Modality-Aware Distillation Network for Microvascular Invasion Prediction of Hepatocellular Carcinoma From MRI Images

Yinghao Zhang, Hong Liu, Lei Zhu, *Member, IEEE*, Huanhuan Chong, Huazhu Fu, *Senior Member, IEEE*, Lequan Yu, *Member, IEEE*, Ping Li, *Member, IEEE*, Jing Qin, *Senior Member, IEEE*, David Dagan Feng, *Life Fellow, IEEE*, and Liansheng Wang, *Member, IEEE*

***Abstract*—Microvascular invasion (MVI) of hepatocellular carcinoma (HCC) is a crucial histopathologic prognostic factor associated with cancer recurrence after liver transplantation or hepatectomy. Recently, clinicoradiologic characteristics are combined with medical images to enhance the HCC prediction. However, compared to medical imaging data, the clinicoradiologic characteristics (e.g., APOe4 genotyping) is not easy to collect or even unavailable, as it requires more efforts of clinicians and more medical instruments for collecting diverse measurements. This work explores how to transfer the knowledge of a teacher network learned from non-image clinical data and image data to a student network with only image data such that the student network can leverage the transferred clinical information to boost HCC classification with only imaging data as input. Specifically, we present a modality-aware distillation network (MD-Net) to transform non-image clinicoradiologic from the teacher network to the student network. The teacher network integrates non-image clinicoradiologic characteristics with two 3D MRI modality images via two MRI-clinical-fusion modules and a symmetric attention (SA) module, while the student network extracts features from two modality MRI data via two MRI-only modules and then refine these two MRI features via a SA module. A classification-level distillation and a feature-level distillation are jointly utilized to transfer the clinical information between teacher and student networks. Furthermore, we design a novel self-supervised task to predict clinicoradiologic characteristics from the imaging data to further enhance the downstream HCC classification. The experimental results from our collected dataset and a multi-modal sarcasm detection dataset have demonstrated the effectiveness of our approach. Specifically, we achieved an AUC score of 71.86% and 75.51% respectively, surpassing the performance of the state-of-the-art classification methods.**

***Index Terms*—Hepatocellular carcinoma (HCC), Microvascular invasion (MVI), Multi-modality, Knowledge distillation.**

Received 10 July 2024; revised 6 December 2024; accepted 18 December 2024. Date of publication 30 December 2024; date of current version 12 May 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62371409, in part by the Fujian Provincial Natural Science Foundation of China under Grant 2023J01005, in part by the Fundamental Research Funds for the Central Universities under Grant NJ2024029, and in part by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007. *(Yinghao Zhang and Hong Liu contributed equally to this work.) (Corresponding authors: Lei Zhu; Liansheng Wang.)*

Yinghao Zhang and Hong Liu are with the Department of Computer Science, School of Informatics, Xiamen University, China.

Lei Zhu is with the Thrust of Robotics and Autonomous Systems (ROAS), The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China, and also with the Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: leizhu@hkust-gz.edu.cn).

Huanhuan Chong is with the Department of Radiology, Zhongshan Hospital, Fudan University, China.

Huazhu Fu is with the Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore.

Lequan Yu is with the Department of Statistics and Actuarial Science, The University of Hong Kong, China.

Ping Li is with the Department of Computing, School of Design, and Research Institute for Sports Science and Technology, The Hong Kong Polytechnic University, Hong Kong.

Jing Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong.

David Dagan Feng is with the Biomedical and Multimedia Information Technology Research Group, School of Computer Science, The University of Sydney, Australia.

Liansheng Wang is with the Department of Computer Science, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: lswang@xmu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2024.3523921, provided by the authors.

Digital Object Identifier 10.1109/TBME.2024.3523921

I. Introduction

HCARCINOMA (HCC) is the fifth most common cancer in the world and the third leading cause of cancer-related death [1]. The 5-year overall survival rate of HCC patients after surgery is only 10-20% [2], [3], [4]. After hepatectomy and liver transplantation, The 5-year recurrence rate can be as high as 50-70% and 35%, respectively [5], [6], [7], [8].

Many literature reports that vascular invasion is one of the important factors that threaten the prognosis of patients [9], [10], [11], [12], which limits the implementation of curable treatment strategies for liver resection, liver transplantation, and radiofrequency ablation [13], [14], [15]. According to its

TABLE I
52 CLINICAL ITEMS (PARAMETERS) WE EXPLORED IN OUR WORK

Groups of Clinical Parameters	Name
pathological parameters	ES, cirrhosis, Ki-67
clinical parameters	Gender, Age, BCLC, Child-pugh
clinical/radiological parameters	size, number, multifocality
clinical/hematological index	HBVDNA, HBV_OR_HCV, HCV, HBV, AFP, CEA, CA199, PLT, TBIL, DBIL, TP, ALB, ALT, AST, AKP, GGT, TBA, LDH, APLB, PT
radiological parameters	Nodule-in-nodule, IP/OP, Necrosis or cystic component, T2WI, rim APHE, T1WI, Mosaic architecture, Targetoid, HBP hypo, Margin, Peritumoral enhancement, Capsule enhancement, Peritumoral hypointensity, hemorrhage, ascites, APHE, washout, Lirads, AP enhancement intensity, non-enhancement capsule, Peritumoral enhanced shape, Enhanced capsule

detection methods, vascular invasion can usually be classified into Macrovascular invasion (MaVI) and Microvascular invasion (MVI) [11], [12], [16]. MaVI refers to the macrovascular tumor thrombi visible to the naked eye (for example, the tumor thrombus in the main portal vein). It is a key risk factor that affects the survival of HCC patients after hepatectomy or liver transplantation. MVI refers to the presence of nests of cancer cells in the vascular cavity lined by endothelial cells under the microscope [2], [17], [18], [19], which is present in 15-57.1% of postoperative liver cancer specimens [18]. Similar to MaVI, MVI is also a risk factor for poor outcomes after liver resection or liver transplantation in patients with liver cancer [12], [20], [21], [22]. However, unlike MaVI, MVI is only visible under the postoperative pathology microscope [17], [18] and requires extensive sampling [23]. Its relatively lagging gold standard for pathology severely limits the timely and effective adjustment of surgical treatment strategies. Therefore, the accurate stratification of MVI grades before surgery can be used as an important evaluation reference index for the formulation of treatment plans for patients with liver cancer and the follow-up monitoring after surgery. According to the number and distribution of microvessels involved, MVI can be further divided into M0 (no MVI), M1 (MVI $<=$ 5 and within 1 cm of the tumor edge) and M2 (MVI $>$ 5 or $>$ 1 cm from the tumor surface) [17].

Pathologically, the peritumoral tissue is the first infiltration area to be invaded by MVI, therefore, compared with the tumor itself, the macroscopic image features of the adjacent liver tissue (for example, the peritumoral enhancement on the arterial phase image, the peritumoral low signal on the hepatobiliary-specific phase image) [23] and the microscopic features (for example, high-dimensional images) Image peritumoral heterogeneity) [24] may be directly related to MVI. This argument has been confirmed by different imaging omics comparison models of tumor areas constructed in previous studies [24].

Recently, magnetic resonance imaging (MRI) has played an import part in the study of MVI prediction [2], [20], [24], [25], [26], [27]. However, there are still some small HCC patients in the clinic, and the dynamic enhancement of Gadolinium-enhanced MRI (GD-DTPA, a widely used MRI contrast agent that enhances image quality by leveraging the paramagnetic properties of gadolinium) is not typical due to the small lesions or small hepatic arterial blood supply, so the detection and characterization of the lesions are difficult [28]. Gadoxetate disodium-enhanced (Gd-EOB-DTPA) MRI offers excellent identifiability of small or early HCC and the information of tumor heterogeneity and vascularization [29], as shown in Fig. 1, so it is reasonable to extract features from multi-parametric images of GD-EOB-DTPA MRI to predict MVI of HCC.

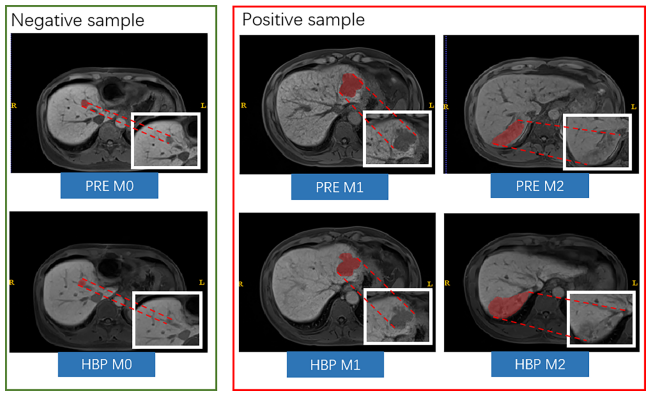

Fig. 1. Examples of tumors in MR images, with the tumor area highlighted in red and further details visible in the magnified section of the figure. In most instances, the tumors are relatively small, but there are cases where the tumor volume is large, indicating a higher risk of MVI. The term 'HBP' refers to the 3D Hepatobiliary phase MRI image, while '3D PRE' denotes the pre-contrast MRI image.

