Representation, Alignment, Fusion: A Generic Transformer-Based Framework for Multi-modal Glaucoma Recognition

You Zhou[1], Gang Yang[1,2(✉)], Yang Zhou[3], Dayong Ding[3], and Jianchun Zhao[3]

[1] School of Information, Renmin University of China, Beijing, China
[2] MOE Key Lab of DEKE, Renmin University of China, Beijing, China
yanggang@ruc.edu.cn
[3] Vistel AI Lab, Visionary Intelligence Ltd., Beijing, China

Abstract. Early glaucoma can be diagnosed with various modalities based on morphological features. However, most existing automated solutions rely on single-modality, such as Color Fundus Photography (CFP) which lacks 3D structural information, or Optical Coherence Tomography (OCT) which suffers from insufficient specificity for glaucoma. To effectively detect glaucoma with CFP and OCT, we propose a generic multi-modal Transformer-based framework for glaucoma, MM-RAF. Our framework is implemented with pure self-attention mechanisms and consists of three simple and effective modules: Bilateral Contrastive Alignment (BCA) aligns both modalities into the same semantic space to bridge the semantic gap; Multiple Instance Learning Representation (MILR) aggregates multiple OCT B-scans into a semantic structure and downsizes the scale of the OCT branch; Hierarchical Attention Fusion (HAF) enhances the cross-modality interaction capability with spatial information. By incorporating three modules, our framework can effectively handle cross-modality interaction between different modalities with huge disparity. The experimental results demonstrate that the framework outperforms the existing multi-modal methods of this task and is robust even with a clinical small dataset. Moreover, by visualizing, OCT can reveal the subtle abnormalities in CFP, indicating that the relationship between various modalities is captured. Our code is available at https://github.com/YouZhouRUC/MM-RAF.

Keywords: Glaucoma recognition · Multi-modal learning · Multiple instance learning · Contrastive learning

1 Introduction

Glaucoma is the second leading ophthalmic blindness disease, with nearly 70 million patients worldwide. Early glaucoma can be detected by color fundus

The computer resources were provided by Public Computing Cloud Platform of Renmin University of China.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_66.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 704–713, 2023.
https://doi.org/10.1007/978-3-031-43990-2_66

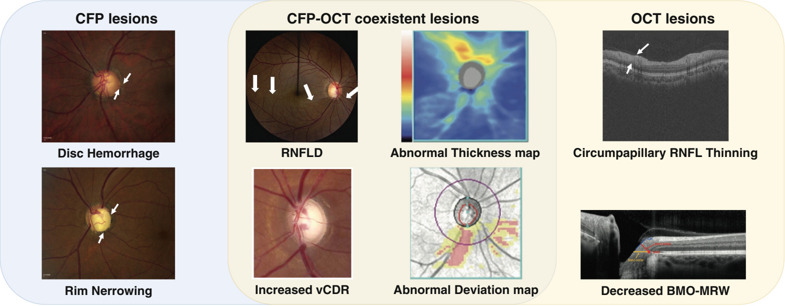

Fig. 1. Some Lesions of glaucoma in clinical examination. Lesions in the blue box appear in CFP. Yellow for OCT. (Color figure online)

photography (CFP) and optical coherence tomography (OCT). The gold standard of glaucoma in CFP includes Cup to Disk Ratio (CDR) enlargement and Retinal Nerve Fiber Layer Defects (RNFLD). Diagnosing with CFP has earned superb performance, but it only captures flat information. OCT scans the fine-grain 3D structure of the fundus, and the quantitative analysis conducted by OCT can help diagnosis for junior clinicians. Many crucial lesions coexist in both CFP and OCT images. We list some critical lesions correlated to glaucoma in Fig. 1.

Early research on automated glaucoma recognition focuses on CFP. Based on the ISNT rule [8], the vertical Cup-to-Disc rate (vCDR) is a feature with high specificity [15]. Besides, the RNFLD is an important clinical diagnostic evidence in glaucoma [6]. Previous research on OCT mainly discusses the RNFL thickness map [2], GCL thickness map [2,9], and cube scan [16]. Combining CFP and OCT for multi-modal diagnosis shows promise in providing accurate diagnostic performance and additional stereo information on retinal structure. However, multi-modal methods on glaucoma recognition have rarely been investigated. Simply combining CFP and the thickness map from the OCT report [1] will waste most information in OCT. An image pair contains merely one CFP image and up to 256 OCT B-scan frames. Interaction between two modalities with unbalanced amounts poses challenges for the existing methods. COROLLA [3] introduces supervised contrastive learning and uses a dual-stream network without modality interaction. Mehta et al. [13] use two different convolutional branches to extract the features of OCT cube and CFP respectively and then ensembles to make the diagnosis. MM-MIL [11] implements two ResNet50 to extract the features from different modalities and adopts Multiple Instance Learning to automatically discover crucial instances. However, the fusion stage is restricted by the Global Average Pooling operation, which limits the spatial information interaction.

