Small Lesions-aware Bidirectional Multimodal Multiscale Fusion Network for Lung Disease Classification

Jianxun Yu[1] *, Ruiquan Ge[2]*, Zhipeng Wang[2],, Cheng Yang[2], Chenyu Lin[2], Xianjun Fu[3], Jikui Liu[4] (✉), Ahmed Elazab[5] (✉), and Changmiao Wang[6]

[1] Xidian University [2] Hangzhou Dianzi University
[3] Zhejiang College of Security Technology, School of Artificial Intelligence
[4] Shenzhen Polytechnic University [5] Shenzhen University
[6] Shenzhen Research Institute of Big Data
liujikui007@gmail.com ahmedelazab@szu.edu.cn

Abstract. The diagnosis of medical diseases faces challenges such as the misdiagnosis of small lesions. Deep learning, particularly multimodal approaches, has shown great potential in the field of medical disease diagnosis. However, the differences in dimensionality between medical imaging and electronic health record data present challenges for effective alignment and fusion. To address these issues, we propose the Multimodal Multiscale Cross-Attention Fusion Network (MMCAF-Net). This model employs a feature pyramid structure combined with an efficient 3D multi-scale convolutional attention module to extract lesion-specific features from 3D medical images. To further enhance multimodal data integration, MMCAF-Net incorporates a multi-scale cross-attention module, which resolves dimensional inconsistencies, enabling more effective feature fusion. We evaluated MMCAF-Net on the Lung-PET-CT-Dx dataset, and the results showed a significant improvement in diagnostic accuracy, surpassing current state-of-the-art methods. The code is available at `https://github.com/yjx1234/MMCAF-Net`.

Keywords: Multimodal Learning · Cross-Modality · Multiscale · Inter-dimensional Uncertainty.

1 Introduction

Computer-aided diagnosis is crucial in modern clinical applications, improving disease detection efficiency and accuracy. Deep learning methods are widely used for medical diagnosis, utilizing information from various modalities such as medical imaging and electronic health records to deliver comprehensive patient assessments [7][3]. However, detecting certain lesions, like squamous cell carcinoma and pulmonary embolism, remains challenging due to their small size in images, which can be easily missed by conventional models [2][11]. Additionally, discrepancies between modalities complicate effective multimodal fusion for

* Co-first authors.

clinical decision-making. These challenges highlight the need for more robust methodologies to enhance multimodal data integration and improve diagnostic accuracy.

Multi-scale methods are effective for small lesion detection by capturing local and global information, but traditional approaches often incur high computational costs. To improve efficiency, Rahman *et al.*. introduced the Multi-scale Convolutional Attention Module (MSCAM) [16], which, while beneficial, is limited to 2D image processing and lacks applicability to 3D medical imaging, highlighting a gap in multi-scale feature extraction for volumetric data. Aligning features across modalities is a persistent challenge in multimodal learning. Yao *et al.*. minimized Jensen–Shannon divergence for feature alignment [18], while Sanjeev *et al.*. used supervised contrastive loss to integrate tabular and image data [17]. However, these strategies often compromise the integrity of complex modalities to align with simpler ones, resulting in suboptimal performance and underscoring the need for more adaptive fusion techniques. Late fusion, known for its simplicity, is widely used but fails to capture complementary relationships between multimodal features [6][17]. To address this, Hayat *et al.*. proposed using LSTMs for aggregating multimodal representations as sequences of single-modal tokens [4], and Yao *et al.*. introduced a disease-aware attention fusion module to weigh modalities based on disease profiles [18]. Other strategies, such as Pölsterl *et al.*.'s dynamic affine feature transformation [14] and Reza *et al.*.'s MMTM module for knowledge exchange between CNNs [8], have also advanced intermediate fusion. However, they often overlook ambiguity in certain feature dimensions, which can degrade representation quality and limit downstream task effectiveness. Thus, more adaptive fusion mechanisms are necessary to enhance meaningful features while mitigating ambiguity.

To overcome these challenges, we propose a novel framework that enables simultaneous multimodal fusion while effectively capturing small lesions in medical images. The key contributions of this work are summarized as follows: 1) We introduce an E3D-MSCA module to enhance lesion detection, particularly for small lesions. 2) We propose MSCA module designed to effectively integrate multimodal data, addressing the challenges posed by significant differences between modalities. 3) We develop a Bidirectional Scale Fusion (BSF) module, which specifically addresses the challenges of integrating features across two distinct linear scales.

