IEEE | Transactions on Medical Imaging

TDSF-Net: Tensor Decomposition-based Subspace Fusion Network for Multi-modal Medical Image Classification

Journal:	*IEEE Transactions on Medical Imaging*
Manuscript ID	TMI-2023-1806
Manuscript Type:	Regular Paper
Date Submitted by the Author:	11-Sep-2023
Complete List of Authors:	Zhang, Yi; Tianjin University of Technology, School of Computer Science and Engineering Xu, Guoxia ; Nanjing University of Posts and Telecommunications, School of Communications and Information Engineering Zhao, Meng; Tianjin University of Technology, School of Computer Science and Engineering Wang, Hao; Xidian University, School of Cyber Engineering Shi, Fan; Tianjin University of Technology, School of Computer Science and Engineering Chen, Stephen; Tianjin University of Technology, School of Computer Science and Engineering
Keywords:	Multi-modality fusion < General methodology, Microscopy < Imaging modalities, Optical Imaging/OCT/DOT < Imaging modalities

TDSF-Net: Tensor Decomposition-based Subspace Fusion Network for Multi-modal Medical Image Classification

Yi Zhang, Guoxia Xu, *Member, IEEE,* Meng Zhao, Hao Wang, *Senior Member, IEEE,* Fan Shi, and Shengyong Chen, *Senior Member, IEEE,*

***Abstract*— Data from multi-modalities bring complementary information for deep learning-based medical image classification models. However, data fusion methods simply concatenating features or images barely consider the correlations or complementarities among different modalities and easily suffer from exponential growth in dimensions and computational complexity when the modality increases. Consequently, this paper proposes a subspace fusion network with tensor decomposition to heighten multi-modal medical image classification. We first introduce a Tucker low-rank tensor decomposition module to map the high-level dimensional tensor to the low-rank subspace, reducing the redundancy caused by multi-modal data and high-dimensional features. Then a cross-tensor attention mechanism is utilized to fuse features from the subspace into a high-dimension tensor, enhancing the representation ability of extracted features and constructing the interaction information among components in the subspace. Extensive comparison experiments with state-of-the-art methods are conducted on one self-established and two public multi-modal medical image datasets, verifying the effectiveness and generalization ability of the proposed method. Code is available at `https://github.com/1zhang-yi/TDSFNet`.**

***Index Terms*— Deep Learning, Multi-modal Fusion, Tucker Reconstruction, Attention Mechanism, Classification.**

I. Introduction

MULTI-MODAL data fusion integrates multiple image information from different modalities, leading to more comprehensive cognition of the objects, and inducing the possibility of discovering underlying mechanisms, thus assisting doctors to achieve higher diagnostic accuracy [1]. Traditional feature extraction methods use diverse operators, yet fused simply by addition, multiplication or concatenation. In this way, the relevance and complementarity of the extracted features are barely analyzed, and the redundancy among these features is also ignored.

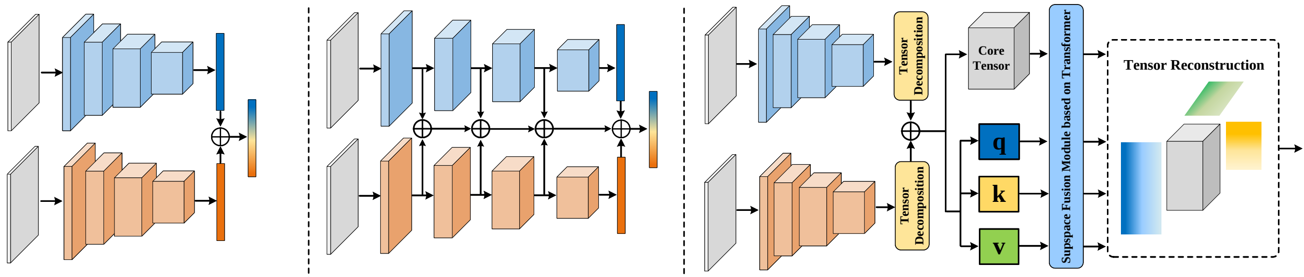
With the wide application of deep learning in computer vision, Convolutional Neural Networks(CNNs) have been introduced for multi-modal image fusion [2]. So far, multi-modal classification architectures based on CNNs can be roughly divided into three categories, early fusion(pixel-level fusion), middle-level fusion(feature-level fusion), and late fusion(decision-level fusion) [3]. Some exiting multi-modal fusion networks adopt early fusion strategy [4], which integrates the minimum granularity of raw data and retains original image information to the greatest extent. However, the early fusion method is relatively coarse. It operates on raw pixels, introducing a lot of noise and requiring a long processing time. Instead, late fusion merges the decision results at the last layer of the network [5], [6]. Images of each modality are fed into a separate network to learn about each modality's information independently. The outputs of individual networks are integrated to get the final prediction result. However, this method only fuses the highest-level features without fully considering the underlying interaction of multi-modal data, which will degrade the performance of the fusion model.

Given the shortcomings of early and late fusion, they will inhibit intra-modal or inter-modal interaction, so most image fusion methods use feature-level fusion [7]–[9] or multi-scale feature aggregation [10]. Similar to the late fusion network structure, the feature-level fusion also takes the image of each modality as an independent input and sends the image into a separate network. Still, the learned individual feature representations will be fused in the middle layers of the network. The network structure of this fusion method is more diverse and flexible. Bi et al. [11] proposed a three branches network named HcCNN to predict the classification results of the 7-point Checklist dataset. HcCNN has a hyper-branch with a multi-scale attention block(MsA) to fuse two modalities across various image scales. The fusion method in hyper-branch is still a simple concatenation followed by convolutions. The simple feature concatenation can easily cause the dimension disaster problem of data processing and increase the difficulty of network training.

To this end, we propose a Tensor Decomposition-based Subspace Fusion Network named TDSFNet for multi-modal

Yi Zhang and Meng Zhao are with are with the Key Laboratory of Computer Vision and System of Ministry of Education, School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, 300384, China. Guoxia Xu is with School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. (e-mail: gxxu.re@gmail.com). Hao Wang is with the School of Cyber Engineering of Xidian University, Xi'an, 710126, China. Yi Zhang and Guoxia Xu contribute equally for this manuscript. Corresponding author: Meng Zhao, Email: zh_m@tju.edu.cn.

This work is supported by the National Natural Science Foundation of China under Grants 62272342, 62020106004 and 92048301.



(a) Multi-modal CNN fusion in last layer **(b) Multi-modal CNN fusion in each layer** **(c) Multi-modal CNN fusion in Feature Subspace**

Fig. 1: Multi-modal fusion architectures.

medical image classification. Our model differs from existing methods that fuse images at the feature level. Inspired by tensor decomposition, we propose a Tucker decomposition module, which can reduce inherent noises caused by data-level input redundancy and excessive feature dimensions by adjusting rank. After obtaining the feature subspace by the Tucker decomposition module, we innovatively combine the subspace with the transformer and use the self-attention mechanism to exchange features in the subspace of different dimensions and exploit the correlations among different parts of subspace, thereby changing the feature space of existing features. Finally, the multiple components of the fused subspace are reconstructed by tensor reconstruction to generate a latent representation, which is used to predict the results of classification. The main contributions of this paper are summarized as follows:

- Unlike most existing multi-modal image fusion networks, we propose the Tucker decomposition module to map features to subspaces instead of fusing directly at the feature level. This feature decomposition process can be regarded as a dimension reduction process that reduces the redundancy caused by high-dimensional features by adjusting the rank on each dimension.
- We use the subspace-based transformer to apply a self-attention mechanism to exchange features in the subspace of different dimensions, which enables the various components of the subspace to be more relevant, fully learn the complementary information, and align the features between different modalities.
- The effectiveness of our method is demonstrated on three multi-modal medical datasets: the Pleural Effusion Cell(PEC) dataset, the Age-related Macular Degeneration(AMD) dataset, and the Seven-Point Checklist(SPC) dataset.

