CP-CLIP: Core-Periphery Feature Alignment CLIP for Zero-Shot Medical Image Analysis

Xiaowei Yu[1], Zihao Wu[2], Lu Zhang[1], Jing Zhang[1], Yanjun Lyu[1], and Dajiang Zhu[1]

[1] The University of Texas at Arlington, Arlington TX 76019, USA
{xxy1302,lu.zhang2,jxz7537,yxl9168}@mavs.uta.edu, dajiang.zhu@uta.edu
[2] University of Georgia, Athens GA 30602, USA
zihao.wu1@uga.edu

Abstract. Multi-modality learning, exemplified by the language and image pair pre-trained CLIP model, has demonstrated remarkable performance in enhancing zero-shot capabilities and has gained significant attention in the field. However, simply applying language-image pre-trained CLIP to medical image analysis encounters substantial domain shifts, resulting in significant performance degradation due to inherent disparities between natural (non-medical) and medical image characteristics. To address this challenge and uphold or even enhance CLIP's zero-shot capability in medical image analysis, we develop a novel framework, Core-Periphery feature alignment for CLIP (CP-CLIP), tailored for handling medical images and corresponding clinical reports. Leveraging the foundational core-periphery organization that has been widely observed in brain networks, we augment CLIP by integrating a novel core-periphery-guided neural network. This auxiliary CP network not only aligns text and image features into a unified latent space more efficiently but also ensures the alignment is driven by domain-specific core information, e.g., in medical images and clinical reports. In this way, our approach effectively mitigates and further enhances CLIP's zero-shot performance in medical image analysis. More importantly, our designed CP-CLIP exhibits excellent explanatory capability, enabling the automatic identification of critical regions in clinical analysis. Extensive experimentation and evaluation across five public datasets underscore the superiority of our CP-CLIP in zero-shot medical image prediction and critical area detection, showing its promising utility in multimodal feature alignment in current medical applications.

Keywords: Zero-Shot · CP-CLIP · Feature Alignment · Multi-Modality.

1 Introduction

Recently, multi-modality learning has emerged as a promising approach to enhance the understanding and analysis of complex data by leveraging information from multiple sources [8, 15, 24, 29, 27, 25]. There has been a growing research interest focused on integrating textual modalities into computer vision models [5].

The synergy between image and text modalities offers mutual benefits, enhancing modeling and reasoning capabilities, and aligns closely with the multimodal perceptual environment of the human brain [30]. One notable advancement in this field is the pre-trained CLIP (Contrastive Language-Image Pre-training) model, which has demonstrated remarkable performance in various tasks by jointly learning from language and image data [10]. The CLIP model aligns image and text embeddings in the latent space through contrastive learning on a dataset comprising 400 million image-text pairs sourced from a diverse range of publicly accessible online platforms. This fusion of modalities has significantly improved zero-shot capabilities, allowing models to generalize to unseen tasks or domains without explicit training [5].

Nevertheless, despite its remarkable accuracy and transfer learning capabilities, CLIP's zero-shot performance heavily relies on large-scale, high-quality image-text paired datasets. Creating such datasets poses significant challenges, especially in specialized domains like healthcare and radiology, where data is not only scarce but often presents distinct patterns in both image and text components compared to natural images and text descriptions that CLIP is trained on. That is, there exists a significant domain shift between natural (non-medical) and medical images [22, 23]. Additionally, CLIP's reliance solely on contrastive loss for extracting image and text features imposes limitations on its ability to align these features effectively [14]. Thus, there is an increasing need to enhance CLIP with additional mechanisms that can not only improve the multimodality feature alignment between image and text features but also leverage CLIP's zero-shot capability on downstream tasks with limited datasets.

