C²MAOT: Cross-modal Complementary Masked Autoencoder with Optimal Transport for Cancer Segmentation in PET-CT Images

Jiaju Huang[1], Shaobin Chen[1], Xinglong Liang[2], Xiao Yang[3], Zhuoneng Zhang[1], Yue Sun[1], Ying Wang[3], and Tao Tan[1⋆]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
taotan@mpu.edu.mo
[2] Netherlands Cancer Institute, Amsterdam, Netherlands
[3] The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China

Abstract. Accurate cancer segmentation in PET-CT images is crucial for oncology, yet remains challenging due to lesion diversity, data scarcity, and modality heterogeneity. Existing methods often struggle to effectively fuse cross-modal information and leverage self-supervised learning for improved representation. In this paper, we introduce C²MAOT, a Cross-modal Complementary Masked Autoencoder with Optimal Transport framework for PET-CT cancer segmentation. Our method employs a novel modality-complementary masking strategy during pre-training to explicitly encourage cross-modal learning between PET and CT encoders. Furthermore, we integrate an optimal transport loss to guide the alignment of feature distributions across modalities, facilitating robust multi-modal fusion. Experimental results on two datasets demonstrate that C²MAOT outperforms existing state-of-the-art methods, achieving significant improvements in segmentation accuracy across five cancer types. These results establish our proposed method as an effective approach for tumor segmentation in PET-CT imaging. Our code is available at https://github.com/hjj194/c2maot.

Keywords: PET-CT Segmentation · Cross-modal Fusion · Self-supervised Learning.

1 Introduction

Positron Emission Tomography-Computed Tomography (PET-CT) is a key imaging modality in oncology, combining the metabolic insights provided by PET with the anatomical details from CT [1]. PET images, using fluorodeoxyglucose (FDG) as a radiotracer, highlights areas of high glucose uptake, which are often indicative of tumors, while CT images provide high-resolution images crucial for precise lesion localization and visualizing organ boundaries [25]. Therefore, accurate tumor segmentation in PET-CT images is essential for enhancing tumor

⋆ Corresponding author

detection, assessing disease staging, and evaluating treatment responses. This capability plays a critical role in clinical applications such as radiation therapy planning, surgical guidance, and ongoing disease monitoring [32,11,5].

Recent advancements in deep learning have shown promising results in automatic tumor segmentation in medical imaging. Models based on convolutional neural networks (CNNs) have demonstrated significant capabilities in feature extraction and segmentation tasks [10,20,19,33,17], while transformer-based architectures have shown exceptional performance in capturing long-range dependencies and contextual information across image regions [28,14,4,6]. Despite these advances, tumor segmentation in PET-CT still faces three major challenges. (1) Data scarcity: The limited availability of labeled PET-CT data and the high cost of manual annotation hinder model training. (2) Lesion diversity: Tumors vary in size, shape, and location, making accurate segmentation difficult. This is particularly problematic for metastatic tumors that can appear in multiple, spatially distributed locations. (3) Modality heterogeneity: PET and CT provide complementary but distinct information, and existing methods struggle to effectively align feature distributions between these modalities and maintain consistent feature representations.

Current approaches have attempted to address these challenges but with limited success. To tackle data scarcity, self-supervised learning (SSL) has gained significant attention for its ability to leverage unlabeled data in medical imaging [18]. Approaches such as contrastive learning (e.g., SimCLR [7], MoCo [16]), masked image modeling (e.g., MAE [15], SimMIM [31]), and self-distillation methods (e.g., DINO [3], BYOL [13]) have shown promise in pre-training models on large datasets without requiring manual annotation. However, most SSL methods focus on individual modalities and often fail to capture the complementary information provided by multiple imaging modalities, such as PET and CT. For the lesion diversity challenge, augmentation techniques and multi-scale architectures have been proposed [19], but these methods often fail to generalize across different tumor types and locations. To address modality heterogeneity, early fusion methods relied on simple strategies like concatenation or weighted averaging of features from each modality [2]. While these approaches provide some benefit, they often struggle to fully exploit the complementary nature of the modalities due to differences in their resolutions, intensities, and noise characteristics [21]. More sophisticated approaches, such as attention mechanism-based models, have been developed to enable more nuanced interactions between modalities [23]. Cross-modal attention mechanisms effectively capture contextual dependencies by dynamically weighting modality-specific features. Despite these efforts, challenges remain in effectively aligning feature distributions between modalities and maintaining consistency in feature representation across different modalities.

