Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.29928

Recent Advances in Deep Multimodal Fusion of PET/CT Imaging for Early Diagnosis and Tumor Segmentation in Lung Cancer

Ruidong Fan

University of New South Whales, Sydney, Australia
frd_unsw@outlook.com

Abstract. Accurate segmentation of lung tumors in PET/CT images plays a critical role in early diagnosis and treatment planning. This study reviews recent advances in multimodal fusion networks for tumor segmentation, with a focus on feature-level and decision-level fusion strategies, uncertainty quantification, and widely used benchmark datasets. Representative models are analyzed, such as ISA-Net, Mirror U-Net, evidential fusion methods, and Mamba-based cross-modal networks. Results show that fusion approaches significantly improve Dice and IoU compared to unimodal baselines, while uncertainty-aware frameworks enhance trustworthiness for clinical adoption. In addition, current limitations are discussed, including data scarcity, generalizability, and clinical translation challenges. This work provides a condensed perspective on the state-of-the-art and future directions in PET/CT tumor segmentation.

Keywords: PET/CT imaging, Lung cancer segmentation, Multimodal fusion networks, Uncertainty quantification, Deep learning

1. Introduction

Lung cancer is one of the most common malignancies and the leading cause of cancer-related deaths worldwide, with an estimated 2.2 million new cases and 1.8 million deaths in 2020 [1]. Early diagnosis and precise tumor segmentation are critical for individualized treatment planning and prognosis evaluation. Medical imaging, especially positron emission tomography (PET) and computed tomography (CT), plays a central role in this process. PET reflects tissue metabolism, helping distinguish benign from malignant lesions, while CT provides high-resolution anatomical detail. However, PET is limited by low spatial resolution and false positives in high-uptake regions, and CT struggles to distinguish tissues with similar densities. Their integration in PET/CT combines functional and anatomical information, improving diagnostic accuracy and staging [2].

Despite these advantages, clinical segmentation still mainly depends on manual annotation, which is both time-consuming and subjective. Traditional threshold- or region-based methods are constrained by heterogeneous uptake and irregular tumor morphology. Deep learning, especially convolutional neural networks such as U-Net and its derivatives, has transformed medical image segmentation, demonstrating robust performance on large multimodal datasets [3]. Building on this,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.29928

deep multimodal fusion methods integrate PET and CT at the input, feature, or decision level, enhanced by attention, cross-modal interaction, and multi-task strategies. These approaches have proven effective in addressing tumor heterogeneity and reducing segmentation uncertainty.

Recent studies show that fusion models significantly improve detection and segmentation of small, low-contrast, or morphologically complex lesions, with higher sensitivity and specificity. Nonetheless, challenges remain, including heterogeneous data standardization, scarcity of labeled data, limited model generalization, and weak interpretability. Future directions include federated learning, cross-domain adaptation, and multi-omics fusion, with the goal of developing scalable, multi-center intelligent diagnostic systems. This review summarizes the theoretical basis, representative models, clinical validation, and future trends of PET/CT-based deep multimodal fusion in lung cancer diagnosis and segmentation.

2. Background and related work

2.1. PET and CT imaging in lung cancer

PET/CT has become a cornerstone in lung cancer diagnosis and treatment, providing complementary functional and anatomical information. Clinically, it supports differentiation between benign and malignant solitary pulmonary nodules (SPNs), staging accuracy for N and M classification, motion management, and comparative imaging. For example, PET/CT-based malignancy models have achieved pooled sensitivity of 88%, specificity of 78%, and AUC of 0.91; staging performance has reached ~0.82/0.88 for N staging and ~0.86/0.89 for M staging. Deep learning–assisted respiratory correction increased SUVmax/SUVmean by ~10%/17% in 906 lesions and improved contour sharpness. Comparative studies with DWI further confirmed PET/CT as competitive and complementary. Evidence from 2022–2025 consistently demonstrates PET/CT as a robust tool for SPN discrimination and staging, with motion compensation and registration control enhancing performance in small or low-contrast lesions [4-7].

