The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

MVCINN: Multi-View Diabetic Retinopathy Detection Using a Deep Cross-Interaction Neural Network

Xiaoling Luo[1, *], **Chengliang Liu**[1, *], **Waikeung Wong**[2, 3], **Jie Wen**[1, †], **Xiaopeng Jin**[4], **Yong Xu**[1, †]

[1]Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China
[2]School of Fashion and Textiles, The Hong Kong Polytechnic University, Kowloon, Hong Kong
[3]Laboratory for Artificial Intelligence in Design, Hong Kong
[4]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China
xiaolingluoo@outlook.com, liucl1996@163.com, calvin.wong@polyu.edu.hk, jiewen_pr@126.com,
jinxiaopengit@gmail.com, laterfall@hit.edu.cn

Abstract

Diabetic retinopathy (DR) is the main cause of irreversible blindness for working-age adults. The previous models for DR detection have difficulties in clinical application. The main reason is that most of the previous methods only use single-view data, and the single field of view (FOV) only accounts for about 13% of the FOV of the retina, resulting in the loss of most lesion features. To alleviate this problem, we propose a multi-view model for DR detection, which takes full advantage of multi-view images covering almost all of the retinal field. To be specific, we design a cross-interaction self-attention based module (CISAM) that interfuses local features extracted from convolutional blocks with long-range global features learned from transformer blocks. Furthermore, considering the pathological association in different views, we use the feature jigsaw to assemble and learn the features of multiple views. Extensive experiments on the latest public multi-view MFIDDR dataset with 34,452 images demonstrate the superiority of our method, which performs favorably against state-of-the-art models. To the best of our knowledge, this work is the first study on the public large-scale multi-view fundus images dataset for DR detection.

Introduction

The IDF Diabetes Atlas 10th edition (Federation 2021) reported that the destructive effects of diabetes are set to continue as a result of the predicted increase in prevalence from 537 million in 2021 to 783 million in 2045. Diabetic retinopathy (DR) is one of the most common complications of diabetes and is a leading cause of preventable blindness in the working-age population. According to the international clinical diabetic retinopathy disease severity scales (Ophthalmoscopy and Levels 2002), the severity of DR can be graded into 0-4 stages: normal, mild, moderate, severe, and proliferative DR (PDR). Over the past decade, computer vision and deep learning-based algorithms have largely contributed to the research in medical image processing. With the successful development of neural network (Huang et al. 2022; Hu, Shi, and Ye 2022; Huang et al. 2021), frameworks such as image classification have been used to predict and analyze the progression of DR (Wang et al. 2017; Luo et al. 2022). However, most previous works are trained on the single-view databases (e.g., MESSIDOR (Decenciere et al. 2014), EyePACS (EyePACS 2015)), which face a risk of losing a large proportion of the lesion features on the retina. The reason is that the single-view data typically has a FOV of only about $45°$-$50°$. Most of the retinal regions are missing in the fundus images used for training, resulting in the poor performance of these single-view models in clinical applications. Furthermore, clinical medical studies (Hu et al. 2019) have also shown that $45°$ single-view Non-mydriatic fundus photography does not meet the technical requirements of DR screening, because of the inability to detect microaneurysms and retinopathy outside a single $45°$ image. Some studies have shown that multi-view fundus imaging has better performance in DR detection (Srihatrai and Hlowchitsieng 2018). The extensive exploration of various multi-view methods (Liu et al. 2022a; Wang et al. 2021; Hu, Lou, and Ye 2022; Wen et al. 2022) indicates that multi-view method has a great performance in solving the problem of data defects. Therefore, we propose a multi-view DR detection model that can simultaneously extract features from four $45°$ views of the same eye, i.e., view $V1$-$V4$: the field

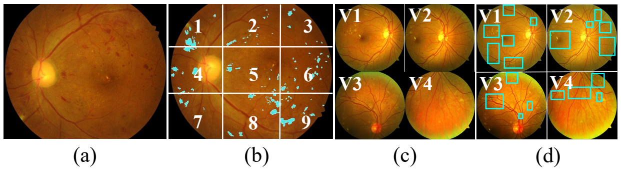


Figure 1: (a) Single-view fundus image. (b) A schematic diagram of long-range global features of the same lesion in different image patches, showing the lesions of hemorrhages. (c) Multi-view fundus images. (d) The related area of the same lesion in different retinal views from one eye, where hemorrhages are marked in the blue boxes.

[*]These authors contributed equally.

[†]Co-corresponding authors: Jie Wen, Yong Xu.

centered on the macula ($V1$), the field centered on the optic disc ($V2$), and the fields tangent to the upper and lower horizontal lines of the optic disc respectively ($V3$ and $V4$). By integrating the features of multi-view fundus images, the deficiency of the single-view method due to information loss is remedied, and the performance of the model in clinical application is improved.

Moreover, with extensive applications of deep learning in DR detection, the methods based on CNN can successfully extract local features of fundus images (Luo et al. 2023). The convolution operation can extract features by constantly learning features in the operation block with the size determined by the convolution kernel, but it does not consider the relationship between different operation blocks (Zhu et al. 2022; Wang et al. 2018). Recently, transformer neural networks (Liu et al. 2023; Vaswani et al. 2017) have been introduced into the field of image processing. In the vision transformer (ViT) algorithm (Dosovitskiy et al. 2021), the whole image is divided into small image blocks with location information. Then the linear embedding sequence of these small image patches is fed into the network as the input of the Transformer to obtain the long-range information between different image patches. A transformer neural network can easily learn global features of images, but it lacks the inductive bias of translation equivariance and locality compared with CNN. In DR detection tasks, not only the feature of local patches can provide important information, but the non-local connections between scattered lesion patches (such as microaneurysms, hemorrhages, and exudations) on fundus images are also useful. As shown in Fig. 1, although the hemorrhages are scattered in different areas on the retinal image, there is a long-distance dependence between the lesions because they belong to the same category. Therefore, the extraction of long-distance global features of fundus images is also important for DR detection tasks.