In this paper, we propose a novel pre-training segmentation framework, C^2MAOT, which integrates cross-modal complementary masking and optimal transport (OT)-guided multi-modal fusion. Our method addresses three key challenges in multi-modal tumor segmentation. First, we tackle data scarcity through a self-supervised pre-training approach that effectively leverages unlabeled PET-

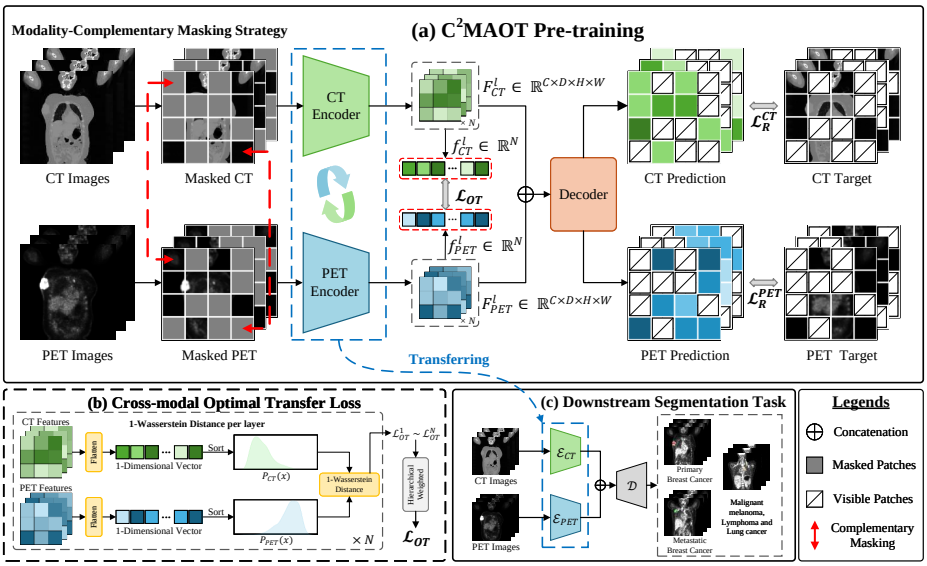

Fig. 1. Overview of the proposed C²MAOT framework. It consists of two main components: pre-training and fine-tuning for downstream segmentation tasks.

CT data. Second, our method addresses lesion diversity via a spatially-aware complementary masking strategy specifically designed for the distributed nature of tumors. By forcing the model to reconstruct complementary regions across modalities, it enhances cross-modal learning within the PET-CT domain, improving the representation of diverse tumor patterns present in the training distribution. Third, we overcome modality heterogeneity by introducing an OT loss to guide the alignment of feature distributions between PET and CT, enabling more effective multi-modal fusion and ensuring consistency in feature representations across modalities.

Our contributions are three-fold: (1) We introduce a novel cross-modal complementary masking strategy for self-supervised pre-training, which explicitly facilitates information exchange between PET and CT encoders while improving the model's ability to handle spatially diverse tumor presentations. (2) An optimal transport loss guides the alignment of feature distributions across modalities, leading to a more effective fusion of PET and CT features. (3) Extensive evaluation of C²MAOT on two datasets demonstrates that our method significantly outperforms SOTA methods, achieving substantial improvements in segmentation accuracy across five tumor types.

2 Methodology

The framework of our proposed C²MAOT is shown in Fig. 1. During the pre-training phase (Fig. 1 (a)), PET and CT images are first masked using modality-

complementary masking. These masked images are then passed through separate encoders, producing feature maps from the intermediate layers. As depicted in Fig. 1 (b), we compute the 1-Wasserstein distance between the corresponding CT and PET features at each layer, resulting in the total OT loss $\mathcal{L}_{OT}$. The features from both encoders are concatenated and fed into a decoder for reconstruction. The reconstruction loss ($\mathcal{L}_R^{CT}$ and $\mathcal{L}_R^{PET}$) is calculated for the masked regions. The overall pre-training loss is defined as:

$$\mathcal{L} = \mathcal{L}_R^{CT} + \mathcal{L}_R^{PET} + \mathcal{L}_{OT}. \tag{1}$$

After pre-training, the encoder weights are transferred to the downstream tumor segmentation task. (Fig. 1 (c)).