II. Related Works

This section reviews related works about Multi-modal Image Fusion, Tensor Decomposition, and Multi-modal Transformers.

A. Multi-modal Image Fusion

Multi-modal fusion refers to the process of integrating the most meaningful information from two or more modalities [1]. The single modality usually cannot contain all the effective information needed to produce accurate results [12], [13]. The multi-modal fusion combines information from two or more modalities to realize information supplement, improve prediction results' accuracy, and improve models' robustness [14].

Medical images have significant remote dependency compared to natural images. All the information in a medical image has potential use value, and subtle structures in medical images are not as irrelevant as in natural images. In addition, human tissue has a high degree of similarity, and a slight change may represent diseased lesions. At the same time, differences in medical images are acquired due to different brands of equipment and many other factors. Furthermore, different staining reagents, imaging devices, device parameters, and photographic angles make multi-modal medical images heterogeneous. In addition, the medical image datasets have the disadvantages of fewer samples, high labeling cost, and no pre-training models [15]. The above characteristics of multi-modal medical images make medical image fusion more challenging. To overcome the challenge of fewer data, Wang et al. [16] proposed a two-stream CNN model to predict the category of Age-related Macular Degeneration and developed two data augmentation methods to improve model accuracy by increasing the amount of data in the training set. He et al. [9] also proposed a Co-Attention Fusion Network named CAFNet to classify the SPC dataset. The CAFNet has a hyper-branch with an attention fusion block at all stages to refine and fuse features from two image modalities.

B. Tensor Decomposition

Tensor decomposition is an important part of tensor analysis. Its basic principle is to use the structural information in tensor data to decompose the tensor into a combination of several tensors with simpler forms and smaller storage scales. In addition, deep wavelet decomposition is also an effective method for feature decomposition [17]. Tucker decomposition is a high-order Principal Component Analysis(PCA) form. It decomposes a tensor into the product of a core tensor and each dimensional matrix. Factor matrices on each dimension can be viewed as principal components, and the core tensor represents the relationship among different factor matrices. The Tucker decomposition of a 3-way tensor $\mathcal{X} \in \mathbb{R}^{c \times h \times w}$ expresses $\mathcal{X}$ as a tensor product between factor matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, and a core tensor $\mathcal{G}$ in such a way that:

$$\mathcal{X} = ((\mathcal{G} \times_1 \mathbf{A}) \times_2 \mathbf{B}) \times_3 \mathbf{C} \tag{1}$$

with $\mathbf{A} \in \mathbb{R}^{c \times c_1}$, $\mathbf{B} \in \mathbb{R}^{h \times h_1}$, $\mathbf{C} \in \mathbb{R}^{w \times w_1}$, $\mathcal{G} \in \mathbb{R}^{c_1 \times h_1 \times w_1}$, and $\times_n$ represents the multiplication of tensors in n-th mode. $\mathcal{X}$ is usually summarized as $\mathcal{X}$ = [$\mathcal{G}$; $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$]. The 3-factor matrices are orthogonal and can be seen as principal components along the corresponding dimension. Each digital element in core tensor $\mathcal{G}$ represents the degree of interaction between the different components. Ben-Younes et al. [18] produced a multi-modal Tucker fusion framework for visual question answering, which enables the network to model rich and accurate correlation information between images and questions.

C. Multi-modal Transformer

Transformer has been widely used in natural language processing, computer vision, and other fields since its inception. Transformer was originally proposed as a sequence-to-sequence model for machine translation [19]. With the rapid development of computer vision, the transformer has made outstanding contributions in the field of computer vision. The inspiration for Vision Transformer(ViT) [20] comes from the self-attention mechanism in natural language processing, in which word embedding is replaced by patch embedding. Transformer also plays an important role in multi-modal fusion [21]–[24]. The attention mechanism in the transformer has the ability to feature aggregation in different feature spaces and global ranges, which is suitable for the alignment and fusion of multi-modal feature expressions. Wang et al. [25] proposed a segmentation network based on U-net, which replaces the skip connection operation in U-net, fuses the features of multiple scales in the encoder through the attention mechanism, and joins them to the features of the decoder for eliminating the ambiguity.

III. Methodology

Multi-modal medical image fusion is to integrate two images $\mathbf{I}_1$ and $\mathbf{I}_2$ acquired from two modalities. Our goal is to predict a specific class ŷ based on the paired input $\{\mathbf{I}_1, \mathbf{I}_2\}$. Fig. 2 gives the structure of our proposed TDSFNet, which includes four main components: Feature Extractor, Tucker Decomposition Module, Subspace Fusion Module, and Tucker Reconstruction Module.

A. Feature Extractor

The images from two modalities are fed into the same network structure, which uses two ResNets as feature extractors to obtain the deep features of two modality images respectively, denoted as $\mathbf{F}_1$ and $\mathbf{F}_2$.

B. Tucker Decomposition Module

Tucker decomposition is widely used in data dimensionality reduction, feature extraction, and tensor subspace learning. We map the 3-dimensional features into the 2-dimensional subspace, reducing the data dimension, which can reduce the complexity of the training network and alleviate the problem of overfitting to a certain extent. High-dimensional features can also be viewed as the sum of high- and low-dimensional noise-free data. We find noise-free data space to remove redundant data while retaining as much useful information as possible.

We use ResNet to extract high dimensional features $\mathbf{F}_1$ and $\mathbf{F}_2$ of images from two modalities. Then, we decompose them into a core tensor and factor matrix on each dimension, respectively. As shown in Fig. 2, $\mathbf{F}_1$ and $\mathbf{F}_2$ are processed by two parallel Tucker decomposition modules, with the aim of learning feature subspace within every single modality. For the sake of simplicity, we only describe the feature of one modality $\mathbf{F}_1$ decomposition. Given $\mathbf{F}_1 \in \mathbb{R}^{c \times h \times w}$, where c, h, and w represent the channel, height, and width of the feature respectively. We first use a series convolution, batch normalization, and ReLU operations to generate a core tensor. The convolution operation can change the channel number and feature size, and the shape of the core tensor represents the rank of Tucker decomposition, denoted as $\mathbf{core}_1$ $\in \mathbb{R}^{rank\_c \times rank\_h \times rank\_w}$. Then, we define three-factor matrix generators in each dimension. The operation of each generator is the same. For simplicity, we only introduce the generator on the channel dimension. The generator comprises pooling operations, convolution operations, and sigmoid operations. Global pooling can effectively capture global information, retaining only one dimension of information. Here, we use global average pooling to obtain global context representation. Afterward, we repeat 1×1 convolution and sigmoid operation $rank\_c$ times to get $rank\_c$ vectors. All vectors are generated using independent convolution kernels. Each of them learns a part of context information along a specific direction. Then, we concatenate $rank\_c$ vectors to obtain a 2-dimensional factor matrix. By using the same operations on height and width dimensions, we can get the corresponding factor matrix:

$$\begin{aligned}
\mathbf{H}_1 &= Concat(Repeat(\sigma(conv(GAP_h(\mathbf{F}_1))))) \\
\mathbf{W}_1 &= Concat(Repeat(\sigma(conv(GAP_w(\mathbf{F}_1))))) \\
\mathbf{C}_1 &= Concat(Repeat(\sigma(conv(GAP_c(\mathbf{F}_1)))))
\end{aligned} \tag{2}$$

where Concat is the concatenate operation, σ is the sigmoid operator, $conv$ is the convolution operation, GAP_h, GAP_w, and GAP_c represent Global Average pooling on h, w, c direction respectively.