Inspired by (Peng et al. 2021), we propose a multi-view DR detection model that can cross-fuse local and global features, including convolutional blocks named Conv-Blocks and transformer blocks named Trans-Blocks for feature interactive learning. Conv-Blocks take advantage of inductive bias in images to obtain key local features of images. The Trans-Blocks retrieve long-range global information through the self-attention mechanism. For the interaction fusion of local features and global features, the cross-interaction self-attention based module (CISAM) is designed. In CISAM, the dimensions of features generated in different branches are adjusted through the Conv-Fusion function and Trans-Fusion function, so that features of Conv-Blocks and Trans-Blocks can be integrated with each other to improve the ability of feature expression. In addition, since our model was trained based on multiple views, there is also long-distance dependence between lesions from different views of the same eye. Thus, we further use the non-local block (NLB) (Wang et al. 2018) and Multilayer Perceptron (MLP) (Dosovitskiy et al. 2021) mechanism to learn long-range information between different views after concatenating the feature maps of multiple views. Compared to the previous methods, this work has the following advantages and contributions:

- We propose a novel multi-view model for DR detection. Compared with the traditional single-view DR detection methods, our multi-view method is more suitable for practical clinical application and can learn more complete features from multi-view retinal images.
- Our network combines both CNN-based and transformer-based learning mechanisms. In the process of feature learning, local and global information is integrated by CISAM to enhance representation learning.
- Our proposed model takes advantage of long (global) dependence between the multiple views from the same eyes to improve the performance of the multi-view model.
- To our knowledge, this work is the first study on the public large-scale multi-view fundus images dataset for DR detection by integrated learning multi-view real fundus image data to improve the performance of the automated DR detection model in clinical applications.

Related Work

CNN–Based Approaches of DR Detection

In order to improve the performance of the CNN model in the DR screening task, Li et al. (Li et al. 2017) pre-trained the model in a large-scale natural image dataset, and then carried out transfer learning in a specified retinal image dataset. On the basis of transfer learning, Wang et al. (Wang et al. 2017) proposed a Zoom-in-net network based on CNN to improve DR diagnosis by highlighting suspicious areas. Zoom-in-net simulated the process of clinicians' examination of fundus images and enlarged the suspicious areas. Pao et al. (Pao et al. 2020) proposed a bichannel CNN that fuses entropy image grayscale and green component features to improve the performance of DR detection through deep learning. However, these DR detection methods are based on single-view fundus images that usually only have a $45°$-FOV view centered on macula without taking the features of multi-view fundus images from the same subject into account. Thus, we propose a novel multi-view DR detection method, which can integrate the features from multiple views to make up for the information loss of the single-view method, by learning the retinal images that contain the most important area for diagnosis including the posterior polar region and some peripheral fundus.

Transformer–Based Approaches of DR Detection

To make up for the deficiency of the CNN-based method in capturing long-distance features, some methods (Dosovitskiy et al. 2021; Wu et al. 2021; Sun et al. 2021) based on transformer have been proposed. Wu et al. (Wu et al. 2021) proposed a DR grading model based on ViT, embedding fundus images into a sequence of image patches with location information. Sun et al. (Sun et al. 2021) proposed a novel lesion-aware transformer (LAT) model for DR prediction and lesion discovery. Kamran et al. (Amit Kamran et al. 2021) presented a ViT-based generative adversarial network, which could generate retinal vascular images from normal or diseased fundus images and simultaneously achieve retinal disease prediction. Although these models can easily obtain long-range information, they have no natural ability to perceive local features. In order to better combine the long-distance dependent features and local features, we propose a cross-Interaction network combining the advantages of CNN and Transformer. Specifically, a cross-interaction self-attention based module (CISAM) is used to transfer and integrate the feature maps generated from the Conv-Blocks and Trans-Blocks.

Method

Overview

In the clinical diagnosis of DR, ophthalmologists usually find the location, size, and number of lesions by observing fundus images taken from multiple directions of the same eye, and determine the DR grade according to the observation results. Our proposed model can simultaneously perform a four-view analysis of fundus images, in turn, can fully emulate the ophthalmologists' reading practice. Since fundus images of the four views are similar, the four-view images are first input into the Sharing Unit for feature extraction to reduce model parameters. In this training process, the images of the four views will share network parameters. Meanwhile, to obtain the precision feature of each view, after rough feature extraction of Sharing Unit, the network is split into four subnets to learn features of the four views respectively.

It is doubtless that humans would do context reasoning, i.e., to judge an ambiguous region by observing not only the local information around the region but also the associated global information. This inspires us to focus on both local features and long-range global features of fundus images when extracting features. In our proposed network, Conv-Blocks and Trans-Blocks can be used to extract local features and long-distance dependent features respectively, and the two kinds of extracted features can complement each other. Further, CISAM is designed for the fusion of features generated by Conv-Blocks and Trans-Blocks, which carries out dimension reconstruction and feature map addition.

To realize the feature fusion of multiple views, we propose Muti-View Fusion Unit. Since each view is fed into Conv-Blocks and Trans-Block for learning, each view obtains a CNN feature map and a transformer embedded sequence. We splice the feature maps and embedded sequences of the four views into new feature maps and sequences respectively and input them into the corresponding special classifier. Finally, the prediction scores of the two classifiers is fused to obtain the final prediction result. An overview of the proposed model is depicted in Fig. 2.

