MPMNet: Modal Prior Mutual-support Network for Age-related Macular Degeneration Classification

Yuanyuan Li[1,2], Huaying Hao[1], Dan Zhang[3], Huazhu Fu[4], Mengting Liu[5], Caifeng Shan[6], Yitian Zhao[1,*], and Jiong Zhang[1,*]

[1] Laboratory of Advanced Theranostic Materials and Technology, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
jiong.zhang@ieee.org; yitian.zhao@nimte.ac.cn
[2] Cixi Biomedical Research Institute, Wenzhou Medical University, Ningbo, China
[3] School of Cyber Science and Engineering, Ningbo University of Technology, Ningbo, China
[4] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore 138632
[5] School of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China
[6] School of Biomedical Engineering, Nanjing University, Nanjing, China

Abstract. Early screening and classification of Age-related Macular Degeneration (AMD) are crucial for precise clinical treatment. Currently, most automated methods focus solely on dry and wet AMD classification. However, the classification of wet AMD into more explicit type 1 choroidal neovascularization (CNV) and type 2 CNV has rarely been explored, despite its significance in intravitreal injection. Furthermore, previous methods predominantly utilized single-modal images for distinguishing AMD types, while multi-modal images can provide a more comprehensive representation of pathological changes for accurate diagnosis. In this paper, we propose a **M**odal **P**rior **M**utual-support **N**etwork (MPMNet), which for the first time combines OCTA images and OCT sequences for the classification of normal, dry AMD, type 1 CNV, and type 2 CNV. Specifically, we first employ a multi-branch encoder to extract modality-specific features. A novel modal prior mutual-support mechanism is proposed, which determines the primary and auxiliary modalities based on the sensitivity of different modalities to lesions and makes joint decisions. In this mechanism, a distillation loss is employed to enforce the consistency between single-modal decisions and joint decisions. It can facilitate networks to focus on specific pathological information within individual modalities. Furthermore, we propose a mutual information-guided feature dynamic adjustment strategy. This strategy adjusts the channel weights of the two modalities by computing the mutual information between OCTA and OCT, thereby mitigating the influence of low-quality modal features on the network's robustness. Experiments on private and public datasets have demonstrated that the proposed MPMNet outperforms existing state-of-the-art methods.

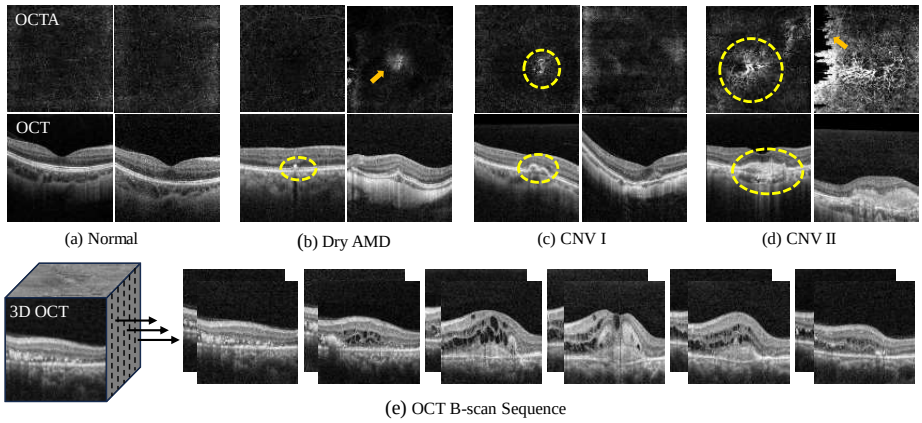

Fig. 1. OCTA (top row) and OCT (second row) images showing a normal eye and eyes affected by AMD. Noise interference is indicated by arrows, while lesions are circled. A sequence of OCT B-scans from an eye with type 2 CNV type (bottom row).

Keywords: Age-related macular degeneration · CNV · OCT · OCTA · Multimodal

1 Introduction

AMD is a major cause of blindness worldwide [1]. It can be classified into dry AMD, characterized by drusen and geographic atrophy, and wet AMD, distinguished by the growth of CNV beneath the retina. The wet AMD can be again divided into type 1 CNV and type 2 CNV depending on whether the CNV breaks through the retinal pigment epithelium (RPE) layer [2]. Treatment approaches for AMD vary across different types. For instance, clinicians tailor treatments to the corresponding CNV types targeting the RPE layer, which aims to minimize cellular damage and postoperative visual function impairment [3]. Thus, precise classification of AMD is crucial for disease analysis and surgical intervention.