C. Subspace Fusion Module

Multi-modal fusion aims to obtain a fusion representation to complete the specific task better. However, in existing multi-modal fusion methods, the fusion operation is just a simple concatenation, add, or multiplication without considering the correlations and complementarity among different modalities. The fusion performance may be limited due to data heterogeneity and misalignment of different modalities. To this end, we design the Subspace Fusion Module with an attention mechanism to generate a more representative fusion tensor. In this module, we aggregate the subspaces from two modalities using average summation on each component of the feature

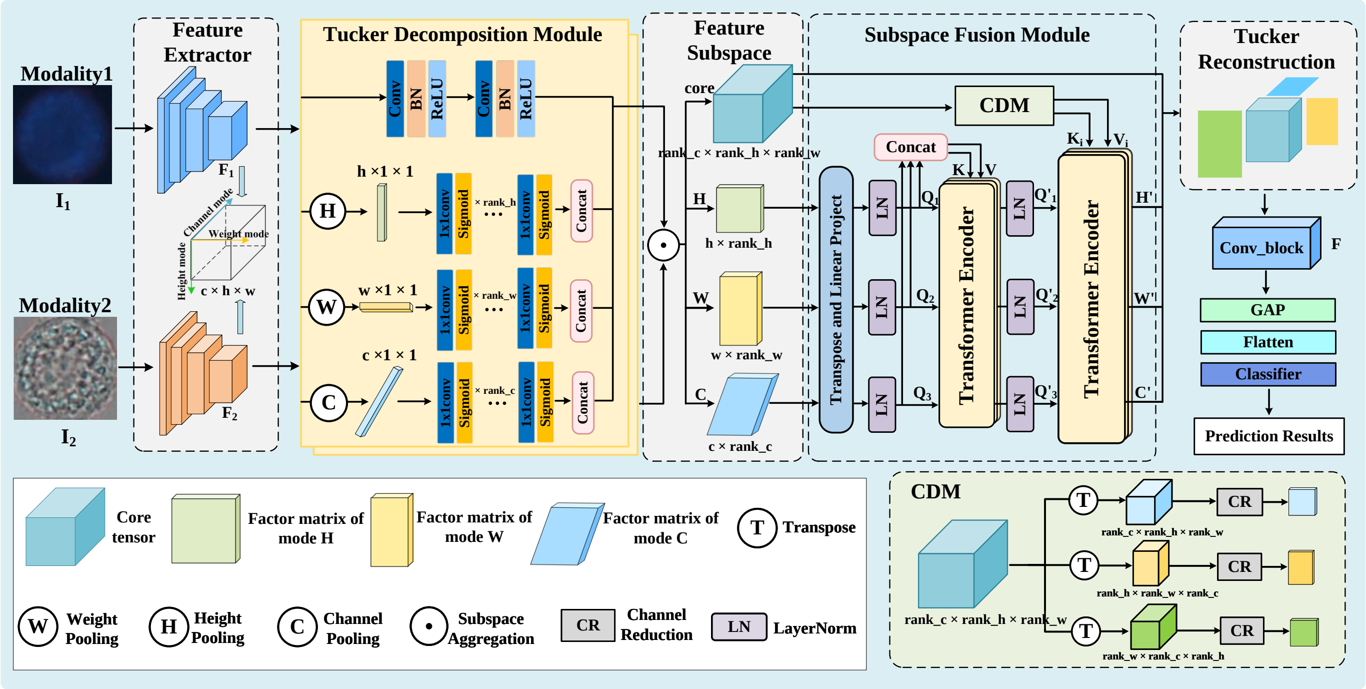


Fig. 2: Overview of the proposed TDSF-Net: Tensor Decomposition-based Subspace Fusion Network for multi-modal medical image classification.

subspace of the two modalities:

$$\begin{aligned}\mathbf{core} &=(\mathbf{core}_1+\mathbf{core}_2)/2\\ \mathbf{C} &=(\mathbf{C}_1+\mathbf{C}_2)/2\\ \mathbf{H} &=(\mathbf{H}_1+\mathbf{H}_2)/2\\ \mathbf{W} &=(\mathbf{W}_1+\mathbf{W}_2)/2\end{aligned}\tag{3}$$

where $\mathbf{core}_1$ and $\mathbf{core}_2$ are core tensors from two modalities, $\mathbf{C}_1$ and $\mathbf{C}_2$ are factor matrices on channel dimension, $\mathbf{H}_1$ and $\mathbf{H}_2$ are factor matrices on height dimension, $\mathbf{W}_1$ and $\mathbf{W}_2$ are factor matrices on width dimension.

Given the outputs of Tucker Decomposition Module $\mathbf{core} \in \mathbb{R}^{rank\_c \times rank\_h \times rank\_w}$, $\mathbf{C} \in \mathbb{R}^{c \times rank\_c}$, $\mathbf{H} \in \mathbb{R}^{h \times rank\_h}$, $\mathbf{W} \in \mathbb{R}^{w \times rank\_w}$, the 3 factor matrices $\mathbf{C}$, $\mathbf{H}$, $\mathbf{W}$ are first transposed and scaled to facilitate subsequent fusion operation and then fed into a multi-head cross-attention module, followed by Feed-Forward Networks(FFN) with residual structure. As shown in Fig. 2, we concatenate the three-factor matrices as the key and value:

$$\begin{aligned}\mathbf{Q}_1 &= LN(T_c(\mathbf{C}W_{q1}))\\ \mathbf{Q}_2 &= LN(T_h(\mathbf{H}W_{q2}))\\ \mathbf{Q}_3 &= LN(T_w(\mathbf{W}W_{q3}))\\ \mathbf{K}=\mathbf{V} &= Concat(\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3)\end{aligned}\tag{4}$$

where T_c, T_h and T_w are transposed operations that transpose c, h, w to the first dimension, LN is layer normalization operation, W_{q1}, W_{q2} and W_{q3} are weights of $\mathbf{Q}_1$, $\mathbf{Q}_2$ and $\mathbf{Q}_3$.

As shown in Fig. 2, the attention mechanism is the fundamental component in our proposed Subspace Fusion Module. The attention function is defined as:

$$Attention(\mathbf{Q_i},\mathbf{K},\mathbf{V}) = softmax(\frac{\mathbf{Q_i}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V} \tag{5}$$

where d_k is the key dimensionality, i = 1, 2, 3.

Here, we often extend the attention mechanism into multiple heads to enable the mechanism to consider various attention distributions and make the model pay attention to different aspects of information:

$$CA_i = (Attention_i^1 + ... + Attention_i^N)/N \tag{6}$$

where N in the number of heads. i = 1, 2, 3. Afterwards, we apply layer normalization and Feed-Forward Network(FFN) and obtain the outputs of transformer module $\mathbf{Q_1^{'}}$, $\mathbf{Q_2^{'}}$ and $\mathbf{Q_3^{'}}$.

$$\mathbf{Q_i^{'}} = CA_i + FFN(\mathbf{Q_i} + LN(CA_i)) \tag{7}$$

where CA_i is the output of the multi-head cross-attention module, and LN is the layer normalization operation.