Sharing Unit

The details of Sharing Unit are shown in Fig. 2. We define the input multi-view data as $V = [V_1, V_2, ..., V_i, ..., V_N]$, where N is the number of views. Assuming that the input image is $V_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, feature extraction of V_i is carried out through the convolutional branch and transformer branch. V_i goes through 8 Conv-Blocks in convolutional branch to obtain ${I_i}^{c'} \in \mathbb{R}^{C' \times H' \times W'}$. Meanwhile, after V_i is input into transformer branch, an embedded patch $E_i \in \mathbb{R}^{L_i \times D_i}$ can be obtained through flatten function first, and then E_i is sent to extract features by 8 Trans-Blocks and output feature ${I_i}^{t'} \in \mathbb{R}^{L' \times D'}$. Among them, the features calculated by Conv-Block and Trans-Block conduct information interaction and fusion through CIASM. The internal structure of Conv-Block and Trans-Block can be seen in Fig 3. In addition, in order to ensure the specificity of the feature captured from each view, after Sharing Unit, the network is split into N branches to learn the features of N views.

Algorithm 1: The training process of MVCINN

Input: Multi-view fundus images $V = [V_1, V_2, ..., V_i, ..., V_N]$ and corresponding labels y.
Parameters: Hypeparameters learning rate and γ, training epochs Epo, embedding dimension D_a.
Initialization: Randomly initialize the network weights.
Output: A trained model.

```
for $k$=1 to $Epo$ do
    Preprocess the input image to the size of 224×224.
    Extract feature map of 3-channel images by Sharing Unit to obtain ${I_i}^{t'}$ and ${I_i}^{t'}$.
    Input ${I_i}^{t'}$ and ${I_i}^{t'}$ into $N$ blocks to learn the features of $N$ views.
    Fuse $N$ view features in MVFU.
    Compute the final prediction scores $P_f = P_c \oplus P_t$ to obtain the DR-grade results.
    Compute focal loss and update gradient.
end for
```

Cross-Interaction Self-Attention Based Module

As shown in Fig. 3, cross-interaction self-attention based module (CISAM), as the main structure to realize the interaction and fusion of local features and long-distance global features, contains two important network partitions, namely Conv-Block and Trans-Block. Inspired by (Peng et al. 2021), the intersectional network design ensures that either Conv-Block or Trans-Block in CISAM can learn features generated from the last Conv-Block and Trans-Block simultaneously. Let $I_i^{c_o} \in \mathbb{R}^{C_o \times H_o \times W_o}$ and $I_i^{t_o} \in \mathbb{R}^{(L_o+1) \times D_o}$ represent the feature maps and embedded patches that are input to CISAM respectively. Firstly, in Conv-Block, the input features $I_i^{c_o}$ and $I_i^{t_o}$ are fused by Conv-Fusion box. In Conv-Fusion box, the input embedded patches $I_i^{t_o}$ can adjust the feature dimension to $C_o \times H_o \times W_o$ by up-sampling, and then add with the input feature maps $I_i^{c_o}$ to obtain the feature map X_i^c. Through convolutional calculation $\mathcal{F}_{conv}(\cdot)$, which contains three convolutional layers with convolution kernel size of 1×1, 3×3, 1×1 respectively, the feature $\mathcal{F}_{conv}(X_i^c)$ is obtained. In order to improve network performance and avoid network degradation caused by the high complexity of the deep network, residual learning of ResNet (He et al. 2016) is used as reference. Thus, the final output of Conv-Block is defined as $\hat{I_i^c} \in \mathbb{R}^{\hat{C} \times \hat{H} \times \hat{W}}$:

$$\hat{I_i^c} = \mathcal{F}_{conv}(X_i^c) \oplus X_i^c, \tag{1}$$

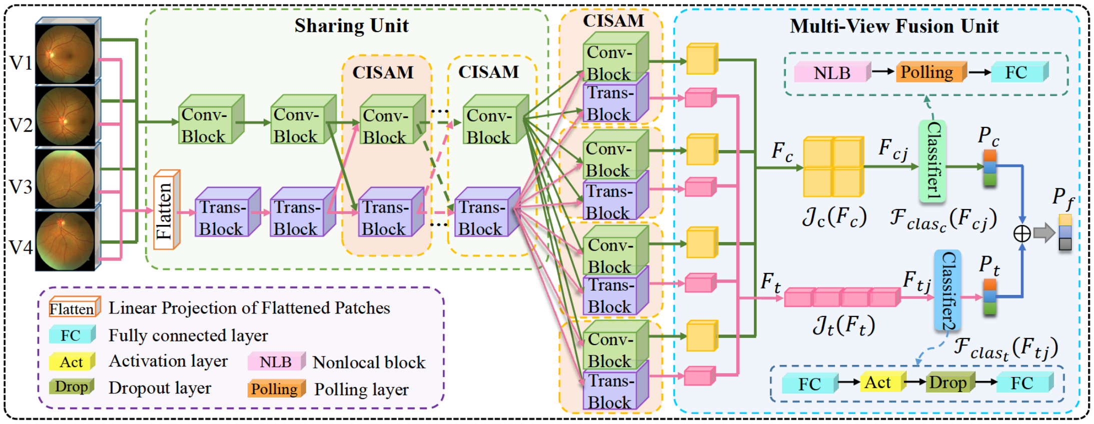


Figure 2: The flow diagram of our proposed MVCINN.

where the ⊕ operation is performed by a residual connection of matrix addition.

For Trans-Block, similar to Conv-Block, it receives features $I_i^{c_o}$ and $I_i^{t_o}$ at the same time and conducts feature fusion through the Trans-Fusion box. In the Trans-Fusion box, $I_i^{c_o}$ is firstly expanded into embedded patches with dimension $L_o \times D_o$, and then the classification token of $I_i^{t_o}$ is combined with it into feature X_i^t with dimension $(L_o + 1) \times D_o$. The fundus image of DR is scattered with a variety of lesions that have a long-distance relationship. We use Multi-Feature Self-Attention Class (MFSAC) to extract multiple global correlation features, so as to capture more accurate features of lesions.