To address the zero-shot performance degradation of the CLIP model in the medical imaging domain, our strategy is to significantly improve the efficiency when aligning multimodal features in latent space by developing a novel information exchange mechanism in a neural network. This mechanism is inspired by brain science research, where the Core-Periphery (CP) organization universally exists in the brain networks [16, 19, 20, 28, 17, 1]. It has been widely confirmed that the CP structure can effectively promote the efficiency of information transmission and communication for biologically integrative processing [18]. In general, CP organization is composed of two qualitatively distinct components: a dense "core" of nodes strongly interconnected with one another, allowing for integrative information processing to facilitate the rapid transmission of messages and a sparse "periphery" of nodes sparsely connected to the core. In this work, we integrate the CP principle into our model to guide neural networks to share weights when processing the information extracted from CLIP, consolidating them into a unified latent space. In this way, our CP-CLIP can align the features from multimodal data more efficiently, thereby alleviating the performance degradation of CLIP and improving performance in downstream tasks, such as disease classification and explainability [7, 21, 6, 26, 24]. We applied our proposed CP-CLIP to five public medical datasets, and the experimental results show that CP-CLIP consistently improves CLIP's zero-shot performance. Additionally, it effectively identifies critical areas for the diseases.

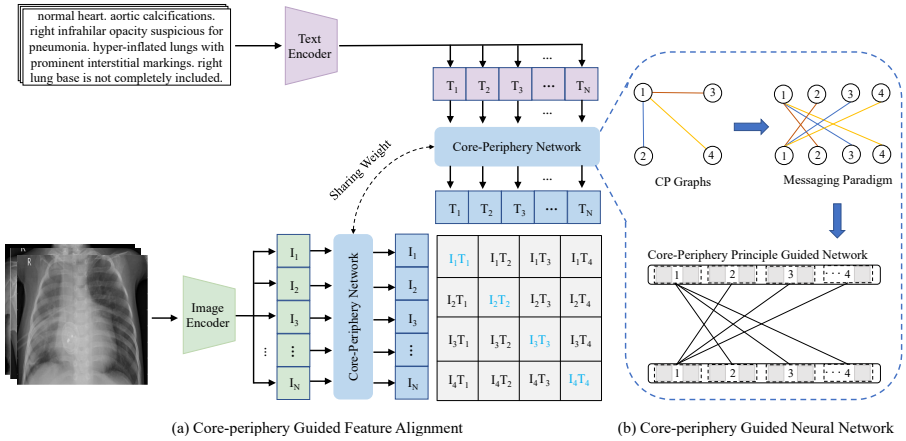

Fig. 1. The CP-CLIP framework. Part (a): The core-periphery network, implemented through a core-periphery principle guided multilayer perceptron neural network. The features extracted from images and texts by CLIP are mapped to a unified space via the core-periphery network. Part (b): The information communication and neuron connection is guided by the generated core-periphery graphs.

2 Method

The CP-CLIP framework is shown in Fig. 1. The CP-CLIP comprises three essential components: the generation of core-periphery graphs (CP graph), the CP graph guided neural network, and the integration of the CP-guided neural network into CLIP. We discuss the details in the following sections.

2.1 Core-Periphery Graph Generation

The core-periphery neural network in CP-CLIP is controlled by Core-Periphery graphs (CP graphs). We introduce the CP graph generation process, which generates a diverse range of CP graphs within the graph space defined by core ratios. It is worth mentioning that in a vanilla multilayer perceptron neural network, neuron connections are fully connected, meaning each neuron is connected to all other neurons. Therefore, the connections in a vanilla multilayer perceptron can be represented by complete graphs, with a core ratio of a complete graph defined as 1.0. To generate graphs with a Core-Periphery (CP) property [19, 22], we define CP graphs as having nodes categorized into core and periphery nodes. Core nodes, acting as information integration hubs, are connected to all nodes, while periphery nodes are solely connected to core nodes. Denoting the total number of nodes as N and the core ratio as $p, p \in (0, 1]$, we calculate the number of core nodes as $n = N \times p$ and the number of periphery nodes as $m = N \times (1 - p)$. Note that when the core ratio equals 1.0, the CP graphs degrade to the complete graph, implying that the CP-guided multilayer perceptron network reverts to its vanilla form.

