mathematics

MDPI

Article

Progressive Reliability-Guided Refinement with Cross-Modal Consistency and Iterative Correction for Weakly Supervised PET/CT Tumor Segmentation

Zhaoshuo Diao [1] **and Huiyan Jiang** [2,*]

[1] School of Software, Shenyang University of Technology, Shenyang 110870, China
[2] Software College, Northeastern University, Shenyang 110819, China
* Correspondence: hyjiang@mail.neu.edu.cn

Abstract

Accurate tumor segmentation in Positron Emission Tomography/Computed Tomography (PET/CT) is essential for clinical diagnosis, staging, and treatment planning. However, the reliance on time-consuming pixel-level annotations limits the practical deployment of deep learning-based segmentation methods. Weakly supervised segmentation using image-level classification labels provides a promising alternative, yet generating accurate supervision remains highly challenging. Specifically, PET provides critical metabolic information but is susceptible to physiological hypermetabolism and background noise, while CT offers anatomical structure but suffers from low soft-tissue contrast and boundary uncertainty. Consequently, static or single-modality pseudo-labels derived from Class Activation Maps (CAMs) often suffer from boundary ambiguity and false-positive activations, failing to reliably guide segmentation. To address the core problem of dynamically evaluating and correcting supervision reliability, we propose a Progressive Reliability-Guided Refinement (PRGR) framework. Specifically, pixel-level supervision reliability is modeled by integrating PET metabolic consistency and CT structural gradients, leveraging complementary cross-modal information without introducing additional trainable parameters. Based on the estimated reliability, the image domain is adaptively partitioned into regions with different supervision credibility levels, enabling progressive optimization with differentiated learning constraints. Furthermore, pseudo-labels are iteratively updated using reliability estimates and segmentation predictions to progressively correct erroneous supervision. Experiments on two public PET/CT datasets demonstrate that our method effectively suppresses false positives and improves boundary delineation. Quantitatively, the proposed PRGR achieves Dice scores of 0.673 and 0.596 on the STS and HeadNeck datasets, respectively, yielding absolute improvements of 1.8% and 8.4% over the second-best baseline. Furthermore, it achieves superior Precision scores of 0.776 and 0.684, demonstrating its effectiveness in narrowing the performance gap between weakly and fully supervised segmentation while reducing the dependence on pixel-level annotations.

Keywords: biomedical segmentation; PET/CT; weakly supervised segmentation; reliability modeling

MSC: 92C50

Check for updates

Academic Editors: Nan Sheng, Qiujie Lv and Yan Wang

Received: 20 August 2026
Revised: 6 October 2026
Accepted: 23 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Positron Emission Tomography/Computed Tomography (PET/CT), as a representative multimodal medical imaging modality, provides an essential foundation for accurate

tumor localization and quantitative analysis by integrating functional metabolic information from PET with anatomical structural information from CT. Specifically, PET captures the abnormal metabolic activity associated with tumor tissues, whereas CT provides high-resolution anatomical details for precise spatial delineation [1,2]. Owing to this complementary nature, PET/CT has been widely applied in various clinical tasks, including tumor detection, staging, treatment response evaluation, and radiotherapy planning. Among these applications, accurate tumor segmentation plays a fundamental role in PET/CT-based quantitative analysis, as it directly determines the reliability of critical imaging biomarkers, such as metabolic tumor volume (MTV) and standardized uptake value (SUV), and consequently affects subsequent clinical decision-making processes. However, in high-throughput clinical environments, manual delineation of tumor boundaries is highly susceptible to inter-observer variability, which can lead to biased diagnostic outcomes. Therefore, reliable automated segmentation is of great clinical significance.

In recent years, with the rapid development of deep learning techniques, automated segmentation methods based on Convolutional Neural Networks and Transformer architectures have attracted increasing attention in PET/CT tumor analysis. Most existing approaches formulate tumor segmentation as a pixel-wise prediction task, employing encoder–decoder architectures to learn discriminative representations and multimodal fusion strategies to integrate PET metabolic information with CT anatomical structures. For example, multi-branch networks extract modality-specific features and enhance cross-modal interaction through fusion modules, while attention-based and Transformer-based methods further improve long-range dependency modeling for complex tumor regions [3–5]. However, these fully supervised methods rely heavily on large-scale pixel-level annotations, which are difficult to obtain due to the time-consuming annotation process, ambiguous tumor boundaries, heterogeneous lesion characteristics, and inter-observer variability [6]. In contrast, weakly supervised learning methods utilize low-cost supervision signals, such as image-level labels, point annotations, or coarse regional annotations, and optimize segmentation models through strategies including pseudo-label generation and consistency regularization. Therefore, weakly supervised learning has emerged as a promising solution to alleviate annotation dependency and enhance the scalability of medical image segmentation [7].

However, compared with fully supervised learning, weakly supervised PET/CT tumor segmentation remains highly challenging due to the limited and indirect nature of supervisory signals. Most weakly supervised approaches rely on Class Activation Maps (CAMs) or their variants to generate initial pseudo-labels from image-level annotations. However, since classification models tend to highlight only the most discriminative regions rather than the complete tumor areas, the generated pseudo-labels often suffer from incomplete coverage, inaccurate boundaries, and false-positive activations. Moreover, the metabolic heterogeneity of PET images and the structural similarity of CT images further increase the uncertainty of pseudo-labels, resulting in spatially varying supervision reliability. Therefore, effectively characterizing pseudo-label reliability and developing robust optimization strategies to alleviate noise propagation under limited supervision conditions remain critical mathematical challenges in weakly supervised PET/CT tumor segmentation. Solving these challenges is not only a mathematical necessity but also a prerequisite for ensuring the robustness and trustworthiness of AI-based diagnostic tools in clinical settings.

To address the aforementioned challenges, this study re-examines weakly supervised PET/CT tumor segmentation from the perspectives of reliability modeling and progressive refinement. Unlike existing approaches that treat pseudo-labels as fixed and uniformly reliable supervision, we observe that pseudo-labels generated from image-level classification

signals exhibit substantial spatial heterogeneity, with varying levels of reliability across different tumor regions. Based on this observation, we propose a Progressive Reliability-Guided Refinement (PRGR) framework, which formulates weakly supervised PET/CT tumor segmentation as a closed-loop optimization process consisting of reliability estimation, adaptive region partitioning, and iterative pseudo-label refinement.

Specifically, this study first exploits PET metabolic information, CT anatomical features, and pseudo-label representations to construct a reliability modeling module, which generates pixel-level reliability maps to quantify the credibility of supervisory signals across different spatial locations. Based on the statistical distribution of reliability scores, an adaptive reliability partition strategy is further introduced to automatically determine optimization regions without manually defined thresholds, separating reliable, moderately reliable, and global regions through inter-class variance maximization. Subsequently, a reliability-aware progressive optimization strategy is developed to incorporate supervision information from different regions according to their reliability levels, enabling the model to learn robust tumor representations from highly reliable regions before gradually adapting to ambiguous and low-reliability areas. In addition, a reliability-guided pseudo-label refinement strategy is proposed to formulate pseudo-label updating as a reliability-constrained optimization process by jointly considering current predictions, historical pseudo-labels, and reliability maps. Through these mechanisms, the proposed framework reformulates pseudo-label noise mitigation as a reliability-aware optimization problem, allowing progressive improvement of supervision quality and segmentation performance without pixel-level annotations.