2.2. Deep learning in medical image segmentation

Deep learning, especially U-Net and nnU-Net, has become the standard for medical image segmentation. nnU-Net introduces adaptive pipelines across preprocessing, augmentation, and post-processing, providing strong baselines across multi-center datasets. In PET/CT tasks, U-Net/nnU-Net often serve as backbones, enhanced with cross-modal attention, uncertainty modeling, or interaction modules. Key challenges include registration errors, class imbalance due to sparse tumor voxels, false positives in high-uptake regions, domain shifts across institutions, and costly voxel-level annotations. Addressing these issues involves focal or composite loss, re-weighted sampling, federated and weakly supervised learning, and leveraging challenging public datasets. Overall, research is increasingly focused on improving robustness, generalization, and efficiency [2, 3, 5, 7-11].

2.3. Multimodal fusion strategies

PET/CT fusion approaches are mainly categorized into early (pixel-level), intermediate (feature-level), and late (decision-level) fusion, with hybrid and multi-task networks also emerging. Each strategy presents trade-offs in expressive power, robustness, interpretability, and computational complexity, forming the methodological foundation for current multimodal PET/CT segmentation research [8-10, 12-14], as summarized in Table 1.

DOI: 10.54254/2755-2721/2025.29928

Table 1. Multimodal fusion strategies in PET/CT imaging for tumor segmentation

Integration level	Typical structure/operation	Advantages	Limitations	Applicable scenarios	Ref
Pixel-level (early)	Channel cascading, input concatenation	Simple structure, end-to-end	Sensitive to registration/noise and poor cross-domain robustness	Limited data, fast baseline	[2]
Feature level	Co-modal/cross-modal attention, graph attention, Transformer interaction	Strong expressiveness and good robustness	Design complex, and large quantities	Complex anatomy, low contrast/small lesions	[2, 6-8, 10]
Decision level	Weighted/voting, evidence theory	uncertainty quantification good interpretability	End-to-end training is restricted and dependent on the quality of sub-models	Robust and interpretable output is required	[4，5, 14]
Hybrid/ multi tasking	Multi-branch + task parallelism	Comprehensive advantages, difficult and robust	Training/parameter tuning is complex and consumes a lot of resources	SOTA/ Clinical implementation	[9, 12，15]

3. Deep multimodal fusion methods for PET/CT

3.1. Early fusion (input-level fusion)

Early fusion concatenates registered PET and CT at the input layer after SUV/HU normalization and resampling, then processes them through a single encoder-decoder. To compensate for PET's low resolution, native CT details are often preserved via skip connections. This approach is simple, lightweight, and fast, thus widely used as a baseline or ablation control. However, it remains sensitive to registration errors and noise, with PET blurring and breathing mismatch amplified in shallow layers, and it cannot explicitly model modality complementarity or uncertainty [2, 7].

3.2. Feature-level fusion

Feature-level fusion introduces cross-modal interactions in the encoder, bottleneck, or decoder to address the sensitivity of early fusion. Strategies include concatenation with channel weighting, spatial and channel attention (e.g., ISA-Net), co-attention for functional–anatomical alignment, graph-based aggregation, multi-scale fusion, and cross-modal Transformers. Multi-task connected U-Net further leverages PET signals to guide CT segmentation. These approaches enhance recall for low-contrast and small lesions, demonstrate greater robustness across centers, and have become the mainstream direction [2, 6-8, 10, 12].

3.3. Decision-level and evidential fusion

Decision-level fusion trains PET-only and CT-only streams (or multi-scale features) and integrates them through weighting or voting, reducing the risk of single-modality failure. Evidential fusion extends this by outputting Dirichlet parameters to provide both class probabilities and uncertainty, thereby improving reliability and supporting clinical review. Bayesian modeling further aids in threshold adaptation and workflow prioritization [4,16].

DOI: 10.54254/2755-2721/2025.29928

3.4. Hybrid and multi-task architectures

Hybrid methods combine feature-level interaction with decision-level robustness while incorporating auxiliary tasks such as boundary or quality estimation. Representative examples include Mirror U-Net (modal fission with multitasking), adaptive multi-task weighting via uncertainty, and multi-branch fusion frameworks. These architectures enhance consistency in challenging cases but demand great computational resources and more intensive parameter tuning [9,15,17].

