

## Probability-Based Lesion-Aware FN/FP Loss for Whole-Body PET/CT Tumor Segmentation

Mobin Mohebi<sup>1,2,3</sup>[0000-0003-4023-8753] (✉), Mathilde MASSE<sup>1,4</sup>, Alexandra FERRAN<sup>1</sup>,  
Maria A. Zuluaga<sup>3</sup>, Olivier HUMBERT<sup>1,4</sup>

<sup>1</sup> Department of Nuclear Medicine, Centre Antoine Lacassagne, Nice, France

<sup>2</sup> Institut de Biologie Valrose (iBV), Université Côte d’Azur, CNRS, Inserm, Nice, France

<sup>3</sup> EURECOM, 06410, Biot, France

<sup>4</sup> Université Côte d’Azur, INSERM, CNRS, iBV, Nice

✉mobin.mohebi@univ-cotedazur.fr

**Abstract.** Accurate tumor segmentation on whole-body 3D PET/CT requires both high voxel-wise overlap and reliable lesion-wise detection. Standard CE+Dice nnU-Net training can achieve strong Dice while still missing individual lesion components. We propose a Probability-Weighted Error-Gated False-Negative/False-Positive (PWEF-FNFP) loss for binary whole-body tumor segmentation. The loss extends CE+Dice with error-gated connected-component terms that apply additional supervision only to current false-negative and false-positive lesion components. The false-negative term increases foreground support inside missed lesions, whereas the false-positive term suppresses spurious predicted components according to mean foreground probability. We further evaluate a center-regularized variant by adding the  $\mathcal{L}_{Center}$  term from Instance-wise and Center-of-Instance (ICI) to incorporate lesion-center localization. A 3D full-resolution nnU-Net was trained on 835 scans from Centre Antoine Lacassagne longitudinal FDG PET/CT lung cancer cohort and evaluated on 92 held-out internal scans and the external AutoPET dataset. On the internal test set, conventional non-lesion-aware losses achieved lesion-wise F1-scores of 0.536–0.599. The best performance was obtained when the PWEF-FNFP loss was combined with  $\mathcal{L}_{Center}$ , suggesting that error-gated FN/FP supervision and center localization provide complementary optimization signals. Size-stratified analysis showed the strongest gains for lesions below 1 cm<sup>3</sup>. On AutoPET, the PWEF-FNFP loss achieved the highest Dice among lesion-aware variants, while the center-regularized version achieved the highest global lesion-wise F1 score. These results support probability-weighted connected-component optimization for improving small-lesion detection while highlighting the remaining challenge of false-positive control and domain generalization.

**Keywords:** PET/CT segmentation, Lesion-aware loss, nnU-Net, Oncology.

## 1 Introduction

Accurate tumor segmentation is essential for quantitative oncological image analysis, supporting tumor-burden estimation and therapy response assessment [1, 2]. In metastatic disease, however, segmentation quality is not fully captured by voxel-wise overlap alone. A model may obtain a reasonable Dice score while still missing small lesions, which typically have low signal and contribute few voxels to the global objective. Such missed lesions may lead to underestimation of disease burden, whereas unsupported false-positive (FP) components increase manual correction burden and may reduce clinical trust. Therefore, lesion-wise false-negatives (FN) and FP are important failure modes for automated tumor segmentation [2, 3]. Recent PET/CT benchmarks such as autoPET and HECKTOR have shown the potential of nnU-Net-like architectures [2, 4, 5]. However, the standard nnU-Net training objective is based mainly on cross-entropy and Dice loss. These losses optimize voxel-wise classification and global overlap rather than lesion-wise detection.