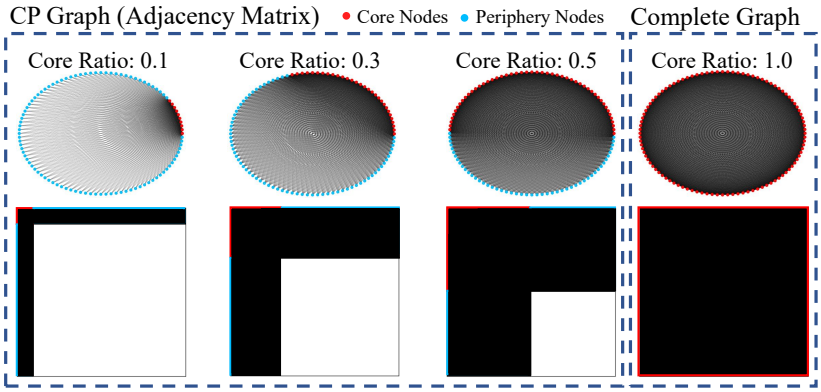

Fig. 2. Examples of Core-Periphery Graphs with different core ratios. The first row displays the graphs, while the second row shows their corresponding adjacency matrices. In the adjacency matrices, the white area denotes 0, indicating no connection, while the black area denotes 1, representing connections between nodes.

The adjacency matrix $A_{N\times N}$ of the generated CP graph can be expressed as:

$$A\left(i,j\right)=\begin{cases}1 & \text{if } \exists(i,j)\in n\\ 0 & \text{if } \forall(i,j)\in m\end{cases} \tag{1}$$

where 1 signifies the presence of an edge between nodes i and j, and 0 indicates no edge between the nodes. By employing various core ratios, denoted by different combinations of n and m, a wide range of candidate graphs can be generated within the graph space. Examples of CP graphs and complete graph are shown in Fig. 2.

2.2 Core-Periphery Principle Guided Neural Network

In CP graphs, core nodes maintain connections to all other nodes, while periphery nodes only connect to the core nodes. To integrate the CP principle into the organization of the multilayer perceptron neural network, we reschedule the neuron connections based on the generated CP graphs. Here, neurons are considered as nodes, and connections between neurons are regarded as edges. This approach allows us to represent neural networks as graphs and utilize the generated Core-Periphery (CP) graph to guide connections. Following this representation paradigm, a complete graph can represent vanilla multilayer perceptron networks. Similarly, we incorporate the Core-Periphery principle into the multilayer perceptron architecture by substituting the complete graph with the generated CP graphs. The new connection rules can then be redefined: CP graph can be represented by $\mathcal{G}=(\mathcal{V},\mathcal{E})$, with nodes set $\mathcal{V}=\{\nu_1,...,\nu_n\}$, edges set $\mathcal{E}\subseteq\{(\nu_i,\nu_j)|\nu_i,\nu_j\in\mathcal{V}\}$, and adjacency matrix A. The information exchange

in the CP graph guided multilayer perceptron network for a specific node i at r-th layer is defined as:

$$z_i^{(r+1)} = \sigma^{(r)}(\{\sum w_{ij}^{(r)} z_j^{(r)}, \forall j \in E(i)\}) \tag{2}$$

where $\sigma(\cdot)$ is the activation function, $z_i^{(r)}$ and $z_j^{(r)}$ are the features stored in the nodes, $w_{ij}^{(r)}$ is the weight of the edge connecting the node i and j, $E(i) = \{i | i \vee (i,j) \in \mathcal{E}\}$ are the neighborhood nodes of node i. We can rewrite the Eq. 2 in matrix form as:

$$\mathbf{Z}^{(r+1)} = \sigma^{(r)}\{(A \odot \mathbf{W}^{(r)})\mathbf{Z}^{(r)}\} \tag{3}$$

where $\mathbf{Z}$ is the feature matrix, and $\mathbf{W}$ is the weight matrix, and $\odot$ is the element-wise matrix multiplication.