3.5. Interactive segmentation

Human-in-the-loop approaches introduce expert priors via clicks or scribbles to refine model predictions, particularly under weak or semi-supervision. In PET/CT, interactions constrain PET hotspots and CT boundaries, thereby reducing false positives. Multimodal interactive frameworks, such as those evaluated on AutoPET, demonstrate that limited interaction can approach full supervision quality, improve Dice/IoU, and reduce annotation cost. Incorporating uncertainty further enables selective review and cross-center adaptation through federated learning [3–5, 14].

4. Representative networks and key advances

4.1. Baseline networks and variants

In recent years, PET/CT tumor segmentation has progressed from baseline networks to multimodal fusion frameworks, achieving notable improvements. The encoder–decoder architecture remains the core foundation. Table 2 presents representative baseline performance on AutoPET-II and HECKTOR2022, showing that while these models achieve reasonable Dice scores, they remain limited in handling complex lesions and cross-center generalization.

Table 2. Performance of representative baseline models on different datasets [13]

Model	Dataset	Dice (%)	HD95 (mm)	Sens (%)	Precision (%)
UNet-3D	AutoPET-II	27.50 ± 16.51	121.35 ± 49.91	51.42 ± 20.38	25.51 ± 20.77
ResUNet-3D	AutoPET-II	39.78 ± 20.53	69.11 ± 42.95	60.70 ± 17.15	33.16 ± 22.53
VNet	Hecktor2022	51.36 ± 8.12	85.69 ± 30.28	58.50 ± 12.13	54.60 ± 10.66
nnUNet	Hecktor2022	50.04 ± 6.04	140.80 ± 23.77	57.11 ± 6.02	56.37 ± 6.89

As illustrated in Fig. 1, ResUNet-3D achieved about a 12% Dice improvement over U-Net-3D on AutoPET-II, along with a marked reduction in boundary error (HD95). On HECKTOR2022, V-Net performed comparably to nnU-Net but still lagged in boundary consistency. These findings indicate that although baseline networks provide a necessary benchmark, their limitations in complex lesions and boundary precision highlight the need for more advanced fusion approaches.

DOI: 10.54254/2755-2721/2025.29928

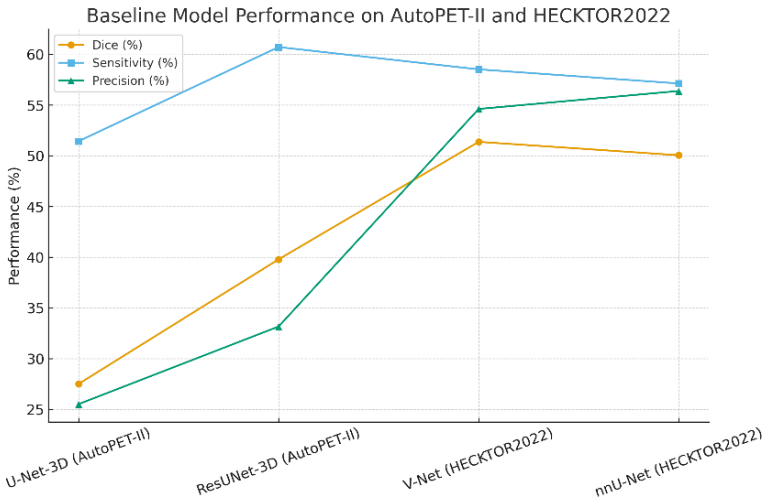

Figure 1. Baseline model performance comparison on AutoPET-II and HECKTOR2022 datasets [14,17]

4.2. Novel multimodal fusion networks

To address the limited expression ability of single-modal features, a large number of studies have explored the multimodal fusion strategy of PET/CT. Representative methods have been summarized in Table 3.

