CDFP-Net: Cross-Modal Dynamic Fusion with Diffusion Priors for PET-CT Tumor Segmentation

Minqin Wei[1], Shangqian Wu[2], Chang Xu[2], Yurong Qian[1], Yongjun Tang[3], Jinmiao Song[1], and Lei Deng[1,2,*]

[1] Xinjiang Key Laboratory of Intelligent Computing and Smart Applications, School of Software, Xinjiang University, Urumqi, China

[2] School of Computer Science and Engineering, Central South University, Changsha, China

leideng@csu.edu.cn

[3] Department of Pediatrics, Xiangya Hospital, Central South University, Changsha, China

* Corresponding Author

Abstract. Accurate automated tumor segmentation in PET-CT imaging is critical for clinical diagnosis and radiotherapy planning. However, limited annotated data, the high computational cost of volumetric modeling, and intrinsic physical discrepancies between metabolic PET and anatomical CT modalities pose substantial challenges. To address these issues, we propose **CDFP-Net**, a dual-stream framework that integrates generative diffusion priors with cross-modal dynamic graph fusion. Specifically, we employ a frozen, self-supervised pre-trained Denoising Diffusion Probabilistic Model (DDPM) as a robust feature prior, augmented with lightweight adapters to enhance modality-specific representations under limited supervision. To effectively bridge cross-modal discrepancies, we design a modality-asymmetric dynamic graph fusion mechanism in which metabolically active PET regions act as spatial anchors to query and aggregate complementary anatomical boundary cues from CT. To alleviate the computational burden of full 3D processing while preserving global semantic topology, we further introduce a Multi-Angle Maximum Intensity Projection (MA-MIP) strategy for efficient volumetric context modeling. Extensive experiments on three multi-center datasets (AutoPET, HECKTOR, and PCLT20K) demonstrate that CDFP-Net consistently outperforms state-of-the-art methods in both accuracy and robustness, highlighting its strong potential for precise biological target volume (BTV) delineation in clinical practice. Source code is available at https://github.com/WeiMinqin/CDFP-Net.

Keywords: PET-CT segmentation · Diffusion Model · Cross-modal Fusion.

1 Introduction

Multimodal positron emission tomography-computed tomography (PET-CT) imaging, combining the high-resolution anatomical structure of CT with the

abnormal metabolic activity of PET, has become the gold standard for clinical tumor diagnosis and radiotherapy planning[20]. However, achieving automated and precise tumor segmentation still faces severe challenges[11,21]. First, data-sharing barriers and prohibitive expert-annotation costs lead to a scarcity of high-quality multi-modal samples, hindering robust model generalization. Second, conventional 3D convolutions incur massive memory overhead, and the processing of numerous background slices inherently leads to computational redundancy [11,2]. Finally, inherent inter-modal gaps and multi-centric intra-modal heterogeneities [18,17] pose significant barriers to synergistic PET-CT feature fusion and accurate segmentation.

Existing mainstream frameworks are mainly based on supervised learning and are susceptible to the inherent statistical noise of PET, leading to jagged boundary distortions or isolated false-positive results [4,6]. At the modality fusion level, existing channel-wise concatenation strategies ignore the physical heterogeneity and spatial non-homogeneity between PET and CT. Such "shallow coupling" frequently triggers modal feature repulsion[2,3], causing strongly-weighted anatomical structural information to overshadow weakly-weighted metabolic functional signals, thereby limiting the depth of cross-modal synergistic representation.

To address these challenges, we propose CDFP-Net, a dual-stream architecture integrating generative diffusion priors with cross-modal dynamic graph fusion. We employ a frozen dual-stream Denoising Diffusion Probabilistic Model (DDPM) [8] as a feature backbone, leveraging self-supervised denoising priors for semantic-level feature regularization. To reduce 3D computational cost while preserving global context, we introduce a Multi-Angle Maximum Intensity Projection (MA-MIP) strategy [19], transforming volumetric data into semantically enriched 2.5D multi-view representations. Cross-modal heterogeneity is addressed through lightweight adapters composed of Input Learnable Normalization (ILN) and Dual-node Squeeze-and-Excitation (Dual-SE) modules. Furthermore, we design a modality-asymmetric dynamic graph fusion mechanism that enables PET-dominant, structure-assisted cross-modal interaction. Extensive evaluations on AutoPET, HECKTOR, and PCLT20K demonstrate consistent improvements over state-of-the-art methods across DSC, HD95, and Recall.