In clinical practice, AMD diagnosis relies on retinal imaging techniques such as color fundus photography (CFP), fluorescein fundus angiography (FFA), and optical coherence tomography (OCT). Compared to FFA and CFP, OCT and OCTA offer a non-invasive approach to provide clear visualization of different layers and blood flow, avoiding potential side effects and risks associated with dye injection [4]. Therefore, the combined use of OCT and OCTA for AMD diagnosis is not only safer but also more precise. Several automated methods have been proposed to employ OCT and OCTA for AMD classification. For example, multi-task networks [5,6,7] were utilized to obtain segmentation results of key regions on individual OCT images, guiding the network to focus on sensitive areas for AMD lesions. Multi-scale convolutional neural networks (CNN) [8,9] were employed to extract rich feature representations for AMD classification.

Recently, Zhang *et al.* [10] achieved high accuracy in AMD classification by utilizing a 2D CNN that was trained with additional supervision on 3D OCTA volumes.

Although progress has been made in AMD classification, there are still several challenges that require investigation. Firstly, existing works focus on the dry and wet AMD classification, with little in-depth research on the CNV-type classification. In addition, many studies solely rely on a single-modal to differentiate AMD types, failing to take full advantage of the OCT layer structure and blood flow information in OCTA. This may cause unreliable clinical diagnosis. In practice, ophthalmologists consider both OCT and OCTA images and combine their modality-specific information to make a more accurate diagnosis. Therefore, to extract modality-specific pathological features and integrate them into deep networks, we propose to combine OCT sequences and OCTA images for AMD classification, and CNV differentiation.

Specifically, we propose a novel **M** modal **P**rior **M**utual-support **N**etwork (MPMNet) for the classification of normal, dry AMD, type 1 CNV, and type 2 CNV. The MPMNet makes full use of the prior pathological features of OCT sequences and OCTA images, where a distillation loss is introduced to enhance the consistency and complementarity of multi-modal classification. The main contributions can be summarized as follows:

(a) To our knowledge, this work is the first attempt to classify normal, dry AMD, type 1 CNV, and type 2 CNV using both OCT sequences and OCTA images.
(b) A new modal prior mutual-support mechanism is proposed to boost the network to focus on the pathology information that is sensitive to each modality.
(c) A mutual information-guided feature dynamic adjustment strategy is designed to reduce the impact of low-quality images and enhance the robustness of the network.

2 Proposed Method

The proposed MPMNet comprises three primary components: a multi-branch encoder, a modal prior mutual-support mechanism, and a mutual information-guided feature adjustment strategy, as depicted in Fig. 2.

2.1 Multi-branch Encoder

The multi-branch encoder consists of two symmetrical CNN branches and a Vision Transformer (ViT) [11] branch.

To better extract features of OCT sequences stacked in the channel direction, we present a lesion region self-attention to replace self-attention in the ViT branch. Then the global information of ViT and the local information of CNN are integrated to obtain the comprehensive high-dimensional features of OCT sequences, as shown in Fig. 2-(a).