Overlap-based losses are dominated by large foreground regions, while small lesions have limited influence on the total loss. Prior losses address this problem through voxel rebalancing, instance balancing with BlobLoss, and instance/center supervision with Instance-wise and Center-of-Instance (ICI) loss [6–11]. However, BlobLoss averages loss uniformly over all instances regardless of detection status and has no explicit false-positive term, while ICI suppresses spurious components only indirectly through center-of-mass competition rather than by directly scoring the model’s confidence within each spurious region. Neither formulation, therefore, explicitly separates the current lesion-level error components that require opposite corrections: missed ground-truth (GT) lesions and spurious predicted components. Existing component-based lesion losses often rely on hard decisions about whether a predicted and ground-truth component overlap. Small overlap changes can therefore alter whether a lesion is treated as detected, missed, or false positive, providing limited information about error severity, especially for small or low-confidence lesions. This leaves a gap for an error-gated loss that identifies current FN/FP components by thresholding and uses foreground probabilities to assign a severity to each error.

To address this gap, we propose a probability-based lesion-aware FN/FP (PWEG-FNFP) loss for binary tumor segmentation. The loss operates on connected components rather than only individual voxels. The FN term increases foreground probability inside missed GT lesions, while the FP term decreases foreground probability inside unsupported predicted components. Thus, the PWEG-FNFP term is not only instance-aware; it gives separate optimization directions to the two clinically relevant detection errors. After threshold-based component selection, probabilities determine how strongly each selected component contributes. This provides a differentiable training signal for lesion-wise errors, since metrics such as F1 and sensitivity depend on discrete component matching and cannot be optimized directly by gradient descent. Unlike generic instance reweighting, the PWEG-FNFP loss allocates additional supervision only to currently missed or unsupported components. Because probabilities are averaged within each selected component, the correction is component-normalized rather than dominated by

lesion volume. We integrate this objective into nnU-Net and also evaluate a center-regularized variant using the center-of-instance term from ICI.

## 2 Methods

### 2.1 Optimization Rationale and Design Principles

CE and Dice distribute gradient across all foreground voxels in proportion to volume; the PWEG-FNFP loss instead restricts gradient to voxels inside currently misclassified lesion components. The PWEG-FNFP loss instead defines two current error sets: missed GT components (FN lesions) and unsupported predicted components (FP lesions).

Instead of applying additional supervision to all lesion instances, as in generic instance reweighting approaches such as BlobLoss and ICI,  $\mathcal{L}_{\text{FN}}$  is activated only for missed GT components and  $\mathcal{L}_{\text{FP}}$  only for unsupported predicted components. This error-gating mechanism gives the two errors opposite optimization directions:  $\mathcal{L}_{\text{FN}}$  increases foreground support inside missed lesions, whereas  $\mathcal{L}_{\text{FP}}$  suppresses foreground support inside spurious predictions. A missed lesion should receive a correction proportional to how confidently it is missed, not to how many voxels it contains. We achieve this by averaging probabilities within each selected component before combining across components, so that each missed lesion receives comparable lesion-level correction independent of its volume — a property we refer to as component normalization. Similarly, the severity of each error should reflect the model's own confidence, not merely whether a component crossed a detection threshold. Thresholding defines which components are FN or FP, while mean foreground probability defines how severe each selected error is: a missed lesion with  $\bar{p} = 0.05$  is penalized more than one with  $\bar{p} = 0.45$ , and an unsupported component with  $\bar{p} = 0.90$  more than one with  $\bar{p} = 0.20$ .

### 2.2 Probability-Based Lesion-Aware FN/FP Loss

Let the model output a foreground probability map  $p$  and let  $g$  denote the corresponding binary GT lesion mask.

**False-negative term.** GT lesion components are first extracted from  $g$  using 3D connected-component labeling. A GT component is considered missed if it has no overlap with the thresholded prediction  $p \geq \tau_{\text{FN}}$ . Let  $\text{FN}$  denote the set of missed GT components in the image, and let  $N_{\text{FN}} = |\text{FN}|$  be the number of such components. For each missed lesion  $n$ ,  $\bar{p}_n$  is the mean predicted foreground probability inside that GT component. The FN term is defined as:

$$\mathcal{L}_{\text{FN}} = \frac{1}{N_{\text{FN}} + \epsilon} \sum_{n \in \text{FN}} (1 - \bar{p}_n) \quad (1)$$

where  $\varepsilon$  is a small constant used for numerical stability. If no missed lesion is present,  $\mathcal{L}_{FN}$  is set to zero. Because probabilities are averaged within each missed component, small selected lesions receive stronger per-voxel correction, while the total correction per selected lesion remains component-normalized.