Our main contributions are summarized as follows:

1. Diffusion-prior-guided representation learning for segmentation. We introduce a parameter-efficient framework that transfers self-supervised diffusion priors from a frozen DDPM backbone to PET-CT tumor segmentation via lightweight adapters. This design enables prior-regularized feature adaptation under limited supervision and improves robustness to multi-center distribution shifts.
2. Modality-asymmetric dynamic graph fusion. We propose a directed cross-modal interaction mechanism formulated as dynamic graph message passing, where metabolically salient PET regions actively query and aggregate complementary structural information from CT. This asymmetric formulation mitigates modal feature competition and enables function-dominant, structure-aware fusion beyond conventional symmetric concatenation.

3. Efficient 2.5D global semantic modeling via MA-MIP. We leverage MA-MIP as a data-processing tool to preserve global topology while substantially reducing 3D overhead for volumetric segmentation.

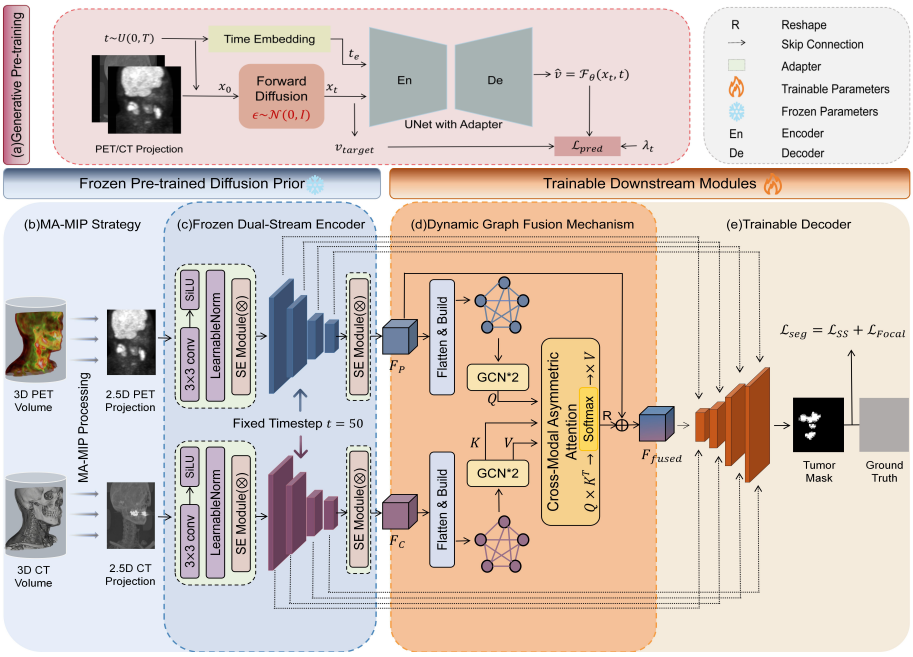

Fig. 1. Overview of the proposed CDFP-Net framework. The pink, blue, and orange regions denote the upstream pre-training process, frozen diffusion priors, and trainable downstream modules, respectively.

