Robust Multimodal Learning for Ophthalmic Disease Grading via Disentangled Representation

Xinkun Wang[1*], Yifang Wang[1*], Senwei Liang[1*], Feilong Tang[2,1†], Chengzhi Liu[3], Ming Hu[2], Chao Hu[4], Junjun He[5], Zongyuan Ge[2✉], and Imran Razzak[1✉]

[1] MBZUAI, United Arab Emirates
[2] Monash University, Australia
[3] Liverpool University, United Kingdom
[4] China Unicom (Shanghai) Industrial Internet Co., Ltd., China
[5] Shanghai AI Lab, China
Email: (imran.razzak@mbzuai.ac.ae; zongyuan.ge@monash.edu)

Abstract. Ophthalmologists often rely on multimodal data to improve diagnostic precision. However, data on complete modalities are rare in real applications due to a lack of medical equipment and data privacy concerns. Traditional deep learning approaches usually solve these problems by learning representations in latent space. However, we highlight two critical limitations of these current approaches: *(i)* Task-irrelevant redundant information existing in complex modalities (*e.g.,* massive slices) leads to a significant amount of redundancy in latent space representations. *(ii)* Overlapping multimodal representations make it challenging to extract features that are unique to each modality. To address these, we introduce the **E**ssence-Point and **D**isentangle **R**epresentation **L**earning (**EDRL**) strategy that integrates a self-distillation mechanism into an end-to-end framework to enhance feature selection and disentanglement for robust multimodal learning. Specifically, Essence-Point Representation Learning module selects discriminative features that enhance disease grading performance. Moreover, the Disentangled Representation Learning module separates multimodal data into modality-common and modality-unique representations, reducing feature entanglement and enhancing both robustness and interpretability in ophthalmic disease diagnosis. Experiments on ophthalmology multimodal datasets demonstrate that the proposed EDRL strategy outperforms the state-of-the-art methods significantly. Code is available at GitHub Repository.

Keywords: Missing Modality · Multi Modality · Ophthalmic Disease

1 Introduction

In recent years, using multimodal data sources has become a common method to enhance diagnostic accuracy for ophthalmic diseases [25,22,8]. In these methods, Optical Coherence Tomography (OCT) and Retinal Fundus Imaging are

[*]These authors contribute equally; [†] Project leader; [✉]Corresponding author

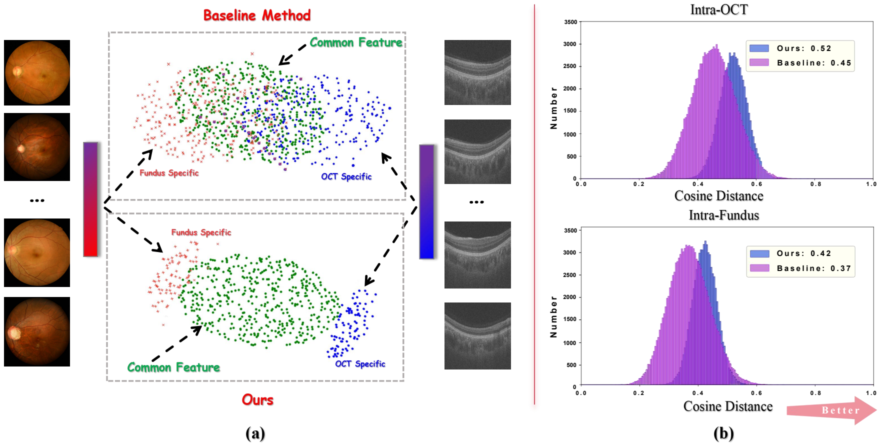


Fig. 1: Overview of feature representation analysis. Baseline methods employ Vision Transformer [3] to extract and concatenate features from both modalities. (a) t-SNE [12] visualization illustrates the distribution of modality-specific and modality-common features, comparing a baseline method with our strategy. (b) Cosine distance quantifies feature separability by measuring how effectively feature from different samples are distinguished within each modality.

typically used modalities [14,13]. Existing methods primarily focus on modality feature fusion, employing spatial and channel attention [30,26,20] or evidence fusion models with the inverse gamma prior distribution [31].