In MFSAC (as shown in Fig. 4), adopting the mechanism of multi-layer perceptron (MLP) (Dosovitskiy et al. 2021), the input embedded patches X_a is initially and randomly divided into multiple heads $X_b \in \mathbb{R}^{L_a \times D_b \times H}$, $X_b = [X_{b1}, X_{b2}, \dots, X_{bm}, \dots, X_{bH}]$ to learn multiple features. The heads' number H can be regarded as the number of feature groups:

$$H = D_a / D_b. \tag{2}$$

The three generators $\mathcal{Q}(\cdot)$, $\mathcal{K}(\cdot)$, and $\mathcal{V}(\cdot)$ are employed to convert X_b to query $\mathcal{Q}(X_b)$, key $\mathcal{K}(X_b)$, and value $\mathcal{V}(X_b)$, respectively. We consider the operations of three generators which are defined as:

$$\mathcal{Q}(X_b) = X_b \cdot w_Q,\ \mathcal{K}(X_b) = X_b \cdot w_K,\ \mathcal{V}(X_b) = X_b \cdot w_V, \tag{3}$$

where w_Q, w_K, and w_V are learnable parameters. Specifically, the vector $\mathcal{Q}(X_b)$ can be regarded as a feature selector for channels of matrix $\mathcal{K}(X_b)$.

In the process of self-attention calculation, we define the pairwise function of $\mathcal{Q}(X_b)$ and $\mathcal{K}(X_b)$ as a matrix multiplication:

$$\mathcal{G}(X_b) = \mathcal{Q}(X_b)\mathcal{K}(X_b)^T, \tag{4}$$

where T operation means matrix transpose. Moreover, the generated $\mathcal{G}(X_b) \in \mathbb{R}^{L_a \times L_b \times H}$ also plays the role of feature selector for value $\mathcal{V}(X_b)$. Then, global attention can be defined as:

$$\mathcal{A}(X_b) = softmax(\mathcal{G}(X_b))\mathcal{V}(X_b),\ \mathcal{A}(X_b) \in \mathbb{R}^{L_a \times D_b \times H}, \tag{5}$$

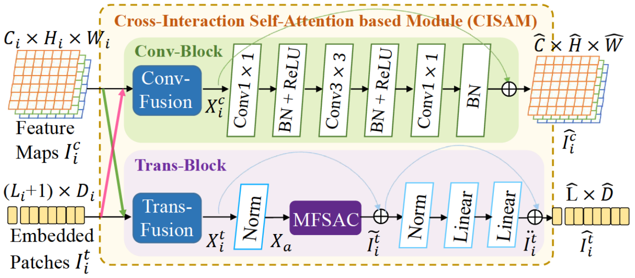


Figure 3: The architecture of our Cross-Interaction Self-Attention Based Module (CISAM).

where the goal of $softmax$ function is to normalize the $\mathcal{G}(X_b)$.

Next, the output $\widetilde{I}_i^t \in \mathbb{R}^{L_a \times D_a}$ of the MFSAC can be roughly described as the splicing of attention maps of multi-feature groups:

$$\widetilde{I}_i^t = Linear(reshape(\mathcal{A}(X_b))). \tag{6}$$

Specifically, the $Linear$ and $reshape$ functions are designed to ensure that the output is concatenated from the group of the obtained attention maps and has dimension $L_a \times D_a$.

After the MFSAC, in order to improve the feature expression ability of the network, two linear layers and residual connections are added to Trans-Block. The final output $\hat{I}_i^t \in \mathbb{R}^{\hat{L} \times \hat{D}}$ of Trans-Block can be regarded as:

$$\begin{aligned} \hat{I}_i^t &= \mathcal{F}_{tr2}(\mathcal{F}_{tr1}(X_i^t) \oplus X_i^t) \oplus (\mathcal{F}_{tr1}(X_i^t) \oplus X_i^t) \\ &= \ddot{I}_i^t \oplus (\widetilde{I}_i^t \oplus X_i^t). \end{aligned} \tag{7}$$

In detail, the specific calculations of $\mathcal{F}_{tr1}(\cdot)$ and $\mathcal{F}_{tr2}(\cdot)$ are as follows:

$$\begin{aligned} \widetilde{I}_i^t &= \mathcal{F}_{tr1}(X_i^t) \\ &= MFSAC(LayerNorm(X_i^t)) \\ &= MFSAC(X_a), \end{aligned} \tag{8}$$

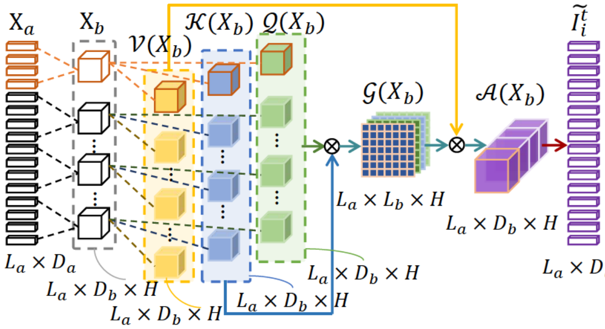


Figure 4: The details of Multi-Feature Self-Attention Class (MFSAC).

$$\begin{aligned}\ddot{I}_i^t &= \mathcal{F}_{tr2}(\mathcal{F}_{tr1}(X_i^t) \oplus X_i^t) \\ &= \mathcal{F}_{tr2}(\widetilde{I}_i^t \oplus X_i^t) \\ &= Linear^2(LayerNorm(\widetilde{I}_i^t \oplus X_i^t)),\end{aligned} \tag{9}$$

where the $\oplus$ operation can be regarded as the element-wise addition, the $LayerNorm$ function stands for normalization of features, and the $Linear^2$ function consists of two linear layers.