2 Method

2.1 Overall Architecture

As illustrated in Fig.1, the proposed CDFP-Net is a novel dual-stream multi-modal segmentation framework designed to leverage large-scale generative pre-training priors to enhance the accuracy and robustness of PET-CT tumor segmentation. As depicted in Fig.1(a), generative pre-training is conducted based on a UNet architecture equipped with adapters, adding noise through a forward diffusion process and predicting the target velocity (v-prediction). During data preprocessing (Fig.1(b)), the 3D images are processed via multi-angle maximum intensity projection (MA-MIP) and converted into 2.5D multi-angle projection maps rich in global spatial semantics. In the downstream fine-tuning stage, the

model utilizes the pre-trained and frozen dual-stream encoder for feature extraction (Fig.1(c)). Subsequently, as shown in Fig.1(d), deep interaction and dynamic fusion of PET and CT features are achieved through graph convolutional networks (GCN) and a cross-modal asymmetric attention mechanism. Finally, Fig.1(e) utilizes the fused features to reconstruct the tumor regions within a trainable decoder via multi-scale upsampling and skip connections.

2.2 Pre-training and Architectural Adapters

To balance computational efficiency and the preservation of spatial topological information, we introduce MA-MIPs as the 2.5D input strategy. By performing rotational projection along a specific axis with a fixed step size for PET and CT, the maximum voxel values along the projection rays (maximum standardized uptake value for PET and maximum Hounsfield Unit for CT) are preserved, effectively compressing the 3D spatial structure into a series of 2D projection maps rich in global semantics.

During the self-supervised pre-training phase, the denoising diffusion probabilistic model (DDPM) is trained independently on a large corpus of unlabeled PET and CT data. Given the original image $\mathbf{x}_0$ and a random time step t, the forward process adds Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to obtain $\mathbf{x}_t$. To improve the stability and convergence speed of multi-level feature learning, we employ a pre-training loss function based on target velocity (v-prediction)[15] and signal-to-noise ratio (SNR) adaptive weighting:

$$\mathcal{L}_{pre} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\lambda_t \left\|\mathcal{F}_\theta(\mathbf{x}_t, t) - \mathbf{v}_{target}\right\|_2^2\right] \tag{1}$$

where the target velocity is defined as $\mathbf{v}_{target} = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t}\mathbf{x}_0$, and the time step smoothing weight is $\lambda_t = (1 + \mathrm{SNR}(t))^{-1}$. This objective forces the encoder $\mathcal{F}_\theta$ to deeply capture the intrinsic data manifold of real medical images.

To enhance the perceptual capability of specific modalities without destroying the fundamental generative manifold of the diffusion model, we introduce lightweight architectural adapters[1] during the pre-training phase.

Input Learnable Normalization (ILN): Composed of a 3×3 convolution, SiLU activation, and a Learnable Normalization (LearnableNorm) layer, this module is situated at the encoder's input stage. It aims to smoothly transition image pixels into the diffusion model's shallow feature space while compensating for the physiological contrast loss induced by mandatory $[-1, 1]$ input scaling. Given the input feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, its channel-wise normalization formula is defined as:

$$\hat{\mathbf{X}}_c = \lambda_c \frac{\mathbf{X}_c - \mu_c}{\sqrt{\sigma_c^2 + \epsilon_{num}}} + \delta_c \tag{2}$$

where μ_c and σ_c^2 are the mean and variance in the channel dimension; λ_c and δ_c are learnable affine parameters; and ϵ_{num} is a small constant utilized for numerical stability.

Dual-node Squeeze-and-Excitation (Dual-SE): This module is embedded after the data input layer and at the bottleneck layer of the encoder.

Given the input $\mathbf{Z}$, the channel weight calibration process can be represented as $\mathbf{s} = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\text{GAP}(\mathbf{Z})))$, where $\text{GAP}(\cdot)$ denotes the global average pooling operation, while σ and δ represent the Sigmoid and ReLU activation functions, respectively. The shallow SE module is utilized to recalibrate low-level textures, while the bottleneck SE module dynamically strengthens deep channels associated with high metabolism and anatomical boundaries.

2.3 Downstream Fine-Tuning and Dynamic Graph Fusion

In downstream tasks, the pre-trained dual-stream diffusion encoder remains completely frozen. Given a fixed time step t, the encoder extracts multi-scale deep features from PET and CT, respectively. Inspired by the Cross-Modality Heterogeneous Graph Fusion (CMHGF) network[23], we design a lightweight modality-asymmetric adaptive graph fusion mechanism.