**False-positive term.** To control spurious predictions, predicted components are extracted from the thresholded probability map  $p \geq \tau_{FP}$ . A predicted component is considered false-positive if it has no overlap with the GT mask  $g$ . Let FP denote the set of false-positive predicted components in the image, and let  $N_{FP} = |\text{FP}|$  be the number of such components. For each false-positive component  $m$ ,  $\bar{p}_m$  is the mean predicted foreground probability inside that predicted component. The FP term is defined as:

$$\mathcal{L}_{FP} = \frac{1}{N_{FP} + \epsilon} \sum_{m \in \text{FP}} \bar{p}_m \quad (2)$$

High-confidence unsupported components receive larger loss values than weak uncertain components. Thus, probability weighting turns the discrete count objective ( $\#FN + \#FP$ ) into a differentiable severity-aware surrogate.

**Total training objective.** During training,  $\mathcal{L}_{FN}$  and  $\mathcal{L}_{FP}$  are computed per image and averaged over the mini-batch, following standard practice for the CE and Dice terms. The total training objective then becomes:

$$L = \alpha L_{CE} + \beta L_{Dice} + \gamma L_{FN} + \delta L_{FP} \quad (3)$$

In the final implementation, the CE, Dice, and FN terms were assigned equal weights. The FP term was disabled during the warm-up period and then linearly increased to its final weight by the end of fine-tuning. This curriculum first encourages lesion recovery and then gradually strengthens FP control.

### 2.3 Complementarity with Center-of-Instance Regularization (Ablation)

The PWEG-FNFP loss (Section 2.2) is the sole proposed contribution of this work. To test whether it provides information complementary to existing lesion-localization mechanisms, we conduct an ablation combining PWEG-FNFP with the center-of-instance term  $\mathcal{L}_{Center}$  from ICI loss [10], which transforms each ground-truth and predicted component into a fixed-size cube centered at its center of mass and computes a Dice-based loss between the two resulting center maps; we refer to [10] for the full formulation. The combined objective used in this ablation is:

$$\mathcal{L}_{ours} + \eta \mathcal{L}_{Center} \quad (4)$$

where  $\eta$  controls the contribution of the center term. This experiment tests whether error-gated FN/FP correction and center-of-instance localization are complementary: PWEG-FNFP determines which error components require correction, whereas  $\mathcal{L}_{Center}$  encourages accurate localization of lesion centers. Results for this ablation are reported alongside the main comparisons, and are not treated as part of the proposed method.

### 3 Experiments and Results

#### 3.1 Datasets

Experiments were conducted at Centre Antoine Lacassagne using longitudinal FDG PET/CT lung cancer cohort and the public AutoPET dataset. The institutional cohort contained 927 scans, including baseline and follow-up examinations from patients with primary lung cancer and metastatic disease. The cohort was split at the scan level into 835 scans for training/validation and 92 held-out scans for internal testing, with each patient constrained to a single set to prevent cross-set leakage. Tumor annotations included both primary tumors and metastatic lesions and were generated from expert-defined PET/CT regions of interest. The AutoPET dataset was used only for external evaluation and was not used during training. The task was formulated as binary whole-body tumor segmentation, with tumor lesions as foreground and all non-tumor tissue as background.