Each node corresponds to one or multiple neurons. We propose the following neuron assignment pipeline to map the original neurons to the nodes: for a CP graph with N nodes, each node will be assigned either $\lfloor M/N \rfloor + 1$ or $\lfloor M/N \rfloor$ neurons, where M is the dimension at a specific layer. For example, if we utilize a CP graph with 5 nodes for a layer with 196 dimensions, the 5 nodes will have 40, 39, 39, 39, and 39 neurons, respectively. Conversely, if we employ a CP graph with M nodes, each node will correspond to 1 neuron.

2.3 Core-Periphery Feature Alignment for CLIP

We integrate the CP network into the CLIP model to enhance the fusion of modality information from both images and texts, mapping them into a unified embedding space. This framework, visually represented in Figure 1, is termed CP-CLIP. After extracting features from images and texts using CLIP, their embeddings are passed through the CP network with shared weights. This process encourages the embeddings of both modalities to converge into a unified latent space, thereby facilitating the alignment of features.

For a mini-batch of images and texts $\mathbf{I}$ and $\mathbf{T}$, the embeddings extracted from CLIP are represented as $\mathbf{Z_I}$ and $\mathbf{Z_T}$. We refer to the core-periphery network as f_{cp}. Then, the CP network maps the text and image embeddings to a unified space as follows:

$$\begin{cases} \mathbf{Z_I'} = f_{cp}(\mathbf{Z_I}) \\ \mathbf{Z_T'} = f_{cp}(\mathbf{Z_T}) \end{cases} \tag{4}$$

The logits are obtained from the cosine similarity between the embeddings of images and texts, which have been aligned by the CP network. This can be formulated as follows:

$$\mathbf{s} = \mathbf{Z_I'} \cdot (\mathbf{Z_T'})^T \tag{5}$$

where T means transpose. For an image i, the scaled cosine similrity is obtained by normalizing across the logits, which represent the scores or probabilities associated with the image's relevance to various text descriptions:

$$y_{ij}^{\mathbf{I}\to\mathbf{T}} = \frac{exp(s_{ij}/\tau)}{\sum_{j=1}^{N_{batch}} exp(s_{ij}/\tau)} \tag{6}$$

Table 1. Comparison of zero-shot classification among CLIP, MedCLIP, and CP-CLIP models on five medical image datasets. The results of CP-CLIP were selected as the best results across various core ratios. The balanced accuracy is shown in percentage.

Model	ChestXray (2 classes)	SIIM-ACR (2 classes)	INbreast (3 classes)	CheXpert5×200 (5 classes)	TMED (3 classes)
CLIP	49.50	47.50	34.67	21.80	33.00
MedCLIP	**68.31**	50.00	33.56	12.90	33.33
CP-CLIP	58.51	**50.05**	**38.66**	**24.90**	**34.00**

where τ is the learnable temperature, similar to CLIP [10], $j \in [1, N_{batch}]$ correspond to the batch of texts. Likewise, we can computer the $y_{ji}^{\mathbf{T}\to\mathbf{I}}$, and reach the loss function:

$$\mathcal{L} = -\frac{1}{2}\sum \log y_{ij}^{\mathbf{I}\to\mathbf{T}} - \frac{1}{2}\sum \log y_{ji}^{\mathbf{T}\to\mathbf{I}} \tag{7}$$

3 Results

In this section, we conduct extensive experiments to evaluate the performance of the proposed CP-CLIP model on various medical image datasets under zero-shot scenarios.

3.1 Datasets and Training Details

- **MIMIC-CXR**: The MIMIC-CXR database [4], comprising 377,110 images from 227,835 radiographic studies accompanied by clinical reports, is a vital resource for medical text-image pair analysis.
- **ChestXray**: The ChestX-ray dataset [13] comprises 112,120 chest X-ray images from 30,805 distinct patients, each annotated with disease labels, focusing on pneumonia and normal states.
- **SIIM-ACR**: The SIIM-ACR dataset [12] comprises 12,047 radiographic images, with 2,669 annotated to distinguish between normal lung function and collapsed lung.
- **INbreast**: The INbreast database [9] consists of 115 cases comprising 6154 images captured from various views and slices. These images are categorized into three classes: malignant, benign, and normal.
- **CheXpert5x200**: CheXpert5x200 [3] includes 5 classes (atelectasis, cardiomegaly, consolidation, edema, pleural effusion), with 200 chest radiographs per class.
- **TMED**: The TMED dataset [2] onsists of ultrasound heart images from routine TTE scans, featuring 599 scans each labeled with a diagnosis for Aortic Stenosis (AS), categorized into severe, early, or none