Although numerous representation learning methods have been developed to address missing modality scenarios, two major issues still exist. **(1) Task-irrelevant Redundant Information**: In the absence of precise annotations, such as patch-wise labeling for regions affected by ophthalmic diseases in fundus and OCT images [7,15], feature representations often contain both task-relevant and irrelevant information relevant to the task [6,21,19,18]. As shown in Fig. 1 (b), the baseline method exhibits lower cosine distance between distinct samples, indicating an insufficient ability to capture distinguishable features and leading to lower grading performance [17,23]. **(2) Overlapping multimodal representations:** Most methods [2,28,29] that focus on cross-modality common representation extraction lead to feature representations of different modalities that have a substantial amount of cross-modal shared information. As shown in Fig. 1 (a), there exists a significant overlap between features of different modalities, hindering the model from utilizing the modality-unique information for diagnosis [24,27].

To this end, we propose the **E**ssence-point and **D**isentangle **R**epresentation **L**earning (EDRL) framework. The Essence-point Representation Learning (EPRL) module identifies essence-points that highlight discriminative information within each modality, reducing task-irrelevant redundancy. For feature disentanglement, the Disentangle Representation Learning (DiLR) module decomposes embed-

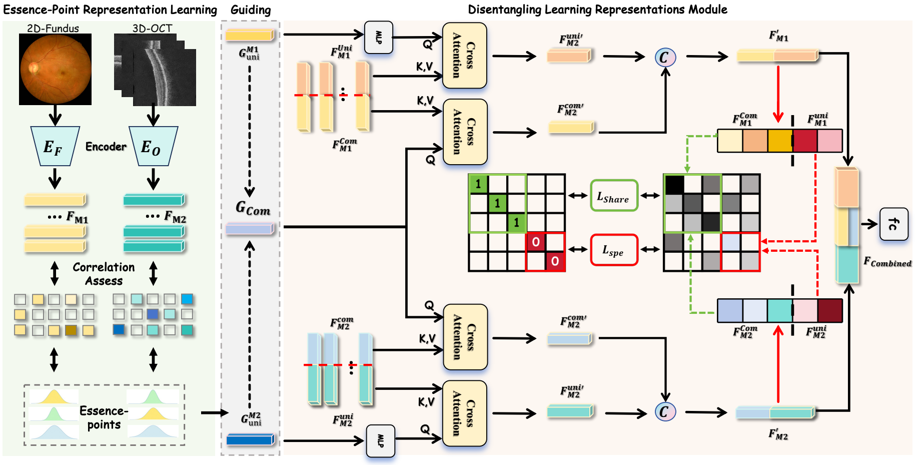


Fig. 2: Illustrates our proposed **EDRL** framework, comprising two key modules: **EPRL** and **DiLR**. The EPRL module maintains series of essence-points to extract discriminative features (e.g., F_{M1} and F_{M2}) from each modality. The DiLR module disentangles these features into independent modality-common (F_{Com}) and modality-unique (F_{Uni}) representations, leveraging attention mechanisms to align shared information while preserving modality-specific characteristics. F_{Com} and F_{Uni} are then concatenated into ($F_{Combined}$) for grading tasks.

dings into modality-common and modality-unique parts. It encourages cross-correlation alignment toward identity for shared features while minimizing correlation across unique components. We also apply self-distillation between two pipelines (complete vs. missing modalities), where the complete pipeline guides missing modality reconstruction to enhance robustness. EDRL thus minimizes redundancy, reduces inter-modality overlap, and improves multimodal discrimination and generalization.

Overall, our contributions are threefold. *(i)* We propose EPRL framework for discriminative instance selection with self-distillation. *(ii)* We introduce DiLR to disentangle features into modality-unique and modality-common representations. *(iii)* We demonstrate effectiveness on three ophthalmology datasets.