To address the problems discussed above, We propose MM-RAF. By following the paradigm of multi-modal learning, i.e., representation, alignment and fusion, we construct three effective modules. To alleviate the semantic gap, BCA introduces bilateral contrastive loss to improve the intra- and inter-modal alignment

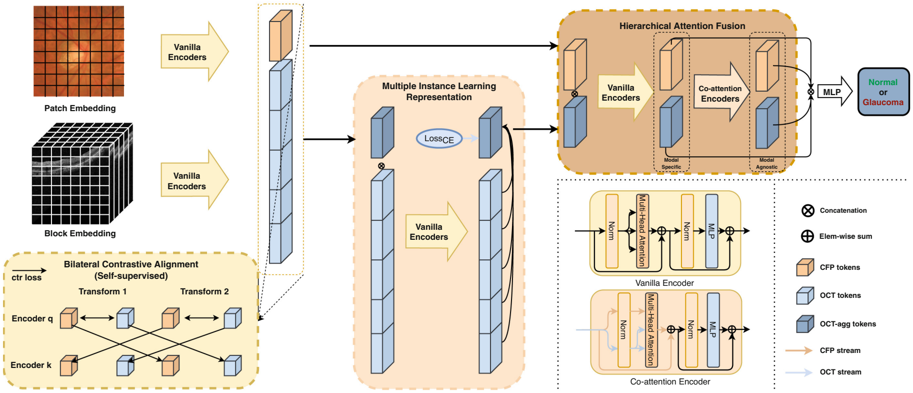

Fig. 2. Overview of our proposed MM-RAF framework.

capability. In the representation stage, MILR extracts the glaucoma-relevant OCT B-scans(instance) from the OCT volume (bag) to assemble a saturated and semantic structure and reduce the scale of the OCT branch. To solve the inability of cross-modality interaction and loss of spatial information, Hierarchical Attention Fusion (HAF) utilizes two strategies to extract modal-specific features and modal-agnostic features to make a final diagnosis. Two classical Transformer encoders are implemented throughout our framework.

Through the experiments on the private dataset, we illustrate that MM-RAF outperforms the existing multi-modal methods on glaucoma recognition. Also, it is demonstrated that MILR and HAF effectively enhance performance in the ablation study. Moreover, extensive experiments on GAMMA dataset [19] prove that the BCA module can promote robustness. We implement relevance-based methods [4] for visualization. The heatmaps illustrate that the framework manifests indistinguishable lesions like RNFLD.

2 Method

As shown in Fig. 2, MM-RAF is a two-phase framework consisting of three modules, i.e., BCA, MILR, and HAF. Inspired by Vision Transformer [7] for modeling long-range dependency, we incorporate two classical transformer encoders throughout our framework to make comprehensive interaction between different modalities. In the contrastive phase, pretraining on unlabeled multi-modal data enables BCA to align the features from different modalities into the same semantic space, diminishing the semantic gap. In the following phase, MILR employs Multiple Instance Learning to refine the cumbersome OCT branch to a semantic structure. Then, with balanced streams from two modalities, HAF renders two cross-modality interaction strategies to make inter- and intra-modal diagnoses. In the following sections, we will clarify each module specifically.

Vanilla Encoder and Co-attention Encoder. Two basic encoders enable effective intra- and inter-modal interaction. As shown in Eqs. 1–2, the Vanilla encoder duplicates the input stream into query, key and value (Q_m, K_m and V_m) components and concentrates on intra-modal interaction, while the Co-attention receives two input streams from different modalities and focuses on inter-modal interaction, with the primary stream acting as query, Q_m, and the subordinate stream replicated as key and value ($K_{\bar{m}}$, $V_{\bar{m}}$). Due to the high computational cost of the self-attention mechanism in the OCT branch, we propose block embedding, partitioning the OCT volume into n OCT blocks and each block will be embedded as $T^k_{O,blk} \in \mathbb{R}^{N\times dim}$($N$ is 196, dim is 768, $k \in \{1,\ldots,n\}$).

$$Vanilla\ Encoder(Q_m, K_m, V_m) = softmax\left(\frac{Q_m K_m^T}{\sqrt{d_k}}\right)V_m \qquad (1)$$

$$Co\text{-}attention\ Encoder(Q_m, K_{\bar{m}}, V_{\bar{m}}) = softmax\left(\frac{Q_m K_{\bar{m}}^T}{\sqrt{d_k}}\right)V_{\bar{m}} \qquad (2)$$

where m and $\bar{m}$ denote different modalities, CFP or OCT in this task. d_k denotes dimension of self-attention.