The main contributions of this study are summarized as follows:

(1) A progressive reliability-guided refinement framework is proposed for weakly supervised PET/CT tumor segmentation. The proposed framework transforms the conventional pseudo-label-based training paradigm into a reliability-driven iterative optimization process, improving robustness against the noise and uncertainty introduced by weak supervision.

(2) A multimodal reliability modeling method is developed by jointly incorporating PET metabolic information, CT anatomical features, and pseudo-label priors to estimate pixel-level supervision reliability. In addition, an adaptive region partition strategy based on reliability distributions is introduced to achieve automatic supervision quality assessment without manually defined thresholds.

(3) A reliability-guided progressive learning and pseudo-label refinement mechanism is designed to optimize supervision signals in a coarse-to-fine manner. By progressively incorporating information from high-reliability regions to low-reliability regions and updating pseudo-labels under reliability constraints, the proposed mechanism improves tumor region completeness and boundary delineation in weakly supervised settings.

2. Related Work

2.1. Deep Learning-Based PET/CT Tumor Segmentation

In recent years, deep learning methods have become the mainstream approach for PET/CT tumor segmentation due to their strong feature representation capabilities. Compared with traditional methods based on handcrafted features, deep neural networks can automatically extract metabolic and anatomical representations from PET and CT images through end-to-end learning. Early CNN-based approaches mainly employed encoder–decoder architectures with multi-scale feature extraction and skip connections to integrate semantic and spatial information.

With the development of multimodal learning, researchers have focused on exploiting the complementary information between PET and CT through dual-branch networks,

feature interaction modules, and attention-based fusion strategies. For example, SCL-Net enhances cross-modal feature interaction through structural collaborative learning, while H2ASeg improves multimodal fusion through hierarchical adaptive interaction and modality weighting mechanisms [3,8]. Recently, Transformer-based approaches, such as Swin-Cross, have further improved segmentation performance by capturing long-range dependencies between multimodal features [5]. In addition, self-supervised learning, transfer learning, and foundation model-based methods have been explored to improve model generalization across diverse datasets [9–11].

Despite these advances, existing deep learning methods still rely heavily on accurate pixel-level annotations, which are expensive to obtain and affected by tumor heterogeneity and inter-observer variability. Therefore, reducing annotation dependency while maintaining segmentation performance remains a critical challenge in PET/CT tumor segmentation.

2.2. Classification Label-Based Weakly Supervised Medical Image Segmentation

Since pixel-level annotation requires extensive clinical expertise, Classification Label-Based Weakly Supervised Segmentation (CL-WSSS) has been widely studied to reduce annotation costs. These methods typically employ classification networks and Class Activation Maps (CAMs) to generate pixel-level pseudo-labels from image-level supervision. Early studies focused on improving CAM localization, such as GradCAM with gradient-based response estimation [12] and WILDCAT with weighted spatial pooling [13]. Subsequent approaches further enhanced activation completeness and pseudo-label quality through consistency learning, regularization, multi-scale enhancement, self-organizing learning, and region reactivation strategies, including SEAM [14], Belharbi et al. [15], WSS_CMER [16], ESOL_WSSS [17], and ReCAM [18].

In medical image analysis, recent advancements in weakly supervised methods have specifically addressed boundary ambiguity and pseudo-label noise through advanced refinement strategies. For instance, recent approaches have improved activation completeness via region consistency constraints [19] and alleviated background interference using feature decoupling [20]. Furthermore, state-of-the-art studies have increasingly focused on dynamic pseudo-label refinement; UM-CAM enhances pseudo-label reliability via uncertainty-guided fusion [21], and DMSPS improves training stability through dynamic soft pseudo-labeling and uncertainty constraints [22]. The recent literature has also explored boundary-aware iterative correction and cross-modal progressive regeneration to further refine initial CAMs [23,24]. For PET/CT tumor segmentation, our previous UCWS-Net introduced uncertainty-guided collaborative learning with a dual-branch architecture to optimize pseudo-labels under classification supervision [25]. However, despite these recent advances, most existing methods either treat pseudo-label reliability as a static property or rely on single-stage refinement, largely ignoring the dynamic spatial variation in supervision reliability during the training process. Therefore, dynamic reliability-driven progressive refinement remains a key challenge in weakly supervised PET/CT tumor segmentation.

3. Method

3.1. Problem Definition

This study investigates weakly supervised PET/CT tumor segmentation under image-level classification supervision. Given a training dataset $\mathcal{D} = \{(X_i^{pet}, X_i^{ct}, c_i)\}_{i=1}^{N}$, where X_i^{pet} and X_i^{ct} denote the PET and CT images, respectively, and c_i represents the corresponding classification label. Unlike fully supervised segmentation, pixel-level annotations Y_i^* are unavailable during training. Therefore, the model learns a segmentation function based only on image-level supervision: $P_i = f_{\theta_{seg}}(X_i^{pet}, X_i^{ct})$, where $P_i \in [0,1]^{H \times W}$ denotes the predicted tumor probability map and $f_{\theta_{seg}}(\cdot)$ denotes the segmentation network (e.g., U-Net in this paper).

Since classification labels provide limited spatial constraints, the reliability of supervision varies across different regions. Accordingly, this study formulates weakly supervised PET/CT segmentation as a reliability-driven optimization problem by learning the segmentation prediction P based on the pixel-level reliability representation R:

$$P = \arg\min_{\theta_{seg}} \mathcal{L}(P, R|\mathcal{D}), \tag{1}$$

where $R \in [0,1]^{H\times W}$ represents the spatial reliability of supervisory signals. The objective is to progressively improve tumor segmentation performance under limited supervision by exploiting reliability modeling and optimization strategies.

The overall architecture of the proposed method is illustrated in Figure 1, which consists of four main components: Weak Supervision and Pseudo Label Construction, Reliability Modeling, Reliability-Aware Progressive Optimization, and Reliability-Guided Pseudo Label Refinement. Each component is detailed below.

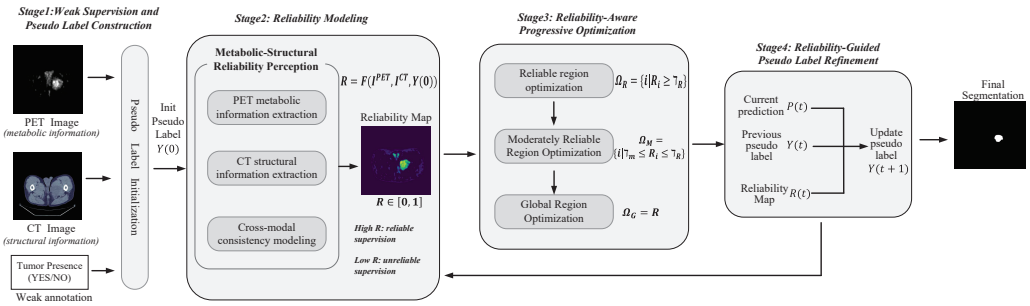


Figure 1. Overview of the proposed Progressive Reliability-Guided Refinement framework for weakly supervised PET/CT tumor segmentation.