This mechanism first flattens the spatial dimensions of the bottleneck feature maps $\mathbf{F}_P, \mathbf{F}_C \in \mathbb{R}^{C\times H\times W}$ into $N = H \times W$ nodes to construct graph representations $\mathbf{V}_P, \mathbf{V}_C \in \mathbb{R}^{N\times C}$. Within their respective modalities, a two-layer GCN is employed for the smooth aggregation of topological features:

$$\tilde{\mathbf{V}}^{(l+1)} = \text{ReLU}(\mathbf{A}\tilde{\mathbf{V}}^{(l)}\mathbf{W}_{gcn}^{(l)}) \tag{3}$$

where $\mathbf{A}$ is the adjacency matrix dynamically learned by calculating the similarities between nodes. Subsequently, non-local cross-attention is utilized to complete the heterogeneous interaction. The enhanced PET nodes are projected as the query matrix $\mathbf{Q} = \tilde{\mathbf{V}}_P\mathbf{W}_Q$, and the corresponding CT nodes are projected as the key $\mathbf{K} = \tilde{\mathbf{V}}_C\mathbf{W}_K$ and value $\mathbf{V} = \tilde{\mathbf{V}}_C\mathbf{W}_V$. The cross-modal feature reconstruction is formulated as follows:

$$\mathbf{V}_{cross} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_{out} \tag{4}$$

This operation explicitly prompts the highly metabolic PET signals to serve as spatial anchors, actively querying the anatomical boundaries within the CT modality. Finally, the fused graph features are mapped back to the image space and residually added to the original PET bottleneck features:

$$\mathbf{F}_{fused} = \text{Reshape}(\mathbf{V}_{cross}) + \mathbf{F}_P \tag{5}$$

The network is optimized via backpropagation by minimizing the joint loss $\mathcal{L}_{seg} = \lambda_1\mathcal{L}_{SS} + \lambda_2\mathcal{L}_{Focal}$, where $\mathcal{L}_{SS}$ is a spatially standardized hard-mining loss adapted from Purma et al.[13]. This loss explicitly forces the network to focus on highly heterogeneous and difficult-to-segment tumor regions by dynamically weighting the standard voxel-wise cross-entropy ($\mathcal{L}_{CE}$). Specifically, we compute an absolute error map $e = |e_1 - e_2|$, where e_1 and e_2 are the spatially standardized ground truth and prediction, defined as $e_x = (x - \mu_x + C)/(\sigma_x + C)$ with a stability constant $C = 0.01$. To perform hard mining, a binary mask $f = \mathbb{I}(e > 0.1 \cdot e_{max})$

isolates pixels exceeding 10% of the batch maximum error e_{max}. $\mathcal{L}_{SS}$ is then computed as:

$$\mathcal{L}_{SS} = \frac{\sum(e \cdot f \cdot \mathcal{L}_{CE})}{\sum f} \tag{6}$$

Synergizing $\mathcal{L}_{SS}$ with $\mathcal{L}_{Focal}$ efficiently handles extreme class imbalance while strictly penalizing boundary prediction errors, realizing a comprehensive optimization.

3 Experiments and Results

3.1 Datasets and Preprocessing

Datasets. We evaluate the proposed method on three public datasets. **AutoPET**[5] contains 1,014 3D whole-body PET-CT scans covering melanoma, lymphoma, and lung cancer, with a balanced distribution of positive and negative samples. **HECKTOR**[14] aggregates approximately 1,200 3D registered images from 11 different centers, focusing on the segmentation of primary oropharyngeal cancer lesions and lymph nodes in the head and neck. **PCLT20K**[12] is a 2D lung tumor PET-CT dataset consisting of 21,930 pairs of slice data from 605 patients. We adopt patient-level partitioning to prevent data leakage.