2.1 Pre-training Backbone

Our pre-training backbone employs an adapted 3D U-Net [10] architecture, which continues to offer superior performance for volumetric medical imaging compared to Transformer-based methods when considering the critical balance between model effectiveness, parameter efficiency, and computational demands. Departing from conventional multimodal approaches like nnU-Net [20] that channel-concatenate modalities for single-encoder processing, our architecture implements dual independent encoders specifically optimized for PET-CT imaging. This allows the model to explicitly extract modality-specific features in the encoder stage, enhancing the representation of each modality's unique characteristics. The decoder receives skip connections and bottleneck features from both encoders, which are concatenated along the channel dimension to effectively combine the modality-specific information.

2.2 Modality-Complementary Masked Autoencoder

Inspired by He et al. [15], our pre-training employs a mask-and-reconstruction strategy. While vanilla MAE randomly masks patches in a single modality, PET and CT images—acquired simultaneously for the same patient—share similar anatomical and semantic structures. Therefore, we propose a modality-complementary masking strategy by randomly masking 50% of the 3D patches ($8 \times 8 \times 8$) in a complementary manner: for corresponding regions, if a PET patch is masked, the CT patch is kept intact, and vice versa. This design strategically compels the network to learn and integrate complementary information from the paired modality to complete the masked modality's missing data during reconstruction, effectively fostering deep cross-modal representation learning. Consistent with MAE, the reconstruction loss is computed solely over the masked regions using the standard L2 (MSE) loss formulation. The reconstruction losses for the PET and CT are defined as follows:

$$\mathcal{L}_{\mathrm{R}}^{PET} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{P}_i^{PET} - \hat{\mathbf{P}}_i^{PET}\|_2^2, \quad \mathcal{L}_{\mathrm{R}}^{CT} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{P}_i^{CT} - \hat{\mathbf{P}}_i^{CT}\|_2^2, \tag{2}$$

where $\mathbf{P}_i^{PET}$ and $\hat{\mathbf{P}}_i^{PET}$ represent the original and reconstructed voxel intensities for the i-th voxel of the PET image, respectively, and similarly for CT. These separate losses allow the model to optimize each modality independently, ensuring better reconstruction accuracy for both PET and CT images during the training process.

2.3 Optimal Transport Loss for Cross-Modal Feature Alignment

In order to effectively fuse features from PET and CT images while mitigating modality discrepancies, we introduce an Optimal Transport (OT) loss. OT is a method for measuring the difference between two modality feature distributions, with the goal of minimizing the transportation cost between them [26]. PET and CT images often exhibit differences in intensity distributions, resolution, and noise characteristics, making direct comparison or fusion of their features challenging. The OT loss helps to align the features from both modalities, facilitating better fusion while preserving each modality's unique characteristics. The core idea behind the OT distance is that, given two distributions, we aim to find an optimal mapping that minimizes the transportation cost between the two distributions. The cost is determined by calculating the distance between corresponding features. To achieve this, we use the 1-Wasserstein distance [27], which provides an effective way to measure the minimal transportation cost between two distributions. In our implementation, the features from PET and CT are extracted using two separate encoders, resulting in feature maps denoted as $\mathbf{F}_{PET}$ and $\mathbf{F}_{CT}$, respectively. The 1-Wasserstein distance is defined as:

$$W_1(\mathbf{F}_{PET}, \mathbf{F}_{CT}) = \inf_{\gamma \in \Gamma(\mathbf{F}_{PET}, \mathbf{F}_{CT})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}[\|\mathbf{x} - \mathbf{y}\|], \tag{3}$$

here, $\mathbb{E}$ represents the expectation, which refers to the average distance between the feature pairs $(\mathbf{x}, \mathbf{y})$. inf is the infimum, representing the minimum transportation cost across all possible transport plans. In simpler terms, this equation means that we seek to find a mapping γ that minimizes the cost of transferring features from the PET distribution to the CT distribution. To prioritize deeper features in the alignment process, we introduce a layer-specific weighting factor w_l, which increases exponentially with depth. Deeper layers tend to capture more abstract, modality-invariant semantic information that is critical for downstream tasks like tumor segmentation. In practice, we set the weights to follow an exponential growth, i.e., $w_l = \alpha^l$, where $\alpha > 1$ is a constant that controls the rate of growth. The final OT loss $\mathcal{L}_{OT}$ is computed as the weighted sum of 1-Wasserstein distances across all layers:

$$\mathcal{L}_{OT} = \sum_{l=1}^{N} w_l \cdot W_1(\mathbf{F}_{PET}^{(l)}, \mathbf{F}_{CT}^{(l)}), \tag{4}$$

where N is the number of layers in the encoder, $\mathbf{F}_{PET}^{(l)}$ and $\mathbf{F}_{CT}^{(l)}$ represent the feature maps at the l-th layer for PET and CT, respectively, and w_l is the weight associated with the l-th layer.