2 Methods

2.1 Problem Formulation

We represent $A = \{\mathbf{a}_j, b_j\}_{j=1}^{K}$ as a multimodal dataset with K patient samples. Each ophthalmological sample $\mathbf{a}_j$ consists of L inputs from different modalities, written as $\mathbf{a}_j = \{\mathbf{a}_j^l\}_{l=1}^{L}$, where L denotes the number of modalities and $b_j \in \{1, 2, \ldots, D\}$ is the label for $\mathbf{a}_j$, with D being the number of grading categories.

We propose an EDRL framework addressing missing modalities through: (1) Inter-modality missing (complete modality absence) and (2) Intra-modality missing (natural noise addition). To reduce task-irrelevant redundancy, we introduce EPRL for task-relevant selection. For overlapped representations, we propose DiLR to generate independent modality-unique and modality-common features. The framework is shown in Fig. 2.

2.2 EPRL: Essence-Point Representation Learning

We propose EPRL to filter out information in the feature map that is indiscriminative to the ohthalmic disease grading task. Since the task-discriminative information follows conditional distributions given modality type m and class label c, EPRL maintains $m \times c$ learnable essence-points E_m^c for each m and c, aiming to model discriminative information distribution given m and c. To guide essence-point learning during training process, we need to match these essence-points with the feature representation based on m and c. Such process can be implemented by the matching loss function L_{Matching}. For each modality m, the loss encourages the feature representation F_M^c to be aligned with their corresponding essence-points E_m^c, while simultaneously minimizing their similarity with essence-points from other classes. Suppose N is the batch size and K is the total number of classes, L_{Matching} with cosine similarity is defined as:

$$L_{\text{Matching}} = -\frac{1}{B}\sum_{i=1}^{N}\left(\text{Sim}(\mathbf{F}_M^c, \mathbf{E}_M^c) - \frac{1}{2K-1}\sum_{j\neq c}^{2K-1}\text{Sim}(\mathbf{F}_M^c, \mathbf{E}_M^j)\right). \quad (1)$$

During the inference, due to the lack of guidance by the label, EPRL will conduct the correlation assessment and select the highest similarity essence-point.

The ***Guiding*** process aims to generate guiding tokens G_{uni}^M that direct the multi-modal representations $\{F_{M1}, F_{M2}\}$ to focus on task-relevant regions while eliminating unrelated information. Assuming that the essence-points follow a Gaussian distribution in each modality, we first employ an MLP to predict the mean and variance of the distributions for the essence-points in label c, denoted as N_{oct}^c and N_{fundus}^c. The guiding tokens G_{uni}^{M1} and G_{uni}^{M2} are sampled from them, respectively. Subsequently, to obtain the cross-modality shared representation, we use the Product-of-Experts [5] to generate the joint distribution N_{Joint}^c based on the two individual distributions N_{oct}^c and N_{fundus}^c by assuming independence. Then, guiding token G_{com} is randomly sampled from N_{Joint}^c.

2.3 DiLR: Disentangling Learning Representations Module

To decouple the representation into independent modality-unique and modality-common features, we introduce the DiLR module. We first decompose the feature embeddings in EPRL $\mathbf{F}_{M1}, \mathbf{F}_{M2} \in \mathbb{R}^D$ into two distinct parts: $\mathbf{F}_M^{com} \in \mathbb{R}^{D_c}$, $\mathbf{F}_M^{uni} \in \mathbb{R}^{D_u}$, where $D_c + D_u = D$. We assume D_c represents the common features across the modalities, while D_u captures the modality-specific features. Subsequently, the guiding tokens G_{uni}^{M1}, G_{uni}^{M2}, and G_{com} from EPRL are used to instruct the task-discriminative information selection in F_{M1} and F_{M2} through cross-attention. Its output, with task-unrelated information removed, $\mathbf{F}_{M1}^{com'}$ and $\mathbf{F}_{M2}^{com'}$, should remain highly similar, while $\mathbf{F}_{M1}^{uni'}$ and $\mathbf{F}_{M2}^{uni'}$ are expected to be decorrelated from each other.