3.2 Implementation Details and Metrics

Implementation. CDFP-Net is implemented in PyTorch and trained on a single NVIDIA RTX A6000 (48GB) GPU. The Adam optimizer is used with an initial learning rate of 10^{-5}. The total diffusion timesteps are set to $T = 1000$, while the timestep for feature fusion is empirically fixed at $t = 50$. **Evaluation.** Segmentation performance is quantitatively evaluated using the Dice Similarity Coefficient (DSC), the 95% Hausdorff Distance (HD95), and Recall.

3.3 Quantitative and Qualitative Segmentation Results

We compared CDFP-Net with seven baseline and SOTA methods across three datasets. Table 1 presents the comprehensive evaluation results.

Quantitative Analysis. Table 1 demonstrates that CDFP-Net achieves the best average performance across all datasets (DSC: 0.7583, HD95: 19.41 mm, Recall: 0.8334). While nnUNet shows competitive HD95 scores on AutoPET, its performance fluctuates significantly across different datasets. In contrast, CDFP-Net maintains superior robustness across diverse anatomical regions.

Qualitative Evaluation. Fig. 2 compares CDFP-Net with representative state-of-the-art methods on three challenging HECKTOR cases. Overall, CDFP-Net yields more complete delineations with fewer false positives and false negatives, resulting in higher Dice scores. Row 2 illustrates a particularly difficult case with extremely low PET tracer uptake. Competing methods under-segment the tumor (e.g., SwinUNETR, DSC = 0.52), whereas CDFP-Net achieves a DSC

Table 1. Comparative experimental results on three datasets. The unit for HD95 is millimeters (mm), and DSC and Recall are reported as percentages (%). Best and second-best results are **bold** and underlined.

Method	AutoPET			HECKTOR			PCLT20K		
	DSC↑	HD95↓	Recall↑	DSC↑	HD95↓	Recall↑	DSC↑	HD95↓	Recall↑
nnUNet[10]	62.34	**14.34**	78.07	68.05	19.56	76.20	77.91	53.98	**87.40**
SwinUNETR[7]	51.43	74.77	44.53	65.69	12.22	70.83	80.56	46.17	85.34
STU-Net[9]	57.70	71.40	78.84	67.45	18.75	77.55	57.25	63.84	55.52
UNETR++[16]	46.38	97.15	70.74	67.46	19.60	76.54	81.66	54.29	83.02
MedCoSS[22]	53.55	43.42	79.97	63.82	22.31	75.52	31.38	66.18	33.77
CIPA[12]	55.49	19.48	72.67	68.68	16.52	78.33	76.15	40.88	77.57
AAHN[23]	61.63	38.19	78.67	65.84	20.05	69.42	74.02	80.17	72.20
CDFP-Net	**63.33**	18.47	**84.14**	**78.44**	**8.26**	**79.72**	**85.72**	**31.51**	86.16

of 0.89 and produces a more continuous boundary. This improvement can be attributed to the modality-asymmetric fusion mechanism, which enables PET-guided aggregation of complementary CT structural cues, and the diffusion priors that enhance morphological consistency in low-contrast regions. These findings are consistent with the ablation results in Sec. 3.4 and are relevant for reliable Biological Target Volume (BTV) delineation.

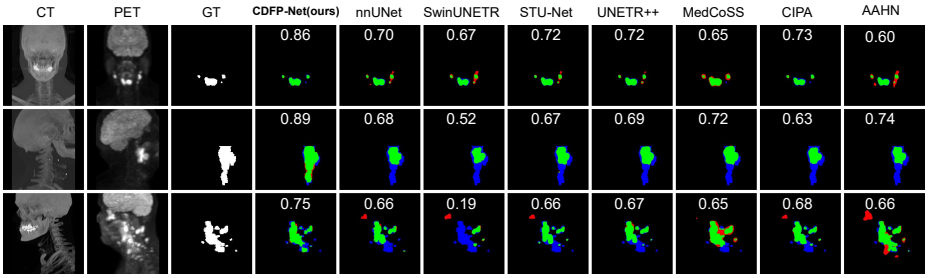

Fig. 2. Visual comparison of our method against SOTA on three sample cases from the HECKTOR dataset. Green: true positive; Red: false positive; Blue: false negative. The numbers denote the corresponding Dice scores.