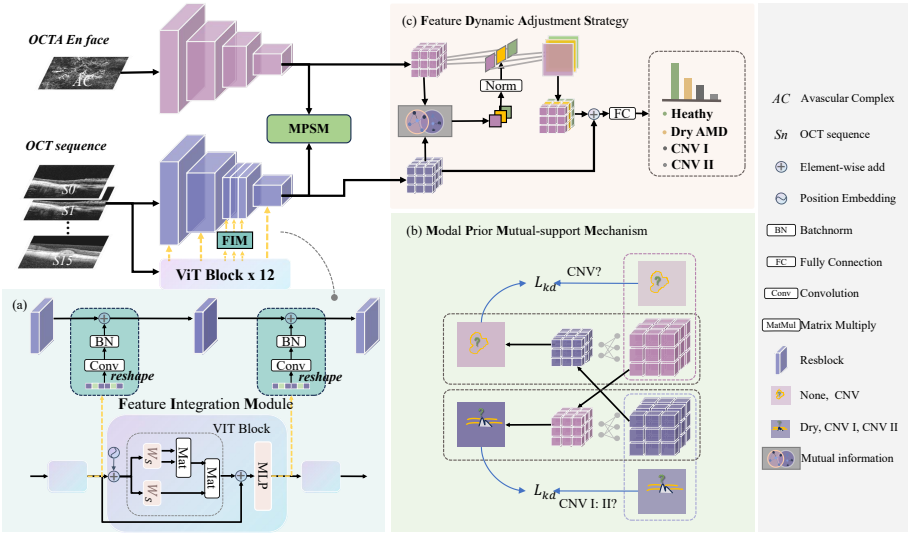

Fig. 2. The architecture of MPMNet and the specific design of its modules.

Specifically, the CNN branch contains four stages, each containing three residual blocks. The ViT branch contains 12 ViT blocks, and the current block receives the output of the previous block as input. While the pathological information in the OCT sequence changes frame by frame, the spatial position of the lesion is roughly the same. To exploit this characteristic, the features fed into the ViT block are projected only once to establish pairwise correlations among the visual markers, as expressed in Eq. 1. This enables each feature to pay attention to features with similar positions while focusing on itself. The current ViT features are processed by 1×1 convolution and upsampling to ensure their spatial scales and channel dimensions are consistent with the local features in the corresponding CNN branches. The integrated CNN features and ViT features with the same scale are used as input to the next residual block. The process is illustrated by the feature integration module shown in Fig. 2-(a). Finally, the high-dimensional features of the OCT sequence are obtained from the last residual block of the CNN branch. This multi-branch encoder is designed based on the characteristics of each modality, which can fully capture rich semantic information on each modality.

$$f_t = Softmax(\frac{XW_sW_s^TX^T}{\sqrt{d}})W_vX, \tag{1}$$

where f_t stands for high-dimensional representation, X is a collection of visual tokens with d dimensions, $W_s, W_v \in \mathbf{R}^{d\times d}$ are projection parameters.

2.2 Modal Prior Mutual-support Mechanism

To enhance feature representation for specific modalities, we proposed a modal prior mutual-support mechanism (MPMM), inspired by a segmentation work [12]. Clearly, OCTA images are highly sensitive to the presence of CNV, while OCT sequences allow for clear observation of drusen and changes in retinal layers. Based on this fact, the MPMM consists of two branches, with each branch considering OCT or OCTA as the primary modal, respectively. As shown in Fig. 2-(b), when utilizing OCTA as the primary modal, two specific features extracted from the multi-branch encoder are separately input into the classifiers to calculate prediction probabilities P_a and P_b for the presence of CNV or not CNV. Where P_a represents the prediction probability for the primary modal OCTA, and P_b for OCT.

When P_a exceeds the predetermined probability threshold T_a, that class will be regarded as the final decision. Otherwise, the probabilities predicted by the two modalities will be used for voting as an auxiliary decision, and the final probabilities will be normalized as the joint decision K_t^c, which can be expressed in Eq. 2 and Eq. 3.

$$K_a^i = \begin{cases} P_a^i, \ P_a^i >= T_a \\ \\ Average\left(P_a^i, P_b^i\right), \ P_a^i < T_a \end{cases} \tag{2}$$

$$K_t^c = \frac{K_a^i}{\sum_i^N K_a^i} \tag{3}$$

where c stands for category, i represents a specific category (i.e. CNV and not CNV), N denotes the total number of categories, and K_t^c denotes the normalized soft label of category c.

Similarly, OCT is employed as the primary modal for discriminating among Dry AMD, type 1 CNV, and type 2 CNV.

Finally, the KL dispersion is employed to compute the distillation loss of the decision result K_a for the primary modal and the joint decision result K_t for the primary modal and auxiliary modal, as shown in Eq. 4.

$$\mathcal{L}_{kd} = KL\left(Softmax\left(\frac{K_a}{\tau}\right) \parallel Softmax\left(\frac{K_t}{\tau}\right)\right) \tag{4}$$

where τ represents the temperature coefficient.