As verified by previous studies [26], [30], multi-sequence multi-parameter MRI can provide complementary information. Therefore, integrating the CNN features of different MRI sequences is helpful for the prediction of MVI. This is also proved in our experiments, in which we find that the prediction effect of multi-sequence data fusion is better than that of single-sequence data. Further, the information provided by the clinical radiological characteristics (As shown in Table I) of patients is more relevant to the prediction of MVI. Experiments have also proved that the introduction of clinical radiological characteristics have greatly improved the prediction effect of MVI. However, clinical radiology characteristics are not always available in practice, and there are many experimental indicators that are not fully obtained. So we thought of transferring the knowledge of the teacher network that introduced clinical radiology characteristics to the student network that only had image data.

Early studies [19], [31], [32] utilized many radiologic features to predict MVI of HCC. Due to the sensitivity of imaging radiologic features to acquisition methods and reconstruction parameters, the imaging radiologic features are very unreliable to be widely used in clinical practice. Instead of relying on these hand-crafted radiologic features, several convolutional neural networks (CNNs) have been developed to take input contrast-enhanced MRI or ultrasound imaging data and then learn discriminative features from these input imaging data for MVI prediction of HCC. More recently, non-image clinicoradiologic

characteristics are introduced to improve the network prediction with only the imaging data to leverage the complementary information between them, such as a pathological complete response(PCR) prediction of breast cancer patients [33]. However, compared to imaging data, clinical information is difficult to be collected due to several reasons: First, more medical instruments and clinicians are required to compute clinical data. such as molecular and demographical data. Second, some patients only have the medical data in some cases, where normal patients are not required to obtain.

In this paper, we present a modality-aware distillation network (MD-Net) for predicting MVI of HCC to distill a teacher network with a combination of imaging and clinical data to a student network with only imaging data. By doing so, the inference stage does not require any clinical data and the network performance with only imaging data in the inference stage can be further improved due to the clinical information transferred from the teacher network in our method. Further, a regression task to predict clinical radiological characteristics from the image data is proposed to transfer the clinical knowledge to the student model, to the best of our knowledge, this is the first study that attempts to to extract additional supervision from the image information by leveraging the complementary information between the imaging modality and the clinical modality. The task specifically designed for imaging-clinical datasets and has a strong pertinence. Here, multiple MRI data and 52 clinical items are utilized in our work. Our main contributions are summarized below:

- We propose a novel modality-aware distillation network (MD-Net) for MVI prediction of HCC. It introduces a new way of transferring knowledge from a teacher network, which uses both image modality and non-image clinical data, to a student network that only uses image modality.
- The architecture of our MD-Net, including the use of two MRI-only modules and a symmetric attention (SA) module in the student network, and two MRI-clinical-fusion modules in the teacher network, is a unique design that refines and fuses features in a novel way. This design contributes to the methodology by offering a new approach to feature extraction and refinement.
- Apart from the original classification-level result distillation, our MD-Net also devise a feature-level distillation to better transfer the clinical data from the teacher network to the student network. Moreover, we devise a regression task to predict clinical data from the image data for further enhancing the MVI prediction.
- We collect a dataset with annotations for testing different classification methods, and experimental results on the collected dataset and a multi-modal sarcasm detection dataset have verified the effectiveness of the developed modality-aware distillation network.

II. Related Work

A. Microvascular Invasion of Hepatocellular Carcinoma

Early MVI prediction works mainly examined radiologic features at the local lesion area of an MR volume [2], [20], [24], [25], and these features included non-smooth tumor margin, peritumoral enhancement on arterial phase (AP), peritumoral hypointensity on hepatobiliary phase (HBP), and so on [24]. However, these hand-crafted features are sensitive to the acquisition methods and reconstruction parameters, thereby suffering from limited capability in handle diverse clinical usage. Motivated by the superior performance of deep features over hand-crafted features in diverse medical image analysis tasks, convolutional neural network (CNNs) have been developed to classify MVI of HCC patients. Jiang et al. [34] utilized eXtreme Gradient Boosting (XGBoost) and deep learning from CT images to predict MVI preoperatively. Zhang et al. [30] developed a 3D CNN prediction model to fuse features from multiple MR sequences. Men et al. [26] embedded long short-term memory (LSTM) into a CNN to fuse multi-modal MR volumes for predicting MVI of HCC patients. Xiao et al. [35] proposed a task relevance driven adversarial learning framework (TrdAL) for simultaneous HCC detection, size grading, and multi-index quantification using multi-modality MRI. However, only MR images are involved to predict MVI status in those CNN-based methods. To boost the MVI prediction accuracy, our work leverages both imaging modality and clinical modality within a knowledge distillation learning framework.

B. Knowledge Distillation

Knowledge distillation techniques [36] often transferred the knowledge from a teacher network (e.g., large complex models) to a student network (e.g., small simple models). In medical imaging, the potential of the knowledge distillation technique [37], [38], [39] is promising yet relatively underexplored as far as we know. Wang et al. [40] employed KD for efficient neuronal structure segmentation from 3D optical microscope images with a teacher-student network. Kats et al. [41] borrowed the concept of KD to perform brain lesion segmentation with soft labels by dilating mask boundaries. Christodoulidis et al. [42] utilized KD for multi-source transfer learning on the task of lung pattern analysis. Li et al. [43] presented a Mutual Knowledge Distillation (MKD) scheme to thoroughly exploit the modality-shared knowledge to facilitate the target-modality segmentation. Xing et al. [44] formulated a Class-guided Contrastive Distillation module to pull closer positive image pairs from the same class in the teacher and student models, while pushing apart negative image pairs from different classes. Ju et al. [45] leveraged relevant retinal disease labels in both semantic and feature space as additional signals and trained the model in a collaborative manner using knowledge distillation. Javed et al. [46] proposed a knowledge distillation algorithm to improve the performance of shallow networks for tissue phenotyping in histology images.

Although achieving superior performance, these distillation networks almost considered diverse imaging data and transferred information from the input imaging data. Unlike this, the teacher network in our method takes multiple imaging data and non-image clinical data, while the student network only utilizes imaging data. Hence, our modality-aware distillation network transfers the clinical information from the teacher network and the student network to enhance the classification accuracy of the student network with only imaging data.

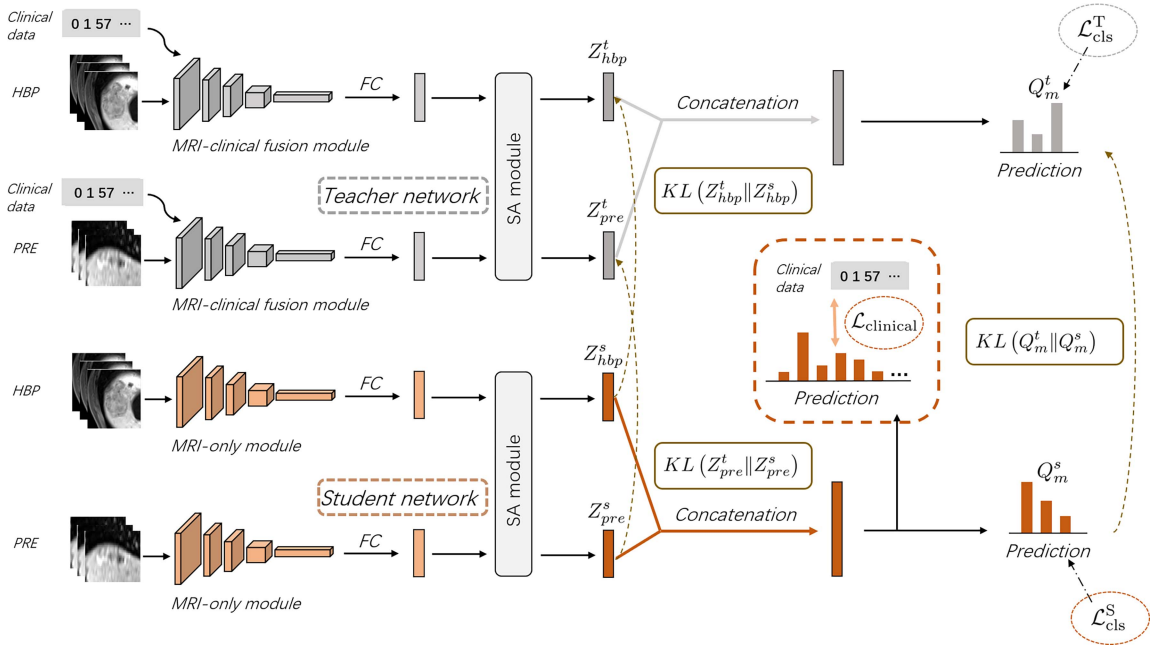

Fig. 2. Illustration of our modality-aware distillation network, which transfers knowledge from a teacher network with clinical data to a student network without clinical data. The student network is trained with a supervised learning loss, along with both classification-level and feature-level distillation losses, to align the student network's predictions and feature distribution with those of the teacher network.