Multi-View Fusion Unit

After N view images are input into the network for feature extraction, each view can obtain one convolutional feature and one transformer feature. Convolutional features $F_c = [F_1^c, F_2^c, \ldots, F_i^c, \ldots, F_N^c] \in \mathbb{R}^{N \times C_c \times H_c \times W_c}$ for N views splice multi-view feature maps into $F_{cj} \in \mathbb{R}^{C_c \times H_c \cdot \sqrt{N} \times W_c \cdot \sqrt{N}}$ through feature-jigsaw function $\mathcal{J}_c(\cdot)$. Subsequently, we take the obtained feature jigsaw F_{cj} as the input of classifier $\mathcal{F}_{clas_c}(\cdot)$, and obtain the output prediction scores $P_c \in [0,\ 1]^C$ of the convolutional branch, where C is the total number of classes. Similarly, the N-view features of transformer branch $F_t = [F_1^t, F_2^t, \ldots, F_i^t, \ldots, F_N^t] \in \mathbb{R}^{N \times L_t \times D_t}$ are spliced into $F_{tj} \in \mathbb{R}^{L_t \times D_t \cdot N}$ by the feature-jigsaw function $\mathcal{J}_t(\cdot)$. Then, the feature jigsaw F_{cj} is input into the classifier $\mathcal{F}_{clas_t}(\cdot)$ to obtain the transformer-branch prediction scores $P_t \in [0,\ 1]^C$. The final output prediction scores of the network can be defined as:

$$P_f = P_c \oplus P_t, \tag{10}$$

where the $\oplus$ operation here represents the corresponding addition of the prediction scores for each category. The classification result can be obtained according to the prediction scores $P_f = [P_1^f, P_2^f, \ldots, P_{ii}^f, \ldots, P_C^f] \in \mathbb{R}^{1 \times C}$.

Since there are global features between the lesions of multiple views, inspired by Wang et al.(Wang et al. 2018), nonlocal-block (NLB) is added to the convolutional classifier $\mathcal{F}_{clas_c}(\cdot)$ to learn the global features of the feature jigsaw. Besides, since long-range connections also exist in multi-view feature jigsaws, two fully connected layers are added to transformer classifier $\mathcal{F}_{clas_t}(\cdot)$, which can be roughly treated as the MLP structure.

Method	Acc.	Prec.	Spec.	F_1	Time
Inception_ResNet_V2	70.53	65.69	67.13	65.43	78.41
Inception_V3	71.78	67.38	69.07	67.51	53.68
Inception_V4	71.32	67.58	71.86	68.13	65.91
MobileNet_V2	70.01	64.14	64.03	63.30	60.04
ResNet101	73.08	69.81	71.61	69.38	53.16
ResNet50	71.92	68.10	71.25	68.50	49.01
ResNext50_32x4d	72.01	67.84	70.07	67.99	58.17
VGG19_bn	74.11	71.27	73.59	70.71	66.84
Swin-B	74.06	71.13	72.36	70.35	43.18
Swin-S	73.04	69.33	71.90	69.37	41.00
Swin-T	71.73	67.71	70.33	67.75	41.32
ConvNeXt-B	74.43	71.97	73.92	71.06	42.47
ConvNeXt-S	73.59	70.60	72.55	69.92	40.01
ConvNeXt-T	73.41	70.29	72.68	69.82	**38.88**
MVCINN	**80.10**	**78.90**	**83.32**	**78.86**	49.59

Table 1: Comparison of single-view models that use V1-view data and our proposed MVCINN method that use four-view data. Quantitative results of accuracy(Acc., %), precision (Prec., %), specificity (Spec., %), F_1 score (F_1, %) and the total elapsed time (Time, s) of testing. The best results are highlighted in bold.

In addition, owing to the fundus image data obtained in real life which is collected based on the actual proportion of patients, there is a problem of category imbalance. To solve this problem, we employ the focal loss (Lin et al. 2017) here:

$$loss = -\sum_{ii=1}^{C} (1 - \hat{P}_{ii})^{\gamma} log(\hat{P}_{ii}), \tag{11}$$

$$\hat{P}_{ii} = [y]_{ii} \cdot [P]_{ii}, \ \ P = softmax(P_f), \tag{12}$$

where $\gamma > 0$ is the adjustable factor, $P_f \in \mathbb{R}^{1 \times C}$ is the predicted score from the model for C classes, and $y \in \mathbb{R}^{1 \times C}$ denotes the sample label in the form of one-hot vector. The $[\cdot]_{ii}$ stands for the ii-th element in array. The training process of MVCINN can be seen in Algorithm 1.

Experiments

Experimental Setups

Dataset. We conducted experiments on the multi-field imaging dataset for DR detection (MFIDDR[1]), which is the only publicly available large-scale dataset of multi-view fundus images on DR so far. The dataset contains 34,452 color retinal images captured by Zeiss Visucam NM/FA camera. Four images per sample of eyes, for example, the image focused on the macula, and the images revolved around the optic disc, which matches the center, top horizon, and bottom horizon of the optic disc respectively. Seven ophthalmologists used multi-view images to classify DR grades of each subject in accordance with international standards. All records are anonymous and retrieved with the patient's consent and with the permission of the hospital to disseminate the information, and there is no ethics conflict. Training and testing sets have been distributed on the MFIDDR, including 25,848 training images and 8,604 test images.