Given the outputs of attention mechanism $\mathbf{Q_1^{'}}$, $\mathbf{Q_2^{'}}$ and $\mathbf{Q_3^{'}}$, we only consider the correlation between three-factor matrices, the core tensor obtained by Tucker decomposition as high-frequency information is equivalent to the main component of the feature tensor, so we still use the attention mechanism to fuse the core tensor and the three-factor matrices. As shown in Fig. 2, we decompose the core tensor into three matrices in three dimensions through the proposed Core Decomposition Module(CDM). The core tensor $\mathbf{core} \in \mathbb{R}^{rank\_c \times rank\_h \times rank\_w}$, we first transpose $\mathbf{core}$, using convolution operations to reduce the number of channels to 1. In this way, we change the three-dimensional tensor to a two-dimension matrix. Following by convolution, batch normalization, and activation operation to refine the feature.

$$\mathbf{K_i} = \mathbf{V_i} = \sigma(BN(conv(T_i(\mathbf{core})))) \tag{8}$$

where σ is the activation function, BN is batch normalization, conv is the convolution operation, T_i is transposed operations that transpose c, h, w to the first dimension respectively, i = c, h, w. The three-factor matrices $\mathbf{Q_1^{'}}$, $\mathbf{Q_2^{'}}$ and $\mathbf{Q_3^{'}}$ in three dimensions have their corresponding $\mathbf{K}_i$ and $\mathbf{V}_i$, the attention function is defined as

$$\begin{aligned} Attention(\mathbf{Q_1^{'}}, \mathbf{K_c}, \mathbf{V_c}) &= softmax(\frac{\mathbf{Q_1^{'}}\mathbf{K_c^{\top}}}{\sqrt{d_c}})\mathbf{V_c} \\ Attention(\mathbf{Q_2^{'}}, \mathbf{K_h}, \mathbf{V_h}) &= softmax(\frac{\mathbf{Q_2^{'}}\mathbf{K_h^{\top}}}{\sqrt{d_h}})\mathbf{V_h} \\ Attention(\mathbf{Q_3^{'}}, \mathbf{K_w}, \mathbf{V_w}) &= softmax(\frac{\mathbf{Q_3^{'}}\mathbf{K_w^{\top}}}{\sqrt{d_w}})\mathbf{V_w} \end{aligned} \tag{9}$$

where d_c, d_h, d_w is key dimension of $\mathbf{Q_1^{'}}$, $\mathbf{Q_2^{'}}$, $\mathbf{Q_3^{'}}$, respectively. The equation of the multi-head attention mechanism is the same as equation(6). Afterward, we apply layer normalization and FFN with a residual structure that is the same as equation(7).

D. Tucker Reconstruction Module

Finally, we obtain the three-factor matrices $\mathbf{C}^{'}$, $\mathbf{H}^{'}$, $\mathbf{W}^{'}$ that fully consider the degree of correlation between the components of the feature of different modalities and a core tensor which represents the main component of the fusion feature. By equation(10), we reconstruct the tensor $\mathbf{F}$, which is used for subsequent classification tasks.

$$((\mathbf{core} \times_1 \mathbf{C}^{'}) \times_2 \mathbf{H}^{'}) \times_3 \mathbf{W}^{'} = \mathbf{F} \tag{10}$$

The classifier includes average pooling, a series of fully connected layers, and dropout operations. The whole network can be trained by minimizing the overall loss between the predicted label and the ground truth label, which can be defined as:

$$arg\,\underset{\theta}{min} = \sum_{i=0}^{N} \mathcal{L}(I_1^i, I_2^i; \hat{y}^i, y^i) \tag{11}$$

where $\mathcal{L}$ calculates the cross-entropy loss of predicted label ŷ and real label y. N represents the number of training data, I_1 and I_2 represent the images from two modalities. θ is the network parameters.

IV. Experiments and Results

In this section, we will first describe the three datasets used to validate the proposed method. Then, we present the implementation details, results of comparison methods, results of the ablation study, and some related discussion.

A. Datasets

To validate the effectiveness of TDSFNet, we use three multi-modal medical image classification datasets: Pleural Effusion Cell(PEC) dataset, Age-related Macular Degeneration(AMD) dataset, and Seven-Point Checklist(SPC) dataset.

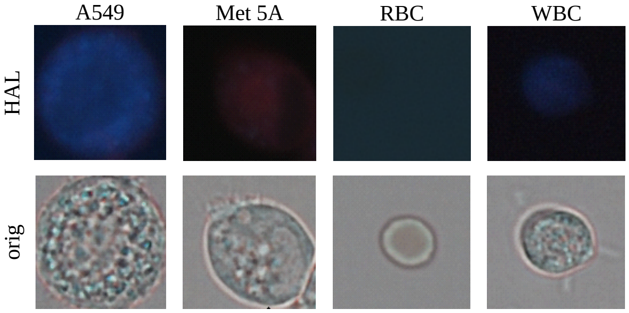


Fig. 3: Cell images stained with HAL reagent named HAL(first row) and original images of unstained cells named orig(second row) with specific classes.

1) Pleural Effusion Cell dataset: We collected the Pleural Effusion Cell dataset from Tianjin Medical University. The dataset contains two modalities of cell images, HAL stained and unstained original cell images. Each modal has four categories of cells: A549, Met 5A, WBC(White Blood Cell) and RBC(Red Blood Cell). Examples of HAL and original cell images are shown in Fig. 3. In this dataset, two paired cell images come from the same specific cell. Eventually, we obtained an expert-labeled multi-modal pleural effusion cell dataset of 400 HAL stained and 400 original cell images.

2) Age-related Macular Degeneration dataset: This dataset consists of 1094 CFP and 1289 OCT images of 829 subjects. Each pair of images is categorized into normal, dryAMD, PCV, or wetAMD. CFP is a common examination method in ophthalmology, which can directly and objectively present fundus lesions. However, CFP has some limitations in evaluating AMD, which needs to be supplemented by higher-resolution imaging techniques. OCT is three-dimensional sliced data that provides cross-sectional structural images of retinal tissue without damaging the body and can show the thickness change of the retinal nerve fiber layer. However, it cannot determine whether microaneurysms exist or not. Given the mentioned characteristics of OCT and CFP, which can reflect the retina from distinct aspects, combining two modalities provides a more reliable basis for the clinical diagnosis of AMD. More details can be referred to [16].

3) Seven-Point Checklist dataset: This dataset is a multi-classification task dataset and contains 1011 multi-modal skin cancer cases. Each case includes clinical and dermoscope images, labels of the seven-point criteria, and diagnosis labels. Clinical images usually present the lesion's structure, color, and shape but not the details. Dermoscope is essentially a skin microscope that can be magnified tens of times and is a powerful tool for observing pigmentary disorders of the skin. Compared with clinical images, it can more clearly describe the distribution of blood vessels and structure under the skin of the lesion. More details of the SPC dataset can be referred to [26].

TABLE I: Quantitative evaluation results of the state-of-the-art multi-modal classification methods on PEC dataset in terms of Precision, Sensitivity, F1-score, and Accuracy based on different ResNet backbones. The highest performance is highlighted in boldface.