C. Multi-Modal Learning

Early multi-modal methods focused on a early-fusion strategy to utilize the multi-modal information by concatenating the input multi-modal images, but these methods are not effective to integrate non-linear relationships among input modalities. For example, OM-Net [47] concatenated multi-modal images along the channel dimension as the input. Later, many recent works adopt a feature-level fusion mechanism to fuse early, middle, or later modality-specific features extracted from different encoders. Xia et al. [48] added clinical features and features from cardiac magnetic resonance (CMR) image for mortality risk prediction in dilated Cardiomyopathy. Duanmu et al. [33] integrated CNN features learned from the input 3D MRI imaging data, molecular data and demographic data. Tadesse et al. [49] presented a fusion of multi-modal physiological data to predict the severity of ANSD with a hierarchy of resource-aware decision making. Fang et al. [50] proposed a multi-modal brain tumor segmentation framework that adopts the hybrid fusion of modality-specific features using a self-supervised learning strategy. Cai et al. [51] devised a graph transformer geometric learning framework to learn the multimodal brain network constructed by structural MRI (sMRI) and diffusion tensor imaging (DTI) for estimating brain age. Xing et al. [52] leveraged transformer modules to learn the intra-modality and inter-modality relationships of multi-modal MRIs for brain tumor segmentation. Lin et al. [53] proposed a multi-modal sensing framework for activity monitoring, it can automatically identify human activities based on multi-modal data, and provide help to patients with moderate disabilities. Giri et al. [54] presented a multi-modal approach for predicting protein functions by utilizing two different kinds of information, namely protein sequence and the protein secondary structure. Unlike these methods focusing on developing techniques to fuse multi-modal features, our work addressed the problem of transferring knowledge of a teacher network learned from non-image clinical data and image data to a student network with only image data. By doing so, the student network can leverage the transferred clinical features to boost HCC classification with only imaging data as the input, since non-image clinical data is not easy to collect or even unavailable when compared to medical image data.

III. Method

Fig. 2 shows the schematic illustration of the proposed Modality-aware Distillation Network (MD-Net) for MVI prediction of HCC. As a distillation network, our MD-Net consists of a teacher network and a student network, but it distills a teacher network with diverse clinical information to a student network without any clinical data. The student network takes two MRI sequences as the input, pass each MRI image into a MRI-only module to extract MRI features, and then develops a symmetric attention (SA) module to refine two MRI features for final MVI prediction. On the other hand, the teacher network presents two MRI-clinical-fusion module to first extract integrated features of the MRI image and the clinical data, and then refine these two obtained features via another SA module for generating a MVI classification result. After that, we devise a distillation scheme by considering both class-level distillation and feature-level distillation. The class-level distillation makes the two predictions of the student network and the teacher network to be similar, while the feature-level distillation transfers the clinical-guided

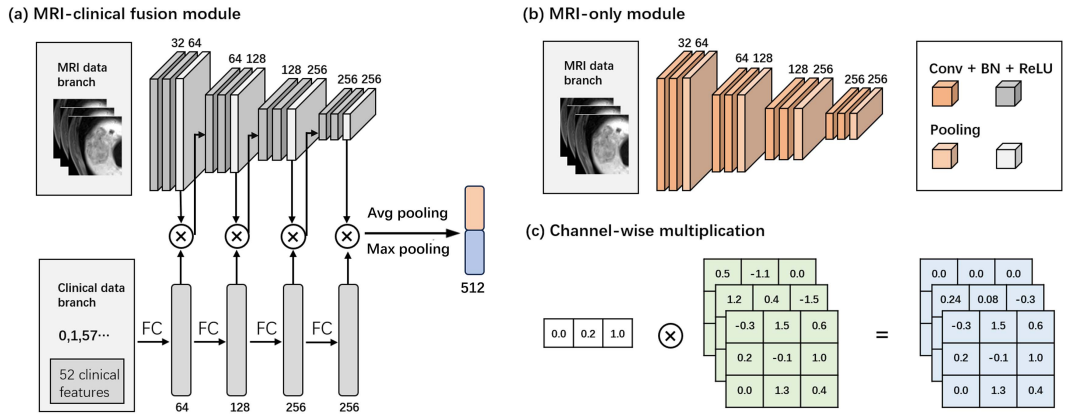

Fig. 3. (a) MRI-clinical-fusion module: fusing MRI data with non-imaging clinical data, (b) MRI-only module: using MRI imaging only, and (c) one example of channel-wise multiplication.

features of the teacher network to student's features, which do not consider any clinical data. Moreover, we devise a regression task to predict the clinical modality data from multiple image modalities and use this task to benefit the downstream the MVI prediction.

A. Teacher Network

Non-image Clinical data has shown its capability of providing complementary information for the classification task with only image data [33]. Motivated by this, we integrate clinical data, the Hepatobiliary phase (HBP) MRI image, and the pre-contrast (PRE) MRI image as the input of the teacher network of our MD-Net for the prediction of MVI in HCC and training the student network with only image data. Specifically, the teacher network first passes the HBP image and clinical data into a MRI-clinical-fusion CNN to extract a 512-dimensional vector Z^t_{hbp}, and the PRE image and clinical data are then feed into another MRI-clinical-fusion CNN for obtaining another 512-dimensional vector Z^t_{pre}. Then, we concatenate Z^t_{hbp} and Z^t_{pre} to produce Z^t, which is passed into two fully-connected layers to predict a classification result P^t with three elements of the teacher network.

MRI-clinical-fusion module: Similar to [33], our MRI-clinical-fusion module integrates MRI data and non-imaging clinical data for a HCC prediction. As shown in Fig. 3(a), taking a 3D MRI data and a vectorized clinical data as the inputs, the image-clinical fusion module first applies four fully-connected (FC) layers on the input clinical data to obtain four feature maps, which feature channels are 64, 128, 256, and 256. Meanwhile, we utilize four convolutional blocks on the input MRI image to obtain another 3D feature maps, and the feature channels are also set as 64, 128, 256, and 256. Each convolutional block consists of two 3×3 convolutional layers, with each of these layers is succeeded by a batch normalization layer and a ReLU activation layer, as indicated in [55]. At the end of the block, a max-pool layer is used to downsample the feature map by a factor of 2. And then we channel-wisely multiply four feature maps from the clinical data and the corresponding four features

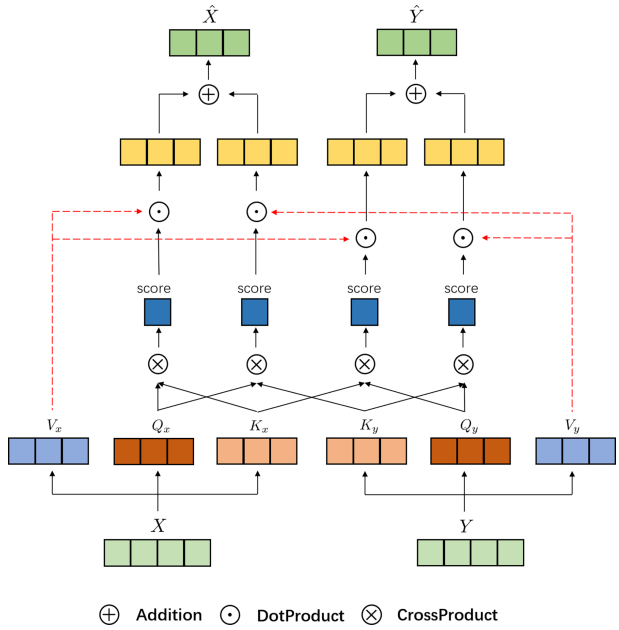

Fig. 4. The schematic illustration of our symmetric attention (SA) module.

from the MRI data for integrating them together. The specific calculation method is shown in Fig. 3(c). The output Y of the multiplication operation is a feature map with the same shape of the input MRI feature X. Specifically, let C denote the clinical features (a vector); X denote the MRI image features (3D); and Y denote the output 3D feature map. Note that the number of elements of the vector C is the same as the number of channels of 3D feature map X. Then, the channel-wise calculation (between C and X) done in Fig. 3(c) is summarized as follows: (a) for i-th element of C, we multiply it with the i-th channel of X and take the multiplication results as the i-th channel of Y:

$$Y_i(u, v) = C_i * X_i(u, v) \tag{1}$$

where C_i denote the i-th element of C. X_i and Y_i represent the i-th channel of the 3D feature map X and Y, respectively. (u, v)

denotes the pixel coordinates at the X_i and Y_i. (b) Then, we conduct such operation for all elements of the clinical feature vector C and thus can generate the multiplication result Y. Regarding i-th feature multiplication operation, we first obtain the multiplication result Y_i from C_i and X_i, and then pass Y_i to following CNN block to produce a new feature $X_{(i+1)}$ for the next $(i+1)$-th feature multiplication. When reaching the last feature multiplication operation of the MRI-clinical-fusion module, we apply average pooling and max pooling separately to the feature multiplication result to obtain two feature vectors, which are then concatenated to produce a feature vector with 512 elements, which is the final output feature of the MRI-clinical-fusion module, see Fig. 3(a).

B. Student Network

Although a fusion of the clinical data and the image data can improve the HCC classification result, the clinical data are not often available when compared to the MRI images for classifying HCC patients in clinical diagnosis. In order to improve the flexibility of our model in clinical application, we devise a modality-aware knowledge distillation network to transfer the knowledge learned by a teacher network with a fusion of a clinical data modality and the image modality to a student network with only the image modality. By doing so, the clinical data knowledge can be distilled from the teacher network to the student network, and thus the classification performance of the student work can be enhanced even though the student network does not get any clinical data in the testing stage.