2.1 Bilateral Contrastive Alignment

The BCA module aims to align the extracted features with the self-supervised strategy before interaction. The semantic gap between CFP and OCT is huge, and direct interaction between different modalities without alignment will lead to a mismatch. ALBEF [10] adopts a similar strategy with the momentum encoders and negative queues, but the weakly-correlated phenomenon and strongly-correlated negative sample in the medical area are so common that ALBEF can hardly reach convergence. To simplify the proxy task, BCA employs the theory of MoCov3 [5] and redesigns the "ctr loss" to adapt to multi-modal tasks. Considering preserving the stereo information of OCT, each block-level token, $T^k_{O\_blk}$, is averaged in the token dimension before being projected. To equally align both branches, 4 projectors are followed symmetrically in both branches to map the tokens to contrastive space. Different augmentation will cause huge discrepancies, especially for OCT images. Therefore, as shown in Fig. 2 and Eq. 3 (Bilateral loss), to mitigate the alignment difficulty for multi-modal tasks, we align both modalities by concentrating on the same modality m with different augmentation $\bar{u}$ and different modalities $\bar{m}$ with the same augmentation u.

$$\mathcal{L} = \textstyle\sum_{m\in\{CFP,OCT\}} \sum_{u=1}^{2} (\mathcal{L}_{inter,m,u} + \mathcal{L}_{intra,m,u}); \qquad (3)$$

$$\mathcal{L}_{intra,m,u} = \mathbb{I}_{u\neq\bar{u}}\ \ ctr(preditor_{m,m,u}(q_{m,u}), k_{m,\bar{u}}); \qquad (4)$$

$$\mathcal{L}_{inter,m,u} = \mathbb{I}_{m\neq\bar{m}} ctr(preditor_{m,\bar{m},u}(q_{m,u}), q_{\bar{m},u}); \qquad (5)$$

where $q_{m,u} \in \mathbb{R}^{dim}$($dim$ is 256 by default) denotes modality m with augmentation u in contrastive space with the query encoder. $k_{m,\bar{u}}$ denote the vectors from the key encoder. $preditor_{m,\bar{m},u}$ denotes the predictor to map modality m to another modality $\bar{m}$. $ctr()$ denotes "ctr loss" [5]. Intuitively, each token behaves like a "query", and BCA aligns the corresponding "values" by projection.

2.2 Multiple Instance Learning Representation

Direct interaction between different modalities with unbalanced amounts is computational-consuming, and the cross-modality relationship is difficult to build. Therefore, we conjecture that the OCT block-level tokens(features) can be formulated as an embedding-level MIL problem in two aspects: (1) In an OCT volume, only certain salient slices are related to glaucoma. (2) High-level embedding features after the BCA module are more distinguishable. In MILR, by defining the i^{th} OCT block, namely $T^{i}_{O,blk}$, as an embedding instance, the integral OCT volume is taken as the bag $\mathbb{B} = \left\{T^{i}_{O,blk} | 1 \leq i \leq n\right\}$. Then we concatenate the OCT bag, $\mathbb{B}$, with an aggregated tokens $T_{O,agg} \in \mathbb{R}^{N \times dim}$. As shown in Fig. 2, several Vanilla encoders render the interaction among $\mathbb{B}$ and $T_{O,agg}$, and eventually, $T^{i}_{O,agg}$ get semantic information from OCT block instances to form a bag prediction. To ensure that MILR aggregates glaucoma-related instances, a supervised signal is incorporated. For efficiency and effectiveness, we only pass semantic $T^{depth}_{O,agg}$ to the subsequent fusion module. MILR can be formulated as:

$$\left\{T^{i+1}_{O,agg} | \mathbb{B}^{i+1}\right\} = Vanilla_{i}\left(T^{i}_{O,agg} \otimes \mathbb{B}^{i}\right), i \in \{1, \ldots, depth\} \tag{6}$$

where $depth$ is the depth of the Vanilla Encoder in MILR.