3.2. Weak Supervision and Pseudo-Label Construction

In classification label-based weakly supervised segmentation, only image-level category labels are available, while pixel-level tumor annotations are absent. Therefore, a classification network is first utilized to generate initial pseudo-labels, providing preliminary supervision for subsequent reliability modeling and progressive refinement.

Given an input PET/CT image pair $X_i = (X_i^{pet}, X_i^{ct})$, the classification network $f_{\theta_{cls}}(\cdot)$ predicts the image-level category probability: $p_i = f_{\theta_{cls}}(X_i)$, where p_i denotes the predicted category probability. The Class Activation Map (CAM) is then generated to identify category-related spatial regions:

$$M_i^{cam} = \Phi(f_{\theta_{cls}}(X_i)), \tag{2}$$

where $\Phi(\cdot)$ represents the CAM generation function and $M_i^{cam} \in [0,1]^{H\times W}$ denotes the activation map. Following standard protocols in weakly supervised learning, the initial pseudo-label $Y(0)_i$ is obtained by thresholding the CAM map with an empirical value of 0.5 to provide a coarse initial seed. This thresholding serves only as a rough initialization, as the subsequent reliability modeling will dynamically correct the inherent noise:

$$Y(0)_i = \begin{cases} 1, & M_i^{cam} \geq 0.5, \\ 0, & M_i^{cam} < 0.5. \end{cases} \tag{3}$$

3.3. Reliability Modeling

The initial pseudo-label $Y(0)_i$ provides preliminary spatial supervision; however, its reliability is affected by tumor metabolic heterogeneity, anatomical complexity, and

boundary ambiguity due to the indirect nature of image-level supervision. Therefore, a reliability modeling module is proposed to construct a pixel-level reliability map without additional network training. Theoretically, in PET/CT oncology imaging, statistical consistency serves as a robust surrogate for supervision reliability. Tumors typically exhibit highly homogeneous metabolic hyper-uptake in PET; therefore, pixels deviating significantly from the statistical core distribution are highly likely to be background noise. For PET images, based on the initial pseudo-label $Y(0)_i$, we compute the mean μ_i^{pet} and standard deviation σ_i^{pet} of PET intensities within the pseudo-label region. The metabolic reliability at pixel p is defined as $R_i^{pet}(p) = \exp\left(-|X_i^{pet}(p) - \mu_i^{pet}|/(\sigma_i^{pet} + \epsilon)\right)$.

For CT images, the theoretical objective is to identify abrupt anatomical boundaries to prevent pseudo-labels from erroneously spreading across distinct tissues. Structural information is characterized by the local gradient magnitude $G_i^{ct}(p) = \sqrt{(\partial_x X_i^{ct}(p))^2 + (\partial_y X_i^{ct}(p))^2}$. Because the magnitude is strictly directionless, it reliably highlights structural edges regardless of the inconsistent gradient directions of internal organs. We compute the mean μ_G and standard deviation σ_G of the gradient distribution within the pseudo-label region, yielding the structural reliability $R_i^{ct}(p) = \exp\left(-|G_i^{ct}(p) - \mu_G|/(\sigma_G + \epsilon)\right)$. A cross-modal reliability term is further introduced as $R_i^{cm}(p) = 1 - |R_i^{pet}(p) - R_i^{ct}(p)|$, where higher values indicate stronger agreement between PET and CT. Finally, the overall pixel-level reliability map is obtained by integrating the three components:

$$R(0)_i(p) = R_i^{pet}(p) \cdot R_i^{ct}(p) \cdot R_i^{cm}(p), \tag{4}$$

where $R(0)_i(p) \in [0, 1]$ represents the reliability of the supervisory signal at pixel location p. This design directly estimates supervision reliability without introducing additional learnable parameters or extra annotations, providing an interpretable reliability representation.

3.4. Reliability-Aware Progressive Optimization

3.4.1. Adaptive Reliability Partition

After obtaining the reliability map $R(0)$, an adaptive reliability partition strategy is introduced to divide pixels according to their supervision reliability. Since reliability distributions vary across patients due to differences in tumor characteristics and imaging conditions, manually predefined thresholds may result in inaccurate region assignment. Therefore, the region boundaries are adaptively determined based on the statistical distribution of the current reliability map.

Let the reliability map be denoted as $R(0) = \{r_1, r_2, \cdots, r_N\}$, where r_i represents the reliability value of the i-th pixel. The reliability space is divided into three regions: $\Omega = \Omega_R \cup \Omega_M \cup \Omega_L$, where Ω_R, Ω_M, and Ω_L represent the reliable, moderately reliable, and low-reliability regions, respectively. To avoid dependence on empirical thresholds, the partition problem is formulated as a three-class clustering problem on the reliability distribution, where two adaptive boundaries are obtained by maximizing the separability among different reliability groups:

$$(\lambda_R^*, \lambda_M^*) = \arg\max \frac{\sigma_{between}^2}{\sigma_{within}^2}, \tag{5}$$

where $\sigma_{between}^2$ and σ_{within}^2 denote the inter-class and intra-class variances of reliability values, respectively. Based on the optimized thresholds, the pixel regions are defined as:

$$\Omega_R = \{p|R(0)(p) \geq \lambda_R^*\}, \quad \Omega_M = \{p|\lambda_M^* \leq R(0)(p) < \lambda_R^*\}, \quad \Omega_L = \{p|R(0)(p) < \lambda_M^*\}. \tag{6}$$

The reliable region Ω_R contains pixels with stable supervisory information and provides strong optimization constraints. The moderately reliable region Ω_M mainly includes ambiguous boundaries and uncertain tumor regions, which require joint optimization with model predictions. The low-reliability region Ω_L contains highly uncertain or noisy pseudo-label regions and is mainly involved in weak constraints and subsequent pseudo-label refinement.

3.4.2. Progressive Optimization with Differentiated Supervision

For the reliable region Ω_R, the supervision signals exhibit high credibility; therefore, the corresponding pseudo-labels are directly utilized to impose strong optimization constraints. The reliable region loss is defined as:

$$\mathcal{L}_R = \frac{1}{|\Omega_R|} \sum_{p \in \Omega_R} \ell(P(p), Y(0)(p)), \tag{7}$$

where $\ell(\cdot)$ denotes the segmentation loss function. In this study, to ensure full methodological reproducibility, we clarify that the base loss function $\ell(\cdot)$ is strictly formulated as the standard Binary Cross-Entropy (BCE) loss.