[1] https://github.com/mfiddr/MFIDDR

Method	Grade 0			Grade 1			Grade 2			Grade 3			Grade 4		
	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1	Prec.	Sens.	F_1
Inception_ResNet_V2_A	74.58	96.14	84.00	42.41	12.75	19.52	46.46	37.16	41.20	62.84	58.78	60.73	78.77	19.23	30.84
Inception_V3_A	76.04	95.13	84.52	41.37	17.68	24.76	48.11	34.70	40.21	62.50	62.00	62.20	76.25	23.72	35.95
Inception_V4_A	78.09	94.83	85.65	45.70	22.82	30.37	50.23	33.74	40.36	61.50	70.27	65.53	68.18	25.00	35.78
MobileNet_V2_A	74.61	96.25	84.05	40.71	11.07	17.15	47.31	37.30	41.40	61.97	64.36	62.93	71.08	15.39	24.99
ResNet101_A	78.14	96.09	86.19	49.01	22.15	30.46	50.73	42.49	46.20	63.16	62.00	62.51	71.11	18.59	29.42
ResNet50_A	78.24	95.30	85.93	48.26	22.60	30.68	50.62	41.26	45.37	63.79	66.72	65.12	79.59	22.44	34.53
ResNext50_32x4d_A	77.99	95.00	85.66	44.87	22.48	29.87	49.25	36.48	41.81	63.90	65.71	64.74	73.44	24.36	36.47
VGG19_bn_A	79.72	95.63	86.94	52.63	24.33	33.17	50.76	45.36	47.64	62.27	69.94	65.85	76.13	25.64	37.52
Swin-B_A	80.62	96.10	87.67	56.12	27.29	36.45	53.91	45.36	48.98	62.39	**74.66**	67.96	78.29	21.16	33.05
Swin-S_A	79.47	96.48	87.15	53.10	25.00	33.91	53.54	43.44	47.81	63.95	69.93	66.79	81.69	22.44	35.14
Swin-T_A	79.93	96.42	87.40	54.19	26.01	35.09	50.79	44.95	47.55	64.30	67.74	65.90	82.51	19.87	31.79
ConvNeXt-B_A	79.54	96.57	87.23	53.05	24.50	33.45	53.64	45.90	49.37	65.68	68.75	67.15	**85.73**	28.85	42.82
ConvNeXt-S_A	78.48	**96.76**	86.66	50.48	21.70	30.26	52.08	42.62	46.78	65.15	67.74	66.39	79.33	21.80	33.91
ConvNeXt-T_A	78.29	96.03	86.26	48.60	20.30	28.58	49.76	46.99	48.29	65.11	64.19	64.64	77.75	19.87	31.52
MVCINN	**86.71**	96.33	**91.26**	**68.25**	**48.10**	**56.43**	**57.44**	**61.20**	**59.26**	**70.00**	66.22	**68.06**	68.42	**33.33**	**44.83**

Table 2: Comparison of multi-view methods and our proposed MVCINN method. Quantitative results of precision(Prec.), sensitivity (Sens.), and F_1 score in DR grades 0-4. The best results are highlighted in bold. (Unit: %)

Method	Acc.	Prec.	Spec.	F_1	Time
Inception_ResNet_V2_A	69.83	64.77	65.31	64.39	78.89
Inception_V3_A	70.32	65.53	67.78	65.91	60.42
Inception_V4_A	71.71	67.67	71.07	68.02	69.65
MobileNet_V2_A	69.87	64.31	65.36	64.00	50.88
ResNet101_A	72.41	68.59	71.00	68.54	57.85
ResNet50_A	72.30	68.69	71.29	68.63	57.00
ResNext50_32x4d_A	71.65	67.61	70.85	68.00	49.68
VGG19_bn_A	73.50	70.36	73.55	70.07	67.97
ConvNeXt-B_A	74.65	71.96	74.85	71.39	57.09
ConvNeXt-S_A	73.94	70.76	73.04	70.39	56.06
ConvNeXt-T_A	74.05	71.07	73.73	70.65	55.74
Swin-B_A	74.14	70.99	73.14	70.64	56.62
Swin-S_A	73.20	69.51	71.46	69.20	57.32
Swin-T _A	72.55	68.78	71.30	68.56	55.85
MVCINN	**80.10**	**78.90**	**83.32**	**78.86**	**49.59**

Table 3: Comparison of multi-view methods and our proposed MVCINN method. The best results are highlighted in bold. (Unit: %)

Method	Acc.	Prec.	Spec.	F_1	Time
$\mathcal{B}_\alpha$	78.34	77.74	78.50	75.16	52.81
$\mathcal{B}_\beta$	66.71	57.61	57.63	58.97	52.08
$\mathcal{B}_\alpha + \mathcal{B}_\beta$	75.59	72.26	72.81	71.84	**48.42**
$\mathcal{B}_\alpha + \mathcal{B}_\beta + \mathcal{M}_\phi$	**80.10**	**78.90**	**83.32**	**78.86**	49.59
w/o MVFU	75.17	72.77	76.25	72.82	436.03
w/o NLB	78.85	77.65	82.84	77.65	57.39

Table 4: Ablation studies in MVCINN. The overall model is denoted as '$\mathcal{B}_\alpha + \mathcal{B}_\beta + \mathcal{M}_\phi$', where '$\mathcal{B}_\alpha$', '$\mathcal{B}_\beta$', and '$\mathcal{M}_\phi$' indicate the baseline convolutional branch, transformer branch, and CISAM respectively. 'w/o MVFU' and 'w/o NLB' denote that MVFU and NLB are removed from overall model, respectively. The best results are highlighted in bold. (Unit: %)

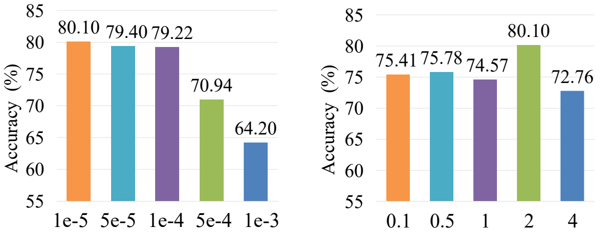
Implementation Details. The backbone of the convolutional branch with Conv-Block in our network is initialized by ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009). The transformer branch with Trans-Block is composed of a 12-layers transformer encoder with 9 heads, which is first pre-trained on the ImageNet dataset. During training, we adopt Mixup (Zhang et al. 2017) strategy for data enhancement to improve the generalization performance and robustness of the model. Furthermore, our model is achieved on PyTorch, and we use a random gradient coefficient with a base learning rate $1e^{-5}$ to improve our model.

Evaluation Metric. For the five DR categories, we adopt the commonly-agreed evaluation metrics (Trevethan 2017; Sasaki 2007) including accuracy (Acc.), precision(Prec.), sensitivity (Sens.), specificity (Spec.), F_1 score, and the total elapsed time. Notably, F_1 score provides a comprehensive evaluation of model performance to avoid the assessment bias of the model caused by sample imbalance.