2 Proposed method

As illustrated in **Fig. 1**, MMCAF-Net consists of three main components. The first component is an image encoder that integrates a feature pyramid with E3D-MSCA module, designed to capture both local and global features of imaging data. This allows for the effective differentiation of challenging cases. The second component employs Kolmogorov–Arnold Networks (KAN) [12] to encode tabular features. The advantages of KAN lie in its faster scaling laws, meaning that as the model size increases, its efficiency improves significantly. Additionally, KAN

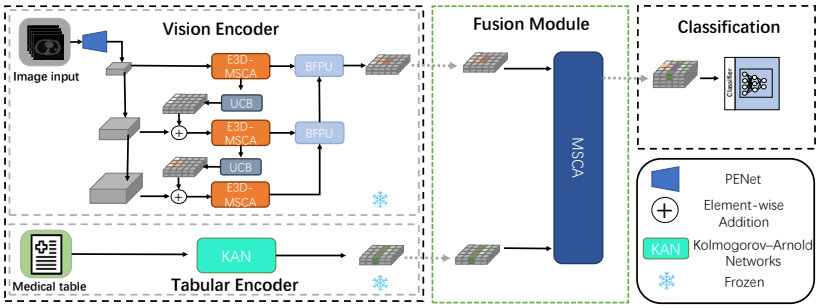

Fig. 1: Overview of the Proposed MMCAF-Net. The framework consists of an encoding module with Vision and Tabular encoders, a fusion module utilizing the MSCA module, and a classification module for final detection.

demonstrates strong expressive capability even with fewer parameters. Initially, KAN was applied to tasks such as function approximation and solving partial differential equations (PDEs). Subsequent research has expanded its use to time series forecasting tasks. In addressing our problem, we flexibly apply KAN to encode our tabular data. Lastly, the third component utilizes the MSCA module to align and fuse the features from both modalities.

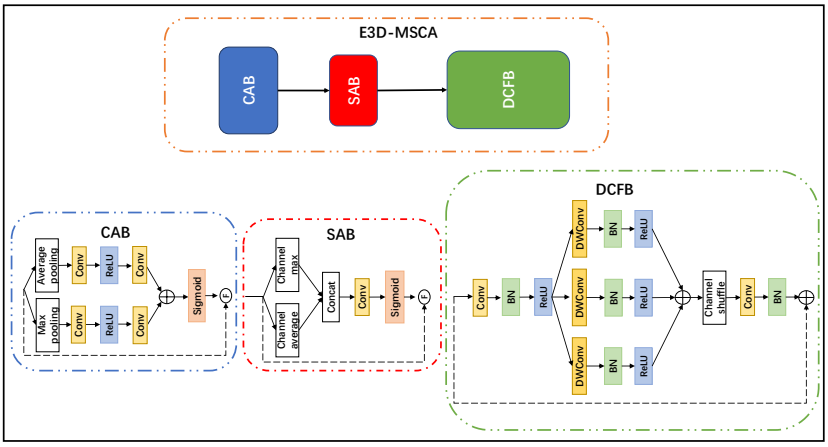

Fig. 2: Detailed structure of E3D-MSCA. It comprises 3D CAB, 3D SAB, and 3D DCFB.

2.1 Efficient 3D Multi-scale PENet

Directly using the conventional lung encoder, PENet [5] has proven insufficient for effectively distinguishing challenging cases. We found that these cases are particularly difficult to differentiate due to the small size of the lesion regions, which occupy only a small portion of the entire image. Traditional encoders often overlook these lesions. Previous studies have typically employed feature pyramids to focus on both local and global features [13][20]. Inspired by these works, we propose an improvement based on the MSCAM method [16]. Specifically, we introduce a novel E3D-MSCA method, which combines the classical encoder with enhanced attention mechanisms. As shown in **Fig. 2**, E3D-MSCA consists of three modules: 3D Channel Attention Block(CAB), 3D Spatial Attention Block(SAB), and 3D Depth-Wise Convolution Fusion Block(DCFB). The formulation of E3D-MSCA is described as follows.