For the moderately reliable region Ω_M, pseudo-labels generally contain uncertainty due to ambiguous boundaries and low-response regions. Instead of directly applying the original pseudo-labels, a progressive supervision strategy is introduced by combining historical pseudo-labels with current model predictions:

$$\hat{Y}(0)(p) = \alpha Y(0)(p) + (1 - \alpha)\mathrm{sg}[P(0)(p)], \quad p \in \Omega_M, \tag{8}$$

where α is a weighting coefficient. Here, $P(0)(p)$ represents the initial probability prediction output by the unrefined segmentation network. To prevent the "cycle of supervision" where the prediction acts as both the optimization variable and the target, a stop-gradient operation, denoted as $\mathrm{sg}[\cdot]$, is strictly applied to $P(0)(p)$ when constructing the target $\hat{Y}(0)(p)$. This ensures that gradients are only backpropagated through the active prediction network. The corresponding loss is:

$$\mathcal{L}_M = \frac{1}{|\Omega_M|} \sum_{p \in \Omega_M} \ell(P(0)(p), \hat{Y}(0)(p)). \tag{9}$$

Although low-reliability regions contain noisy supervision, completely discarding global constraints can cause the network to overfit local high-confidence regions, leading to excessive boundary shrinkage and catastrophic forgetting of the initial lesion localization. Therefore, $\mathcal{L}_G$ serves as a weak global regularization anchor. The overall reliability-aware progressive optimization loss integrates the regional and global objectives:

$$\mathcal{L}_G = \frac{1}{|\Omega|} \sum_{p \in \Omega} \ell(P(0)(p), Y(0)(p)). \tag{10}$$

The overall reliability-aware progressive optimization loss integrates the regional and global objectives:

$$\mathcal{L}_{PRGR} = \mathcal{L}_R + \lambda_M \mathcal{L}_M + \lambda_G \mathcal{L}_G, \tag{11}$$

where λ_M and λ_G are weighting coefficients. This strategy progressively incorporates supervision from reliable regions to uncertain regions and from local constraints to global consistency, thereby improving segmentation stability.

3.5. Reliability-Guided Pseudo-Label Refinement

After reliability-aware progressive optimization, the model obtains the current prediction $P(0)$. However, in classification label-based weakly supervised segmentation, pseudo-label quality remains a key factor affecting segmentation performance. Since initial pseudo-labels generally cover only discriminative tumor regions while missing boundaries and low-response areas, and inaccurate pseudo-labels may accumulate during iterative optimization, a reliability-guided pseudo-label refinement strategy is introduced to progressively update pseudo-labels by integrating prediction results and reliability information.

At the t-th iteration, the model obtains the current pseudo-label $Y(t)$, prediction map $P(t)$, and reliability map $R(t)$. Theoretically, this refinement process can be formulated as a reliability-constrained Exponential Moving Average (EMA) update in the label space. Instead of applying a uniform replacement, the update process dynamically adjusts the contribution of historical supervision and current model predictions according to the pixel-level spatial reliability:

$$Y(t+1) = R(t) \odot Y(t) + (1 - R(t)) \odot P(t), \tag{12}$$

where $\odot$ denotes element-wise multiplication. The reliability map $R(t)$ acts as a spatial weighting factor: high-reliability regions retain more original supervisory information, while low-reliability regions rely more on model predictions for correction. This enables adaptive pseudo-label evolution.

To prevent excessive iterations, the process is terminated when the pseudo-labels become stable between consecutive iteration steps:

$$\frac{\|Y(t) - Y(t-1)\|_1}{|\Omega|} < \epsilon, \tag{13}$$

where ϵ is a predefined convergence threshold. Once satisfied, the final prediction is output. The overall algorithmic procedure is summarized in Algorithm 1.

Algorithm 1 Progressive Reliability-Guided Refinement (PRGR)

Require: Training dataset $\mathcal{D}$, hyperparameters $\alpha, \lambda_M, \lambda_G$, convergence threshold ϵ.
Ensure: Segmentation network $f_{\theta_{seg}}$ and final prediction P.
Stage 1: Weak Supervision & Initialization
Generate CAM $M^{cam} = \Phi(f_{\theta_{cls}}(X))$
Initialize pseudo-label $Y(0) = \mathbb{I}(M^{cam} \geq 0.5)$
Set iteration $t \leftarrow 1$
repeat
Stage 2: Reliability Modeling
  Compute reliability map $R(t-1)$ as in Equation (4)
Stage 3: Reliability-Aware Progressive Optimization
  Compute adaptive boundaries $(\lambda_R^*, \lambda_M^*)$ via Equation (5)
  Partition space into $\Omega_R, \Omega_M, \Omega_L$ using Equation (6)
  Formulate progressive supervision $\hat{Y}(t-1)$ for region Ω_M using Equation (8)
  Optimize $f_{\theta_{seg}}$ using $\mathcal{L}_{PRGR}$ in Equation (11)
  Obtain current prediction $P(t-1) = f_{\theta_{seg}}(X^{pet}, X^{ct})$
Stage 4: Reliability-Guided Pseudo Label Refinement
  Update pseudo-label via Equation (12): $Y(t) = R(t-1) \odot Y(t-1) + (1 - R(t-1)) \odot P(t-1)$
  $t \leftarrow t + 1$
until $\frac{\|Y(t-1) - Y(t-2)\|_1}{|\Omega|} < \epsilon$
return Final segmentation prediction $P \leftarrow P(t-1)$

4. Experiments

4.1. Datasets

In this paper, we conduct experiments using two publicly available PET/CT imaging datasets: the Soft Tissue Sarcoma (STS) dataset and the Head and Neck (HeadNeck) tumor dataset. Detailed descriptions of each dataset are provided below.

STS: The STS dataset, originally introduced by Vallières et al. [26], contains PET/CT scans from 51 patients with soft tissue sarcoma. The PET images have a spatial resolution of 128×128, while the corresponding CT images are acquired at a resolution of 512×512. Tumor annotations are obtained using the publicly available radiomics processing pipeline provided by the original authors.

HeadNeck: The HeadNeck dataset, originally introduced by Vallières et al. [27], contains PET/CT scans from 400 patients with head-and-neck tumors collected from four medical centers. Following the evaluation protocol of Andrearczyk et al. [28], 202 patients with oropharyngeal cancer are selected for experiments. The ground-truth tumor annotations are obtained from expert-delineated radiotherapy contours.

For the STS dataset, 36 cases are used for training, 5 cases for validation, and 10 cases for testing. For the HeadNeck dataset, 141 cases are used for training, 10 cases for validation, and 41 cases for testing. During training, pixel-level segmentation annotations are not utilized; instead, only image-level classification labels indicating tumor presence are adopted for weakly supervised learning. The ground-truth segmentation masks are exclusively used for quantitative evaluation on the test set.

4.2. Implementation Details

In the proposed Progressive Reliability-Guided Refinement framework, two backbone networks are employed for classification and segmentation tasks, respectively. In the weak supervision and pseudo-label construction stage, ResNet is utilized as the classification backbone to extract image-level representations and generate initial Class Activation Maps. Subsequently, a standard U-Net is adopted as the segmentation backbone to perform dense prediction during the reliability-aware progressive optimization and pseudo-label refinement stages.