	A549			Met 5A			WBC			RBC			overall	
Methods	PRE	SEN	F1	PRE	SEN	F1	PRE	SEN	F1	PRE	SEN	F1	F1	ACC
DeepGuided [27]	0.640	0.500	0.650	0.670	0.800	0.730	0.670	0.800	0.727	0.810	0.850	0.829	0.737	0.737
CAFNet-18 [9]	0.533	0.800	0.640	0.600	0.300	0.400	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.760	0.775
CAFNet-34 [9]	0.650	0.650	0.650	0.650	0.650	0.650	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.820	0.825
CAFNet-50 [9]	0.629	0.850	0.720	0.769	0.500	0.610	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.830	0.837
MIMO [28]	0.608	0.700	0.651	0.727	0.400	0.516	0.769	**1.000**	0.869	**1.000**	**1.000**	**1.000**	0.775	0.775
HcCNN-18 [11]	0.666	0.500	0.570	0.600	0.750	0.670	**1.000**	0.800	0.890	0.833	**1.000**	0.910	0.760	0.763
HcCNN-34 [11]	0.700	0.700	0.700	0.700	0.700	0.700	**1.000**	0.800	0.890	0.833	**1.000**	0.910	0.800	0.800
HcCNN-50 [11]	0.692	0.450	0.550	0.571	0.800	0.670	**1.000**	0.700	0.820	0.800	**1.000**	0.890	0.730	0.738
PT-DCN [29]	0.608	0.700	0.651	0.647	0.550	0.594	0.950	0.950	0.950	**1.000**	**1.000**	**1.000**	0.800	0.800
CentralNet-18 [8]	0.580	**0.900**	0.710	0.777	0.350	0.480	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.800	0.813
CentralNet-34 [8]	0.705	0.600	0.650	0.652	0.750	0.700	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.840	0.838
CentralNet-50 [8]	0.551	0.800	0.650	0.636	0.350	0.450	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.780	0.788
MMNet-18 [30]	0.592	0.800	0.680	0.692	0.450	0.550	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.810	0.812
MMNet-34 [30]	0.640	0.800	0.710	0.733	0.550	0.630	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.830	0.837
MMNet-50 [30]	0.666	0.800	0.730	0.750	0.600	0.670	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.850	0.850
TDSFNet-18	0.700	0.700	0.700	0.700	0.700	0.700	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.850	0.850
TDSFNet-34	0.714	0.750	0.730	0.737	0.700	0.720	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.860	0.863
TDSFNet-50	**0.857**	**0.900**	**0.880**	**0.894**	**0.850**	**0.870**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.937**	**0.937**

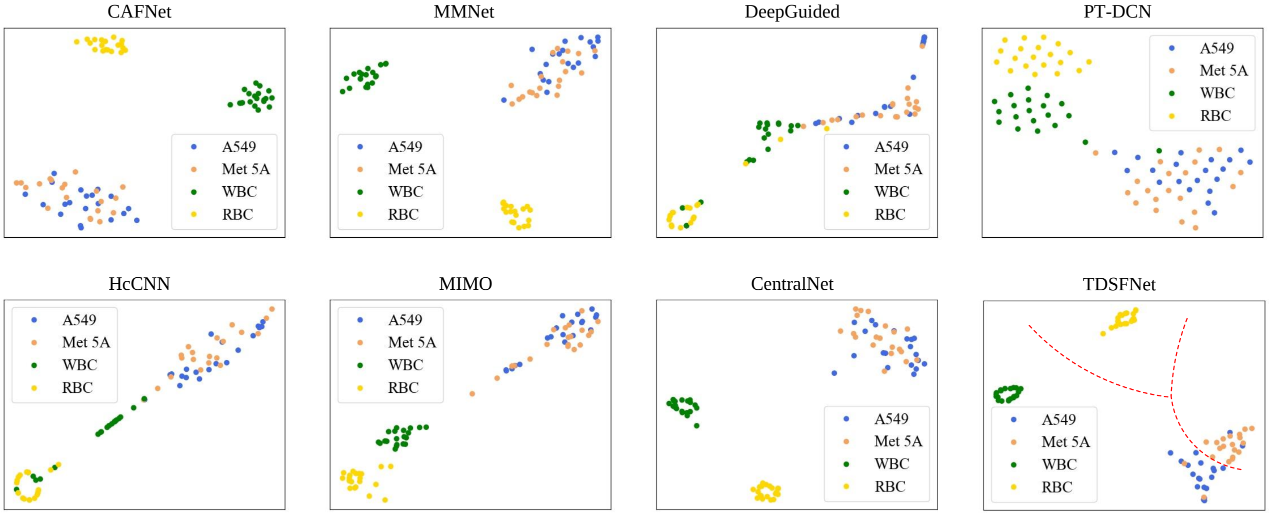


Fig. 4: The visualization distribution of testing set images on PEC dataset in TDSFNet and other different comparison experimental feature spaces.

B. Implementation Details

1) Pleural Effusion Cell dataset: In PEC, all the images are 120 × 120, the division ratio of the training set, validation set, and testing set is 6: 2: 2. For training, we perform common data augmentations.

2) Age-related Macular Degeneration dataset: In AMD, all the images are resized to 224 × 224. We randomly split the data set into the training, validation, and testing sets on eye identities. All the images from the same eye will only appear in one of the three sets. Common data augmentations, including random rotating, random vertical flipping, and random horizontal flipping, are used in the training set to avoid overfitting and enhance the model's generalization performance.

3) Seven-Point Checklist dataset: SPC is a public dataset with a divided training, validation, and testing set. The training, validation, and testing sets are 413, 203, and 395, respectively. The image sizes in this dataset are inconsistent, so all the images are resized to 224 × 224.

In each dataset, we use different ResNet backbones to evaluate our approach. In the Tucker decomposition module, we select the rank of each dimension based on the shape of the extracted feature. For lower dimensions of features, the rank we choose is consistent with the corresponding dimension. For higher dimensions of features, we choose the optimal rank through experiments. All the networks are trained using Adam optimizer, and the learning rate is set to 0.0001, and the batch

TABLE II: Results of methods with different comparison methods on PEC in terms of AUROC. The highest performance is highlighted in boldface.

Category	PT-DCN	MIMO	DeepGuided	CentralNet-50	MMNet-50	CAFNet-50	HcCNN-50	TDSFNet-50
A549	0.896	0.883	0.880	0.896	0.878	0.900	0.848	**0.954**
Met 5A	0.887	0.891	0.933	0.884	0.914	0.872	0.874	**0.952**
WBC	0.984	**1.000**	0.887	**1.000**	**1.000**	**1.000**	0.925	**1.000**
RBC	**1.000**	0.995	0.961	**1.000**	**1.000**	**1.000**	0.974	**1.000**
Mean	0.954	0.937	0.919	0.967	0.971	0.969	0.891	**0.985**

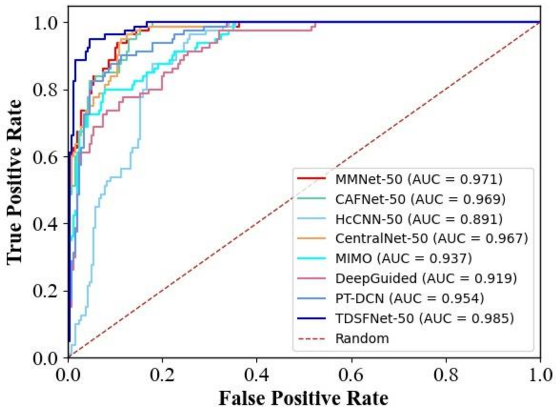


Fig. 5: ROC curves of TDSFNet and other comparison methods.

size is 16. We implemented our model with PyTorch.