$$E3D - MSCA(x) = DCFB(SAB(CAB(x))), \tag{1}$$

where x represents the output features from the feature pyramid. By utilizing depth-wise convolution within the *DCFB*, we significantly enhance the performance of PENet in all metrics without incurring substantial computational costs. Finally, we incorporate the Bidirectional Feedback Propagation Unit (BFPU) proposed by Chen *et al.*. [1] to effectively fuse features from multiple scales. The formulation of BFPU is provided using:

$$F_{mid} = S\left(\mathrm{Conv}_{3\times3}(F_a) \otimes \mathrm{Conv}_{3\times3}(F_b)\right), \tag{2}$$

$$F_{out} = \left[F_a + F_{mid} \otimes F_a,\ F_b + F_{mid} \otimes F_b\right], \tag{3}$$

where F_a and F_b represent features extracted from two different scales. The functions S and $\otimes$ denote the sigmoid activation function and element-wise multiplication, respectively, while the notation $[\cdot]$ signifies channel-wise concatenation.

2.2 Multiscale Cross Attention

Due to the significant dimensional disparity between medical images and clinical data tables, we performed feature dimensionality reduction in vision encoder to align the dimensions with those of the tabular data. However, the differing distributions of image and tabular features make direct cross-fusion ineffective for leveraging the information from both sources. Moreover, dimensional conflicts can degrade the fusion performance. In prior work, Yu *et al.*. utilized a multi-scale group aggregation bridge for feature fusion [19]. Inspired by their approach, we propose a MSCA fusion module. As illustrated in **Fig. 3**, we applied consecutive inverted pyramid dimensionality reductions to both image and tabular features, resulting in three pairs of features with different dimensions. Next, the feature pairs from each scale are fed into the Cross Attention module for feature fusion. In this module, we use the image and table features as queries (Q), while another feature is utilized as keys (K) and values (V). Following this, Q, K, and V undergo a multi-head operation. We compute the attention scores between

the two modalities using Q and K, normalize these attention scores, and then multiply them with V to derive the final output. The details of this process are outlined as follows:

$$Q' = QW_q \quad (W_q \in \mathbb{R}^{D_q \times D_q}), \tag{4}$$

$$K' = KW_k \quad (W_k \in \mathbb{R}^{D_k \times D_q}), \tag{5}$$

$$V' = VW_v \quad (W_v \in \mathbb{R}^{D_v \times D_q}), \tag{6}$$

where Q has a shape of [B, N, D_q] while the keys K and values V have a shape of [B, M, $D_k v$].

$$Q_h = \text{reshape}(Q', B, H, N, C), \tag{7}$$

$$K_h = \text{reshape}(K', B, H, M, C), \tag{8}$$

$$V_h = \text{reshape}(V', B, H, M, C), \tag{9}$$

where C represents the focus dimension of each attention head, H denotes the number of attention heads.

$$\text{Attention}(Q_h, K_h) = \frac{Q_h K_h^T}{\sqrt{C}}, \mathrm{O_h} = \text{Attention} * \mathrm{V_h}, \tag{10}$$

Attention represents the attention weights, and O_h denotes the final output features. Subsequently, we introduce the BSF module to integrate features across multiple scales. The structure of this module is illustrated in **Fig. 3**. Initially, we apply a linear transformation to align the two features to a common scale. Next, we compute the dimensional importance of both features by performing an element-wise multiplication, followed by calculating the Softmax scores. This process effectively reduces the significance of the uncertain dimensions. Finally, we merge the features from the two scales based on their dimensional importance.

2.3 Loss Function

We employ a binary cross-entropy loss to quantify the distance between the predicted outcomes and the ground truths. Mathematically, the classification loss $\mathcal{L}_{cls}$ is expressed as follows:

$$\mathcal{L}_{cls}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{11}$$

where N represents the number of samples, y_i signifies the real labels for sample i, $\hat{y}_i$ is the probability values of the model predicting sample i range between (0,1).