During preprocessing, the 3D PET and CT volumes are decomposed into 2D axial slices. These 2D slices are then resized to a unified spatial resolution of 512×512 pixels using bilinear interpolation to ensure voxel-level alignment between modalities. Although interpolation introduces slight spatial smoothing, the macroscopic anatomical boundaries in CT inherently possess high Hounsfield Unit (HU) contrast. Consequently, the gradient magnitude at these critical structural boundaries remains dominant, ensuring that the CT gradient-based reliability calculation is not invalidated. The PET Standardized Uptake Values and CT Hounsfield Units are then normalized using standard normalization strategies.

The proposed framework is implemented with the PyTorch 2.14.1 framework. All experiments are conducted on a server equipped with an Intel Core E7-2680 CPU and two NVIDIA GeForce RTX 2080 Ti GPUs. The networks are optimized using the Adam optimizer with a weight decay of 1×10^{-4}. The initial learning rate is set to 1×10^{-4} and adjusted using a cosine annealing scheduler. Due to GPU memory limitations, the batch size is set to 16, and the models are trained for 200 epochs.

For the reliability-aware progressive optimization, the loss weights λ_M and λ_G are set to 0.5 and 1.0, respectively. The progressive supervision coefficient α in the moderately reliable region optimization is set to 0.5. The convergence threshold for pseudo-label refinement is set to $\epsilon = 0.05$.

4.3. Evaluation Metrics

To quantitatively evaluate the segmentation performance of the proposed PRGR framework, four commonly used metrics are adopted: the Dice, Recall, Precision, and the 95th percentile Hausdorff Distance (HD95). Let $P \in \{0,1\}^{H\times W}$ denote the predicted binary segmentation mask and $Y \in \{0,1\}^{H\times W}$ represent the ground-truth annotation. The Dice coefficient is defined as $2|P \cap Y|/(|P| + |Y|)$, Recall as $|P \cap Y|/|Y|$, and Precision as $|P \cap Y|/|P|$. While Dice, Recall, and Precision evaluate regional overlap and detection accuracy, HD95 rigorously evaluates the spatial distance between the predicted and ground-truth boundaries, providing a strict penalty for local over-segmentation and boundary anomalies. All metrics are computed for each test case, and the final performance is reported as the average value over the entire test set. Furthermore, to rigorously assess the statistical significance of the performance improvements, a paired Wilcoxon signed-rank test was conducted between the proposed PRGR framework and the baseline models. A p-value of less than 0.05 is considered statistically significant, demonstrating that the segmentation improvements are consistent across the test cohort

4.4. Results

4.4.1. Comparison Experiment

To evaluate the effectiveness of the proposed PRGR framework, we comprehensively compare it with several mainstream weakly supervised segmentation approaches, including WILDCAT [13], ESOL_WSSS [17], ReCAM [18], Belharbi et al. [15], SEAM [14], WSS_CMER [16], and UCWS-Net [25]. Among these selected baselines, WILDCAT directly utilizes classification networks to obtain the initial CAMs for object localization. Building upon this foundation, SEAM incorporates rotation consistency constraints to enhance the activation regions. Approaches including ESOL_WSSS, ReCAM, and Belharbi et al. employ multi-step iterative optimization strategies to progressively refine the CAMs and improve pseudo-label quality. Specifically designed for multimodal medical images, WSS_CMER and UCWS-Net introduce cross-modality consistency constraints to guide weakly supervised segmentation. In addition to the aforementioned weakly supervised methods, we also train a standard U-Net using fully supervised pixel-level annotations to serve as the theoretical performance upper bound for this task. The quantitative comparative results of all methods are summarized in Table 1.

As shown in the results, PRGR achieves Dice similarity coefficients of 0.673 on the STS dataset and 0.596 on the HeadNeck dataset. Compared to the second-best performing baseline, UCWS-Net, this method yields a Dice improvement of 1.8% on the STS dataset, which is statistically significant ($p = 4.89 \times 10^{-3} < 0.05$) based on the paired Wilcoxon signed-rank test. On the HeadNeck dataset ($N = 41$), the method achieves an 8.4% margin over UCWS-Net, which is also statistically significant ($p = 2.69 \times 10^{-2} < 0.05$). Furthermore, the segmentation performance of PRGR narrows the gap with the fully supervised U-Net upper bound. On the STS dataset, the Dice score difference between the weakly supervised framework score of 0.673 and the fully supervised model score of 0.684 is 1.1%, indicating the capacity of the method to learn tumor representations using image-level classification labels. Furthermore, regarding the computational complexity, it is important to note that the additional computational overhead introduced by the PRGR framework (including reliability modeling and iterative pseudo-label refinement) is strictly confined to the training phase. During the inference phase, the classification network and refinement modules are completely discarded, and segmentation is performed solely by the trained U-Net backbone. Consequently, the inference time and computational cost of our method are exactly identical to those of a standard U-Net, ensuring that the substantial

performance improvements do not entail any extra computational burden during practical clinical deployment.

Table 1. Comparison of Weakly Supervised Segmentation Results on STS and HeadNeck Dataset.

Dataset	Method	Dice	Recall	Precision	HD95	*p*-Value
STS	WILDCAT	0.255	0.697	0.173	19.236	6.10×10^{-5}
	ESOL_WSSS	0.310	**0.853**	0.207	18.818	3.52×10^{-5}
	ReCAM	0.249	0.851	0.152	12.632	5.63×10^{-5}
	Belharbi et al. [15]	0.365	0.731	0.262	12.105	5.61×10^{-5}
	SEAM	0.575	0.725	0.524	9.832	7.20×10^{-3}
	WSS_CMER	0.391	0.284	0.720	15.31	1.22×10^{-4}
	UCWS-Net	0.655	0.631	0.724	8.602	4.89×10^{-3}
	PRGR (Ours)	**0.673**	0.656	**0.776**	**8.321**	-
	U-Net (Upper bound)	0.684	0.797	0.643	7.668	-
HeadNeck	WILDCAT	0.450	0.609	0.433	14.946	1.22×10^{-2}
	ESOL_WSSS	0.284	**0.925**	0.172	18.167	4.88×10^{-4}
	ReCAM	0.294	0.922	0.180	17.918	2.81×10^{-4}
	Belharbi et al. [15]	0.156	0.921	0.086	21.562	2.50×10^{-4}
	SEAM	0.371	0.772	0.258	26.038	1.46×10^{-3}
	WSS_CMER	0.354	0.786	0.243	18.935	4.85×10^{-4}
	UCWS-Net	0.512	0.560	0.515	10.961	2.69×10^{-2}
	PRGR (Ours)	**0.596**	0.609	**0.684**	**10.593**	-
	U-Net (Upper bound)	0.625	0.726	0.618	8.220	-

Analysis of the Precision and Recall metrics provides information on the characteristics of PRGR regarding over-segmentation and pseudo-label noise. Multi-step iterative methods, such as ESOL_WSSS and ReCAM, achieve Recall scores exceeding 0.85 on STS and 0.92 on HeadNeck, alongside Precision scores below 0.21 on both datasets. This metric imbalance suggests that these baseline methods activate a broader range of background tissues, leading to false positives. In contrast, PRGR achieves a Precision of 0.776 on STS and 0.684 on HeadNeck. The Precision of this weakly supervised method surpasses that of the fully supervised U-Net, which scores 0.643 and 0.618 respectively. These precision scores indicate that the reliability modeling and progressive optimization mechanisms penalize uncertain background regions and restrict the network focus to confident tumor areas, yielding precise anatomical boundaries.