As shown in Fig. 2, our student work takes a 3D HBP MRI image and a 3D PRE MRI image as the input, and then passes the HBP data into a MRI-only module to obtain features Z^s_{hbp} and the PRE data into another MRI-only module to obtain the feature map Z^s_{pre}. Note that Z^s_{hbp} and Z^s_{pre} are two vectors with 512 elements. After that, we concatenate Z^s_{hbp} and Z^s_{pre} to obtain Z^s and feed Z^s into a fully-connected layer to predict a HCC classification result P^s, which has three elements.

MRI-only module: In our student network, the MRI-only module extracts a 512-dimensional feature vector from an 3D MRI image. As shown in Fig. 3(b), the Image-only module's architecture is the same as the image feature extraction part in the teacher network, which consists of Four convolutional blocks, and one fully-connected (FC) layers.

Moreover, to balance the efficiency and computational burden, we set the channel number of output features of nine convolutional blocks to be different. The feature channels of the first five layers are 32,32,64,64 and 128, while the feature channels of the last four layers are empirically set as 256, 128, 256, and 256.

C. Symmetric Attention (SA) Module

Our SA module is to refine two features from different image modalities by leveraging their complementary information based on self-attention frameworks [56], [57], [58]. As shown in Fig. 4, specifically, let X and Y to denote the input two feature maps of the SA module. Then, the SA module first applies a linear transformation layers on X to obtain three feature maps, including query Q_x, key K_x, and value V_x. Meanwhile, we apply a linear transformation layers on Y to generate a key feature map K_y and a value feature map V_y. After that, we generate a score map S_x by multiplying Q_x and the transpose of K_x, and another score map S_y by multiplying Q_x and the transpose of K_y. Then, we multiple the obtained score maps S_x with the value feature map V_x, and multiply S_y with V_y to produce two resultant feature maps, which are then added together to generate the output refined feature map $\hat{X}$:

$$\hat{X} = V_x \times (Q_x \times {K_x}^T) + V_y \times (Q_x \times {K_y}^T). \tag{2}$$

Similarly, the SA module applies another transformation layer on Y to obtain a feature map Q_y. After that, the refined feature map $\hat{y}$ is computed by:

$$\hat{Y} = V_x \times (Q_y \times {K_x}^T) + V_y \times (Q_y \times {K_y}^T). \tag{3}$$

D. Regression Learning for Clinical Data Prediction

This work present a novel pretext task to train our modality-aware distillation network. Note that we have the clinical data, the HBP MRI image, the PRE MRI image, and the underlying HCC label for each patient of the training set. The pretext task in our MD-Net aims to predict the clinical data from the HBP and PRE image modalities for learning generic knowledge to benefit a downstream HCC classification. The clinical data prediction task takes the medical image data as the input to predict the clinical data, which is one of the inputs of the teacher network of our method in the training stage.

As shown in Fig. 2, the student network feeds the concatenated features from the input HBP image and the input PRE image into two fully-connected (FC) layers to predict a 52-dimensional vector P^c for estimating the underlying clinical information. And we utilize the input clinical data as the ground truth of the predicted P^c.

E. Modality-Aware Distillation

We apply the knowledge distillation strategy to transform the clinical information of the teacher network to the student network. Apart from the straightforward classification result-level distillation, we present an auxiliary feature-level distillation loss to distill features fused from clinical data and MRI image of the teacher network to features from only MRI image.

Classification-level distillation: Let $q^s_m(x_i)$ denote the class probabilities for the class of the MRI x_i data produced from the student network, while $q^t_m(x_i)$ represent the class probabilities for the class of the MRI x_i data produced from the teacher network network. Then, the classification-level distillation loss L^d_{class} is simply defined to push make the class probabilities from the teacher network as targets for training the student network. To do so, we utilize the Kullback Leibler (KL) divergence to measure the difference of two distribution:

$$\begin{aligned} L^d_{class} &= DKL\left(q^t_m(x_i) \| q^s_m(x_i)\right) \\ &= \sum_{i=1}^{N} \sum_{m=1}^{M} p^m_2(x_i) \log \frac{p^m_2(x_i)}{p^m_1(x_i)} \end{aligned} \tag{4}$$

where N and M denote the number of training sample and the number of total class. $DKL(\cdot)$ represents the Kullback-Leibler divergence between two probabilities.

Feature-level distillation: Apart from the classification-level knowledge distillation, we also transfer the intermediate features of the teacher network with the clinical information to that of the student network. In this regard, we devise a feature-level distillation strategy. Specifically, we distill the output features of two Interactive Models of the teacher network, since these two features integrate the clinical data and the HBP image and the clinical data and the PRE image respectively. Hence, we compute a feature-level distillation loss $L^d_{feature}$ as the combination of the Kullback Leibler (KL) divergence between Z^t_{hbp} and Z^s_{hbp} and the Kullback Leibler (KL) divergence between Z^t_{pre} and Z^s_{pre}:

$$\begin{aligned} L^d_{feature} &= DKL\left(Z^t_{hbp} \| Z^s_{hbp}\right) + \beta_1 DKL\left(Z^t_{pre} \| Z^s_{pre}\right) \\ &= \sum_{i=1}^{N} \sum_{m=1}^{M} p_2^m\left(x_i\right) \log \frac{p_2^m\left(x_i\right)}{p_1^m\left(x_i\right)} \end{aligned} \tag{5}$$

where β_1 is to weight Kullback-Leibler divergence terms, and the weight β_1=1. $DKL(Z^t_{hbp} \| Z^s_{hbp})$ denote the Kullback-Leibler divergence between two features Z^t_{hbp} and Z^s_{hbp}. $DKL(Z^t_{pre} \| Z^s_{pre})$ represents the Kullback-Leibler divergence between two features Z^t_{pre} and Z^s_{pre}.

Our loss function: The loss function of our network consists of two supervised losses on the teacher network and the student network, the regression loss for the clinical data prediction, and distillation loss between the student network and the teacher network. The definition of our loss function is given by:

$$L_{total} = L^s_T + L^s_S + L_{clinical} + L^d_{class} + L^d_{feature} \tag{6}$$

where L^s_T and L^s_S denote the supervised loss of the teacher network prediction and the supervised loss of the student network prediction, respectively. Here, we utilize focal loss [59] to compute the prediction loss of L^s_T and L^s_S. $L_{clinical}$ represents the regression loss for the clinical data prediction, and we use the cross-entropy loss to compute the prediction error of the P^c and the underlying ground truth of the clinical data. L^d_{class} denotes the classification-level distillation loss of (4) and $L^d_{feature}$ is the feature-level distillation loss of (5) between the teacher network and the student network. We utilize the loss function of (6) to train our modality-aware distillation network for a MIV prediction. Pseudo code of our proposed method is shown in Algorithm 1.

Algorithm 1:

Require: The multimodal image data $\{X, Y\}$, the clinical feature C, and the number of training epochs N_e.
Output: The trained teacher model T and student model S.
1: **Server executes**:
2: Initialize T and S
3: **for** epoch $r = 1$ to N_e **do**
4: **for** $x, y, c \in X, Y, C$ **do**
5: $q^t, Z^t_{hbp}, z^t_{pre} = T(x, y, c)$ ▷ Input the multimodal image data x, y and clinical feature c into the teacher model.
6: $q^s, Z^s_{hbp}, z^s_{pre}, P^c = S(x, y)$ ▷ Input the multimodal image data x, y into the student model.
7: Update T and S according to (6) using Adam

F. Technical Details

Data Processing: Note that different patients have different tumor sizes, and the tumors are often smaller in proportion to the image of a patient. If the entire image of the patient is directly passed into the network as the input, there are a large number of non-tumor pixel values, which is not conducive to the training of the network. To avoid this, we find the largest circumscribed cube for the three-dimensional tumors of all patients, and then remove other non-tumor regions outside this cube. The cube size is empirically set as $80 \times 80 \times 20$. Moreover, we randomly extract $(64 \times 64 \times 16)$ volumes from the selected cube region $(80 \times 80 \times 20)$) of each patient's 3D data for data augmentation in the network training. It is important to note that we use the tumor mask solely to identify the tumor's location. The input cube fed into the network contains not only the tumor region but also surrounding background areas (e.g., peritumoral liver tissue or other organs). During model deployment, to avoid the need for manually labeling the cancerous region, we can train a region-cutting network to approximate the tumor's location, which can then be used for predictions.

Inference stage: Given a $80 \times 80 \times 20$ HBP MRI image and a $80 \times 80 \times 20$ PRE MRI image, we employ a center cropping operation on two input volumes to obtain two $64 \times 64 \times 16$ volumes and pass them into the student network of our MD-Net to produce a HCC classification result P^s and a clinical data prediction result P^c. And we directly take P^s as the final classification result of our MD-Net.

Implementation Details: We implement our MD-Net with deep learning framework "Pytorch". Random affine transformation, and a horizontal flip, and a vertical flip are employed to augment the training data. Adam optimizer was used to minimize the total loss function of the deep learning framework. The total epoch number and the batch size are set as 80 and 16. The initial learning rate is 0.0003, and we adjust the learning rate by a decay rate of 0.9 in every 2 epochs. For reproducible research, our code and the collected dataset are available at: https://github.com/lianjizhe/MD-NET.