3 Experiments and Results

3.1 Datasets and Evaluation Metric

Datasets The study utilized PET-CT data from two sources: the AutoPET III Dataset [12] and an In-house Breast Dataset. The AutoPET III Dataset provided 1,014 cases (501 tumor-positive, 513 control) covering lung cancer, lymphoma, and melanoma. The In-house Breast Dataset contributed 393 cases (287 tumor-positive, 106 control) of primary and metastatic breast cancer (BC). We allocated 50% of the AutoPET III Dataset for pre-training, with the remaining data used alongside the In-house Breast Dataset for downstream segmentation evaluation.

Evaluation Metric Model performance was assessed using four metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), recall, and precision. DSC and IoU measure spatial overlap between predicted and ground truth segmentations, while recall and precision evaluate detection accuracy and reliability respectively.

3.2 Implementation Details

Experiments were conducted using Python 3.11, PyTorch 2.4.1, and Ubuntu 22.04 on dual NVIDIA 4090 GPUs (24GB memory each), with a patch size of $128 \times 128 \times 128$. Pre-training employed the Adam optimizer (initial learning rate: 1×10^{-4}, batch size: 2) for 500 epochs, including a 10-epoch warm-up period. Learning rate decay followed a polynomial strategy:

$$\text{lr} = \text{initial lr} \times \left(1 - \frac{\text{epoch}}{\text{max epoch}}\right)^{0.9}. \tag{5}$$

For both pre-training and tumor segmentation fine-tuning, we applied comprehensive data augmentation including geometric transformations, intensity modifications, and random noise. Fine-tuning maintained the initial learning rate and decay strategy, with training combining cross-entropy and dice loss over 1000 epochs with 250 iterations each.

3.3 Quantitative and Qualitative Results

Comparison with Existing SOTA Methods To evaluate the effectiveness of our method, we compared it with advanced segmentation methods including nnUNet [20], UNETR [28], 3DUX-Net [22], and U-Mamba [24], as well as SSL methods such as MoCov3 [8], MAE3D [9], Swin-MM [29], and VoCo [30] that utilize 3D UNet as the backbone. As shown in Table 1, our method achieved the highest or near-highest scores across all metrics including DSC, IoU, recall, and precision. Quantitatively, the average DSC surpasses the second-best method by 1.15% and the average IoU improves by 1.04%. Notably, with the exception of MoCov3 (which may be attributed to its requirement for large batch

sizes to obtain sufficient negative samples), self-supervised pretraining generally enhanced the performance of baseline models, demonstrating the potential of self-supervised methods in improving segmentation performance. The qualitative results in Fig. 2 further illustrate that our method generates segmentation results closely aligned with ground truth, with fewer instances of over-segmentation and under-segmentation, particularly for dispersed lesions such as metastatic breast cancer, lymphoma, and melanoma.

Table 1. Comparison for tumor segmentation. The evaluation metrics include DSC, IoU, Recall, and Precision in (%). Best results are shown in **bold**, while second-best results are underlined.

Category	Metrics	*From Scratch*				*With General SSL*		*With Medical SSL*		
		nnUNet	UNETR	3DUX-Net	U-Mamba	MoCov3	MAE3D	Swin-MM	VoCo	Proposed
Primary Breast Cancer	DSC	83.32	79.13	78.88	82.80	82.17	84.43	85.25	87.02	**88.18**
	IoU	71.41	65.47	65.13	70.65	69.74	73.28	74.46	77.49	**78.86**
	Recall	83.64	78.79	79.36	**90.27**	82.04	85.44	87.52	88.27	89.39
	Precision	84.28	82.85	81.60	83.48	81.89	85.23	86.62	**87.84**	87.37
Metastatic Breast Cancer	DSC	51.56	54.38	55.49	44.50	49.72	55.85	57.65	57.13	**59.47**
	IoU	34.73	37.34	38.40	28.62	33.08	38.74	40.50	40.27	**42.32**
	Recall	55.35	62.56	59.64	49.89	52.41	57.83	64.72	71.46	**72.43**
	Precision	54.25	54.74	58.28	50.38	51.28	56.27	63.39	72.11	**73.66**
AutoPET III Lesions	DSC	50.33	46.58	46.59	34.85	50.45	53.48	55.76	**57.40**	57.34
	IoU	33.63	30.36	30.37	21.10	33.73	36.50	38.66	**40.50**	40.19
	Recall	38.85	37.09	36.68	34.15	37.72	40.53	39.48	41.29	**42.50**
	Precision	45.24	43.18	44.10	37.72	43.63	46.85	**47.32**	46.43	46.22
Average	DSC	61.74	60.03	60.32	54.05	60.78	64.59	66.22	67.18	**68.33**
	IoU	46.59	44.39	44.63	40.12	45.52	49.51	51.21	52.75	**53.79**
	Recall	59.28	59.48	58.56	58.10	57.39	61.27	63.91	67.01	**68.11**
	Precision	61.26	60.26	61.33	57.19	58.93	62.78	65.78	68.79	**69.08**