Furthermore, in terms of boundary-sensitive evaluation, the PRGR framework demonstrates clear superiority. As shown in Table 1, PRGR achieves the lowest HD95 scores among all weakly supervised methods, yielding 8.321 on the STS dataset and 10.593 on the HeadNeck dataset. This significant reduction in boundary distance compared to baselines like ESOL_WSSS and ReCAM mathematically validates our claim that the reliability-guided refinement mechanism effectively suppresses false-positive boundary leakage and refines anatomical tumor delineation.

4.4.2. Ablation Experiment

To deeply investigate the functional necessity of each module within the PRGR framework, we conduct exhaustive ablation studies dissecting the loss components, the reliability formulations, and the refinement mechanism. The comprehensive results on the STS and HeadNeck datasets are summarized in Table 2.

Table 2. Ablation Study of the Proposed PRGR Framework on STS and HeadNeck Datasets.

Dataset	Method	Dice	Recall	Precision
STS	CAM-Baseline	0.261	**0.756**	0.164
	S2C	0.514	0.533	0.607
	PRGR-Single	0.555	0.542	0.632
	PRGR-Single w/o $\mathcal{L}_M + \mathcal{L}_G$	0.435	0.305	**0.791**
	PRGR-Single w/o $\mathcal{L}_G$	0.510	0.406	0.685
	PRGR-Single w/o $\mathcal{L}_M$	0.515	0.545	0.590
	PRGR-Single w/o $R^{ct} + R^{cm}$	0.477	0.726	0.436
	PRGR-Single w/o R^{cm}	0.542	0.598	0.486
	PRGR-Full (Ours)	**0.673**	0.656	0.776
HeadNeck	CAM-Baseline	0.273	**0.933**	0.185
	S2C	0.452	0.499	0.628
	PRGR-Single	0.482	0.490	0.655
	PRGR-Single w/o $\mathcal{L}_M + \mathcal{L}_G$	0.327	0.398	**0.732**
	PRGR-Single w/o $\mathcal{L}_G$	0.419	0.374	0.690
	PRGR-Single w/o $\mathcal{L}_M$	0.455	0.539	0.453
	PRGR-Single w/o $R^{ct} + R^{cm}$	0.410	0.773	0.368
	PRGR-Single w/o R^{cm}	0.465	0.704	0.437
	PRGR-Full (Ours)	**0.596**	0.609	**0.684**

We sequentially evaluate the impact of the regional loss components. When relying solely on the reliable region loss ($\mathcal{L}_R$, denoted as PRGR-Single w/o $\mathcal{L}_M + \mathcal{L}_G$), the model achieves exceptionally high Precision (0.791 on STS and 0.732 on HeadNeck) but suffers from severe boundary shrinkage, resulting in poor Recall scores. Integrating the moderately reliable region loss (PRGR-Single w/o $\mathcal{L}_G$) improves the Dice score to 0.510 on STS. However, without the global regularization loss ($\mathcal{L}_G$), the model occasionally suffers from local overfitting. The complete single-stage optimization ($\mathcal{L}_R + \mathcal{L}_M + \mathcal{L}_G$, denoted as PRGR-Single) achieves a balanced Recall and Precision, demonstrating that $\mathcal{L}_G$ is indispensable as a structural anchor to prevent catastrophic lesion forgetting.

We further validate the multimodal reliability modeling. Using only PET metabolic reliability (PRGR-Single w/o $R^{ct} + R^{cm}$) yields a high Recall but low Precision (0.436 on STS), as physiological uptake introduces false positives. Multiplying by CT structural reliability (PRGR-Single w/o R^{cm}) filters out background noise by enforcing anatomical boundaries, improving Precision. Finally, the full formulation incorporating the cross-modal consistency term yields the optimal PRGR-Single performance, proving that multimodal alignment provides mathematically superior supervision quality compared to single-modality estimations.

To validate the superiority of our dynamic pseudo-label refinement against external approaches, we replaced our refinement module with S2C [29], a recent state-of-the-art pseudo-label refinement approach leveraging the Segment Anything Model (SAM) for weakly supervised segmentation. As shown in Table 2, while S2C improves upon the CAM-Baseline, it struggles to fully resolve the complex metabolic-structural mismatches in PET/CT, achieving Dice scores of 0.514 and 0.452 on the STS and HeadNeck datasets, respectively. In contrast, our complete iterative framework (PRGR-Full) dynamically updates the reliability map in a closed-loop EMA manner, achieving the highest overall Dice scores of 0.673 and 0.596 by effectively suppressing cross-boundary leakage while recovering uncertain tumor areas.

4.4.3. Qualitative Evaluation

To visually demonstrate the segmentation performance of the proposed PRGR framework, Figure 2 presents a qualitative comparison of partial slice segmentation results. Slices a and b represent cases from the STS dataset, while slices c and d correspond to cases from the HeadNeck dataset. In the visual representations, the green areas indicate true positives, the blue areas represent false positives, and the red areas denote false negatives. The visualization reveals a common limitation among the baseline weakly supervised methods, which frequently fail to balance activation coverage and boundary restriction, resulting in either extensive false positive regions or severe missing tumor activations. The PRGR framework delineates tumor boundaries more accurately, producing fewer false positive and false negative pixels across both datasets.

The segmentation results on the STS dataset in slices a and b illustrate the susceptibility of baseline methods to severe over-segmentation. In slice a, methods including ESOL_WSSS, ReCAM, and UCWS-Net exhibit extensive blue regions. This visual pattern indicates a severe precision degradation where the models mistakenly classify a large portion of the surrounding normal tissues as tumors, despite successfully covering the main tumor body. The proposed PRGR method effectively suppresses these false activations, maintaining a tight and accurate boundary around the target lesion. This over-segmentation issue among the baselines is even more pronounced in slice b, where the majority of the evaluated baseline methods show massive false positive areas spreading widely across the background. In contrast, the PRGR framework eliminates background noise almost entirely, maximizing the green true positive regions while minimizing the blue false positive areas.

The visualizations for the HeadNeck dataset in slices c and d further validate the stability of the proposed method in complex anatomical regions. In slice c, methods like ESOL_WSSS and ReCAM continue to suffer from severe over-segmentation, characterized by prominent blue regions. Conversely, WSS_CMER fails to activate the target region almost completely, resulting in extensive red false negative areas, which indicates a severe under-segmentation problem. The PRGR framework successfully maintains structural integrity, accurately covering the tumor with minimal red or blue regions. Similarly, in slice d, methods such as Belharbi et al., ESOL_WSSS, and ReCAM exhibit a tendency to aggressively expand the predicted boundaries, leading to significant false positive predictions. The PRGR model outputs a highly balanced prediction, successfully localizing the tumor boundaries without erroneously activating the adjacent healthy tissues. These purely qualitative results confirm that the progressive optimization strategy effectively constrains pseudo-label expansion and preserves anatomical details.