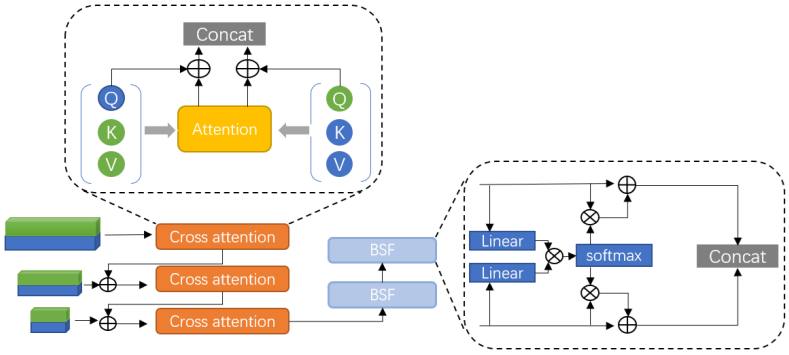

Fig. 3: Detailed structure of MSCA module.

3 Experiments

3.1 Dataset and Implementation

Dataset. We used a publicly available dataset provided by The Cancer Imaging Archive, called Lung-PET-CT-Dx [10], which includes CT or PET) scans and tabular clinical information for 355 cases. The dataset provides tumor classification labels for each patient. Most of the PET/CT data is stored in DICOM format, and all data has undergone de-identification. The dataset includes 251 adenocarcinoma samples and 61 squamous cell carcinoma samples. Since the adenocarcinoma class and squamous cell carcinoma class samples are very imbalanced in the original dataset, we decided to oversample the squamous cell carcinoma class. We use a random oversampling strategy to supplement the training set from 34 samples to 198 samples, and the number of verification sets and test sets remains the same, which are 12 and 15 respectively. The tabular data we use includes attributes such as gender, age, weight, TNM stage, and smoking history.

Implementation details. We implemented MMCAF-Net on a device equipped with Tesla V100S-PCIE-32GB using torch-2.4.1. During the training phase, we utilized binary cross-entropy loss in conjunction with the stochastic gradient descent optimizer. The model was trained for 50 epochs, with an initial learning rate set at 0.0001 and a weight decay of 0.01. We employed a batch size of 4 and processed 12 slices per input sample. Data augmentation techniques, including rotation, sharpening, and normalization, were applied to the input images, which were subsequently reshaped to a dimension of 192×192 pixels.

3.2 Experiment Result

The Lung-PET-CT Result. In our study, we compared MMCAF-Net against the six state-of-the-art multi-modal approaches using the Lung-PET-CT dataset. As shown in **Table 1**, MMCAF-Net outperforms all comparative approaches across several key metrics, including accuracy (ACC), F1 score (F1), specificity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV). While MMCAF-Net falls short of mmtm by 1.6% in area under the receiver operating characteristic curve(AUROC), it surpasses mmtm by approximately 10% in both ACC and F1 scores, and exhibits a remarkable 15% improvement in PPV. This indicates that MMCAF-Net achieves a lower rate of false positives, demonstrating its high accuracy and reliability. Overall, it is noteworthy that all models exhibit relatively poor performance on the F1, suggesting the presence of noise or outliers in the dataset. Additionally, we visualized the classification results for challenging cases. As shown in **Fig. 4**, each case has 12 slices, and we selected the most representative slice of the lesion for display. The results indicate that our model outperforms the others in distinguishing subtle and difficult cases.

Ablation study. We conducted ablation experiments on the Lung-PET-CT dataset to evaluate the effectiveness of image encoding and multi-modal fusion techniques. As demonstrated in **Table 2**, we compared our proposed Efficient 3D Multi-scale PENet with the Segment Anything Model (SAM) [9] combined with multi-scale techniques and the original PENet. The results indicate that our method outperforms the other two approaches across all evaluated metrics, achieving improvements of 10% and 15% in AUROC, and 10% and 14% in F1 scores, respectively. Furthermore, our method shows significant enhancements in both AUROC and NPV compared to the standalone PENet, suggesting that the improved architecture enhances the reliability of negative class sample identification. Additionally, we performed ablation experiments on different fusion methods, as presented in **Table 3**, comparing our proposed MSCA with Cross-Attention, Late_Fusion, and clip aligned fusion [15] techniques. The results illustrate that MSCA outperforms the other three fusion methods across most metrics, with an increase in ACC of 5%, 10%, and 12%, respectively. This indicates that our fusion method demonstrates superior overall performance compared to the other approaches.