IV. Experiments

A. Dataset and Evaluation Metric

Dataset: Currently, there is no public annotated dataset for a MVI prediction in HCC. To evaluate the effectiveness of the developed MD-Net, we collected a dataset (denoted as "XMU-HCC", Ethic code: B2021-113) consisting of 270 pathologically confirmed HCC patients with preoperative Gd-EOB-DTPA MRI met the inclusion criteria. The HCC MRI data were taken by a 7-point baseline sample collection protocol [17]. Gd-EOB-DTPA is short for Gadolinium ethoxybenzyl-diethylenetriaminepentaacetic acid. Specifically,

TABLE II
QUANTITATIVE RESULTS (MEAN ± VARIANCE) OF OUR NETWORK AND STATE-OF-THE-ART METHODS ON OUR DATASET

Method	F1-score (%)	Accuracy (%)	AUC (%)	Kappa	p-value
Concat [61]	58.82 ± 1.28	60.75 ± 0.99	67.13 ± 0.73	0.311 ± 0.063	1.30e-3
3DCNN [34]	43.54 ± 6.20	53.93 ± 2.76	67.00 ± 1.02	0.304 ± 0.075	7.27e-4
LSTM [34]	58.20 ± 4.32	61.37 ± 3.24	67.97 ± 1.58	0.339 ± 0.067	5.73e-3
M^2Net [62]	57.45 ± 3.24	59.12 ± 2.87	66.79 ± 1.66	0.326 ± 0.059	1.71e-3
Concat_2S [63]	58.50 ± 0.69	61.13 ± 0.99	68.84 ± 1.34	0.367 ± 0.064	1.02e-2
AdaMSS [64]	57.89 ± 2.77	60.63 ± 3.01	68.48 ± 1.89	0.344 ± 0.061	8.33e-3
XSurv [65]	59.03 ± 3.01	61.67 ± 3.40	69.12 ± 1.56	0.363 ± 0.068	3.03e-2
KD_ours [36]	61.69 ± 3.08	63.14 ± 2.02	71.01 ± 1.60	0.392 ± 0.058	4.50e-2
SP_ours [66]	58.55 ± 4.89	61.57 ± 3.65	68.43 ± 2.64	0.358 ± 0.059	4.90e-4
Our method	$\mathbf{62.53 \pm 3.05}$	$\mathbf{63.78 \pm 3.06}$	$\mathbf{71.86 \pm 1.88}$	$\mathbf{0.402 \pm 0.047}$	

Gd-EOB-DTPA is a novel hepatobiliary-specific MRI contrast agent that can provide functional and structural information of hepatobiliary lesions. Several studies [60] reported that Gd-EOB-DTPA-enhanced MRI could reflect some of the biological features of HCC, including histological grade and microvascular invasion (MVI). In this retrospective study, each patient with preoperative Gd-EOB-DTPA MRI met the inclusion criteria: (a) solitary HCC with the longest diameter ≤ 5 cm; (b) without gross vascular invasion, bile duct tumor thrombosis or extrahepatic metastasis upon preoperative imaging; (c) without previous history of HCC-related treatments (hepatectomy, liver transplantation, chemotherapy, radiotherapy, transarterial chemoembolization, radiofrequency ablation, and immunosuppressive therapy); (d) complete histopathologic description of HCC; (e) MRI with sufficient image quality scanned within 1 mo before surgery [25]. According to the high-risk factors of adverse outcomes, all 270 patients were classified into M_0 (no MVI), or M_1 (invaded vessels were no more than five and located at the peritumoral region adjacent to the tumor surface within 1 cm), or M_2 (MVI of >5 or at >1 cm away from the tumor surface), respectively. The collected dataset consists of 128 M_0 patients, 93 M_1 patients, and 49 M_2 patients. Pre phase images (denoted as "PRE"), hepatobiliary phase images (denoted as "HBP"), and clinical data are collected for each patient. We do not require any registration operation between HBP and PRE images. Moreover, we utilize a five-fold cross-validation strategy to test our network and state-of-the-art classification methods. Specifically, following the stardard steps of a leave-one-out five-fold cross-validation scheme, we split the whole datasets with 270 cases (128 M_0 patients, 93 M_1 patients, and 49 M_2 patients) into five folds. In each round of the cross-validation, we take one fold as the testing set and the other four folds as the training set. Then, we compute the mean and variance value of five rounds for all evaluation metrics, which are F1-score, AUC, accuracy and kappa values, in order to conduct the comparisons between our network and compared methods. We conduct cross-validation five times using different random seeds, resulting in 25 results for each model. We then report the mean and standard deviation of these results. To determine statistically significant differences in comparisons with our proposed method, we calculate p-values using two-sided Wilcoxon signed-rank test to compare the AUC of the 25 results between our network and the compared methods, with the corresponding p-values presented in Table II.

Clinical Data: The clinical data consists of 52 Preoperative laboratory indexes and is reported in Table I. Non-image clinical data in our work are collected and obtained from the report of blood tests, the patient's medical record report, as well as the MRI hallmarks by the radiologists' reviews. Note that a wide resection margin is recommended to improve the prognosis of MVI-positive patients. However, MVI is defined as the cancer cell nest in vessels lined with endothelium, which is visible only on microscopy and poses a challenge for non-invasive diagnosis [25]. Hence, the preoperative laboratory indicators, Child-Pugh score and Barcelona Clinic Liver Cancer (BCLC) stages is of vital importance for the non-invasive stratification of MVI grades before hepatectomy or liver transplantation. The main clinical data was concluded as follows: 1) the preoperative laboratory indicators contain serum tumor markers (e.g., alpha-fetoprotein, carcinoembryonic antigen, carbohydrate antigen 19-9, etc.), hepatitis virus (e.g., hepatitis B virus, anti-hepatitis C virus, HBVDNA loads, etc.), liver function indexes (e.g., alanine aminotransferase, aspartate aminotransferase, total bilirubin, prothrombin time, etc.). 2) Child-Pugh score considers five factors, three of which assess the synthetic function of the liver (i.e., total bilirubin level, serum albumin, and international normalized ratio, or INR) and two of which are based on clinical assessment (i.e., degree of ascites and degree of hepatic encephalopathy). 3) The BCLC classification includes information related to the extent of disease, liver function, and patient performance status to define the disease stage. Please refer to the supplementary material for the detailed meaning of all 52 clinical items.

B. Comparisons Against STATE-of-The-Art Methods

Compared Methods: We evaluate the effectiveness of our classification network by comparing it against seven state-of-the-art methods, including concatenation-based feature fusion method [61] (denoted as "Concat"), "3DCNN" [34], LSTM-based multi-modality fusion method [26] (denoted as "LSTM"), M^2Net [62], stage wise multi-modality fusion network [63] (denoted as "Concat_2S"), AdaMSS [64], XSurv [65], traditional knowledge distillation [36] with our module (denoted as "KD_ours"), and similarity-preserving knowledge distillation [66] with our module (denoted as "SP_ours"). For a fair comparison, we obtain the classification results of all competitors by exploiting its public implementations or implementing them by ourselves, and the network parameters of each network are fine-turned to obtain the best classification results for comparisons.

Quantitative Comparisons: Table II reports the mean ± variance results of three metrics for our method and seven compared networks under a five-fold cross-validation experiment on our "XMU-HCC" dataset. From the results, we can find that "KD_ours" has the best performance on three metrics on all compared methods, and they are the F1-score score of 61.69, the Accuracy score of 63.14, the AUC score of 71.01, and the kappa score of 0.402. More importantly, our method has larger F1-score, Accuracy, and AUC scores than "KD_ours". Specifically, our method has a F1-score improvement of 0.84%, an Accuracy improvement of 0.64%, an AUC improvement of 0.85% and a kappa score improvement of 0.01, when compared

to KD_ours. Moreover, our method outperforms "KD_ours" and "SP_ours" on all three metrics, which demonstrates the superior performance of our distillation method over "KD_ours" and "SP_ours". We compute p-values with two-sided Wilcoxon signed-rank test between our network and compared methods in terms of the AUC metric, and report the corresponding p-values in Table II. Apparently, we can find that all the p-values of our network over compared methods are smaller than 0.05. It indicates that our method has an AUC significant improvement between our network and each compared method.

TABLE III
QUANTITATIVE RESULTS OF OUR METHOD AND BASELINE NETWORKS OF THE ABLATION STUDY

Method	T/S/C	F1-score (%)	Accuracy (%)	AUC (%)	Kappa	p-value
S-PRE	S	57.67 ± 4.36	60.52 ± 2.65	68.35 ± 1.61	0.346 ± 0.058	5.35e-3
S-HBP	S	59.31 ± 1.73	60.83 ± 1.60	69.35 ± 1.43	0.362 ± 0.029	5.72e-4
S-PRE-HBP	S	60.02 ± 0.98	61.95 ± 0.83	69.98 ± 0.96	0.369 ± 0.046	4.22e-2
S-PRE-HBP(Clinical)	S→C	59.92 ± 4.29	62.34 ± 3.06	70.27 ± 1.51	0.382 ± 0.054	4.38e-2
Only-Clinical	C	69.38 ± 4.70	69.93 ± 4.38	79.62 ± 3.31	0.503 ± 0.053	2.54e-5
T-Clinical-PRE	T	74.00 ± 2.41	74.89 ± 2.16	83.61 ± 1.29	0.586 ± 0.068	4.79e-2
T-Clinical-HBP	T	72.38 ± 2.64	72.80 ± 2.57	82.61 ± 1.01	0.558 ± 0.026	1.26e-4
T-Clinical-PRE-HBP	T	75.43 ± 2.71	76.16 ± 2.39	84.69 ± 1.18	0.612 ± 0.078	
ours-w/o-hbp	S→T	58.29 ± 3.95	60.66 ± 2.62	69.04 ± 2.01	0.347 ± 0.069	1.72e-2
ours-w/o-pre	S→T	60.26 ± 3.51	62.39 ± 2.53	70.21 ± 1.47	0.380 ± 0.092	4.78e-2
Our method	S→T	$\mathbf{62.53 \pm 3.05}$	$\mathbf{63.78 \pm 3.06}$	$\mathbf{71.86 \pm 1.88}$	$\mathbf{0.402 \pm 0.047}$	

C. Ablation Study

We also conduct the ablation study experiments to verify the major components in our network design. Here, we construct ten baseline networks, and compare the quantitative results of our method and baseline networks on the "XMU-HCC" dataset.