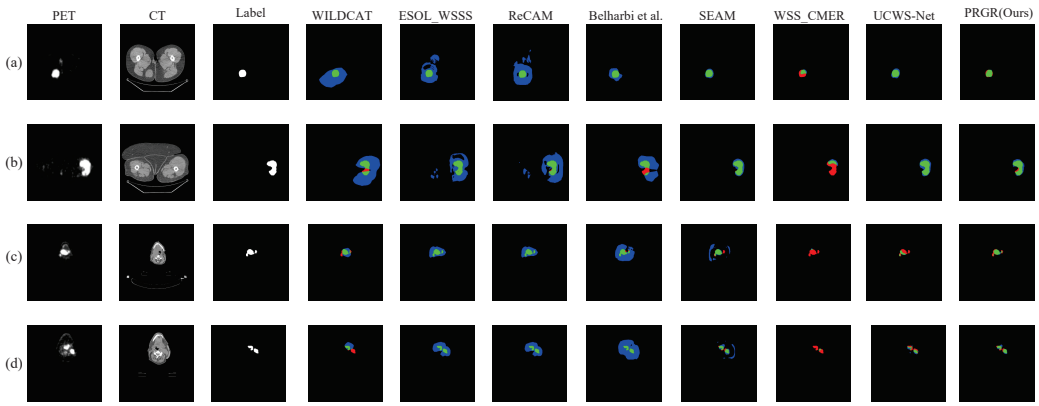


Figure 2. Comparison of partial slice segmentation results, (**a**–**d**) represent four slices. Green area indicates true positive, blue area indicates false positive, red area indicates false negative.

5. Discussion

The experimental results on the STS and HeadNeck datasets provide several key insights into the behavior of weakly supervised learning in multimodal PET/CT imaging. A prominent phenomenon observed in our comparative experiments is the severe Recall–Precision imbalance among conventional iterative CAM-based methods (e.g., ESOL_WSSS and ReCAM), which yield high Recall (>0.85) but extremely low Precision (<0.21) alongside large HD95 errors. This occurs because image-level classification signals inherently encourage activation expansion into surrounding background tissues when lacking explicit spatial boundary constraints. In contrast, the proposed PRGR framework achieves substantial improvements in Precision (0.776 on STS and 0.684 on HeadNeck) and the lowest HD95 distances (8.321 and 10.593, respectively), even surpassing the Precision of the fully supervised U-Net upper bound. This empirical advantage stems from two core mechanisms. First, our parameter-free reliability modeling integrates PET metabolic consistency (R^{pet}) with CT structural gradients (R^{ct}) and cross-modal agreement (R^{cm}), creating an anatomical barrier that penalizes pseudo-label spillover across sharp tissue interfaces. Second, rather than treating noisy pseudo-labels uniformly, the adaptive variance-maximization partitioning decomposes the supervision space into reliable (Ω_R), moderately reliable (Ω_M), and low-reliability (Ω_L) regions. By applying stop-gradient-constrained progressive supervision in Ω_M and closed-loop EMA refinement, the network steadily recovers under-segmented tumor regions while suppressing false-positive noise propagation. Furthermore, the ablation comparison against the SAM-based refinement method (S2C) reveals that general-purpose visual foundation models struggle to resolve the unique metabolic-structural discrepancies in PET/CT, underscoring the necessity of domain-specific cross-modal reliability modeling.

From a clinical and practical perspective, the findings of this study hold several meaningful implications for automated PET/CT image analysis. First, manual pixel-level tumor delineation across multimodal volumes is highly labor-intensive and prone to inter-observer variability, which restricts the construction of large-scale training cohorts. By narrowing the performance gap to within 1.1% Dice of the fully supervised U-Net on the STS dataset using only image-level labels, PRGR offers a cost-effective paradigm for leveraging retrospective clinical archives without exhaustive voxel-wise annotation. Second, in quantitative oncology imaging, false-positive boundary leakage into adjacent normal structures can severely distort critical biomarkers such as Metabolic Tumor Volume (MTV) and Total Lesion Glycolysis (TLG). The high Precision and reduced HD95 achieved by PRGR indicate its potential to provide tighter and more anatomically plausible contour initializations for quantitative biomarker extraction and radiotherapy target volume definition. Third, a critical advantage for practical deployment is the decoupling of training complexity from inference efficiency. Because the classification backbone, reliability estimation, and iterative refinement modules are exclusively executed during the training phase, the final deployed model is a standard U-Net architecture. This guarantees that the improved boundary delineation is achieved with zero additional computational overhead or latency during inference.

Despite the promising performance of the proposed PRGR framework, several limitations remain. First, the reliability modeling heavily depends on cross-modal consistency, which could be compromised by severe imaging artifacts such as patient motion or metal interference. Second, although the model is trained on all slices, weakly supervised segmentation in PET/CT remains challenged by physiological hypermetabolic regions (e.g., brain, heart, or bladder), which can occasionally induce false positive activations due to the lack of pixel-level negative constraints. Third, the current framework processes 2D axial slices independently, which overlooks inter-slice spatial consistency along the z-axis. Furthermore, the multi-stage iterative pseudo-label refinement introduces

additional computational overhead compared to single-stage methods. Future work will focus on developing robust reliability estimation strategies resilient to single-modality degradation, integrating anatomical priors to suppress benign high-uptake regions, exploring 2.5D or full 3D network architectures to incorporate volumetric context, and exploring lightweight progressive optimization algorithms to accelerate the training process.

6. Conclusions

In this paper, we proposed a Progressive Reliability-Guided Refinement framework, denoted as PRGR, to address the challenges of pseudo-label noise and spatially varying supervision quality in weakly supervised PET/CT tumor segmentation. By integrating PET metabolic distribution and CT structural features, a reliability modeling module constructs spatial reliability maps to evaluate the credibility of initial pseudo-labels without requiring additional learnable parameters. Based on adaptive reliability partitioning, a progressive optimization strategy combined with iterative pseudo-label refinement iteratively updates pseudo-labels to improve segmentation performance and enforce anatomical boundary constraints. Experiments on the STS and HeadNeck datasets demonstrate that the proposed framework achieves higher Dice and Precision scores compared to existing weakly supervised segmentation methods. Specifically, PRGR achieves Dice scores of 0.673 on the STS dataset and 0.596 on the HeadNeck dataset, representing absolute improvements of 1.8% and 8.4% over the second-best baseline methods, alongside remarkable Precision scores of 0.776 and 0.684. The quantitative and qualitative results confirm that PRGR effectively mitigates over-segmentation, reduces false positive activations, and generates precise tumor boundaries, thereby narrowing the performance gap with fully supervised approaches. By achieving high-precision segmentation with minimal annotation requirements, the proposed PRGR framework provides a promising technical approach for weakly supervised PET/CT tumor segmentation. Future work will focus on conducting prospective clinical validation and workflow studies to further evaluate its practical applicability in assisting clinical image analysis.

Author Contributions: Z.D.: Methodology, Writing—Original Draft. H.J.: Methodology, Writing—Review & Editing. All authors have read and agreed to the published version of the manuscript.