According to [13], the overall task loss is represented by Eq. 5.

$$\mathcal{L}_{\text{task}} = \alpha * \mathcal{L}_{cls} + \beta * \tau^2 * \mathcal{L}_{kd} \tag{5}$$

where $\mathcal{L}_{cls}$ represents the multimodal classification loss computed using the cross-entropy loss function, while α and β are hyperparameters used to balance the gradients, with assigned values of 0.7 and 0.3, respectively.

2.3 Feature Dynamic Adjustment Strategy

To further enhance the weighting of reliable modalities in the fusion process, we propose a feature dynamic adjustment strategy (FDAS) by employing mutual information to evaluate the significance of OCTA images. OCT provides more categorized information than OCTA and some OCTA images are of low quality. Therefore, we treat each channel of OCTA as a feature, computing the mutual information between each OCTA feature and OCT features to adjust the weights of OCTA features. The mathematical definition of mutual information is given by Eq. 6.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) log \frac{P(x,y)}{P(x)P(y)}, \tag{6}$$

where $P(x,y)$ represents the joint probability of X and Y, while $P(x)$ and $P(y)$ denote the marginal probabilities of X and Y, respectively. However, the joint probability distribution of high-dimensional features is highly complex, rendering direct computation of mutual information challenging. To address this, We adopted the neural network-based mutual information estimation method (MINE) proposed by [14]. The high-dimensional features of OCTA are treated as individual features on a channel-by-channel basis. Subsequently, the mutual information between these features and the entire high-dimensional OCT features is calculated using the MINE method. Utilizing normalized mutual information as dynamic weights, we adjust the OCTA features and integrate them with OCT features. This approach mitigates the impact of low-quality OCTA images on fused features, thereby enhancing the robustness of the network.

3 Experimental Results

3.1 Experimental Setup

Dataset: The proposed MPMNet was evaluated on a private dataset consisting of paired OCT and OCTA images from 384 eyes, obtained using the Heidelberg OCT2 system (Heidelberg, Germany). The enface projection of the avascular complex served as the OCTA images, which were captured within a 3×3 mm^2 area centered on the fovea. OCT sequences were based on central B-scan images, randomly selecting 16 OCT B-scan slices from the macular region. All images were resized to 224×224 resolution for training. The dataset consists of eyes categorized as healthy, dry AMD, type 1 CNV, and type 2 CNV, with respective counts of 79, 74, 83, and 112. For each eye, two ophthalmologists jointly classify its condition as one of these four categories, by examining the corresponding CFP along with OCT, FA, or OCTA images.

Implementation Details: Our method is implemented based on the PyTorch framework with four Nvidia RTX 3090 GPUs. We train the model using an Adam optimizer with an initial learning rate of 0.0001 and a batch size of 8 for 200 epochs, without implementing a learning rate decay strategy. The model

Table 1. Performance comparison of different methods on private dataset and public dataset MMC-AMD.

Method	Private datasets			Public datasets		
	F1	ACC	Kappa	F1	ACC	Kappa
OCTA-CNN	0.5368	0.6029	0.4507	-	-	-
OCT-CNN	0.7556	0.7058	0.5903	-	-	-
Resnet-50 [15]	0.8230	0.7794	0.6927	0.6700	0.8167	0.6133
ConvNext [16]	0.7952	0.7059	0.6014	0.6410	0.8000	0.6131
Vision Transformer [11]	0.5381	0.7107	0.41.68	0.7405	0.7500	0.6531
Swin Transformer [17]	0.7015	0.6764	0.5448	0.7429	0.71875	0.5527
CCDFuse [19]	0.5793	0.5143	0.3532	0.4534	0.6250	0.4181
MMC-AMD [18]	0.8488	0.7813	0.7083	**0.8929**	0.8516	0.8115
Ours	**0.8930**	**0.8676**	**0.8215**	0.8920	**0.8750**	**0.8248**

inputs were subjected to a standard data augmentation pipeline during training, including random horizontal flipping, random rotation, and random cropping. A 5-fold cross-validation method was used to evaluate the performance.

3.2 Comparison with State-of-the-arts:

Single-modal experiments: We conducted extensive single-modal experiments on OCTA and OCT images to determine the optimal input configuration. Two single-modal models were trained: OCTA-CNN and OCT-CNN. As shown in Table 1, the results indicate that the classification performance of the OCT images is superior to that of OCTA images.