With this in mind, we measure the similarity of two embeddings $\mathbf{F}_{M1}, \mathbf{F}_{M2} \in \mathbb{R}^D$ through the corrleation matrix:

$$c_{ij} = \frac{\sum_b \mathbf{F}_{M1,b,i}\mathbf{F}_{M2,b,j}}{\sqrt{\sum_b (\mathbf{F}_{M1,b,i})^2}\sqrt{\sum_b (\mathbf{F}_{M2,b,j})^2}}, \tag{2}$$

where b indexes batch samples, and i, j indexes the dimension of the embeddings. $\mathbf{C}_{ij} \in \mathbb{R}^{D \times D}$ is a square matrix with values ranging from -1 to 1. In $\mathbf{C}_{ij}$, we select the submatrix $\mathbf{C}_{com} \in \mathbb{R}^{D_c \times D_c}$ that only utilizes the common dimensions from $\mathbf{F}_{M1}$ and $\mathbf{F}_{M2}$ to denote the similarity between two common features $\mathbf{F}_{M1}^{com}$ and $\mathbf{F}_{M2}^{com}$. Since $\mathbf{F}_{M1}^{com}$ and $\mathbf{F}_{M2}^{com}$ should remain high in similarity, $\mathbf{C}_{com}$ should approach the identity matrix. $\mathbf{C}_{uni}$ is expected to approximate a target matrix with zero diagonal conversely. Thus, the common loss and unique loss are respectively defined as:

$$L_{com} = \sum_i (1 - c_{cii})^2 + \lambda_c \cdot \sum_i \sum_{j \neq i} c_{cij}^2, \tag{3}$$

$$L_{uni} = \sum_i c_{uii}^2 + \lambda_u \cdot \sum_i \sum_{j \neq i} c_{uij}^2. \tag{4}$$

To calculate these losses, we design a realignment network. $\mathbf{F}_M^{Uni}$ conducts a self-attention process to extract finer-grained features. An average operation is then employed to squeeze $\mathbf{F}_M^{Uni}$. For extracting the common information from both modalities, we utilize the shared features sampled from EPRL network as the guiding token (query), while $\mathbf{F}_{M1}^{com}$ and $\mathbf{F}_{M2}^{com}$ serve as key and value for two cross-attention modules respectively to allow the model to extract task-related common features. Subsequently, $\mathbf{F}_M^{Uni}$ and $\mathbf{F}_M^{Com}$ are concatenated as F_{M1} and F_{M2} for further computation of the correlation matrix and its loss. F_{M1} and F_{M2} are concatenated to form a combined feature $F_{Combined}$.

2.4 Unified Self-Distillation Mechanism

Specifically, feature-level and logits-level consistency are employed to guide the model towards generating more accurate representations for incomplete modalities. For feature distillation, we employ Maximum Mean Discrepancy loss to minimize the discrepancy between combined features $F_{combine}^{miss}$ and $F_{combine}^{complete}$.

$$L_{\text{features}} = \frac{1}{B} \sum_{j=1}^{b} \hat{D}_T(F_{combine}^{miss}, F_{combine}^{complete}), \tag{5}$$

where

$$D_T(x, y) \triangleq \|\mathbb{E}_x[\varphi(X_1)] - \mathbb{E}_y[\varphi(X_2)]\|_T^2, \tag{6}$$

where $\varphi(\cdot)$ is a feature transformation, and T is the Reproducing Kernel Hilbert Space [1,16]. For logits distillation, we apply Jensen-Shannon (JS) divergence [9] to minimize the difference between the logits of different modality-missing cases:

$$D_{\mathrm{JS}}(p_1 \parallel p_2) = \frac{1}{2}\left(D_{\mathrm{KL}}(p_1 \parallel q) + D_{\mathrm{KL}}(p_2 \parallel q)\right), \tag{7}$$

where q represents the average distribution of the logits, and the corresponding logits distillation loss is:

$$L_{\text{logits}} = D_{\mathrm{JS}}(MLP(F^1_{Combibed}) \parallel MLP(F^2_{Combined})). \tag{8}$$