3.4 Ablation Study

Ablation experiments were conducted on the **HECKTOR** dataset to evaluate the core architectural designs and hyperparameter settings of CDFP-Net.

Components and Pre-training. Table 2 demonstrates that utilizing diffusion pre-training (Diff-Base) improves the DSC by 5.38% compared to training from scratch. Integrating the ILN layers and Dual-SE adapters (Proposed Full) yields the optimal results (DSC: 78.44%, HD95: 8.26 mm), validating the efficacy of our lightweight adaptation strategy.

Table 2. Ablation study on pre-training strategies and architectural adapters.

Model	ILN[1]	Dual-SE[2]	DSC (%) ↑	HD95 ↓
Scratch	—	—	65.99	24.48
Diff-Base	×	×	71.37	16.52
+ ILN	✓	×	77.34	9.70
+ Dual-SE	×	✓	77.47	9.37
Full	✓	✓	**78.44**	**8.26**

[1] ILN: Input Learnable Normalization;
[2] Dual-SE: Dual-node SE.

Table 3. Ablation study on fixed downstream time steps t.

t	DSC (%) ↑	HD95 ↓
0	75.92	11.98
50	**78.44**	**8.26**
100	75.28	13.55
200	70.47	17.80
400	63.56	31.69

Fusion and Modality Interaction. (1) **Guidance:** Setting the Query to PET (Q = PET) and the Key/Value to CT (K, V = CT) outperforms the reverse configuration (Table 4(a)). This confirms our hypothesis that highly metabolic PET signals should serve as the primary functional "anchor," while CT provides the necessary geometric "constraints." (2) **Strategy:** As shown in Table 4(b), our modality-asymmetric graph-based fusion surpasses traditional Concatenation and Addition strategies in capturing long-range dependencies and refining boundary details.

Timestep Sensitivity. As depicted in Table 3, setting the diffusion timestep to $t = 50$ achieves an optimal balance between abstract semantic features and precise structural details. Increasing the timestep beyond $t = 100$ introduces excessive Gaussian noise, leading to a sharp degradation in the HD95 metric.

Table 4. Ablation study on cross-modal fusion strategies. (a) Multi-modal Guidance strategies, where Q and K,V denote the Query and Key/Value modalities. (b) Fusion methods.

Aspect	Method	DSC (%) ↑	HD95 (mm) ↓	Params (M)
(a) Guidance	PET-only	75.31	12.16	49.02
	Q=CT, K,V=PET	77.36	10.06	86.64
	Q=PET, K,V=CT	**78.44**	**8.26**	86.64
(b) Fusion	Concat.	76.58	9.97	84.64
	Add.	76.02	10.20	84.37
	Graph	**78.44**	**8.26**	86.64

4 Conclusions

We proposed CDFP-Net, a dual-stream framework for PET-CT tumor segmentation that integrates diffusion-prior-guided representation learning with modality-asymmetric dynamic graph fusion. A frozen DDPM backbone with lightweight

adapters enables prior-regularized feature adaptation under limited supervision, while the MA-MIP strategy provides efficient 2.5D global semantic modeling to reduce 3D computational cost. The directed cross-modal graph mechanism further promotes PET-guided aggregation of complementary CT structural cues, improving boundary consistency. Experiments on three multi-center datasets (AutoPET, HECKTOR, and PCLT20K) show consistent improvements over strong baselines across standard metrics, demonstrating the robustness and generalizability of the proposed approach.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (Grant Nos. U23A20321 and 62272490) and the Natural Science Foundation of Hunan Province of China (Grant No. 2025JJ20062).