4 Conclusion

In this study, we introduce MMCAF-Net, a multimodal multiscale fusion model that effectively focuses on small lesions. MMCAF-Net incorporates the E3D-MSCA module to dynamically adjust feature weights across different scales, enhancing attention to small lesions. The MSCA module addresses the dimensional conflicts that arise when merging two modalities with significantly different characteristics. The BSF module tackles the impact of dimensional uncertainty between different scales on the fusion results. Comparative studies demonstrate

Table 1: Quantitative comparison with other methods on the public dataset. The best and the second-best results are bolded and underlined, respectively.

Method	AUROC ↑	ACC ↑	F1 ↑	Specificity ↑	Sensitivity ↑	PPV ↑	NPV ↑
PECon [17]	0.786	0.744	0.645	0.786	<u>0.667</u>	0.625	<u>0.815</u>
MedFuse [4]	0.786	0.721	0.571	<u>0.821</u>	0.533	0.615	0.767
Drfuse [18]	0.613	0.676	0.353	0.800	0.333	0.375	0.769
MMTM [8]	**0.802**	0.698	0.581	0.750	0.600	0.562	0.778
PEfusion [6]	0.740	0.721	0.600	0.786	0.600	0.600	0.786
daft [14]	0.727	<u>0.729</u>	<u>0.667</u>	0.786	0.650	<u>0.684</u>	0.759
MMCAF-Net (Ours)	<u>0.786</u>	**0.791**	**0.690**	**0.857**	**0.667**	**0.714**	**0.828**

Table 2: Image encoder ablation experiments on the Lung-PET-CT dataset. The best and second-best results are bolded and underlined, respectively.

Method	AUROC ↑	ACC ↑	F1 ↑	Specificity ↑	Sensitivity ↑	PPV ↑	NPV ↑
SAM+E3D-MSCA	0.610	0.651	0.444	0.786	0.400	0.500	0.710
PENet [5]	0.560	0.721	0.400	<u>0.964</u>	0.267	<u>0.800</u>	0.711
PENet+E3D-MSCA	<u>0.644</u>	<u>0.721</u>	<u>0.500</u>	0.893	<u>0.400</u>	0.667	<u>0.735</u>
PENet+E3D-MSCA+drop	**0.712**	**0.767**	**0.545**	**0.964**	**0.400**	**0.857**	**0.857**

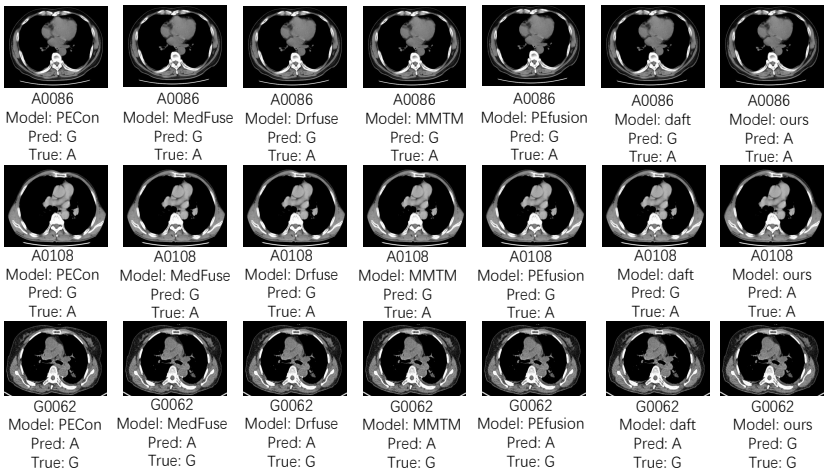

Fig. 4: Visual Representation of Classification Performance for Challenging Cases. The G denotes lung squamous cell carcinoma and A denotes lung adenocarcinoma.

that MMCAF-Net outperforms competing methods on most metrics, indicating its practical effectiveness. Future plans include adapting the model for the diagnosis of lung and other medical diseases in clinical settings.

Table 3: Fusion ablation experiments on the Lung-PET-CT dataset. The best and second-best results are bolded and underlined, respectively.