Multi-modal experiments: To benchmark the performance of our model, we compared it with state-of-the-art AMD classification methods, including Resnet-50 [15], ConvNext [16], Swin Transformer [17], MMC-AMD [18], CDDFuse [19]. The first three classification networks were adapted to dual branches to accommodate multi-modal image inputs. We evaluated the classification performance using metrics such as F1 score (F1), Accuracy (ACC), and Cohen's Kappa Coefficient (Kappa). The results in Table. 1 show that our MPMNet outperforms existing state-of-the-art approaches. Conventional classification networks struggle with redundant information and low-quality images in this dataset, while newer baselines like MMC-AMD are also impacted by their design focus. Similarly, advanced fusion networks like CDD-Fuse overlook pathological priors from both modalities. In contrast, our method considers the contribution of each modality's priors and utilizes their complementarity, yielding superior performance in AMD classification with OCT and OCTA.

Extended experiment on public dataset: To further verify our method's stability and generability, we conducted experiments on a public dataset MMC-AMD [18], which includes paired CFP images and OCT images. The implementation details remain consistent with the experiments on the in-house dataset.

Table 2. Ablation results for AMD classification on private dataset.

Method				**Private dataset**		
Backbone	Multi-brunch	MPSM	FDAS	F1	ACC	Kappa
✓				0.8709	0.8125	0.7467
✓	✓			0.8844	0.8438	0.7872
✓	✓	✓		0.8800	0.8529	0.8022
✓	✓	✓	✓	**0.8930**	**0.8676**	**0.8215**

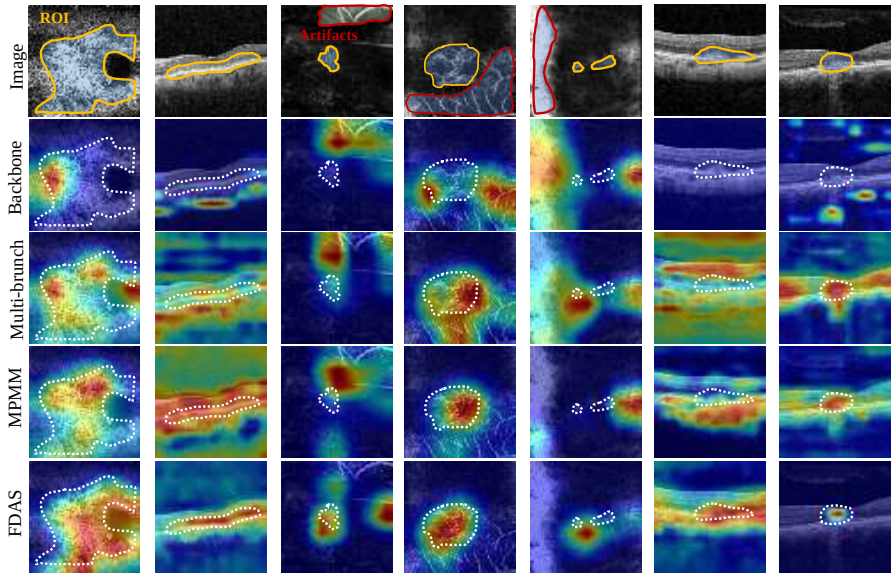

Fig. 3. Visualization of the same dataset's Grad-CAM across different modules. The orange circles indicate ROIs, while the red circles highlight interference from artifacts.

Since no specific dataset division was published, we randomly selected 300 pairs of images, comprising 50 pairs of healthy samples, 75 pairs of dry AMD, 100 pairs of wet AMD, and 75 pairs of polypoidal choroidal vasculopathy. As demonstrated in Table 1, our method achieved the best performance compared to other methods, with an accuracy of 0.8750 and a Kappa coefficient of 0.8215.

3.3 Ablation Study

To demonstrate the effectiveness of MPMNet in AMD classification, we employed a dual-branch CNN network as the backbone and systematically integrated individual components into the training framework. The ablation validation results for each component are summarized in Table. 2, and the respective visualizations are shown column-by-column in Fig. 3. The first two columns reveal the gradual increase in the network's ability to capture regions of interests (ROIs) of com-

plex shapes. The middle three columns demonstrate that the network initially focuses on the highlighted artifact regions but then gradually shifts its attention to the ROIs. The last two columns show that with the introduction of the modules, the network starts to recognize ROIs that were previously ignored. From the above results, the network demonstrates the capability to accurately discern each modality-sensitive region, facilitated by the mutual-support mechanism of modal priors, thereby achieving higher accuracy and robustness.