3 Experiment

3.1 Datasets

We evaluate the proposed framework on three public multimodal datasets from Harvard-30k [11]: Harvard-30k AMD, DR, and Glaucoma, which focus on Age-related Macular Degeneration (AMD), Diabetic Retinopathy (DR), and Glaucoma. The datasets provide four-class grading for AMD and two-class grading for DR and Glaucoma, with fundus images of size 448 × 448 and OCT images of size 200 × 256 × 256 (200 OCT slices).

We compare our model with three state-of-the-art multi-modality fusion methods, as shown in Table 1. For baseline, we use Vision Transformer [3] and UNETR [4] as backbones for Fundus and OCT, respectively, and directly concatenate their feature maps for classification. Compared methods include: (1) B-IF (early fusion); (2) M^{2}LC [26], combining channel and spatial attention; and (3) IMDR [10], which uses mutual information loss for cross-modality decoupling. Evaluations are conducted under three conditions: (1) complete modality, (2) noisy modality, and (3) missing modality.

Complete Modality and noisy modality Setting. In the ideal scenario without any missing or noise, our model achieves the best performance among the models we test. Building upon this, we also test our approach under conditions where various Gaussian noise with different variance is introduced to each modality (In Fig 3). As the noise level increases, a clear performance decline is observed in all models, emphasizing the challenges posed by data loss within a single modality on the stability of multimodal representations. Despite this, our method demonstrates exceptional robustness, particularly in scenarios with high levels of noise, consistently outperforming the other models.

Inter-Modality completely missing. We evaluate our strategy by comparing its performance with that of the other methods under OCT missing or Fundus missing situations. Even a performance decline is observed across all models when a modality is missing, our strategy demonstrates greater robustness. Result proves our strategy has robust ability to separate multimodal features and reconstruct the missing information to serve for the grading task.

Table 1: Our model is benchmarked against existing methods on the Harvard-30k dataset across three conditions: OCT missing, Fundus missing, and complete modality. The top-performing results are emphasized in bold and highlighted.

Method	**Dataset**	**AMD**			**DR**			**Glaucoma**		
	Modality	**OCT**	**Fundus**	**Both**	**OCT**	**Fundus**	**Both**	**OCT**	**Fundus**	**Both**
Baseline	**ACC**	65.07	72.92	70.87	70.53	73.81	74.07	65.69	73.02	73.35
	AUC	69.88	75.38	81.06	69.94	79.11	78.73	69.86	75.35	74.53
	F1	69.64	72.28	70.83	62.01	70.46	71.17	70.91	72.31	71.64
B-IF	**ACC**	69.57	72.35	73.17	69.05	73.62	76.36	69.64	73.39	73.39
	AUC	70.14	71.98	83.82	65.25	67.50	77.95	68.95	76.61	73.32
	F1	67.45	70.03	71.25	67.93	69.68	75.61	67.18	72.47	72.11
M^{2}LC	**ACC**	68.97	73.24	74.93	67.20	73.04	75.21	67.70	72.78	74.98
	AUC	72.23	72.67	82.39	65.05	67.89	79.68	71.22	70.23	76.45
	F1	65.06	73.80	71.20	64.33	74.59	74.39	65.60	71.11	74.23
IMDR	**ACC**	70.62	75.17	79.50	72.62	76.19	78.57	71.16	75.54	77.31
	AUC	72.69	80.48	85.09	74.69	79.07	85.00	75.07	78.47	78.98
	F1	71.90	76.59	72.52	72.90	72.18	77.04	70.37	75.12	78.90
Ours	**ACC**	**71.79**	**76.69**	**81.42**	**74.38**	**77.50**	**79.50**	**72.53**	**76.28**	**78.55**
	AUC	**74.84**	**81.55**	**85.82**	**76.88**	**80.60**	**86.71**	**76.28**	**79.59**	**79.32**
	F1	**72.94**	**76.79**	**78.93**	**74.28**	**76.71**	**79.81**	**72.54**	**76.62**	**80.54**