Funding: The research was was funded by Joint Fund General Project of Liaoning Provincial Department of Science and Technology of grant number 2024-BSLH-206. The APC was funded by 2024-BSLH-206.

Institutional Review Board Statement: IRB is waived because the study did not involve human experiments or new human subjects, and it is a secondary analysis of publicly available, de-identified images.

Informed Consent Statement: ICS is waived because the study did not involve human experiments or new human subjects, and it is a secondary analysis of publicly available, de-identified images.

Data Availability Statement: The data analyzed in this study were obtained from publicly available open-source datasets. The datasets are available at https://doi.org/10.1088/0031-9155/60/14/5471 and https://doi.org/10.1038/s41598-017-10371-5. No new data were created.

Acknowledgments: During the preparation of this work, the authors used DeepSeek-V4.1-Flash-1017 for language polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fallahpoor, M.; Chakraborty, S.; Pradhan, B.; Faust, O.; Barua, P.D.; Chegeni, H.; Acharya, R. Deep learning techniques in PET/CT imaging: A comprehensive review from sinogram to image space. *Comput. Methods Programs Biomed.* **2024**, *243*, 107880. [CrossRef]
2. Islam, M.Z.; Naqvi, R.A.; Haider, A.; Kim, H.S. Deep learning for automatic tumor lesions delineation and prognostic assessment in multi-modality PET/CT: A prospective survey. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106276. [CrossRef]
3. Wang, M.; Jiang, H.; Shi, T.; Yao, Y.d. SCL-Net: Structured collaborative learning for PET/CT based tumor segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *27*, 1048–1059. [CrossRef]
4. Diao, Z.; Jiang, H.; Shi, T. A spatial squeeze and multimodal feature fusion attention network for multiple tumor segmentation from PET–CT Volumes. *Eng. Appl. Artif. Intell.* **2023**, *121*, 105955. [CrossRef]
5. Li, G.Y.; Chen, J.; Jang, S.I.; Gong, K.; Li, Q. SwinCross: Cross-modal Swin transformer for head-and-neck tumor segmentation in PET/CT images. *Med. Phys.* **2024**, *51*, 2096–2107. [CrossRef]
6. Khan, W.; Leem, S.; See, K.B.; Wong, J.K.; Zhang, S.; Fang, R. A comprehensive survey of foundation models in medicine. *IEEE Rev. Biomed. Eng.* **2025**, *19*, 283–304. [CrossRef]
7. Früh, M.; Fischer, M.; Schilling, A.; Gatidis, S.; Hepp, T. Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging. *J. Med. Imaging* **2021**, *8*, 054003. [CrossRef]
8. 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 *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 316–327.
9. Leung, K.H.; Rowe, S.P.; Sadaghiani, M.S.; Leal, J.P.; Mena, E.; Choyke, P.L.; Du, Y.; Pomper, M.G. Deep semisupervised transfer learning for fully automated whole-body tumor quantification and prognosis of cancer on PET/CT. *J. Nucl. Med.* **2024**, *65*, 643–650. [CrossRef]
10. Górka, M.; Jaworek, D.; Wodzinski, M. Deep learning-based segmentation of tumors in pet/ct volumes: benchmark of different architectures and training strategies. *arXiv* **2024**, arXiv:2404.09761.
11. Yang, D.; Wang, Y.; Ma, Y.; Yang, H. A Multi-Scale Interpretability-Based PET-CT Tumor Segmentation Method. *Mathematics* **2025**, *13*, 1139. [CrossRef]
12. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
13. Durand, T.; Mordan, T.; Thome, N.; Cord, M. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 642–651.
14. Zhang, X.; Wei, Y.; Kang, G.; Yang, Y.; Huang, T. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 597–613.
15. Belharbi, S.; Rony, J.; Dolz, J.; Ayed, I.B.; McCaffrey, L.; Granger, E. Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. *IEEE Trans. Med. Imaging* **2021**, *41*, 702–714. [CrossRef]
16. Patel, G.; Dolz, J. Weakly supervised segmentation with cross-modality equivariant constraints. *Med. Image Anal.* **2022**, *77*, 102374. [CrossRef]
17. Li, J.; Jie, Z.; Wang, X.; Wei, X.; Ma, L. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16037–16051. [CrossRef]
18. Chen, Z.; Wang, T.; Wu, X.; Hua, X.S.; Zhang, H.; Sun, Q. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 969–978.
19. Du, W.; Huo, Y.; Zhou, R.; Sun, Y.; Tang, S.; Zhao, X.; Li, Y.; Li, G. Consistency label-activated region generating network for weakly supervised medical image segmentation. *Comput. Biol. Med.* **2024**, *173*, 108380. [CrossRef]
20. Kuang, Z.; Yan, Z.; Yu, L. Weakly supervised learning for multi-class medical image segmentation via feature decomposition. *Comput. Biol. Med.* **2024**, *171*, 108228. [CrossRef]
21. Fu, J.; Wang, G.; Lu, T.; Yue, Q.; Vercauteren, T.; Ourselin, S.; Zhang, S. UM-CAM: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised segmentation. *Pattern Recognit.* **2025**, *160*, 111204. [CrossRef]
22. Han, M.; Luo, X.; Xie, X.; Liao, W.; Zhang, S.; Song, T.; Wang, G.; Zhang, S. DMSPS: Dynamically mixed soft pseudo-label supervision for scribble-supervised medical image segmentation. *Med. Image Anal.* **2024**, *97*, 103274. [CrossRef]
23. Shi, L.; Li, G.; Zhang, J. Uncertainty-Aware Extreme Point Tracing for Weakly Supervised Ultrasound Image Segmentation. *arXiv* **2025**, arXiv:2510.15666.
24. Khan, H.; Chen, W.; Niazi, M.K.K. Weakly Supervised Teacher–Student Framework with Progressive Pseudo-Mask Refinement for Multi-class Gland Segmentation. *J. Clin. Transl. Pathol.* **2026**, *6*, 1. [CrossRef]
25. Diao, Z.; Jiang, H. An uncertainty-based Collaborative Weakly Supervised Segmentation Network for Positron emission tomography-Computed tomography images. *Eng. Appl. Artif. Intell.* **2025**, *148*, 110442. [CrossRef]

26. Vallières, M.; Freeman, C.R.; Skamene, S.R.; El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* **2015**, *60*, 5471–5496. [CrossRef]
27. Vallières, M.; Kay-Rivest, E.; Perrin, L.J.; Liem, X.; Furstoss, C.; Aerts, H.J.; Khaouam, N.; Nguyen-Tan, P.F.; Wang, C.S.; Sultanem, K.; et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci. Rep.* **2017**, *7*, 10117. [CrossRef]
28. Andrearczyk, V.; Oreiller, V.; Vallières, M.; Castelli, J.; Elhalawani, H.; Jreige, M.; Boughdad, S.; Prior, J.O.; Depeursinge, A. Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In *Proceedings of the Medical Imaging with Deep Learning*; PMLR: Cambridge, MA, USA, 2020; pp. 33–43.
29. Kweon, S.; Yoon, S.H.; Kim, H.; Park, J.; Yoon, K.J. From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 3155–3165.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.