Compared Methods. Several open source methods are adopted, which can be coarsely categorized into models based on CNN and Transformer: Inception_Resnet_V2 (Szegedy et al. 2016), Inception_V3 (Szegedy et al. 2016), Inception_V4 (Szegedy et al. 2016), MobileNet_V2 (Sandler et al. 2018), ResNet101 (He et al. 2016), ResNet50 (He et al. 2016), ResNext50_32x4d (Xie et al. 2017), VGG19_bn (Simonyan and Zisserman 2014), Swin-B (Liu et al. 2021), Swin-S (Liu et al. 2021), Swin-T (Liu et al. 2021), ConvNeXt-B (Liu et al. 2022b), ConvNeXt-S (Liu et al. 2022b) and ConvNeXt-T (Liu et al. 2022b). All baseline models have been pre-trained on the ImageNet dataset.

Main Results

Comparisons on Single-View Methods. Due to the data limitation, most of the previous methods only use the single-view $45°$ images centered on the macula (i.e. V1-view im-



(a) Learning Rate

(b) Parameter γ

Figure 5: Evaluation of the hyperparameters. Comparative analysis of (a) learning rate, and (b) γ in $loss$ function.

ages). To verify the effectiveness of the multi-view method for DR detection, our multi-view method using four-view data is compared with the single-view methods using V1-view data. As shown in Table 1, our multi-view method MVCINN obtains the best performance in the comparisons, which indicates the features learning from multi-view images benefit the classification of five DR grades. Specifically, our method obtains an accuracy of 80.10% that outperforms the accuracy of single-view methods by 5.67%-10%. Moreover, MVCINN achieves the best performance of overall precision, specificity, F_1 score in the comprehensive evaluation, proving the superiority of the our multi-view method. The total elapsed time of MVCINN model is 49.59s, which is not the shortest time, but is better than average in terms of performance and network complexity.

Comparisons on Multi-View Methods. To our knowledge, there are few works to conduct DR detection using multi-view data. In order to demonstrate the effectiveness of our MVCINN to multi-view DR detection, we compare our approach to the multi-view methods implemented by fusing the results of multiple views. With the state-of-the-art open source models as the backbone, the models learn the data of each view separately, and finally take the average of the classification results of each view as the results of the multi-view methods. The experimental results are shown in Tables 3 and 2, and the suffix '_A' is added after the name of the baseline method to distinguish it. Compared with the multi-view methods, MVCINN achieves the best performance with the average improvements of 7.66%, 10.29%, 12.34%, 10.40%, and 9.77s in terms of overall accuracy, precision, specificity, F_1 score and elapsed time, respectively. In particular, Table 2 provides the precision, sensitivity, and F_1-score results for each grade in classification. Thus, we can conclude that our method achieves better results through the combination of multi-view features than the methods that simply use multi-view data without considering the correlation of views.

Ablation Studies

Through the experiment, we scrutinize the contribution of each component to the proposed model, as described in Table 4. Our model as a whole mark '$\mathcal{B}_\alpha + \mathcal{B}_\beta + \mathcal{M}_\phi$', where '$\mathcal{B}_\alpha$', '$\mathcal{B}_\beta$', and '$\mathcal{M}_\phi$' indicate that the use of the single convolutional branch, single transformer branch, and the cross-interaction self-attention based module (CISAM), respectively. Furthermore, we also verify the impact of Multi-View Fusion Unit (MVFU) by removing it.

Analysis on Dual Branches. Our network can be roughly regarded as a dual-branch network, which can be divided into a network branch based on CNN ($\mathcal{B}_\alpha$) and a branch based on transformer mechanism ($\mathcal{B}_\beta$). First, we adopt the baselines $\mathcal{B}_\alpha$ and $\mathcal{B}_\beta$ in investigation of DR detection. As shown in the first two lines of Table 4, the performance is unsatisfactory. In '$\mathcal{B}_\alpha + \mathcal{B}_\beta$', two branches are used to extract features independently without information interaction, which balances the results of two single branches.

Effectiveness of CISAM. Furthermore, the CISAM is added into the overall network (i.e., $\mathcal{B}_\alpha + \mathcal{B}_\beta + \mathcal{M}_\phi$). Comparing the results of the third and fourth rows in Table 4, CISAM can make full use of the advantages of dual branches to fuse local and long-distance features so as to improve the performance of the model.

Effectiveness of MVFU. Considering the correlation between multiple views, we carry out fusion learning of multi-view features through MVFU. To verify the effectiveness of MVFU, we removed this unit from the overall model and then implemented the classification using the sum of multi-view results. As shown in Table 4, the model trained without MVFU drops sharply compared to the overall model, which indicates the effectiveness of MVFU. In addition, the degradation of model performance after removing NLB indicates that it is effective to consider long-distance features between views when fusing multi-view features.

Hyperparameter Evaluations. We analyze the influence of the learning rate and the γ parameter in $loss$ function on the model. We find that with the increase of learning rate, the model accuracy tends to decrease. Our model chooses a learning rate of $1e^{-5}$, and use the AdamW optimizer (Kingma and Ba 2015) with the cosine annealing schedule (Loshchilov and Hutter 2017). As shown in Fig. 5, we also evaluate the influence of γ. Our model achieves much better performance when $\gamma = 2$.

Conclusion

In this paper, we present a novel multi-view model for DR detection by interactive learning local and long-distance features of multiple views. Our method makes use of the roughly complete retinal information from multiple views, which makes up for the loss of feature information in the current common single-view methods. Specially, the model utilizes location information to merge multi-view features in MVFU, which is better than simply adding or averaging the results of each view. To enhance the representation of the model, we propose a mechanism to simultaneously learn local and long-range lesion features on fundus images and use CISAM to perform an interactive fusion of these two features. Experimental results prove that CISAM plays a key role in developing local and long-range features and improving model performance. Our future work will expand the number of samples by increasing unlabeled images with semi-supervised learning to improve detection performance of the model.