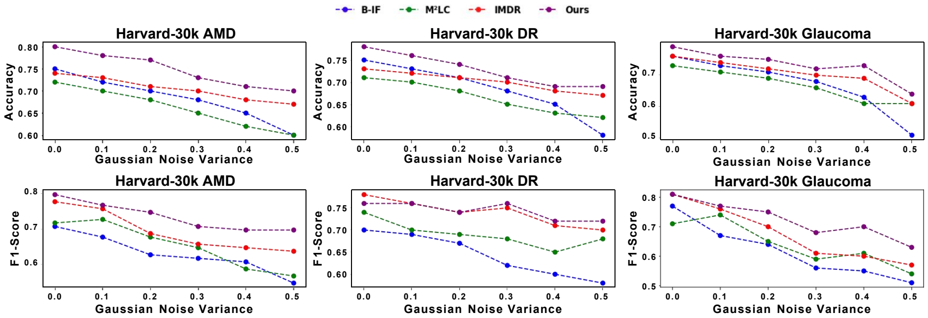


Fig. 3: A comprehensive evaluation of performance across different missing data rates within the context of intra-modality incompleteness.

3.2 Ablation Study

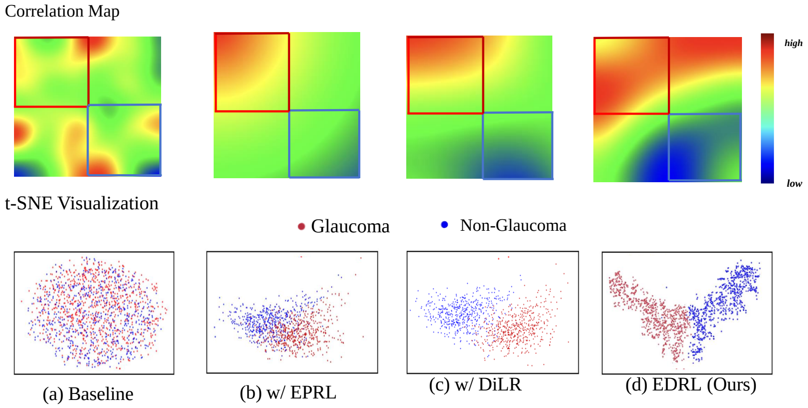
Effectiveness of each component. To assess the effectiveness of EPRL and DiLR, we conducted an ablation study on the Harvard-30k test set with Gaussian noise (variance = 0.5), as shown in Table 2. From Variant I to II, adding EPRL reduces task-irrelevant information and notably improves accuracy. From Variant I to III, DiLR enhances modality disentanglement, boosting accuracy by 5%. Variant IV, integrating both modules, consistently outperforms II and III, high-

Table 2: Baseline: Using transformer backbone to extract two modality data and simply concatenates their features. EPRL: Our Essence-point Representation Learning. DiLR: Our Disentangling Learning Representations.

Variants	Baseline	EPRL	DiLR	**ACC**	**AUC**	**F1**
I	✓			59.51	63.42	53.47
II	✓	✓		66.58	70.95	66.19
III	✓		✓	64.67	66.45	64.88
IV	✓	✓	✓	**69.37**	**66.39**	**57.94**

Table 3: Implementation of a comprehensive hyperparameter sensitivity analysis within the full-modality framework of the Harvard-30k dataset. Percentage *(p)*: the ratio of common dimensions to total dimensionality.