Baseline Network Design: For renaming different ablation study networks, we first utilize a "S", "T" and "C" to denote these networks based on the student network, teacher network and the clinical variables of our distillation framework, and then we add the modality names (i.e., PRE, HBP, or Clinical) to define the names of different methods. Moreover, we utilize a format of "ours-w/o-[component name]" to rename these baseline networks, which is reconstructed by remove one component (it can be a modality, a loss, or a module) from our network. We first construct four baseline networks based on the student network of our modality-aware distillation method, and they are denoted as "S-PRE", "S-HBP", "S-PRE-HBP" and "S-PRE-HBP(Clinical)". Here, "S-PRE-HBP" represents our student network with the PRE image and the HBP image. "S-PRE-HBP(Clinical)" represents our student network with the PRE image and HBP image as inputs and a multi-task outputs that predicts the clinical variables and the class. "S-PRE" denotes our student network with the only PRE image, while "S-HBP" represents our student network with the only HBP image. The next baseline networks (denoted as "Only-Clinical") by taking only clinical data (which is later converted into the class probabilities through one fully-connected layer) for the MVI prediction. Moreover, we construct three baseline networks based on the teacher network of our method. The first baseline network (denoted as "T-Clinical-PRE-HBP") is our teacher network with the PRE image, the HBP image, and the non-image clinical data, while another two baseline networks (denoted as "T-Clinical-PRE" and "T-Clinical-HBP") are our teacher network with the PRE image and the non-image clinical data, respectively. On the other hand, we build two networks to respectively remove the PRE MRI data and the HBP MRI data from the student network and the teacher network of our method, and they are denoted as "ours-w/o-pre", and "ours-w/o-hbp". Lastly, apart from classical two supervised loss functions at the student network and the teacher network, our method includes another four loss functions, which are L^d_{class}, $L^d_{\text{feature(hbp)}}$, $L^d_{\text{feature(pre)}}$ and L_{clinical}. Hence, we construct five baseline networks to evaluate the effectiveness of these four loss functions by removing each loss function respectively. Specifically, the first baseline network ("ours-w/o-class-distill-loss") is reconstructed by removing the classication-level distillation loss L^d_{class} from the total loss of our network. The second baseline network (denoted as "ours-w/o-HBP-feature-distill-loss") and the third baseline network (denoted as "ours-w/o-PRE-feature-distill-loss") is reconstructed by removing the feature-level distillation loss $L^d_{\text{feature(hbp)}}$ on features from HBP image modality and the feature-level distillation loss $L^d_{\text{feature(pre)}}$ on features from the PRE image modality from the total loss of our network. The fourth baseline network (denoted as "ours-w/o-feature-distill-loss") is reconstructed by moving the feature-level distillation loss $L^d_{\text{feature(hbp)}}$ and $L^d_{\text{feature(pre)}}$ on features from two image modalities from the total loss of our network. The last baseline network (denoted as "ours-w/o-clinical-pred-loss") is to remove the regression loss L_{clinical} of the clinical data prediction from the total loss of our network. Tables III and IV reports the mean and variance results of F1-score, Accuracy, and AUC from our network and all fifteen baseline networks. We compute p-values between our network and baseline methods in terms of the AUC metric, excluding "Only-Clinical", "T-Clinical-PRE", "T-Clinical-HBP" and "T-Clinical-PRE-HBP". Apparently, we can find that all the p-values of our network over baseline methods are smaller than 0.05. It indicates that our method has a AUC significant improvement between our network and each baseline method. Moreover, we compute p-values between "T-Clinical-PRE-HBP" and the other three baseline networks ("Only-Clinical", "T-Clinical-PRE" and "T-Clinical-HBP") and found that all the p-values were less than 0.05, indicating a significant improvement of the teacher network compared to the other three baseline networks in terms of AUC value.

Effectiveness of multi-modality in our teacher and student network: According to the quantitative results of Table III, we can find that "S-PRE-HBP" has higher F1-score, Accuracy, AUC values than "S-PRE" and "S-HBP". It shows that combining the PRE and HBP MRI data together can enhance the MVI classification performance of our student network. "S-PRE-HBP" enhances the mean F1-score value from 59.31% to 60.02%, the mean Accuracy value from 60.83% to 61.95%, and the mean AUC value from 69.35% to 69.98%. Moreover, "T-Clinical-PRE-HBP" outperforms "T-Clinical-PRE" and "T-Clinical-HBP" in terms of F1-score, Accuracy, AUC metrics. It indicates that the combination of the PRE and HBP MRI data in the teacher network of our method improves the MVI classification accuracy.

TABLE IV
QUANTITATIVE RESULTS OF OUR METHOD AND BASELINE NETWORKS OF THE ABLATION STUDY

Method	Loss						F1-score (%)	Accuracy (%)	AUC (%)	Kappa	p-value
	L_{T}^{s}	L_{S}^{s}	L_{clinical}	L_{class}^{d}	$L_{\mathrm{feature(hbp)}}^{d}$	$L_{\mathrm{feature(pre)}}^{d}$					
ours-w/o-clinical-pred-loss	✓	✓	×	✓	✓	✓	61.15 ± 2.94	63.13 ± 2.57	70.84 ± 1.48	0.391 ± 0.045	3.09e-2
ours-w/o-class-distill-loss	✓	✓	✓	×	✓	✓	60.73 ± 3.87	63.03 ± 2.82	70.89 ± 1.96	0.383 ± 0.038	4.85e-2
ours-w/o-HBP-feature-distill-loss	✓	✓	✓	✓	×	✓	61.23 ± 2.35	62.48 ± 1.91	70.76 ± 1.16	0.381 ± 0.039	3.66e-2
ours-w/o-PRE-feature-distill-loss	✓	✓	✓	✓	✓	×	60.23 ± 3.36	62.47 ± 1.69	70.62 ± 1.95	0.375 ± 0.039	2.76e-2
ours-w/o-feature-distill-loss	✓	✓	✓	✓	×	×	61.00 ± 4.44	63.11 ± 3.14	70.58 ± 1.89	0.394 ± 0.022	4.80e-2
Our method	✓	✓	✓	✓	✓	✓	$\mathbf{62.53 \pm 3.05}$	$\mathbf{63.78 \pm 3.06}$	$\mathbf{71.86 \pm 1.88}$	$\mathbf{0.402 \pm 0.047}$	

Effectiveness of medical multi-modality in our modality-aware distillation method: According to the quantitative comparisons in Table III, it can be easily observed that our method has a superior performance of F1-score, Accuracy, and AUC over "ours-w/o-pre" over "ours-w/o-hbp". Moreover, compared to the best-performing results of "ours-w/o-pre" and "ours-w/o-hbp", our method improves F1-score from 60.26% to 62.53%, Accuracy from 62.39%, and AUC from 70.21% to 71.86%. It demonstrates that removing the PRE MRI data or HBP MRI data from our network degrades the MVI classification performance of our network.

Effectiveness of the clinical knowledge distillation: On the other hand, the student network "S-PRE-HBP" has the F1-score value of 60.02%, the Accuracy value of 61.95%, and the AUC value of 69.98%; see Table III. And the teacher network the F1-score, Accuracy, AUC values of "T-Clinical-PRE-HBP" are 75.43%, 76.16%, and 84.69%. Apparently, "T-Clinical-PRE-HBP" has a superior F1-score, Accuracy, and AUC performance over "S-PRE-HBP". It shows that the teacher network has successfully leveraged the additional clinical information. Furthermore, our method has the F1-score value of 62.53%, the Accuracy value of 63.78%, and the AUC value of 71.86%. Hence, we can observe that our method has larger F1-score, Accuracy, and AUC scores than "S-PRE-HBP". It indicates that our network has successfully learned the clinical knowledge from the teacher network to enhance the MVI classification performance of the student network, which relies on only MRI data.