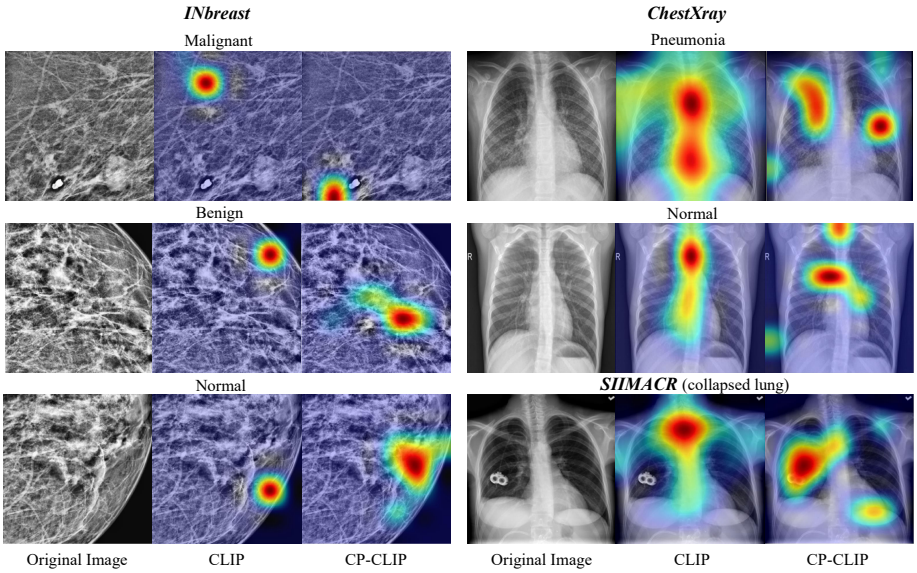

Fig. 3. Visualization comparison of CLIP and CP-CLIP on the ChestXray, SIIMACR, and INbreast datasets. CP-CLIP highlights critical disease-related areas, while CLIP tends to identify shortcuts.

We initialize CP-CLIP with the pre-trained CLIP weights and subsequently train both CP-CLIP and CLIP models on the MIMIC-CXR dataset, which comprises medical text-image pairs. Subsequently, we evaluate its zero-shot performance on other five medical image datasets. Training involves 20 epochs with a batch size of 128 on Titan GPUs, utilizing the AdamW optimizer and cosine learning rate scheduler. Initial learning rates are 1e−6 for the CLIP model and 5e−4 for the CP network with a minimum learning rate of 1e−8.

3.2 Classification Results

We evaluate the trained CP-CLIP model under a zero-shot setting to assess the model's multimodal representation robustness and generalizability. Specifically, we select five unseen medical imaging datasets and compare them to two baselines: the CLIP and MedCLIP. Note that MedCLIP was well trained on the MIMIC-CXR dataset [14]. For fair comparisons, we use the same single text prompt for all experiments with the three models. As depicted in Table 1, CP-CLIP demonstrates superior classification accuracy across the datasets, highlighting the effectiveness of the CP principle guided feature alignment.

3.3 Critical Area Identification

To investigate how the CP mechanism enhances image-text alignment, we utilized Grad-CAM [11] visualization on images from various datasets to analyze