4 Conclusion

In summary, this study presents a novel approach aimed at addressing the challenges of AMD classification. It incorporates a multi-branch encoder, modality prior mutual-support mechanisms, and a modality feature adjustment strategy based on mutual information to achieve an accurate classification of AMD. Experimental results indicate that by integrating modality prior mutual-support to enhance feature extraction and conducting multi-modal feature adjustment, the accuracy and robustness of AMD classification models can be significantly improved. Compared to current state-of-the-art methods, the proposed approach demonstrates superior performance.

Acknowledgments. This work was supported in part by the National Science Foundation Program of China (62371442, 62103398, 62272444, 62350068), in part by the Zhejiang Provincial Natural Science Foundation (LQ23F010002, LZ23F010002, LR24F010002), in part by the Ningbo Natural Science Foundation (2022J143).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Mitchell, P., Liew, G., Gopinath, B., Wong, T.Y.: Age-related macular degeneration. The Lancet **392**(10153) (2018) 1147–1159
2. Dumar, A.T.H., Arias, J.D., Do, D.V.: Choroidal neovascularization: Oct angiography findings. (2020)
3. Ohno-Matsui, K., Ikuno, Y., Lai, T.Y., Cheung, C.M.G.: Diagnosis and treatment guideline for myopic choroidal neovascularization due to pathologic myopia. Progress in Retinal and Eye Research **63** (2018) 92–106
4. Wilde, C., Patel, M., Lakshmanan, A., Amankwah, R., Dhar-Munshi, S., Amoaku, W.: The diagnostic accuracy of spectral-domain optical coherence tomography for neovascular age-related macular degeneration: a comparison with fundus fluorescein angiography. Eye **29**(5) (May 2015) 602–610
5. Diao, S., Su, J., Yang, C., Zhu, W., Xiang, D., Chen, X., Peng, Q., Shi, F.: Classification and segmentation of oct images for age-related macular degeneration based on dual guidance networks. Biomedical Signal Processing and Control **84** (2023) 104810

6. He, X., Deng, Y., Fang, L., Peng, Q.: Multi-modal retinal image classification with modality-specific attention network. IEEE transactions on medical imaging **40**(6) (2021) 1591–1602
7. Moradi, M., Chen, Y., Du, X., Seddon, J.M.: Deep ensemble learning for automated non-advanced amd classification using optimized retinal layer segmentation and sd-oct scans. Computers in Biology and Medicine **154** (2023) 106512
8. Thomas, A., Harikrishnan, P., Ramachandran, R., Ramachandran, S., Manoj, R., Palanisamy, P., Gopi, V.P.: A novel multiscale and multipath convolutional neural network based age-related macular degeneration detection using oct images. Computer methods and programs in biomedicine **209** (2021) 106294
9. Rasti, R., Rabbani, H., Mehridehnavi, A., Hajizadeh, F.: Macular oct classification using a multi-scale convolutional neural network ensemble. IEEE transactions on medical imaging **37**(4) (2017) 1024–1034
10. Zhang, Z., Zhang, H., Heinke, A., Galang, M., Deussen, D., Wen, B., Bartsch, D.U., Freeman, W., Nguyen, T., An, C.: Robust amd stage grading with exclusively octa modality leveraging 3d volume
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Qiu, Y., Chen, D., Yao, H., Xu, Y., Wang, Z.: Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2023) 21317–21326
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv: Machine Learning,arXiv: Machine Learning (Mar 2015)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations,International Conference on Learning Representations (Jan 2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
16. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (Jun 2022)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). (Oct 2021)
18. Wang, W., Li, X., Xu, Z., Yu, W., Zhao, J., Ding, D., Chen, Y.: Learning two-stream cnn for multi-modal age-related macular degeneration categorization. IEEE Journal of Biomedical and Health Informatics **26**(8) (2022) 4111–4122
19. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. (Nov 2022)