Table 3. Comparison of new multimodal fusion network methods

Method	Core mechanism	Dataset/task	Main advantages	limitations	Ref
ISA-Net	Improve spatial attention and cross-modal weighting	PET-CT tumor segmentation	Feature alignment is more precise, enhancing cross-modal complementarity	Depending on the large to small sample of training	[2]
Mirror U-Net	Modal "fission - mapping", multi-task learning	PET-CT segmentation	Enhanced modal interaction and information sharing among tasks	The model has high complexity and high training cost	[14]
Deep Evidential Fusion	Introducing evidence theory and uncertainty quantification	Multimodal segmentation	Provide predictive confidence and enhance robustness	Stability theory and complex, training co., LTD	[4, 16]
Interactive segmentation framework	Human-computer interaction, pseudo-label correction	Small sample/clinical application	Significantly reduce annotation pressure and enhance clinical usability	Human involvement is required and the efficiency is limited	[15]

4.3. Benchmark datasets and evaluation metrics

Commonly used PET/CT segmentation benchmarks mainly fall into two series. The AutoPET dataset covers multi-center, multi-disease whole-body PET/CT scans with SUV normalization and cross-domain registration. It is often used to assess generalization. The HECKTOR dataset focuses on head and neck tumors, providing high-quality PET/CT annotations for evaluating accuracy in complex regions [17].

DOI: 10.54254/2755-2721/2025.29928

Evaluation typically relies on several indicators. The Dice Similarity Coefficient (DSC) measures overlap and is the most widely used; Hausdorff Distance (HD95) evaluates boundary consistency; Sensitivity reflects lesion recall; and Precision measures prediction reliability. In practice, screening tasks emphasize sensitivity, whereas radiotherapy planning places greater importance on boundary precision (HD95) [17].

5. Applications and clinical validation

5.1. Early lung cancer detection

Deep multimodal PET/CT fusion has demonstrated value in early detection. Meta-analyses report stable performance for solitary pulmonary nodules, with pooled sensitivity ~0.88, specificity ~0.78, and AUC ~0.91. In NSCLC staging, sensitivities for N and M staging were ~0.82 and ~0.86, with specificities ~0.88 and ~0.89, thereby reducing the risk of misstaging. Comparative studies also confirm that PET/CT serves as a complementary modality to DWI. Algorithmically, deep learning–based respiratory motion correction improved SUVmax/SUVmean by ~10%/17% across 906 lesions, enhancing the detection of small-volume lesions. Feature-level fusion and attention mechanisms (e.g., ISA-Net, cross-modal Transformers, graph attention) consistently outperformed early fusion and single-modal baselines in both sensitivity and F1 metrics [2, 6-7, 15].

5.2. Automated tumor segmentation

Recent models consistently surpass baseline performance. ISA-Net improved Dice scores by +1.4–6.1 over 3D U-Net; Mirror U-Net gained +3.16 Dice on AutoPET; evidential fusion increased Dice by +4.1 while providing uncertainty estimates. Robustness across centers remains critical. Federated learning enables cross-institutional training without data sharing, while domain adaptation, large-scale pretraining, and weak/few-shot supervision mitigate distribution shifts and annotation scarcity. Public benchmarks, such as AutoPET, provide standardized evaluation. In practice, feature-level fusion, attention, and uncertainty modeling yield the strongest improvements, while federated and adaptive methods are essential for multi-center deployment [2, 5, 11, 14].

5.3. Clinical studies and deployment

As algorithms mature, clinical studies validate PET/CT fusion in real-world workflows. Large-scale meta-analyses confirm its diagnostic value, while federated frameworks demonstrate privacy-preserving robustness. Motion compensation enhances SUV quantification, and interactive systems achieve near-expert segmentation with minimal clicks. A typical workflow integrates: (i) PACS/RIS data access with normalization; (ii) one-pass inference using fusion networks; (iii) uncertainty-driven review requiring ≤5 clicks; (iv) report output including segmentation, SUV, and confidence maps; and (v) continuous fine-tuning under local or federated setups. These strategies demonstrate that PET/CT fusion not only improves precision but also aligns with clinical requirements for safety, compliance, and explainability [2-7, 10, 14, 17].

DOI: 10.54254/2755-2721/2025.29928

6. Challenges and future directions

6.1. Data scarcity and annotation bottlenecks

PET/CT data vary substantially across centers in terms of protocols, reconstruction, and SUV normalization, which hinders model generalization. Additionally, voxel-level annotation is costly and inconsistent, especially for small or low-contrast lesions.