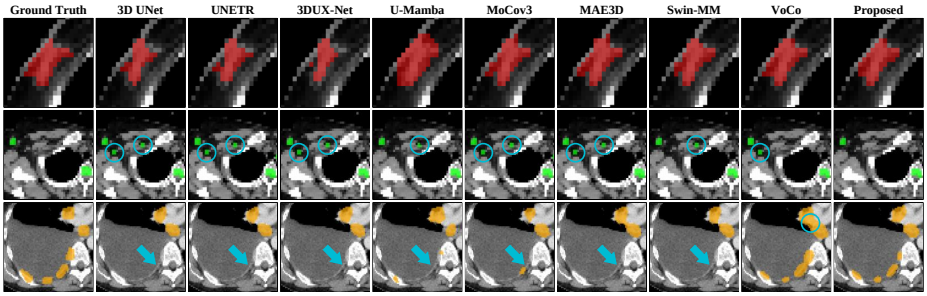

Fig. 2. Visualization of tumor segmentation. Red: primary breast cancer; Green: metastatic breast cancer; Orange: lung cancer, lymphoma and melanoma. Cyan: circles and arrows indicate over- and under-segmentation errors, respectively

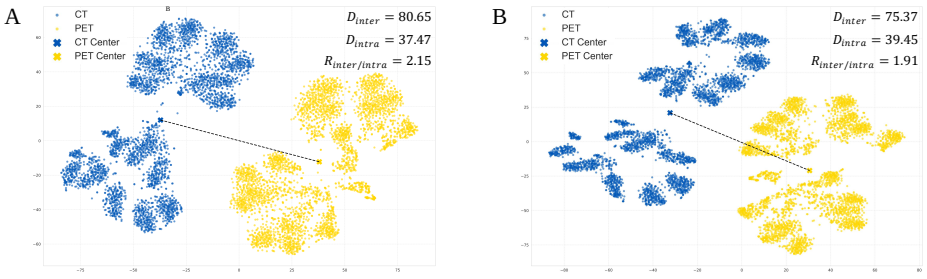

Fig. 3. T-SNE visualization of the CT and PET feature spaces. (A) Without OT loss, (B) With OT loss. Blue represents CT features, yellow represents PET features.

Table 2. Ablation study results on the average DSC (%).

Experimental Setting			Category			
Baseline	OT-Loss	MC-Mask	Primary-BC	Metastatic-BC	AutoPET	AVG
✓			85.12	53.42	52.08	63.54
✓	✓		86.03	55.14	53.78	64.98
✓		✓	85.74	57.13	55.26	66.04
✓	✓	✓	**88.18**	**59.47**	**57.34**	**68.33**

Ablation Study Table 2 demonstrates the contribution of key components to the model's performance. Starting with the baseline model (trained from scratch), we first introduce the OT loss, which leads to improvements in the DSC across all tumor types, with an average DSC increase of 1.44%. Incorporating the Modality-Complementary Mask (MC-Mask) further enhances performance, particularly for more dispersed tumors, resulting in a more substantial average DSC increase of 2.5%. The effect of OT loss can be visualized in Fig. 3, which illustrates its impact on feature distribution alignment. As shown, OT loss reduces the inter-modality distance while increasing intra-modality dispersion, decreasing the inter-to-intra ratio from 2.15 to 1.91, which indicates a more optimal feature space organization where different modalities are better aligned while preserving richer feature representations within each modality. This demonstrates that OT loss effectively promotes feature space consolidation while preserving modality-specific characteristics, facilitating more robust cross-modal fusion.