2.3 Hierarchical Attention Fusion

Before HAF, each modality interacts within its internal modality. To extract modal-specific features and modal-agnostic features respectively, HAF implements a mid-fusion strategy consisting of two fusion stages, Merged-attention and Cross-attention. As shown in Fig. 2, in the Merged-attention blocks, CFP tokens and $T^{depth}_{O,agg}$ will be concatenated, $T_{all} \in \mathbb{R}^{2 * N \times dim}$, to pass through several Vanilla encoders. Except for interaction within their own modality, the salient area will also engage mildly with the other modality, e.g., CFP will leverage intra-modal (CFP) information and inter-modal (OCT) refined knowledge to reinforce the modal-specific features in CFP. Co-attention encoders in the Cross-attention stage render the CFP tokens to interact solely with OCT and thus extract the modal-agnostic features related in both modalities. Eventually, the modal-specific and modal-agnostic features from CFP and OCT will be fed into a projector for joint diagnosis. We will mention the HAF module again in the interpretability experiments in Sect. 3.4.

3 Experiments

3.1 Datasets

For multi-modal glaucoma recognition, the existing public dataset, GAMMA [19] on glaucoma grading, includes **macular** OCT and CFP. GAMMA dataset consists of 100 accessible labeled cases and another 100 unlabeled cases as the benchmark. As the dataset is limited in size for Transformer-based models, we construct a new dataset. 872 multi-modal cases are collected using Topcon Maestro-1 at the outpatient clinic in the Department of Ophthalmology of a state hospital from July 2020 to January 2021. The scan mode is 3D **Optic Disc** with one CFP image and 128 horizontal OCT B-scan images obtained simultaneously. Due to the expensive human annotations, we acquire pseudo labels of CFP by our advanced ensemble model for training. To build a trust-worthy test set for evaluation, we first split the dataset in train/val/test in 6:2:2. A clinician relabels the test set (172 cases) as GON/normal by considering CFP and OCT thickness map. The performance is evaluated on both private test set and GAMMA.

3.2 Experiment Details

Due to the high computational cost, the fixed sampling interval technique is employed to extract 32 OCT images. To avoid over-fitting, we reduce the depths of the encoders to 3 layers. For a fair comparison between all models in this study, we use the following standard setup: initializing with the pre-trained weight of ViT-Base-16 on ImageNet. In the first experiment, the existing multi-modal methods and classical baselines are compared with MM-RAF on the private test set. The baseline includes ResNet, ViT, DeiT [17], and Swin-Transformer [12] (pre-trained weight from timm [18]) with single-modal or early-fusion multi-modal experiments. Multi-modal methods include COROLLA [3], MBT [14]and MM-MIL [11]. The robustness is evaluated on the GAMMA dataset by comparing it with CNNs. The metrics are averaged over three runs. All experiments are implemented in Python 3.7 and Pytorch 1.7 with four NVIDIA TITAN X GPUs and the training configuration is included in Supplementary Material.

3.3 Experimental Results

Compared with Single-Modal and Multi-modal Solutions. As shown in Table 1, ResNet50 for CFP attains the best AP score in single modality, proving that CFP is more sensitive than OCT for glaucoma diagnosis. Besides, the transformer-based method is inferior to ResNet in CFP but surpasses CNN in OCT modality, indicating that the transformer-based methods need sufficient data to learn inductive bias. MM-ViT outperforms CFP-ViT and OCT-ViT, exemplifying that Transformer can benefit from multi-modal learning. Our framework, MM-RAF, outperformed MM-ViT with a 6% improvement in AP score and achieved **SOTA** with F1, AP, and AUC metrics in this study.

Table 1. Results of baseline and existing works comparison in the private test set. In each column, bold text denotes the best results.

lMethod		Metrics				
		Sen	Spe	F1	AP	AUC
Single-modal	CFP-ResNet50	**0.9333**	0.8078	0.8515	0.8925	0.9584
	CFP-ViT	0.6286	0.9756	0.7565	0.8483	0.9585
	CFP-DeiT	0.7348	0.9512	0.8291	0.8531	0.9512
	CFP-Swin-Transformer	0.6335	0.892	0.7729	0.8212	0.9241
	OCT-ResNet50	0.4571	0.9854	0.5994	0.8605	0.9363
	OCT-ViT	0.7333	**0.9854**	0.8397	0.8892	0.944
	OCT-DeiT	0.6286	0.9756	0.7565	0.8897	0.9585
	OCT-Swin-Transformer	0.6111	0.9281	0.7309	0.7925	0.911
Multi-modal	MM-ViT	0.8334	0.8873	0.8579	0.8983	0.934
	MM-DeiT	0.7904	0.9513	0.8629	0.8982	0.9514
	MM-Swin-Transformer	0.6111	0.7961	0.6853	0.5771	0.7639
	MM-MIL [11]	0.781	0.9416	0.8453	0.8837	0.9608
	MBT [14]	0.6667	0.9640	0.7859	0.8287	0.9384
	COROLLA [3]	0.6667	0.9376	0.7735	0.8318	0.942
	MM-RAF	0.9238	0.9027	**0.9081**	**0.9584**	**0.9855**