Effectiveness of the MRI images in our method: We construct three baselines from the teacher model of our network by adding PRE images, HBP images, as well as PRE + HBP images, respectively. And these three baselines are denoted as "T-Clinical-PRE", "T-Clinical-HBP", and "T-Clinical-PRE-HBP". From the quantitative results of Table III, we can find that our network with the clinical variables only has a F1-score of 69.38%, an Accuracy score of 69.93%, and an AUC score of 79.62%. By adding the PRE images or the HBP images with the clinical variables, the teacher model of our network has larger F1-score, Accuracy, and AUC scores. Moreover, by combining PRE images, HBP images, and clinical variables, the teacher network of our method has the best MVI grade prediction performance. It indicates that the PRE and HBP images has the complementary information to the clincial variables for predicting the MVI grade.

Effectiveness of feature-level single-modality distillation: Table IV compares the mean and variance values of F1-score, Accuracy, and AUC of a five-fold cross-validation experiment from our method, "ours-w/o-feature-distill-loss", "ours-w/o-HBP-feature-distill-loss" and "ours-w/o-PRE-feature-distill-loss". Apparently, our method has larger F1-score, Accuracy, AUC scores than "ours-w/o-feature-distill-loss". It indicates that the additional feature-level KL divergence loss (see (5)) between the student network and the teacher network has its contribution to the superior MVI classification of our method. Moreover, our method also outperforms "ours-w/o-HBP-feature-distill-loss" and "ours-w/o-PRE-feature-distill-loss", which indicates the effectiveness of the feature distillation loss functions from the HBP image and the PRE image.

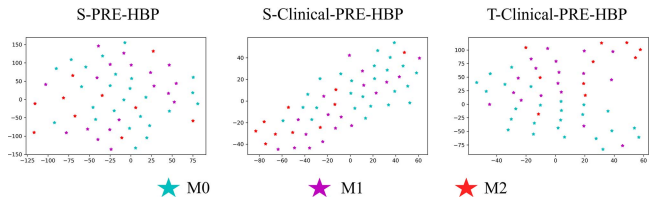

Fig. 5. t-SNE [67] visualization of feature reductions after the SA module for three configurations: S-PRE-HBP, S-Clinical-PRE-HBP, and T-Clinical-PRE-HBP. M0: no MVI, M1: MVI $<= 5$ and within 1 cm of the tumor edge, and M2: MVI > 5 or > 1 cm from the tumor surface.

Effectiveness of the classification-level multi-modality distillation of our method: As shown in Table IV, compared to "ours-w/o-class-distill-loss", our method improves the AUC from 70.89% to 71.86%. It demonstrates that removing the classification-level distillation loss from our network degrades the MVI classification performance of our network.

Effectiveness of the regression task in our method: From the quantitative results shown in Table IV, we can observe that our method also outperforms "ours-w/o-clinical-pred-loss' in terms of F1-score, Accuracy, and AUC. It improve the mean F1-score value from 61.15% to 62.53%, the mean Accuracy value from 63.13% to 63.78%, and the mean AUC value from 70.84% to 71.86%. This is also verified in Table III, where the additional clinical prediction task (S-PRE-HBP (Clinical)) can effectively improve the performance of S-PRE-HBP, increasing the AUC from 69.98% to 70.27%. It shows that the regression loss of predicting the clinical data enables our network to achieve a higher MVI classification accuracy.

D. Qualitative Results of Our MVI Prediction Network

Fig. 5 presents the t-SNE [67] visualization of feature reductions after the SA module for three configurations: S-PRE-HBP, S-Clinical-PRE-HBP, and T-Clinical-PRE-HBP. The

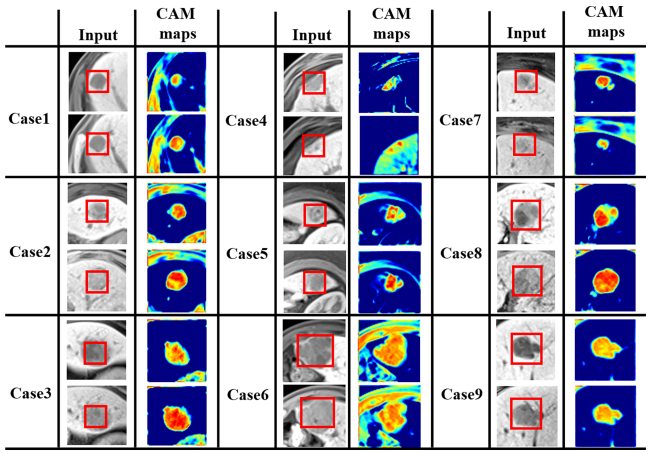

Fig. 6. Visualization of class activation maps for nine cases (Case1 to Case9) with correct MVI predictions.

comparison reveals that the features from S-Clinical-PRE-HBP and T-Clinical-PRE-HBP, which incorporate clinical data (with S-Clinical-PRE-HBP incorporating clinical data indirectly through distillation and supervision), form more compact clusters for each class (i.e., M0, M1, and M2) compared to the features from S-PRE-HBP, which do not use clinical data. It can further demonstrate the effectiveness of clinical features for enhancing the MVI classification of our method.

Fig. 6 shows the middle slices of nine randomly chosen PRE or HBP data (see Case 1 to Case 9), and their corresponding class activation maps (M3D-CAM) [68] for our classification results. From these class activation maps (CAMs), we can find that the important areas for our MVI grade classification task are tumoral areas and peritumoral areas. Hence, when the CAMs for the input data can highlight these tumoral areas and peritumoral areas, our method has the correct MVI grade prediction.

V. Discussion and Conclusion

This work presents a modality-aware knowledge distillation network (MD-Net) for a MVI prediction in HCC. Our MD-Net transfers the teacher network with a non-image clinical modality and a multi-MRI image modality to a student network with only image multi-MRI image by formulating a classification-level distillation and a feature-level distillation. By doing so, with the help of the distilled clinical information, our student network can obtain a superior HCC prediction in the testing stage, which does not have any clinical modality data. In the teacher network, we formulate MRI-clinical-fusion CNNs and a symmetric attention (SA) module to integrate two groups of the MRI data and the clinical data. Then, we formulate two MRI-only module and a SA module to fuse features from two MRI data in the student network of our MD-Net. Moreover, we devise a regression task to predict a clinical data from the MRI images for benefiting the downstream HCC prediction task. Experimental results on our collected dataset and a multi-modal sarcasm detection dataset show that our MD-Net outperforms state-of-the-art methods in terms of a MVI prediction in HCC.

Note that this work only involves the PRE and HBP MRI data into the developed MVI classification network. However, we find that the classification performance degrades when more MRI data is added. The reason behind is likely that our network with more MRI modality images tends to be over-fitted due to the insufficient training data. We argue that this issue can be resolved by devising a model to more efficiently fuse different MRI modality images or enlarging the training dataset. Moreover, we will test our network on more and larger datasets and extend it to handle multi-center data.

References

[1] A. Forner et al., "Treatment of hepatocellular carcinoma," *Crit. Rev. Oncol./ Hematol.*, vol. 60, no. 2, pp. 89–98, 2006.
[2] L. Yang et al., "A radiomics nomogram for preoperative prediction of microvascular invasion in hepatocellular carcinoma," *Liver Cancer*, vol. 8, no. 5, pp. 373–386, 2019.
[3] R. De Angelis et al., "Cancer survival in Europe 1999–2007 by country and age: Results of Eurocare-5—A population-based study," *Lancet Oncol.*, vol. 15, no. 1, pp. 23–34, 2014.
[4] N. Fujiwara et al., "Risk factors and prevention of hepatocellular carcinoma in the era of precision medicine," *J. Hepatol.*, vol. 68, no. 3, pp. 526–549, 2018.
[5] V. Mazzaferro et al., "Metroticket 2.0 model for analysis of competing risks of death after liver transplantation for hepatocellular carcinoma," *Gastroenterology*, vol. 154, no. 1, pp. 128–139, 2018.
[6] O. Miltiadous et al., "Progenitor cell markers predict outcome of patients with hepatocellular carcinoma beyond milan criteria undergoing liver transplantation," *J. Hepatol.*, vol. 63, no. 6, pp. 1368–1377, 2015.
[7] A. Kardashian et al., "Liver transplantation outcomes in a us multicenter cohort of 789 patients with hepatocellular carcinoma presenting beyond milan criteria," *Hepatology*, vol. 72, no. 6, pp. 2014–2028, 2020.
[8] A. Vogel et al., "Hepatocellular carcinoma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 29, pp. iv238–iv255, 2018.
[9] Y. Yang et al., "Patterns and clinicopathologic features of extrahepatic recurrence of hepatocellular carcinoma after curative resection," *Surgery*, vol. 141, no. 2, pp. 196–202, 2007.
[10] N. Nagasue et al., "Incidence and factors associated with intrahepatic recurrence following resection of hepatocellular carcinoma," *Gastroenterology*, vol. 105, no. 2, pp. 488–494, 1993.
[11] S. Roayaie et al., "A system of classifying microvascular invasion to predict outcome after resection in patients with hepatocellular carcinoma," *Gastroenterology*, vol. 137, no. 3, pp. 850–855, 2009.
[12] S. Lee et al., "Preoperative gadoxetic acid–enhanced MRI for predicting microvascular invasion in patients with single hepatocellular carcinoma," *J. Hepatol.*, vol. 67, no. 3, pp. 526–534, 2017.
[13] N. Portolani et al., "Early and late recurrence after liver resection for hepatocellular carcinoma: Prognostic and therapeutic implications," *Ann. Surg.*, vol. 243, no. 2, 2006, Art. no. 229.
[14] H. Imamura et al., "Risk factors contributing to early and late phase intrahepatic recurrence of hepatocellular carcinoma after hepatectomy," *J. Hepatol.*, vol. 38, no. 2, pp. 200–207, 2003.
[15] S. Okada et al., "Predictive factors for postoperative recurrence of hepatocellular carcinoma," *Gastroenterology*, vol. 106, no. 6, pp. 1618–1624, 1994.
[16] S. Sumie et al., "The significance of classifying microvascular invasion in patients with hepatocellular carcinoma," *Ann. Surg. Oncol.*, vol. 21, no. 3, pp. 1002–1009, 2014.
[17] W.-M. Cong et al., "Practice guidelines for the pathological diagnosis of primary liver cancer: 2015 update," *World J. Gastroenterol.*, vol. 22, no. 42, 2016, Art. no. 9279.
[18] Z. Lei et al., "Nomogram for preoperative estimation of microvascular invasion risk in hepatitis b virus–related hepatocellular carcinoma within the milan criteria," *JAMA Surg.*, vol. 151, no. 4, pp. 356–363, 2016.
[19] M. Rodriguez-Peralvarez et al., "A systematic review of microvascular invasion in hepatocellular carcinoma: Diagnostic and prognostic variability," *Ann. Surg. Oncol.*, vol. 20, no. 1, pp. 325–339, 2013.
[20] X. Xu et al., "Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma," *J. Hepatol.*, vol. 70, no. 6, pp. 1133–1144, 2019.