C. Comparison Methods

1) Results on PEC: We first compare the performance of our model with the state-of-the-art multi-modal classification methods on PEC, including DeepGuided [27], CAFNet [9], MIMO [28], HcCNN [11], PT-DCN [29], CentralNet [8] and MMNet [30].

To quantitatively evaluate the classification performance, Precision, Sensibility, and F1-score are adopted as an evaluation metric in each category, and F1-score and Accuracy are adopted as overall evaluation. Experimental results are presented in Table I, where the best results are boldfaced. In CAFNet, HcCNN, CentralNet, MMNet and TDSFNet, ResNet is adopted as the backbone, so we conducted the experiments to verify the effect of different variants of ResNet. Table I shows that our model's overall F1-score and Accuracy exceed that of other comparison experiments. Other methods are directly fused at the feature level. Compared with these methods, we introduce the Tucker Decomposition Module in the feature processing process so that the network can extract fine-grained and more expressive feature subspace. Then, feature fusion with attention mechanism is performed on a lower-dimensional feature subspace. In Fig. 4, the visualization distribution of test images on PEC dataset in our model and other methods by t-SNE is presented. Our model can better distinguish A549 and Met 5A than other methods. In Fig. 5, we draw the ROC curves of several methods. Our model achieves the highest AUROC score of 0.985. More details are shown in Table II.

2) Results on AMD: In this section, we report the performance of our model on AMD. The results are summarized in Table III. We follow the evaluation criterion of [16]. We compare our results with the state-of-the-art methods including CAFNet [9], CentralNet [8], MMNet [30], HcCNN [11] and MM-CNN-da [16]. The best results are shown in bold. Due to the large amount of data in the category of ”normal” and the large difference from other categories, so almost all methods can achieve the best performance in the category ”normal.” For ”dryAMD”, this category's data is small. MM-CNN-da uses the data augmentation method to expand the amount of data in the training set so the specificity of ”dryAMD” reaches the highest. For F1-score metrics in each category, our method is optimal. In the fourteen metrics, TDSFNet achieves nine best values. Our model's overall F1-score and Accuracy arrive at 0.942 and 0.930, respectively. Compared with other methods, our model shows its priority.

3) Results on SPC: We compared the performance of TDSFNet with that of the state-of-the-art multi-modal classification methods. The results of accuracy are shown in Tabel IV. The best results are shown in bold. Among all the comparison methods, our method TDSFNet-50 achieves the highest average accuracy of 74.7% and achieves the highest accuracy on the other four classification criteria. On the SPC dataset, we observe a phenomenon that not all methods work best on models with deeper network layers. For example, the performance of CAFNet-18, CentralNet-18, TripleNet-18, HcCNN-18, and EmbeddingNet-18 is better than the corresponding model with deeper network layers. This shows that not all datasets are suitable for complex networks. In all comparative experiments, the performance of EmbeddingNet and TDSFNet is higher than other methods. The common point of the two methods is that there is no feature interaction in the middle stage of feature extraction, and only fusion is performed in the last layer of the network. Therefore, the interaction of intermediate layer features will introduce some interference to the final fusion performance.

4) Computational Efficiency Analysis: Computational efficiency is a factor that should be considered. We compare the parameters and the corresponding FLOPs on the PEC dataset. The results are summarized in Tabel V. We only compare the methods using the same backbone for fairness, including CAFNet, CentralNet, MMNet, HcCNN, and TDSFNet. Taking

TABLE III: Quantitative evaluation results of the state-of-the-art multi-modal classification methods on AMD in terms of Sensitivity, Specificity, F1-score, and Accuracy based on different ResNet backbones. The highest performance is highlighted in boldface.

	normal			dryAMD			PCV			wetAMD			overall	
Methods	SEN	SPE	F1	SEN	SPE	F1	SEN	SPE	F1	SEN	SPE	F1	F1	ACC
CAFNet-18 [9]	**1.000**	**1.000**	**1.000**	0.833	0.983	0.770	0.881	0.904	0.830	0.832	0.928	0.870	0.867	0.877
CAFNet-34 [9]	**1.000**	**1.000**	**1.000**	0.833	0.996	0.870	**0.955**	0.876	0.840	0.824	**0.976**	0.890	0.899	0.893
CAFNet-50 [9]	**1.000**	**1.000**	**1.000**	0.833	0.991	0.830	0.746	0.972	0.820	0.950	0.856	0.900	0.889	0.898
CentralNet-18 [8]	**1.000**	**1.000**	**1.000**	0.833	**1.000**	0.910	0.657	**0.977**	0.770	**0.966**	0.800	0.890	0.891	0.881
CentralNet-34 [8]	**1.000**	**1.000**	**1.000**	0.750	0.983	0.720	0.806	0.972	0.860	0.924	0.872	0.900	0.869	0.898
CentralNet-50 [8]	**1.000**	**1.000**	**1.000**	0.833	0.983	0.770	0.910	0.932	0.870	0.882	0.952	0.910	0.888	0.910
MMNet-18 [30]	**1.000**	**1.000**	**1.000**	**1.000**	0.938	0.740	0.909	0.888	0.820	0.724	0.950	0.820	0.845	0.836
MMNet-34 [30]	**1.000**	**1.000**	**1.000**	0.750	0.996	0.820	0.672	0.972	0.770	0.950	0.800	0.880	0.867	0.873
MMNet-50 [30]	**1.000**	**1.000**	**1.000**	**1.000**	0.983	0.860	0.925	0.893	0.840	0.807	0.960	0.870	0.892	0.885
HcCNN [11]	0.978	0.985	0.960	0.250	0.845	0.120	0.373	0.944	0.490	0.765	0.752	0.760	0.580	0.672
MM-CNN-da [16]	**1.000**	**1.000**	**1.000**	0.868	**1.000**	0.929	0.794	0.948	0.864	0.868	0.860	0.864	0.914	0.863
TDSFNet-18	**1.000**	**1.000**	**1.000**	0.833	0.982	0.770	0.835	0.960	0.860	0.907	0.896	0.900	0.900	0.901
TDSFNet-34	**1.000**	**1.000**	**1.000**	0.750	0.995	0.820	0.910	0.937	**0.880**	0.907	0.936	0.920	0.900	0.918
TDSFNet-50	**1.000**	**1.000**	**1.000**	**1.000**	0.996	**0.960**	0.866	0.960	**0.880**	0.932	0.928	**0.930**	**0.942**	**0.930**

TABLE IV: Quantitative evaluation results of the state-of-the-art multi-modal classification methods on SPC dataset in terms of Accuracy based on different ResNet backbones. The highest performance is highlighted in boldface.