Method	AUROC ↑	ACC ↑	F1 ↑	Specificity ↑	Sensitivity ↑	PPV ↑	NPV ↑
Cross-Attention	<u>0.752</u>	<u>0.744</u>	<u>0.686</u>	0.714	**0.800**	0.600	**0.870**
clip_Fusion [15]	0.717	0.698	0.629	0.679	<u>0.733</u>	0.550	0.826
Late_Fusion	0.729	0.674	0.462	<u>0.821</u>	0.400	0.545	0.719
MSCA_Fusion	**0.786**	**0.791**	**0.690**	**0.857**	0.667	**0.714**	<u>0.828</u>

Acknowledgments. This work was supported by the Open Project Program of the State Key Laboratory of CAD & CG (No. A2410), Zhejiang University; Zhejiang Provincial Natural Science Foundation of China (No. LY21F020017, 2023C03090); National Natural Science Foundation of China (No. 61702146, 62076084, U20A20386, U22A2033); Shenzhen major science and technology project (KCXFZ20240903093923031), Special projects fund in key areas of the Guangdong Provincial Department of Education (2023ZDZX2085); Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515110570), Futian Healthcare Research Project (No.FTWS002); Medical Scientific Research Foundation of the Guangdong Province of China (A2023158,C2023106).

Disclosure of Interests. The authors declare no competing interests.

References

1. Chen, X., Pan, J., Dong, J.: Bidirectional multi-scale implicit neural representations for image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25627–25636 (2024)
2. Guo, Z., Gan, H.: Cpp-net: Embracing multi-scale feature fusion into deep unfolding cp-ppa network for compressive sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25086–25095 (2024)
3. Hager, P., Menten, M.J., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23924–23935 (2023)
4. Hayat, N., Geras, K.J., Shamout, F.E.: Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In: Machine Learning for Healthcare Conference. pp. 479–503. PMLR (2022)
5. Huang, S.C., Kothari, T., Banerjee, I., Chute, C., Ball, R.L., Borus, N., Huang, A., Patel, B.N., Rajpurkar, P., Irvin, J., et al.: Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging. NPJ Digital Medicine **3**(1), 61 (2020)
6. Huang, S.C., Pareek, A., Zamanian, R., Banerjee, I., Lungren, M.P.: Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. Scientific Reports **10**(1), 22147 (2020)
7. Jiang, J.P., Ye, H.J., Wang, L., Yang, Y., Jiang, Y., Zhan, D.C.: Tabular insights, visual impacts: transferring expertise from tables to images. In: Forty-first International Conference on Machine Learning (2024)

8. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: Mmtm: Multimodal transfer module for cnn fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13289–13299 (2020)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
10. Li, P., Wang, S., Li, T., Lu, J., HuangFu, Y., Wang, D.: A large-scale ct and pet/ct dataset for lung cancer diagnosis (lung-pet-ct-dx). https://doi.org/10.7937/TCIA.2020.NNC2-0461 (2020)
11. Liu, S., Ma, Y., Zhang, X., Wang, H., Ji, J., Sun, X., Ji, R.: Rotated multi-scale interaction network for referring remote sensing image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26658–26668 (2024)
12. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024)
13. Ma, T., Dai, X., Zhang, S., Wen, Y.: Pivit: Large deformation image registration with pyramid-iterative vision transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 602–612. Springer (2023)
14. Pölsterl, S., Wolf, T.N., Wachinger, C.: Combining 3d image and tabular data via the dynamic affine feature map transform. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 688–698. Springer (2021)
15. 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)
16. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11769–11779 (2024)
17. Sanjeev, S., Al Khatib, S.K., Shaaban, M.A., Almakky, I., Papineni, V.R., Yaqub, M.: Pecon: Contrastive pretraining to enhance feature alignment between ct and ehr data for improved pulmonary embolism diagnosis. In: International Workshop on Machine Learning in Medical Imaging. pp. 434–443. Springer (2023)
18. Yao, W., Yin, K., Cheung, W.K., Liu, J., Qin, J.: Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16416–16424 (2024)
19. Yu, X., Elazab, A., Ge, R., Jin, H., Jiang, X., Jia, G., Wu, Q., Shi, Q., Wang, C.: Ich-scnet: Intracerebral hemorrhage segmentation and prognosis classification network using clip-guided sam mechanism. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2795–2800. IEEE (2024)
20. Zhu, V., Ji, Z., Guo, D., Wang, P., Xia, Y., Lu, L., Ye, X., Zhu, W., Jin, D.: Low-rank continual pyramid vision transformer: Incrementally segment whole-body organs in ct with light-weighted adaptation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 371–381. Springer (2024)