[21] J. Shindoh et al., "Microvascular invasion and a size cutoff value of 2 cm predict long-term oncological outcome in multiple hepatocellular carcinoma: Reappraisal of the American joint committee on cancer staging system and validation using the surveillance, epidemiology, and end-results database," *Liver Cancer*, vol. 9, no. 2, pp. 156–166, 2020.

[22] T. Iguchi et al., "New pathologic stratification of microvascular invasion in hepatocellular carcinoma: Predicting prognosis after living-donor liver transplantation," *Transplantation*, vol. 99, no. 6, pp. 1236–1242, 2015.

[23] H.-T. Hu et al., "Peritumoral tissue on preoperative imaging reveals microvascular invasion in hepatocellular carcinoma: A systematic review and meta-analysis," *Abdominal Radiol.*, vol. 43, no. 12, pp. 3324–3330, 2018.

[24] S.-T. Feng et al., "Preoperative prediction of microvascular invasion in hepatocellular cancer: A radiomics model using Gd-EOB-DTPA-enhanced MRI," *Eur. Radiol.*, vol. 29, no. 9, pp. 4648–4659, 2019.

[25] H.-H. Chong et al., "Multi-scale and multi-parametric radiomics of gadoxetate disodium–enhanced MRI predicts microvascular invasion and outcome in patients with solitary hepatocellular carcinoma 5 cm," *Eur. Radiol.*, vol. 31, pp. 4824–4838, 2021.

[26] S. Men et al., "Prediction of microvascular invasion of hepatocellar carcinoma with contrast-enhanced mr using 3D CNN and LSTM," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 810–813.

[27] X. Ma et al., "Preoperative radiomics nomogram for microvascular invasion prediction in hepatocellular carcinoma using contrast-enhanced CT," *Eur. Radiol.*, vol. 29, no. 7, pp. 3595–3605, 2019.

[28] T. Hennedige and S. K. Venkatesh, "Imaging of hepatocellular carcinoma: Diagnosis, staging and treatment monitoring," *Cancer Imag.*, vol. 12, no. 3, 2012, Art. no. 530.

[29] A. Kitao et al., "Gadoxetic acid-enhanced MR imaging for hepatocellular carcinoma: Molecular and genetic background," *Eur. Radiol.*, vol. 30, no. 6, pp. 3438–3447, 2020.

[30] Y. Zhang et al., "Deep learning with 3D convolutional neural network for noninvasive prediction of microvascular invasion in hepatocellular carcinoma," *J. Magn. Reson. Imag.*, vol. 54, no. 1, pp. 134–143, 2021.

[31] E. Ünal et al., "Microvascular invasion in hepatocellular carcinoma," *Diagn. Interventional Radiol.*, vol. 22, no. 2, 2016, Art. no. 125.

[32] J. Zheng et al., "Preoperative prediction of microvascular invasion in hepatocellular carcinoma using quantitative image analysis," *J. Amer. College Surgeons*, vol. 225, no. 6, pp. 778–788, 2017.

[33] H. Duanmu et al., "Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 242–252.

[34] Y.-Q. Jiang et al., "Preoperative identification of microvascular invasion in hepatocellular carcinoma by xgboost and deep learning," *J. Cancer Res. Clin. Oncol.*, vol. 147, no. 3, pp. 821–833, 2021.

[35] X. Xiao, J. Zhao, and S. Li, "Task relevance driven adversarial learning for simultaneous detection, size grading, and quantification of hepatocellular carcinoma via integrating multi-modality MRI," *Med. Image Anal.*, vol. 81, 2022, Art. no. 102554.

[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[37] W. Hou et al., "Early neoplasia identification in Barrett's esophagus via attentive hierarchical aggregation and self-distillation," *Med. Image Anal.*, vol. 72, 2021, Art. no. 102092.

[38] S. Vesal et al., "Domain generalization for prostate segmentation in transrectal ultrasound images: A multi-center study," *Med. Image Anal.*, vol. 82, 2022, Art. no. 102620.

[39] C. Yang et al., "Source free domain adaptation for medical image segmentation with Fourier style mining," *Med. Image Anal.*, vol. 79, 2022, Art. no. 102457.

[40] H. Wang et al., "Segmenting neuronal structure in 3D optical microscope images via knowledge distillation with teacher-student network," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2019, pp. 228–231.

[41] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2019, pp. 1563–1566.

[42] S. Christodoulidis et al., "Multisource transfer learning with convolutional neural networks for lung pattern analysis," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 76–84, Jan. 2017.

[43] K. Li et al., "Towards cross-modality medical image segmentation with online mutual knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 01, pp. 775–783.

[44] X. Xing et al., "Categorical relation-preserving contrastive knowledge distillation for medical image classification," in *Proc. Med. Image Comput. Comput. Assist. Interv.—MICCAI 2021: 24th Int. Conf.*, Strasbourg, France, Sep. 27–Oct. 1, 2021, pp. 163–173.

[45] L. Ju et al., "Synergic adversarial label learning for grading retinal diseases via knowledge distillation and multi-task learning," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3709–3720, Oct. 2021.

[46] S. Javed et al., "Knowledge distillation in histology landscape by multi-layer features supervision," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 4, pp. 2037–2046, Apr. 2023.

[47] C. Zhou et al., "One-pass multi-task convolutional neural networks for efficient brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2018, pp. 637–645.

[48] C. Xia et al., "A multi-modality network for cardiomyopathy death risk prediction with CMR images and clinical information," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2019, pp. 577–585.

[49] G. A. Tadesse et al., "Multi-modal diagnosis of infectious diseases in the developing world," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 2131–2141, Jul. 2020.

[50] F. Fang et al., "Self-supervised multi-modal hybrid fusion network for brain tumor segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5310–5320, Nov. 2022.

[51] H. Cai, Y. Gao, and M. Liu, "Graph transformer geometric learning of brain networks using multimodal MR images for brain age estimation," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 456–466, Feb. 2022.

[52] Z. Xing et al., "Nestedformer: Nested modality-aware transformer for brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 140–150.

[53] F. Lin et al., "Adaptive multi-modal fusion framework for activity monitoring of people with mobility disability," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4314–4324, Aug. 2022.

[54] S. J. Giri et al., "MultiPredGo: Deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1832–1838, May 2021.

[55] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[56] S. W. Oh et al., "Video object segmentation using space-time memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9226–9235.

[57] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[58] X. Wang et al., "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[59] T.-Y. Lin et al., "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[60] O. Clément et al., "Gadolinium-ethoxybenzyl-DTPA, a new liver-specific magnetic resonance contrast agent. kinetic and enhancement patterns in normal and cholestatic rats," *Invest. Radiol.*, vol. 27, no. 8, pp. 612–619, 1992.

[61] D. Nie et al., "Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, 2019.

[62] T. Zhou et al., "M^{2}Net: Multi-modal multi-channel network for overall survival time prediction of brain tumor patients," in *Med. Image Comput. Comput. Assist. Interv.—MICCAI 2020: 23rd Int. Conf.*, Lima, Peru, Oct. 4–8, 2020, pp. 221–231.

[63] T. Zhou et al., "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis," *Hum. Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019.

[64] M. Meng et al., "Adamss: Adaptive multi-modality segmentation-to-survival learning for survival outcome prediction from PET/CT images," 2023, *arXiv:2305.09946*.

[65] M. Meng et al., "Merging-diverging hybrid transformer networks for survival prediction in head and neck cancer," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2023, pp. 400–410.

[66] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1365–1374.

[67] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[68] K. Gotkowski et al., "M3D-CAM: A pytorch library to generate 3D data attention maps for medical deep learning," in *Proc. Bildverarbeitung für Die Medizin 2021: Proc., German Workshop Med. Image Comput.*, Regensburg, Germany, Mar. 7–9, 2021, pp. 217–222.