(p)	0.3	0.4	0.5	0.6	0.7
AMD	79.56	81.42	79.05	80.23	73.47
DR	76.87	78.13	78.75	79.50	77.50
Glaucoma	77.32	78.55	77.44	76.35	77.50



(a) Baseline (b) w/ EPRL (c) w/ DiLR (d) EDRL (Ours)

Fig. 4: Corrleation map and t-SNE on the Harvard-30k Glaucoma dataset. In the ideal scenario, the top-left region of the heatmap should exhibit predominantly red areas, indicating a high correlation between F_{M1}^{com} and F_{M2}^{com}, while the bottom-right region should show more blue areas, signifying lower correlation between F_{M1}^{uni} and F_{M2}^{uni}

.

lighting their complementary strengths in learning decoupled, low-redundancy representations.

Qualitative Results. As shown in Fig. 4, we visualize correlation maps and t-SNE plots for four variants to evaluate feature disentanglement and clustering. The baseline (Fig. 4 (a)) shows weak decoupling and poor cluster separation. Adding EPRL (Fig. 4 (b)) improves feature selection and cluster quality. Incorporating DiLR (Fig. 4 (c)) further disentangles modality-common and unique features, enhancing separation. Our full EDRL model (Fig. 4 (d)) achieves clear modality disentanglement and distinct clusters, validating its effectiveness in learning discriminative, modality-aware representations for grading.

Hyperparameter Sensitivity Analysis. To validate the robustness of our model, we conduct a series of hyperparameter sensitivity analysis in Table 3. In DiLR, the common dimension percentage affects performance: increasing it initially improves results, but excessive sharing impairs modality-specific information expression, causing performance decline.

4 Conclusion

In multimodal ophthalmology diagnosis, two main challenges are intra-modal redundancy due to task-unrelated information and cross-modal entanglement in the latent space. To tackle these, we propose the EPRL framework to reduce redundancy, followed by the DiLR module for disentangling cross-modal features. Extensive experiments on multimodal ophthalmic datasets show that our method outperforms state-of-the-art approaches, improving interpretability.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Berlinet, A., Thomas, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers (2004)
2. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: ICCV (2021)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020), https://arxiv.org/abs/2010.11929
4. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: WACV (2022)
5. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural computation **14**(8), 1771–1800 (2002)
6. Hosseini, M.S., Ehteshami Bejnordi, B., Trinh, V.Q.H., Hasan, D., Li, X., Kim, T., Zhang, H., Wu, T., Chinniah, K., Maghsoudlou, S., et al.: Computational pathology: A survey review and the way forward. arXiv preprint arXiv:2304.05482 (2023), https://arxiv.org/abs/2304.05482
7. Huang, D., Swanson, E., Lin, C., Schuman, J., Stinson, W., Chang, W., Hee, M., Flotte, T., Gregory, K., Puliafito, C., et al.: Optical coherence tomography. Science **254**(5035), 1178–1181 (1991)
8. Lam, C., Wong, Y.L., Tang, Z., Hu, X., Nguyen, T.X., Yang, D., Zhang, S., Ding, J., Szeto, S.K., Ran, A.R., et al.: Performance of artificial intelligence in detecting diabetic macular edema from fundus photography and optical coherence tomography images: a systematic review and meta-analysis. Diabetes Care (2024)