Methods	PN	BWV	VS	PIG	STR	DaG	RS	DIAG	Average
TripleNet-18 [31]	55.7	85.0	78.7	61.6	65.6	53.1	73.5	64.7	67.2
TripleNet-34 [31]	51.6	84.4	78.2	62.0	66.9	53.1	74.1	64.7	66.9
TripleNet-50 [31]	51.9	84.1	79.7	61.0	66.5	50.0	72.6	63.6	66.2
HcCNN-18 [11]	62.0	87.5	78.0	66.6	72.7	57.6	76.0	69.4	71.2
HcCNN-34 [11]	62.2	84.7	79.3	67.1	70.2	54.6	73.7	69.9	70.2
HcCNN-50 [11]	58.2	86.2	79.3	68.1	69.6	56.9	73.7	68.1	70.0
FusionM4Net-FS [32]	64.5	87.3	80.2	65.6	71.3	57.5	78.0	71.2	72.0
CAFNet-18 [9]	64.7	86.7	80.7	68.0	69.6	58.7	77.2	70.3	71.9
CAFNet-34 [9]	65.4	86.5	80.9	67.1	68.2	59.7	77.9	70.2	72.0
CAFNet-50 [9]	64.5	84.9	79.7	66.9	67.2	56.6	79.1	68.9	71.0
CentralNet-18 [8]	65.6	86.5	81.5	64.2	70.5	57.2	77.3	70.2	71.6
CentralNet-34 [8]	64.3	85.8	80.3	62.7	71.3	58.2	76.0	69.3	71.0
CentralNet-50 [8]	64.2	83.8	80.6	64.1	71.3	56.5	79.8	72.3	71.6
EmbeddingNet-18 [33]	66.6	**88.5**	**82.6**	66.9	74.5	**65.5**	78.8	73.0	74.5
EmbeddingNet-34 [33]	67.8	85.7	80.1	69.9	73.2	58.4	81.6	72.0	73.6
EmbeddingNet-50 [33]	67.1	87.5	79.8	69.1	71.2	61.5	79.8	71.4	73.4
TDSFNet-18	65.8	87.5	81.9	68.6	74.0	60.7	**81.9**	71.4	73.9
TDSFNet-34	68.6	87.7	81.9	69.4	72.7	62.0	80.9	73.0	74.5
TDSFNet-50	**69.1**	87.5	81.6	**70.0**	**75.5**	61.2	79.3	**73.2**	**74.7**

TABLE V: Parameters(M) and FLOPs(G) of the compared methods with the image size of 120 x 120, where TDSFNet achieves a relatively low FLOPs.

	CAFNet-50	CentralNet-50	MMNet-50	HcCNN-50	TDSFNet-50
Params(M)	106.20	171.53	96.13	231.72	143.12
FLOPs(G)	16.42	12.79	10.18	26.38	10.97

the 120 × 120 input image as an example, the FLOPs(floating point operations) for TDSFNet is 10.97G. In all methods using the same backbone, the FLOPs of TDSFNet are only a little higher than MMNet. However, the performance of TDSFNet is optimal.

D. Ablation Study

1) Effectiveness of TDSFNet: To investigate the proposed TDSFNet's effectiveness, we conducted ablation experiments on the PEC dataset. The results are shown in Table VI. Precision, Sensibility, and F1-score are adopted as an evaluation metric in each category, and F1-score and Accuracy in the overall evaluation to quantitatively evaluate the classification performance.

First, we compared the performance of TDSFNet with that of the single-modality methods. All the compared methods took ResNet-50 as the backbone. Our single-modal baselines are HAL-CNN and orig-CNN. To verify whether the Tensor Decomposition(TD) Module can extract useful information and more expressive features, we introduce TDM in single-modal baselines named HAL-CNN-TD and orig-CNN-TD. From Table VI, we can see that HAL-CNN-TD improves the Accuracy by 3.8% against the HAL-CNN and orig-CNN-TD improves the Accuracy by 5% against the orig-CNN. This suggests that TDM can better extract representative features of specific modalities.

As for multi-modal methods, we consider the following three methods, LateFusion, TDSFNet without Subspace Fu-

TABLE VI: Ablation studies of TDSFNet on PEC dataset in terms of Precision, Sensitivity, F1-score and Accuracy. The highest performance is highlighted in boldface.

	A549			Met 5A			WBC			RBC			overall	
Methods	PRE	SEN	F1	PRE	SEN	F1	PRE	SEN	F1	PRE	SEN	F1	F1	ACC
HAL-CNN	0.550	0.550	0.550	0.500	0.400	0.440	0.818	0.900	0.860	0.909	**1.000**	0.950	0.700	0.712
orig-CNN	0.576	0.750	0.650	0.692	0.450	0.550	0.952	**1.000**	0.980	**1.000**	**1.000**	**1.000**	0.793	0.800
HAL-CNN-TD	0.666	0.800	0.730	0.769	0.500	0.610	0.823	0.700	0.760	0.769	**1.000**	0.870	0.739	0.750
orig-CNN-TD	0.653	0.850	0.740	0.800	0.600	0.690	**1.000**	0.950	0.970	**1.000**	**1.000**	**1.000**	0.849	0.850
LateFusion	0.722	0.650	0.680	0.681	0.750	0.710	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.849	0.850
TDSFNet (w/o SFM)	0.785	0.550	0.650	0.680	**0.850**	0.760	0.952	**1.000**	0.980	**1.000**	**1.000**	**1.000**	0.844	0.845
TDSFNet (w/o CDM)	0.708	0.850	0.770	0.812	0.650	0.720	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.873	0.875
TDSFNet	**0.857**	**0.900**	**0.880**	**0.894**	**0.850**	**0.870**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.937**	**0.937**

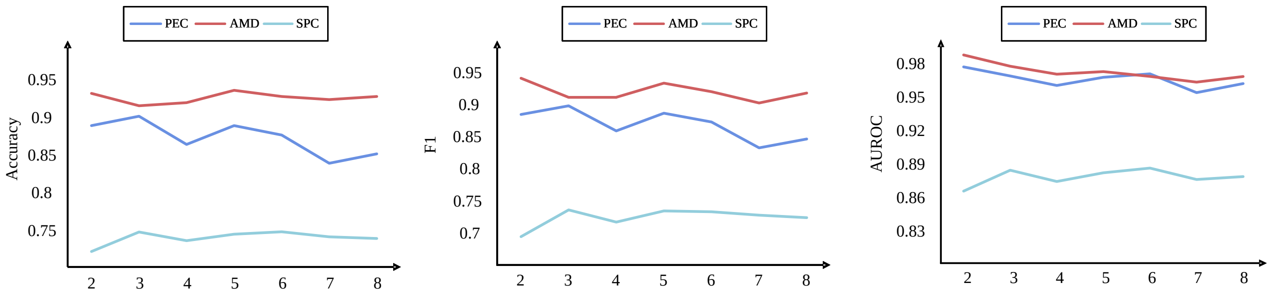


Fig. 6: Performance of TDSFNet by varying hyper-parameter on PEC and AMD dataset. The integer x represents the number of transformer cascade layers.