Solutions include semi-/weakly supervised learning with consistency regularization and teacher–student pseudo-labeling:

$$L_{\text{consistency}} = \frac{1}{N}\sum_{i=1}^{N}\left|f_{\theta}\left(x_i+\delta\right)-f_{\theta}\left(x_i\right)\right|^2 \tag{1}$$

Here, f_{θ} represents the segmented network, and δ denotes a random perturbation or enhancement.

The overall loss function when pseudo-labels are combined with the teacher-student framework is:

$$L_{\text{semi}} = L_{\text{sup}} + \lambda \cdot L_{\text{unsup}} \tag{2}$$

Other approaches include meta-learning for cold starts, self-supervised pretraining, augmentation with synthetic data, and federated learning for multi-institution training. Public benchmarks such as AutoPET provide standardized evaluation. Future work should integrate weak supervision, pretraining, federated frameworks, and limited expert interaction for low-cost usability.

6.2. Model generalizability and robustness

Cross-center deployment is challenged by domain shift, lack of external validation, and susceptibility to noise. Performance stability is often quantified by Drop%:

$$\text{Drop}(\%) = \frac{M_{\text{internal}} - M_{\text{external}}}{M_{\text{internal}}} \times 100\% \tag{3}$$

Here,$M_{internal}$represents the performance metrics on the internal validation set (such as DICE or IOU), and $M_{external}$ represents the metrics on the external test set.

Key directions include domain adaptation (contrastive learning, style transfer), adversarial robust training with uncertainty modeling, federated collaboration, and physical consistency modeling such as motion correction. Unified validation on AutoPET/HECKTOR with Drop% reporting ensures reproducibility.

6.3. Explainability, trust, and clinical integration

Interpretability and safety are essential for clinical adoption. Deep evidential fusion outputs probabilities and uncertainty:

$$p\left(y=cx\right) = \frac{\alpha_c}{S}, S = \sum_{j=1}^{K}\alpha_j \tag{4}$$

Here,α_c represents the evidence strength of category c, and K is the number of categories.

Its uncertainty can be quantified by variance:

DOI: 10.54254/2755-2721/2025.29928

$$\mathrm{Var}\left[p\left(y=cx\right)\right]=\frac{\alpha_c(S-\alpha_c)}{S^2(S+1)} \tag{5}$$

A small S indicates low confidence, prompting manual review. Visualization method (e.g., Grad-CAM, attention maps) aids understanding, while human-in-the-loop correction builds trust. Clinical integration requires outputting segmentation with confidence, case-level metrics, and secure PACS/RIS compatibility.

6.4. Emerging trends

Future directions emphasize: (i) foundation models with large-scale pretraining and efficient fine-tuning; (ii) multimodal and multi-omics fusion for prognosis prediction; (iii) federated learning for privacy-preserving collaboration; and (iv) real-time deployment with pruning, quantization, and online uncertainty. These trends aim to accelerate the transition from research prototypes to clinical standards.

7. Conclusion

This article reviewed recent progress in deep multimodal PET/CT fusion for lung cancer diagnosis and segmentation. Clinically, ^{18}F-FDG PET/CT provides stable sensitivity and specificity for solitary pulmonary nodule risk stratification and NSCLC staging, forming key evidence in diagnosis and treatment. Deep learning–assisted motion correction further improves SUV quantification and boundary sharpness, enabling reliable detection of small, low-contrast lesions.

At the algorithmic level, research has evolved from U-Net/nnU-Net baselines to fusion paradigms emphasizing feature interaction and attention mechanisms, including ISA-Net, cross-modal Transformers, graph and multi-scale fusion. Hybrid and multi-task models (e.g., Mirror U-Net, DMMT), evidential uncertainty, and interactive human-in-the-loop strategies enhance lesion recall, reduce false positives, and improve interpretability.

Future directions emphasize foundation models with multi-center pretraining, multimodal–multiomics integration, federated privacy-preserving collaboration, and real-time deployment with motion correction and uncertainty estimation. Overall, PET/CT fusion is transitioning from "performance improvement" toward "clinical credibility," with standardized multi-center validation expected to accelerate its translation into routine practice for early detection and precise segmentation of lung cancer.

References

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), pp.209-249.