9. Li, M., Yang, D., Lei, Y., Wang, S., Wang, S., Su, L., Yang, K., Wang, Y., Sun, M., Zhang, L.: A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In: AAAI (2024)
10. Liu, C., Huang, Z., Chen, Z., Tang, F., Tian, Y., Xu, Z., Luo, Z., Zheng, Y., Meng, Y.: Incomplete modality disentangled representation for ophthalmic disease grading and diagnosis. AAAI (2025)
11. Luo, Y., Tian, Y., Shi, M., Elze, T., Wang, M.: Eye fairness: A large-scale 3d imaging dataset for equitable eye diseases screening and fair identity scaling (2024), https://openreview.net/forum?id=Lv9KZ5qCSG
12. der Maaten, L.V., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11), 2579–2605 (2008), http://jmlr.org/papers/volume9/VDMaaten08a/VDMaaten08a.pdf
13. Meleppat, R., Roonning, K., Karlen, S., Burns, M., Pugh, E.N., J., Zawadzki, R.: In vivo multimodal retinal imaging of disease-related pigmentary changes in retinal pigment epithelium. Scientific Reports **11**(1), 16252 (2021). https://doi.org/10.1038/s41598-021-95756-3, https://doi.org/10.1038/s41598-021-95756-3
14. Mleppat, R., Zhang, P., Ju, M., Manna, S., Jian, Y., Pugh, E., Zawadzki, R.: Directional optical coherence tomography reveals melanin concentration dependent scattering properties of retinal pigment epithelium. Journal of Biomedical Optics **24**(6), 066011 (2019). https://doi.org/10.1117/1.JBO.24.6.066011, https://doi.org/10.1117/1.JBO.24.6.066011
15. Müller, P., Wolf, S., Dolz-Marco, R., Tafreshi, A., Schmitz-Valckenberg, S., Holz, F.: Ophthalmic diagnostic imaging: retina, pp. 87–106. Springer (2019)
16. Okutmustur, B.: Reproducing Kernel Hilbert Spaces. Master's thesis, Bilkent University (August 2005), http://www.thesis.bilkent.edu.tr/0002953.pdf
17. Rippel, O., Paluri, M., Dollar, P., Bourdev, L.: Metric learning with adaptive density discrimination. arXiv preprint arXiv:1511.05939 (2015)
18. Tang, F., Huang, Z., Liu, C., Sun, Q., Yang, H., Lim, S.N.: Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In: ICLR (2025)
19. Tang, F., Liu, C., Xu, Z., Hu, M., Huang, Z., Xue, H., Chen, Z., Peng, Z., Yang, Z., Zhou, S., Li, W., Li, Y., Song, W., Su, S., Feng, W., Su, J., Lin, M., Peng, Y., Cheng, X., Razzak, I., Ge, Z.: Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In: CVPR (2025)
20. Tang, F., Xu, Z., Qu, Z., Feng, W., Jiang, X., Ge, Z.: Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In: CVPR (2024)
21. Udandarao, V., Gupta, A., Albanie, S.: Sus-x: Training-free name-only transfer of vision-language models. In: ICCV (2023)
22. Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., Carneiro, G.: Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In: MICCAI (2023)
23. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
24. Wang, Y., Albrecht, C.M., Braham, N.A.A., Liu, C., Xiong, Z., Zhu, X.X.: Decoupling common and unique representations for multimodal self-supervised learning. In: ECCV (2024)
25. Watanabe, T., Hiratsuka, Y., Kita, Y., Tamura, H., Kawasaki, R., Yokoyama, T., Kawashima, M., Nakano, T., Yamada, M.: Combining optical coherence tomography and fundus photography to improve glaucoma screening. Diagnostics **12**(5), 1100 (2022)

26. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV (2018)
27. Xiong, Z., Yuan, Y., Wang, Q.: Ask: Adaptively selecting key local features for rgb-d scene recognition. TIP (2021)
28. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: ICCV (2023)
29. Xue, H., Tang, F., Liu, C., Hu, M., Huang, Z., Chen, Z., Peng, Z., Yang, Z., Zhou, S., Li, W., Li, Y., Song, W., Su, S., Feng, W., Su, J., Lin, M., Peng, Y., Cheng, X., Razzak, I., Ge, Z.: Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation (2025)
30. Zheng, J., Liu, H., Feng, Y., Xu, J., Zhao, L.: Casf-net: Cross-attention and cross-scale fusion network for medical image segmentation. Computer Methods and Programs in Biomedicine **229**, 107307 (2023)
31. Zou, K., Lin, T., Han, Z., Wang, M., Yuan, X., Chen, H., Zhang, C., Shen, X., Fu, H.: Confidence-aware multi-modality learning for eye disease screening. MIA (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103214