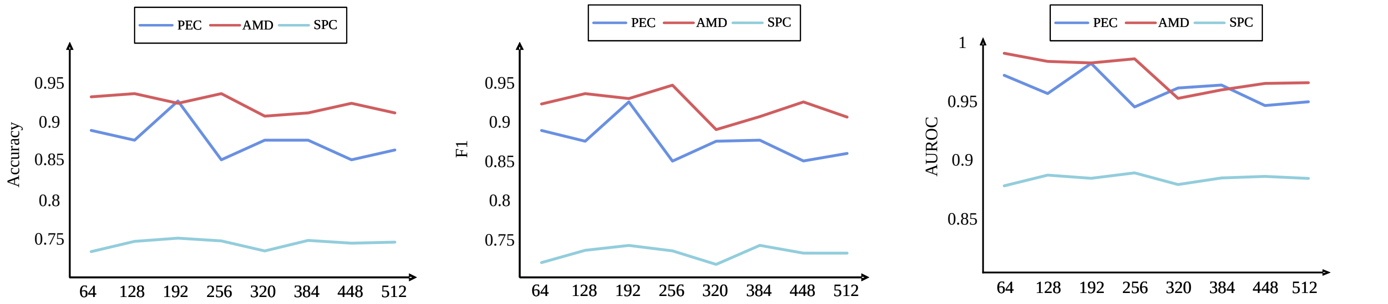


Fig. 7: Performance of TDSFNet by varying hyper-parameter on PEC and AMD dataset. The integer x represents the rank of the channel dimension.

sion Module(SFM), and TDSFNet without Core-tensor Decomposition Module(CDM). LateFusion concatenates the last layer of HAL-CNN and orig-CNN, TDSFNet without SFM adds the feature subspace obtained by TDM of two modalities and then directly reconstructs the final feature for classification. It can be seen from Table VI that the features of multiple modalities that are randomly concatenated or added do not yield better performance. We also compare the results of TDSFNet and TDSFNet without SFM in terms of overall F1-score and Accuracy, TDSFNet with the SFM improves the performance by 9.3% and 9.2%, respectively, and significant improvement can also be observed in the indicators of A549. In experiments without CDM, we removed the CDM and only considered the correlation among the three dimensions. The overall Accuracy is improved by 3% compared with TDSFNet without SFM. To validate the effectiveness of the CDM, we compared the TDSFNet and TDSFNet without CDM. TDSFNet performed best among all the evaluation metrics, especially for the two complex categories of A549 and Met 5A.

2) Impact of Transformer Layers: Transformer is an essential component of the Subspace Fusion Module, and the number of its cascades has a vital influence on the fusion performance. We compared the experimental results under the different number of transformer cascade layers. The results are shown in Fig. 6. As we can see, we experimented with layers from 2 to 8 on each dataset and showed the results of Accuracy, F1-score, and AUROC. The experimental results indicate that the optimal number of layers is 3 in the PEC dataset, 5 in AMD, and 6 in the SPC dataset.

3) Impact of Rank on Channel dimension: In the Tensor Decomposition Module, the most crucial hyper-parameter is the rank of the channel dimension. The results are shown in Fig. 7. We still report Accuracy, F1-score, and AUROC. We adjust the rank from 64 to 512 at an interval of 64. In the PEC dataset, we set the rank to 192, achieving the best performance. In AMD, the optimal rank is 256. In the SPC dataset, the optimal rank is 384.

V. Conclusion

This paper proposed a novel method for multi-modal medical image classification. Our proposed TDSFNet introduces a Tucker Decomposition Module to extract fine-grained features and reduce redundancy caused by high-dimensional features. In addition, we fuse different modalities at the subspace level rather than at the feature level and introduce a cross-attention mechanism into the fusion strategy. We also collected the multi-modal Pleural Effusion Cell dataset containing two cell image modalities. Our model achieves state-of-the-art results on PEC, AMD, and SPC datasets.

References

[1] W. Hu, X. Meng, Y. Bai, A. Zhang, G. Qu, B. Cai, G. Zhang, T. W. Wilson, J. M. Stephen, V. D. Calhoun *et al.*, "Interpretable multimodal fusion networks reveal mechanisms of brain cognition," *IEEE transactions on medical imaging*, vol. 40, no. 5, pp. 1474–1483, 2021. 1, 2

[2] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *2017 20th international conference on information fusion (Fusion)*. IEEE, 2017, pp. 1–7. 1

[3] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019. 1

[4] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical image analysis*, vol. 43, pp. 98–111, 2018. 1

[5] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review*, vol. 33, pp. 1–39, 2010. 1

[6] Y. Guo, W. He, and C. Gao, "Human activity recognition by fusing multiple sensor nodes in the wearable sensor systems," *Journal of Mechanics in Medicine and Biology*, vol. 12, no. 05, p. 1250084, 2012. 1

[7] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-net: hybrid-fusion network for multi-modal mr image synthesis," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2772–2781, 2020. 1

[8] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0. 1, 6, 7, 8

[9] X. He, Y. Wang, S. Zhao, and X. Chen, "Co-attention fusion network for multimodal skin cancer diagnosis," *Pattern Recognition*, vol. 133, p. 108990, 2023. 1, 2, 6, 7, 8

[10] C. He, K. Li, Y. Zhang, G. Xu, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping," *arXiv preprint arXiv:2305.11003*, 2023. 1

[11] L. Bi, D. D. Feng, M. Fulham, and J. Kim, "Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network," *Pattern Recognition*, vol. 107, p. 107502, 2020. 1, 6, 7, 8

[12] G. Muhammad, F. Alshehri, F. Karray, A. El Saddik, M. Alsulaiman, and T. H. Falk, "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems," *Information Fusion*, vol. 76, pp. 355–375, 2021. 2

[13] G. Xu, C. He, H. Wang, H. Zhu, and W. Ding, "Dm-fusion: Deep model-driven network for heterogeneous image fusion," *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[14] S. Ye, T. Wang, M. Ding, and X. Zhang, "F-darts: Foveated differentiable architecture search based multimodal medical image fusion," *IEEE Transactions on Medical Imaging*, 2023. 2

[15] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323–336, 2021. 2

[16] W. Wang, X. Li, Z. Xu, W. Yu, J. Zhao, D. Ding, and Y. Chen, "Learning two-stream cnn for multi-modal age-related macular degeneration categorization," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4111–4122, 2022. 2, 5, 7, 8

[17] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 046–22 055. 2

[18] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2612–2620. 3

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 3

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 3

[21] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558. 3

[22] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4467–4480, 2019. 3

[23] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *arXiv preprint arXiv:2206.06488*, 2022. 3

[24] Q. Zhu, H. Wang, B. Xu, Z. Zhang, W. Shao, and D. Zhang, "Multimodal triplet attention network for brain disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3884–3894, 2022. 3

[25] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2441–2449. 3

[26] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018. 5

[27] M. Mallya and G. Hamarneh, "Deep multimodal guidance for medical image classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*. Springer, 2022, pp. 298–308. 6, 7

[28] L. Yang, D. Mehta, D. Mahapatra, and Z. Ge, "Leukocyte classification using multimodal architecture enhanced by knowledge distillation," in *Medical Optical Imaging and Virtual Microscopy Image Analysis: First International Workshop, MOVI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*. Springer, 2022, pp. 63–72. 6, 7

[29] X. Gao, F. Shi, D. Shen, and M. Liu, "Task-induced pyramid and attention gan for multimodal brain image imputation and classification in alzheimer's disease," *IEEE journal of biomedical and health informatics*, vol. 26, no. 1, pp. 36–43, 2021. 6, 7

[30] Y. Dai, Y. Gao, F. Liu, and J. Fu, "Mutual attention-based hybrid dimensional network for multimodal imaging computer-aided diagnosis," *arXiv preprint arXiv:2201.09421*, 2022. 6, 7, 8

[31] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer, 2017, pp. 250–258. 8

[32] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, and T. Lasser, "Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification," *Medical Image Analysis*, vol. 76, p. 102307, 2022. 8

[33] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Experimental dermatology*, vol. 27, no. 11, pp. 1261–1267, 2018. 8