Acknowledgments

This work was supported in part by the Science and Technology Innovation Committee of Shenzhen Municipality under grant no. GJHZ20210705141812038, the Shenzhen Fundamental Research Fund under grant no. JCYJ20190806142416685, and the Stable Support Projects for Shenzhen Higher Education Institutions under grant no. 20220715183602001.

References

Amit Kamran, S.; Fariha Hossain, K.; Tavakkoli, A.; Zuckerbrod, S. L.; and Baker, S. A. 2021. VTGAN: Semi-supervised Retinal Image Synthesis and Disease Prediction using Vision Transformers. arXiv:2104.06757.

Decenciere, E.; Zhang, X.; Cazuguel, G.; Lay, B.; Cochener, B.; Trone, C.; Gain, P.; Ordonez, R.; Massin, P.; Erginay, A.; Charton, B.; and Klein, J.-C. 2014. Feedback on a publicly distributed database: the Messidor database. *Image Analysis Stereology*, 33(3): 231–234.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. ImageNet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

EyePACS. 2015. Kaggle-EyePACS. https://www.kaggle.com/c/diabetic-retinopathy-detection/data. Accessed: 2022-07-27.

Federation, I. D. 2021. IDF Diabetes Atlas, 10th edn. https://www.diabetesatlas.org. Accessed: 2022-07-27.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Hu, J.; Chen, R.; Lu, Y.; Dou, X.; Ye, B.; Cai, Z.; Pu, Z.; and Mou, L. 2019. Single-Field Non-Mydriatic Fundus Photography for Diabetic Retinopathy Screening: A Systematic Review and Meta-Analysis. *Ophthalmic Research*, 62: 61–67.

Hu, S.; Lou, Z.; and Ye, Y. 2022. View-Wise Versus Cluster-Wise Weight: Which Is Better for Multi-View Clustering? *IEEE Transactions on Image Processing*, 31: 58–71.

Hu, S.; Shi, Z.; and Ye, Y. 2022. DMIB: Dual-correlated Multivariate Information Bottleneck for Multi-view Clustering. *IEEE Transactions on Cybernetics*, 52(6): 4260–4274.

Huang, C.; Wen, J.; Xu, Y.; Jiang, Q.; Yang, J.; Wang, Y.; and Zhang, D. 2022. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Huang, C.; Wu, Z.; Wen, J.; Xu, Y.; Jiang, Q.; and Wang, Y. 2021. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Transactions on Industrial Informatics*, 18(8): 5171–5179.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Li, X.; Pang, T.; Xiong, B.; Liu, W.; Liang, P.; and Wang, T. 2017. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–11.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

Liu, C.; Wen, J.; Luo, X.; and Xu, Y. 2023. Incomplete Multi-View Multi-Label Learning via Label-Guided Masked View- and Category-Aware Transformers. In *AAAI Conference on Artificial Intelligence*.

Liu, C.; Wu, Z.; Wen, J.; Xu, Y.; and Huang, C. 2022a. Localized Sparse Incomplete Multi-view Clustering. *IEEE Transactions on Multimedia*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A ConvNet for the 2020s. arXiv:2201.03545.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Luo, X.; Wang, W.; Xu, Y.; Lai, Z.; Jin, X.; Zhang, B.; and Zhang, D. 2023. A deep convolutional neural network for diabetic retinopathy detection via mining local and long-range dependence. *CAAI Transactions on Intelligence Technology*.

Luo, X.; Zhang, H.; Su, J.; Wong, W.; Li, J.; and Xu, Y. 2022. RV-ESA: A novel computer-aided elastic shape analysis system for retinal vessels in diabetic retinopathy. *Computers in Biology and Medicine*, 152: 106406.

Ophthalmoscopy, D.; and Levels, E. 2002. International clinical diabetic retinopathy disease severity scale detailed table. *American Academy of Ophthalmology*.

Pao, S. I.; Lin, H. Z.; Chien, K. H.; Tai, M. C.; Chen, J. T.; and Lin, G. M. 2020. Detection of Diabetic Retinopathy Using Bichannel Convolutional Neural Network. *Journal of Ophthalmology*, 2020: 1–7.

Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; and Ye, Q. 2021. Conformer: Local Features Coupling Global Representations for Visual Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 367–376.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Sasaki, Y. 2007. The truth of the F-measure. *Teach Tutor Mater*, 1(5): 1–5.

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

Srihatrai, P.; and Hlowchitsieng, T. 2018. Thanita The diagnostic accuracy of single- and five-field fundus photography in diabetic retinopathy screening by primary care physicians. *Indian Journal of Ophthalmology*, 66: 94–97.

Sun, R.; Li, Y.; Zhang, T.; Mao, Z.; Wu, F.; and Zhang, Y. 2021. Lesion-Aware Transformers for Diabetic Retinopathy Grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10938–10947.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Trevethan, R. 2017. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, 5: 307.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, S.; Liu, X.; Zhu, X.; Zhang, P.; Zhang, Y.; Gao, F.; and Zhu, E. 2021. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. *IEEE Transactions on Image Processing*, 31: 556–568.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-Local Neural Networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Wang, Z.; Yin, Y.; Shi, J.; Fang, W.; Li, H.; and Wang, X. 2017. Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 267–275.

Wen, J.; Zhang, Z.; Fei, L.; Zhang, B.; Xu, Y.; Zhang, Z.; and Li, J. 2022. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Wu, J.; Hu, R.; Xiao, Z.; Chen, J.; and Liu, J. 2021. Vision Transformer-based recognition of diabetic retinopathy grade. *Medical Physics*, 48(12): 7850–7863.

Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. arXiv:1710.09412.

Zhu, Y.; Ma, J.; Yuan, C.; and Zhu, X. 2022. Interpretable learning based Dynamic Graph Convolutional Networks for Alzheimer's Disease analysis. *Information Fusion*, 77: 53–61.