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

References

1. Ahmad, I., Anwar, S.J., Hussain, B., et al.: Anatomy guided modality fusion for cancer segmentation in pet ct volumes and images. Scientific Reports **15**(1), 12153 (2025)
2. Aksu, F., Gelardi, F., Chiti, A., et al.: Multi-stage intermediate fusion for multimodal learning to classify non-small cell lung cancer subtypes from ct and pet. Pattern Recognition Letters **193**, 86–93 (2025)
3. Alshmrani, G.M., Ni, Q., Jiang, R., et al.: Hyper-dense_lung_seg: Multimodal-fusion-based modified u-net for lung tumour segmentation using multimodality of ct-pet scans. Diagnostics **13**(22), 3481 (2023)
4. Azimi, M.S., Felfelian, V., Zeraatkar, N., et al.: Deep supervised transformer-based noise-aware network for low-dose pet denoising across varying count levels. Computers in Biology and Medicine **196**, 110733 (2025)
5. Gatidis, S., Kuestner, T.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-ct-lesions) (version 1) [data set] (2022). https://doi.org/10.7937/GKR0-XV29
6. Gong, K., Johnson, K., El Fakhri, G., et al.: Pet image denoising based on denoising diffusion probabilistic model. European Journal of Nuclear Medicine and Molecular Imaging **51**(2), 358–368 (2024)
7. He, Y., Nath, V., Yang, D., et al.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer Nature Switzerland (2023)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
9. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., Qiao, Y.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training (2023)
10. Isensee, F., Jaeger, P.F., Kohl, S.A.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 (Feb 2021)

11. Jeon, Y.S., Yang, H., Fu, H., et al.: No more sliding window: efficient 3d medical image segmentation with differentiable top-k patch sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 376–386. Springer Nature Switzerland (2025)
12. Mei, J., Lin, C., Qiu, Y., et al.: Cross-modal interactive perception network with mamba for lung tumor segmentation in pet-ct images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15653–15662 (2025)
13. Purma, V., Srinath, S., Srirangarajan, S., et al.: Genselfdiff-his: generative self-supervision using diffusion for histopathological image segmentation. IEEE Transactions on Medical Imaging **44**(2), 618–631 (2024)
14. Saeed, N., Hassan, S., Hardan, S., et al.: A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. arXiv preprint arXiv:2509.00367 (2025)
15. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=TIdIXIpzhoI
16. Shaker, A., Maaz, M., Rasheed, H., et al.: Unetr++: delving into efficient and accurate 3d medical image segmentation. IEEE Transactions on Medical Imaging **43**(9), 3377–3390 (2024)
17. Shekari, M., Verwer, E.E., Yaqub, M., et al.: Harmonization of brain pet images in multi-center pet studies using hoffman phantom scan. EJNMMI Physics **10**(1), 68 (2023)
18. Shimovolos, S., Shushko, A., Belyaev, M., et al.: Adaptation to ct reconstruction kernels by enforcing cross-domain feature maps consistency. Journal of Imaging **8**(9), 234 (2022)
19. Toosi, A., Harsini, S., Bénard, F., et al.: How to segment in 3d using 2d models: automated 3d segmentation of prostate cancer metastatic lesions on pet volumes using multi-angle maximum intensity projections and diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 212–221. Springer Nature Switzerland (2024)
20. Trotter, J., Pantel, A.R., Teo, B.K.K., et al.: Positron emission tomography (pet)/computed tomography (ct) imaging in radiation therapy treatment planning: a review of pet imaging tracers and methods to incorporate pet/ct. Advances in Radiation Oncology **8**(5), 101212 (2023)
21. Xing, X., Rafique, M.U., Liang, G., et al.: Efficient training on alzheimer's disease diagnosis with learnable weighted pooling for 3d pet brain image classification. Electronics **12**(2), 467 (2023)
22. Ye, Y., Xie, Y., Zhang, J., et al.: Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11114–11124 (2024)
23. Zheng, S., Ye, X., Yang, C., et al.: Asymmetric adaptive heterogeneous network for multi-modality medical image segmentation. IEEE Transactions on Medical Imaging **44**(4), 1836–1852 (2025)