Table 2. Robustness experiments on GAMMA dataset. The glaucoma grading task (normal/early-glaucoma/progressive-glaucoma) is evaluated by *kappa*.

Method	*kappa*
Dual-ResNet50	0.7352
MM-MIL w/o transfer learning	0.8502
MM-MIL w/ transfer learning	**0.8562**
Ours w/o transfer learning, BCA	0.5289
Ours w/ transfer learning; w/o BCA	0.6277
Ours w/ BCA; w/o transfer learning	0.8072
Ours w/ transfer learning, BCA	**0.8467**

Robustness. Due to the limited size of GAMMA(100 cases), our transformer-based method is likely to get overfitting if training from scratch. To this end, pre-training on our private mid-scale dataset which captures **optic disc** OCT can gain enhancements even when transferring to the cross-domain dataset (GAMMA scans **macular** area). Cohen's *kappa* coefficient is implemented as metrics. As shown in Table 2, our method has better cross-domain generalization after applying BCA and transfer learning strategy. Furthermore, MM-RAF achieves results comparable to CNNs, highlighting our framework's robustness to learn inductive biases from images even with a limited dataset. When providing more domain-related data, our approach has the potential to perform marginally better than CNNs for compensating the lack of inductive bias.

Ablation Study. The ablation study on the private dataset examines the contribution of three modules, the order of HAF, and the depth of each module. From Table.3, MILR and HAF modules bring 0.03 and 0.02 AP increases, respec-

Table 3. Ablation study on private dataset. HAF-R denotes to reverse the order of two stages in HAF modules. Depth controls the encoder depth in all modules.

BCA	MILR	HAF-Merged	HAF-Cross	HAF-R	depth	Sen	Spe	F1	AP	AUC
		✓			3	0.7714	0.8759	0.8195	0.8348	0.9311
		✓	✓		3	0.8476	0.9537	0.8938	0.9275	0.9718
	✓	✓	✓		3	0.8473	**0.9803**	0.9074	0.9554	**0.9862**
✓	✓	✓	✓	✓	3	0.6574	0.904	0.7147	0.8261	0.9401
✓	✓	✓	✓		1	0.8381	0.9538	0.8906	0.9187	0.9678
✓	✓	✓	✓		6	0.8095	0.9732	0.8803	0.938	0.9793
✓	✓	✓	✓		3	**0.9238**	0.9027	**0.9081**	**0.9584**	0.9855

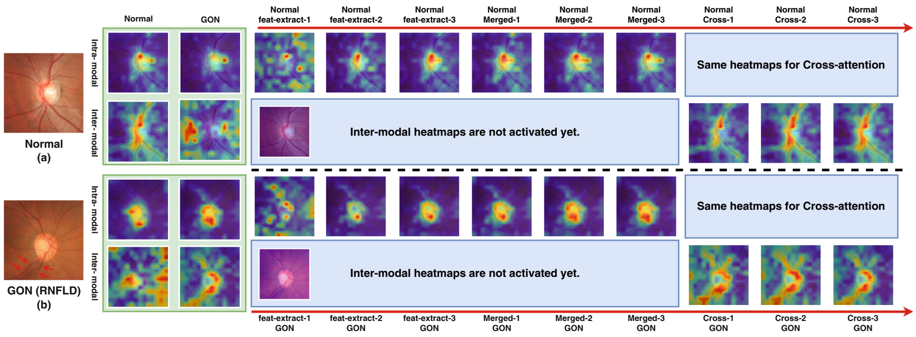

Fig. 3. Visualization. Case (a): Normal; Case (b): GON (Glaucomatous optic neuropathy) with RNFLD. Each column denotes a different class decision or a different stage in our framework. "Merged" denotes Merged-attention, "Cross" denotes Cross-attention. It is recommended to zoom in to view this figure.

tively. Reversing the order of the HAF module brings a decrease, which indicates that the modal-agnostic features should be extracted after the Merged-attention.