[2] Huang, Z., Zou, S., Wang, G., Chen, Z., Shen, H., Wang, H., Zhang, N., Zhang, L., Yang, F., Wang, H. and Liang, D., 2022. ISA-Net: Improved spatial attention network for PET-CT tumor segmentation. Computer Methods and Programs in Biomedicine, 226, p.107129.

[3] Li, Y., Jing, B., Li, Z., Wang, J. and Zhang, Y., 2023. nnsam: Plug-and-play segment anything model improves nnunet performance. arXiv preprint arXiv: 2309.16967.

[4] Liu, J., Xia, X., Zou, Q., Xie, X., Lei, Y., Wan, Q. and Li, X., 2023. Diagnostic performance of diffusion-weighted imaging versus 18F-FDG PET/CT in differentiating pulmonary lesions: an updated meta-analysis of comparative studies. BMC Medical Imaging, 23(1), p.37.

[5] Li, Y., Shi, Y.B. and Hu, C.F., 2024. 18F-FDG PET/CT based model for predicting malignancy in pulmonary nodules: a meta-analysis. Journal of Cardiothoracic Surgery, 19(1), p.148.

DOI: 10.54254/2755-2721/2025.29928

[6] Lu, Y., Kang, F., Zhang, D., Li, Y., Liu, H., Sun, C., Zeng, H., Shi, L., Zhao, Y. and Wang, J., 2024. Deep learning-aided respiratory motion compensation in PET/CT: addressing motion induced resolution loss, attenuation correction artifacts and PET-CT misalignment. European Journal of Nuclear Medicine and Molecular Imaging, 52(1), pp.62-73.
[7] Yu, D. and Chen, C., 2025. [18F] FDG PET/CT versus [18F] FDG PET/MRI in staging of non-small cell lung cancer: a head-to-head comparative meta-analysis. Frontiers in Medicine, 11, p.1517805.
[8] Wan, Z., Zhang, P., Wang, Y., Yong, S., Stepputtis, S., Sycara, K. and Xie, Y., 2025, February. Sigma: Siamese mamba network for multi-modal semantic segmentation. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 1734-1744). IEEE.
[9] Jia, D., Guo, J., Han, K., Wu, H., Zhang, C., Xu, C. and Chen, X., 2024. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. arXiv preprint arXiv: 2406.01210.
[10] Mei, J., Lin, C., Qiu, Y., Wang, Y., Zhang, H., Wang, Z. and Dai, D., 2025. Cross-Modal Interactive Perception Network with Mamba for Lung Tumor Segmentation in PET-CT Images. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 15653-15662).
[11] Hassan, S., Al Hammadi, H., Mohammed, I. and Khan, M.H., 2024, October. Multi-Modal Medical Image Fusion for Non-Small Cell Lung Cancer Classification. In 2024 IEEE International Conference on Image Processing (ICIP) (pp. 3091-3097). IEEE.
[12] Zhou, L., Wu, C., Chen, Y. and Zhang, Z., 2024. Multitask connected U-Net: automatic lung cancer segmentation from CT images using PET knowledge guidance. Frontiers in Artificial Intelligence, 7, p.1423535.
[13] Lu, J., Chen, J., Cai, L., Jiang, S. and Zhang, Y., 2024, October. 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). Cham: Springer Nature Switzerland.
[14] Marinov, Z., Reiß, S., Kersting, D., Kleesiek, J. and Stiefelhagen, R., 2023. Mirror u-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2283-2293).
[15] Hallitschke, V.J., Schlumberger, T., Kataliakos, P., Marinov, Z., Kim, M., Heiliger, L., Seibold, C., Kleesiek, J. and Stiefelhagen, R., 2023, April. Multimodal interactive lung lesion segmentation: a framework for annotating PET/CT images based on physiological and anatomical cues. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.
[16] Vaz SC, Adam JA, Delgado Bolton RC, et al. Joint EANM/SNMMI/ESTRO recommendations: PET/CT in lung cancer RT planning. Eur J Nucl Med Mol Imaging. 2022.
[17] Zhong, S., Mo, J. and Liu, Z., 2022. Autopet challenge 2022: Automatic segmentation of whole-body tumor lesion based on deep learning and fdg pet/ct. arXiv preprint arXiv: 2209.01212.