4 Conclusion

In this study, we proposed C^{2}MAOT, a novel pre-training segmentation framework specifically designed to address the challenges of cancer segmentation in PET-CT images. Our method leveraged a modality-complementary masking strategy within a masked autoencoder architecture to promote explicit cross-modal interaction and learning between PET and CT encoders. By incorporating an OT loss, our proposed method effectively aligned feature distributions

from both modalities, leading to enhanced and robust multi-modal feature fusion. Extensive experimental results on two datasets, covering a range of cancer types, conclusively demonstrated that our method achieved state-of-the-art segmentation performance, surpassing existing supervised and self-supervised approaches. The significant improvements observed underscored the efficacy of our complementary masking and OT guided fusion in capturing and integrating the complementary information inherent in PET-CT images. This work established C^{2}MAOT as a promising approach for advancing accurate and reliable cancer segmentation in multi-modal medical imaging.

Acknowledgements

This work was supported by the Science and Technology Development Fund of Macao (0105/2022/A).

Disclosure of Interests

The authors declare no competing interests.

References

1. Aide, N., Lasnon, C., et al.: Advances in pet/ct technology: an update. In: Seminars in nuclear medicine. vol. 52, pp. 286–301. Elsevier (2022)
2. Azam, M.A., et al.: A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Computers in biology and medicine **144**, 105253 (2022)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis (2024)
5. Chen, S., Wu, Z., Li, M., Zhu, Y., Xie, H., Yang, P., Zhao, C., Zhang, Y., Zhang, S., Zhao, X., et al.: Fit-net: Feature interaction transformer network for pathologic myopia diagnosis. IEEE Transactions on Medical Imaging **42**(9), 2524–2538 (2023)
6. Chen, S., Zhao, X., Wu, Z., Cao, K., Zhang, Y., Tan, T., Lam, C.T., Xu, Y., Zhang, G., Sun, Y.: Multi-risk factors joint prediction model for risk prediction of retinopathy of prematurity. EPMA Journal **15**(2), 261–274 (2024)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
8. 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)
9. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., et al.: Masked image modeling advances 3d medical image analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1970–1980 (2023)

10. Çiçek, Ö., et al.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
11. Gao, Y., Tan, T., et al.: Multi-modal longitudinal representation learning for predicting neoadjuvant therapy response in breast cancer treatment. IEEE Journal of Biomedical and Health Informatics (2025)
12. Gatidis, S., et al.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions sci. Data **9**(1), 601 (2022)
13. Grill, J.B., Strub, F., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
14. 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: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
17. Huang, J., Chen, S., Liang, X., Sun, Y., Hu, M., Tan, T.: All-in-one multi-organ segmentation in 3d ct images via self-supervised and cross-dataset learning. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
18. Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Medicine **6**(1), 74 (2023)
19. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., et al.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)
20. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
21. Khan, S.U., Khan, M.A., Azhar, M., Khan, F., Lee, Y., Javed, M.: Multimodal medical image fusion towards future research: A review. Journal of King Saud University-Computer and Information Sciences **35**(8), 101733 (2023)
22. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022)
23. Lu, J., Chen, J., Cai, L., Jiang, S., Zhang, Y.: H2aseg: Hierarchical adaptive interaction and weighting network for tumor segmentation in pet/ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 316–327. Springer (2024)
24. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
25. Ming, Y., Wu, N., et al.: Progress and future trends in pet/ct and pet/mri molecular imaging approaches for breast cancer. Frontiers in oncology **10**, 1301 (2020)

26. Montesuma, E.F., Mboula, F.M.N., Souloumiac, A.: Recent advances in optimal transport for machine learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
27. Panaretos, V.M., Zemel, Y.: Statistical aspects of wasserstein distances. Annual review of statistics and its application **6**(1), 405–431 (2019)
28. Tang, Y., Yang, D., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
29. Wang, Y., Li, Z., et al.: Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 486–496. Springer (2023)
30. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22873–22882 (2024)
31. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
32. Zhang, T., Tan, T., et al.: Predicting breast cancer types on and beyond molecular level in a multi-modal fashion. NPJ breast cancer **9**(1), 16 (2023)
33. Zhang, Z., Han, L., Sun, Y., Tan, T., et al.: Unimrisegnet: Universal 3d network for various organs and cancers segmentation on multi-sequence mri. IEEE Journal of Biomedical and Health Informatics (2024)