*, vol. 99, no. 4, pp. 332–350, 2016.
[2] A. Marx et al., "The 2021 WHO classification of tumors of the thymus and mediastinum: What is new in thymic epithelial, germ cell, and mesenchymal tumors?," *J. Thoracic Oncol.*, vol. 17, no. 2, pp. 200–213, 2022.
[3] X. Han et al., "Relationship between computed tomography imaging features and clinical characteristics, masaoka–koga stages, and world health organization histological classifications of thymoma," *Front. Oncol.*, vol. 9, 2019, Art. no. 1041.
[4] W. Wang et al., "Learning two-stream CNN for multi-modal age-related macular degeneration categorization," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 4111–4122, Aug. 2022.
[5] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.
[6] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1591–1602, Jun. 2021.
[7] Q. Zuo, J. Zhang, and Y. Yang, "DMC-fusion: Deep multi-cascade fusion with classifier-based feature synthesis for medical multi-modal images," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3428–3449, Sep. 2021.

[8] Y. Lu et al., "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13376–13386.

[9] Z. Xue et al., "Multi-modal co-learning for liver lesion segmentation on PET-CT images," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3531–3542, Dec. 2021.

[10] W. Shao et al., "Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers," *Med. Image Anal.*, vol. 65, 2020, Art. no. 101795.

[11] Z. Ning, D. Du, C. Tu, Q. Feng, and Y. Zhang, "Relation-aware shared representation learning for cancer prognosis analysis with auxiliary clinical variables and incomplete multi-modality data," *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 186–198, Jan. 2022.

[12] Q. Yang, X. Guo, Z. Chen, and Y. Yuan, "D^2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2953–2964, Oct. 2022.

[13] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Latent correlation representation learning for brain tumor segmentation with missing MRI modalities," *IEEE Trans. Med. Imag.*, vol. 30, pp. 4263–4274, 2021.

[14] J. Ouyang, E. Adeli, K. M. Pohl, Q. Zhao, and G. Zaharchuk, "Representation disentanglement for multi-modal brain MRI analysis," in *Proc. 27th Int. Conf. Inf. Process. Med. Imag.*, 2021, pp. 321–333.

[15] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-net: Hybrid-fusion network for multi-modal MR image synthesis," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2772–2781, Sep. 2020.

[16] Y. Li, K. K. Singh, U. Ojha, and Y. J. Lee, "MixNMatch: Multifactor disentanglement and encoding for conditional image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8036–8045.

[17] Y. Pan, M. Liu, Y. Xia, and D. Shen, "Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6839–6853, Oct. 2022.

[18] Z. Yu et al., "MouseGAN++: Unsupervised disentanglement and contrastive representation for multiple MRI modalities synthesis and structural segmentation of mouse brain," *IEEE Trans. Med. Imag.*, vol. 42, no. 4, pp. 1197–1209, Apr. 2023.

[19] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 1789–1819, 2021.

[20] M. Hu et al., "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2020, pp. 772–781.

[21] G. Zhang et al., "Cross-modal prostate cancer segmentation via self-attention distillation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5298–5309, Nov. 2022.

[22] S. Wei, C. Luo, and Y. Luo, "MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20039–20049.

[23] X. Xing, Z. Chen, M. Zhu, Y. Hou, Z. Gao, and Y. Yuan, "Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 636–646.

[24] X. Li, L. Lei, C. Zhang, and G. Kuang, "Dense adaptive grouping distillation network for multimodal land cover classification with privileged modality," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4411114.

[25] X. Xing, Z. Chen, Y. Hou, and Y. Yuan, "Gradient modulated contrastive distillation of low-rank multi-modal knowledge for disease diagnosis," *Med. Image Anal.*, vol. 88, 2023, Art. no. 102874.

[26] Z. Yuan et al., "X-Trans2Cap: Cross-modal knowledge transfer using transformer for 3D dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8553–8563.

[27] Y. Zhang et al., "mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 107–117.

[28] Z. Liu, J. Wei, R. Li, and J. Zhou, "SFusion: Self-attention based N-to-One multimodal fusion block," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2023, pp. 159–169.

[29] H. Liu et al., "ModDrop: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 444–453.

[30] R. Huang et al., "AW3M: An auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound," *Med. Image Anal.*, vol. 72, 2021, Art. no. 102137.

[31] R. Dorent et al., "Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101862.

[32] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology report generation with a learned knowledge base and multi-modal alignment," *Med. Image Anal.*, vol. 86, 2023, Art. no. 102798.

[33] Y. Liu et al., "Multi-modal learning for predicting the genotype of glioma," *IEEE Trans. Med. Imag.*, vol. 42, no. 11, pp. 3167–3178, Nov. 2023.

[34] Y. Liu, L. Fan, C. Zhang, T. Zhou, Z. Xiao, and D. Shen, "Incomplete multi-modal representation learning for Alzheimer's disease diagnosis," *Med. Image Anal.*, vol. 69, 2021, Art. no. 101953.

[35] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8199–8208.

[36] D. Liu, X. Qu, X. Di, Y. Cheng, Z. Xu, and P. Zhou, "Memory-guided semantic learning network for temporal sentence grounding," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1665–1673.

[37] S. Lee, H. G. Kim, D. H. Choi, H.-I. Kim, and Y. M. Ro, "Video prediction recalling long-term motion context via memory alignment learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3053–3062.

[38] J. U. Kim, S. Park, and Y. M. Ro, "Robust small-scale pedestrian detection with cued recall via memory learning," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 3030–3039.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[40] Y. Zhou, G. Wu, Y. Fu, K. Li, and Y. Liu, "Cross-MPI: Cross-scale stereo for image super-resolution using multiplane images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14837–14846.

[41] D. Li et al., "Involution: Inverting the inherence of convolution for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12316–12325.

[42] R. Gu et al., "Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 245–256, Jan. 2023.

[43] Z. Zhu et al., "MuRCL: Multi-instance reinforcement contrastive learning for whole slide image classification," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1337–1348, May 2023.

[44] Y. Tan, K.-F. Yang, S.-X. Zhao, and Y.-J. Li, "Retinal vessel segmentation with skeletal prior and contrastive loss," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2238–2251, Sep. 2022.

[45] Y. Zhou, T. Zhou, T. Zhou, H. Fu, J. Liu, and L. Shao, "Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning," *IEEE Trans. Med. Imag.*, vol. 40, no. 4, pp. 1196–1206, Apr. 2021.

[46] M. A. Gillette et al., "Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma," *Cell*, vol. 182, no. 1, pp. 200–225, 2020.

[47] Y. Zhu and Y. Wang, "Student customized knowledge distillation: Bridging the gap between student and teacher," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 5037–5046.

[48] C.-H. Hua et al., "Convolutional network with twofold feature augmentation for diabetic retinopathy recognition from multi-modal images," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2686–2697, Jul. 2021.

[49] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, 2021, Art. no. 1384.

[50] Y. Chen et al., "Dual polarization modality fusion network for assisting pathological diagnosis," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 304–316, Jan. 2023.

[51] Q. Zhou, Z. Huang, M. Ding, and X. Zhang, "Medical image classification using light-weight CNN with spiking cortical model based attention module," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 1991–2002, Apr. 2023.

[52] K. Chen, W. Lei, S. Zhao, W. Zheng, and X. Zhang, "PCCT: Progressive class-center triplet loss for imbalanced medical image classification," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 4, pp. 2026–2036, Apr. 2023.

[53] S. Lu, J. Liu, X. Wang, and Y. Zhou, "Collaborative multi-metadata fusion to improve the classification of lumbar disc herniation," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3590–3601, Dec. 2023.