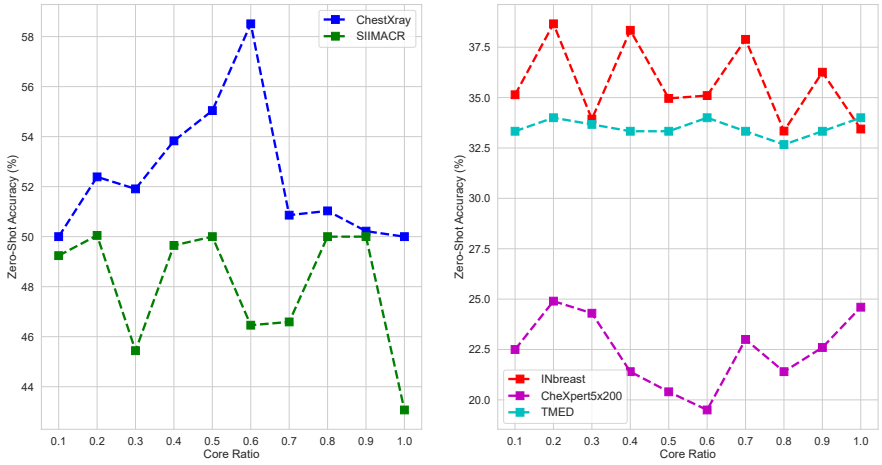

Fig. 4. The impact of core ratios on classification accuracy across the five datasets.

how the model processes core and peripheral elements during inference. As depicted in Fig. 3, the original CLIP struggles to identify critical areas for disease identification. For instance, in the case of malignancy from the INbreast dataset, the area identified by CLIP is unrelated to breast cancer identification. In contrast, the CP-CLIP model not only focuses on lesion regions but also considers minor areas for comprehensive reasoning. Besides, in the ChestXray dataset, CLIP tends to highlight the spine area, whereas our CP-CLIP can identify the areas related to disease identification.

3.4 Ablation Study

We conducted an ablation study on CP-CLIP with different core ratios. The results, as shown in Fig. 4, demonstrate that our CP-CLIP outperforms the baselines across a wide range of core ratios. The best performance appears with a core ratio smaller than 1.0, indicating that CP graphs guided networks perform better than the vanilla form represented in complete graphs. These results reveal an interesting conclusion: the sparse connections of neurons brought by the CP structure can perform better than fully connected neurons.

4 Conclusion

We introduce CP-CLIP, a novel framework that integrates the core-periphery principle into the CLIP model by constructing an auxiliary core-periphery graph guided neural network specifically designed for zero-shot medical image analysis. This auxiliary network improves the fine-grained alignment between image and text embeddings, directing the model's attention towards critical information. Experimental results on five distinct medical image datasets demonstrate the effectiveness of CP-CLIP in medical image analysis.

Acknowledgments. This work was supported by National Institutes of Health (R01AG075582 and RF1NS128534).

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

References

1. Chavoshnejad, P., Chen, L., Yu, X., Hou, J., Filla, N., Zhu, D., Liu, T., Li, G., Razavi, M.J., Wang, X.: An integrated finite element method and machine learning algorithm for brain morphology prediction. Cerebral Cortex **33**(15), 9354–9366 (2023)
2. Huang, Z., Long, G., Wessler, B., Hughes, M.C.: Tmed 2: a dataset for semi-supervised classification of echocardiograms. DataPerf: Benchmarking Data for Data-Centric AI Workshop (2022)
3. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. vol. 33, pp. 590–597 (2019)
4. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
5. Liu, Z., Jiang, H., Zhong, T., Wu, Z., Ma, C., Li, Y., Yu, X., et al.: Holistic evaluation of gpt-4v for biomedical imaging. arXiv preprint arXiv:2312.05256 (2023)
6. Lyu, Y., Yu, X., Zhang, L., Zhu, D.: Classification of mild cognitive impairment by fusing neuroimaging and gene expression data. In: Proceedings of the 15th international conference on PErvasive technologies related to assistive environments. pp. 26–32 (2021)
7. Lyu, Y., Yu, X., Zhu, D., Zhang, L.: Classification of alzheimer's disease via vision transformer. In: Proceedings of the 15th international conference on PErvasive technologies related to assistive environments. pp. 463–468 (2022)
8. Ma, C., Jiang, H., Chen, W., Wu, Z., Yu, X., et al.: Eye-gaze guided multi-modal alignment framework for radiology. arXiv preprint arXiv:2403.12416 (2024)
9. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic radiology **19**(2), 236–248 (2012)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: CVPR. pp. 618–626 (2017)
12. Stephens, K.: Acr, siim name winners of pneumothorax detection machine learning challenge. AXIS Imaging News (2019)
13. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR. pp. 2097–2106 (2017)

14. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
15. Xiao, Z., Chen, Y., Yao, J., Zhang, L., Liu, Z., Wu, Z., Yu, X., et al.: Instruction-vit: Multi-modal prompts for instruction learning in vision transformer. Information Fusion p. 102204 (2024)
16. Yu, X., Hu, D., Zhang, L., Huang, Y., Wu, Z., Liu, T., Wang, L., Lin, W., Zhu, D., Li, G.: Longitudinal infant functional connectivity prediction via conditional intensive triplet network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 255–264 (2022)
17. Yu, X., Scheel, N., Zhang, L., Zhu, D.C., Zhang, R., Zhu, D.: Free water in t2 flair white matter hyperintensity lesions. In: Alzheimer's & Dementia. p. e057398 (2021)
18. Yu, X., Zhang, L., Dai, H., Lyu, Y., Zhao, L., Wu, Z., Liu, D., Liu, T., Zhu, D.: Core-periphery principle guided redesign of self-attention in transformers. arXiv preprint arXiv:2303.15569 (2023)
19. Yu, X., Zhang, L., Dai, H., Zhao, L., Lyu, Y., Wu, Z., Liu, T., Dajiang, Z.: Gyri vs. sulci: Disentangling brain core-periphery functional networks via twin-transformer. arXiv preprint arXiv:2302.00146 (2023)
20. Yu, X., Zhang, L., Lyu, Y., Liu, T., Zhu, D.: Supervised deep tree in alzheimer's disease. In: IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023)
21. Yu, X., Zhang, L., Zhao, L., Lyu, Y., Liu, T., Dajiang, Z.: Disentangling spatial-temporal functional brain networks via twin-transformers. arXiv preprint arXiv:2204.09225 (2022)
22. Yu, X., Zhang, L., Zhu, D., Liu, T.: Robust core-periphery constrained transformer for domain adaptation. arXiv preprint arXiv:2308.13515 (2023)
23. Zhang, L., Liu, Z., Zhang, L., Wu, Z., Yu, X., Holmes, J., Feng, H., Dai, H., Li, X., Li, Q., Wong, W.W., Vora, S.A., Zhu, D., Liu, T., Liu, W.: Generalizable and promptable artificial intelligence model to augment clinical delineation in radiation oncology. Medical Physics (2024)
24. Zhang, L., Na, S., Liu, T., Zhu, D., Huang, J.: Multimodal deep fusion in hyperbolic space for mild cognitive impairment study. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 674–684. Springer (2023)
25. Zhang, L., Wang, L., Gao, J., Risacher, S.L., Yan, J., Li, G., Liu, T., Zhu, D., Initiative, A.D.N., et al.: Deep fusion of brain structure-function in mild cognitive impairment. Medical image analysis **72**, 102082 (2021)
26. Zhang, L., Wang, L., Liu, T., Zhu, D.: Disease2vec: Encoding alzheimer's progression via disease embedding tree. Pharmacological Research **199**, 107038 (2024)
27. Zhang, L., Wang, L., Zhu, D.: Jointly analyzing alzheimer's disease related structure-function using deep cross-model attention network. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 563–567. IEEE (2020)
28. Zhang, L., Yu, X., Lyu, Y., Liu, T., Zhu, D.: Representative functional connectivity learning for multiple clinical groups in alzheimer's disease. In: IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023)
29. Zhang, L., Zaman, A., Wang, L., Yan, J., Zhu, D.: A cascaded multi-modality analysis in mild cognitive impairment. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI, Proceedings 10. pp. 557–565. Springer (2019)
30. Zhao, L., Zhang, L., Wu, Z., Chen, Y., Dai, H., Yu, X., Liu, Z., Zhang, T., Hu, X., Jiang, X., et al.: When brain-inspired ai meets agi. Meta-Radiology p. 100005 (2023)