3.4 Visualization

For visualization, we employ a class-dependent relevance-based method [4] that captures inter- and intra-modal relevance. Since MILR has aggregated the OCT into high-level $T_{O,agg}^{depth}$ features which are complex to visualize, we choose CFP images to interpret the mechanism of how different modalities interact. For each case presented in Fig. 3, intra-modal and inter-modal heatmaps are calculated by CFP tokens and OCT tokens, respectively. In intra-modal maps, the salient area is centralized on the optic disc, while the inter-modal maps are sensitive to the temporal region where OCT can provide fine-grain stereo information of the optic nerve, indicating that our framework incorporates multi-modal information. The sparsely distributed situation, e.g., case(a) GON's inter-modal view, may be attributed to the lack of significant lesions in the image, causing a random selection of tokens. Also, we visualize how the framework considers the correct prediction with the network going deeper. In case (b), deeper layers concentrate intra- and inter-modal features on lesion areas. The intra-modal Merged-attention focuses on the optic disc(modal-specific feature), and the inter-modal Cross-attention maps are more sensitive to the temporal region (modal-agnostic feature), demonstrating the effectiveness of HAF in extracting and combining modal-agnostic and modal-specific features to make the multi-modal decision.

4 Conclusion

The challenges in multi-modal glaucoma recognition include the huge discrepancies, the unbalanced amounts, and the lack of spatial information interaction between different modalities. To this end, we propose MM-RAF, a pure self-attention multi-modal framework consisting of three modules dedicated to the problems: BCA fills the semantic gap between CFP and OCT and promotes robustness. MILR and HAF complete semantic aggregation and comprehensive relationship probing with better performance. While MM-RAF outperforms other solutions in multi-modal glaucoma recognition, the performance can be further improved with sufficient data. Our next direction is to utilize a lightweight transformer to leverage more information from both modalities. Besides, addressing the issue of uncertainty measurement and preventing the bias of any specific modality from influencing the overall decision in the multi-modal recognition scenario is crucial, especially when diagnosing glaucoma using OCT for its limited specificity. Cross-modal uncertainty measurement is also our further research direction.

References

1. An, G., et al.: Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. J. Healthcare Eng. **2019** (2019)
2. Asaoka, R., et al.: Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. Am. J. Ophthalmol. **198**, 136–145 (2019)
3. Cai, Z., Lin, L., He, H., Tang, X.: Corolla: an efficient multi-modality fusion framework with supervised contrastive learning for glaucoma grading. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–4. IEEE (2022)
4. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 397–406 (2021)
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9640–9649 (2021)
6. Ding, F., Yang, G., Ding, D., Cheng, G.: Retinal nerve fiber layer defect detection with position guidance. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 745–754. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_72
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)

8. Harizman, N., et al.: The isnt rule and differentiation of normal from glaucomatous eyes. Arch. Ophthalmol. **124**(11), 1579–1583 (2006)
9. Lee, J., Kim, Y.K., Park, K.H., Jeoung, J.W.: Diagnosing glaucoma with spectral-domain optical coherence tomography using deep learning classifier. J. Glaucoma **29**(4), 287–294 (2020)
10. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: vision and language representation learning with momentum distillation. Adv. Neural. Inf. Process. Syst. **34**, 9694–9705 (2021)
11. Li, X., et al.: Multi-modal multi-instance learning for retinal disease recognition. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2474–2482 (2021)
12. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
13. Mehta, P., et al.: Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. Am. J. Ophthalmol. **231**, 154–169 (2021)
14. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Adv. Neural. Inf. Process. Syst. **34**, 14200–14213 (2021)
15. Raghavendra, U., Bhandary, S.V., Gudigar, A., Acharya, U.R.: Novel expert system for glaucoma identification using non-parametric spatial envelope energy spectrum with fundus images. Biocybernetics Biomed. Eng. **38**(1), 170–180 (2018)
16. Ran, A.R., et al.: Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis. The Lancet Digital Health **1**(4), e172–e182 (2019)
17. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers; distillation through attention. In: International Conference on Machine Learning, vol. 139, pp. 10347–10357, July 2021
18. Wightman, R.: Pytorch image models (2019). https://github.com/rwightman/pytorch-image-models. https://doi.org/10.5281/zenodo.4414861
19. Wu, J., et al.: Gamma challenge: Glaucoma grAding from Multi-Modality imAges (2022)