#### 3.2 Implementation Details

All experiments used the 3D full-resolution nnU-Net framework. PET and CT were used as two-channel input, with isotropic spacing of  $(1.65 \times 1.65 \times 1.65)$  mm<sup>3</sup> and patch size  $(192 \times 192 \times 192)$ . The learning rate followed the default nnU-Net polynomial schedule, and no loss-specific learning-rate tuning was performed. A CE+Dice baseline model was first trained for 4000 epochs. This checkpoint was then used to initialize all fine-tuning experiments, which were trained for 200 epochs under identical settings. This two-stage design ensured that differences between methods were mainly attributable to the loss formulation rather than initialization or training duration.

For the proposed PWEG-FNFP loss, we used  $\alpha = 0.5$ ,  $\beta = 0.5$ ,  $\gamma = 0.5$  for the CE, Dice, and FN terms, respectively. The FP term was disabled during the first seven fine-tuning epochs and then linearly increased to  $\delta_{\max} = 0.1$ . FN and FP components were identified using thresholds  $\tau_{\text{FN}} = 0.6$  and  $\tau_{\text{FP}} = 0.3$ , respectively. As an additional reference, we also implemented an ICI-paper-style patch-based baseline using a MONAI 3D Residual U-Net with PET/CT two-channel input [10]. This model (PET+CT ICI Net) followed the original ICI patch-based setting as closely as possible, using  $96^3$  patches, Adam optimization, center cube size  $7^3$ , weights: global:  $a = 1$ , blob:  $b = 1$ , center:  $c = 1$ ), sigmoid thresholding at 0.5, and the full ICI loss with global, instance-wise, and center-of-instance terms.

#### 3.3 Compared Losses

The PWEG-FNFP loss was compared with both conventional voxel-wise losses and lesion-aware losses. Conventional losses included generalized Dice loss (GDL), generalized Dice focal loss (GDFL), DiceFocal loss, Tversky loss, and Hausdorff-distance-based loss. These losses mainly address voxel-wise class imbalance, hard voxels, asym-

metric FN/FP weighting, or boundary errors. Lesion-aware baselines included BlobLoss, ICI, and LfalseInstance. BlobLoss addresses instance imbalance by averaging the loss over connected components [9], ICI combines global, instance-wise, and center-of-instance terms [10, 11], and LfalseInstance penalizes missed and false predicted instances using connected-component matching [8]. All compared losses were fine-tuned from the same CE+Dice checkpoint for 200 epochs. Because this baseline differs from the nnU-Net loss-only comparisons in both architecture and training protocol, it was reported separately as an ICI-paper-style reference model.

### 3.4 Evaluation Metrics

Performance was evaluated using both voxel-level and lesion-level metrics. Voxel-level segmentation quality was assessed using Dice score. Lesion-level performance was assessed by extracting 3D connected components from the predicted and GT masks. A predicted component and a GT component were considered matched if at least 50% of the GT component overlapped the prediction, or at least 50% of the predicted component overlapped the GT. Matched GT lesions were counted as true positives, unmatched GT lesions as false negatives, and unmatched predicted components as FP. Lesion-wise sensitivity, precision, and F1-score were computed from these component-level counts. Macro lesion-wise F1 was computed as the mean of per-scan F1 values, whereas global lesion-wise F1 was computed from pooled TP, FP, and FN counts across all scans. To assess small-lesion behavior, sensitivity was stratified by GT lesion volume and precision by predicted component volume. Four size categories were used:  $<0.1$  cubic centimeters ( $\text{cm}^3$ ),  $0.1\text{--}1$   $\text{cm}^3$ ,  $1\text{--}5$   $\text{cm}^3$ , and  $>5$   $\text{cm}^3$ . Paired statistical comparisons of per-scan lesion-wise F1 were performed using two-sided Wilcoxon signed-rank tests, with Holm correction for multiple comparisons.

### 3.5 Internal Test Performance

Table 1 summarizes the internal test performance of the CE+Dice baseline, conventional voxel-wise losses, lesion-aware baselines, and the PWEG-FNFP loss variants.

**Table 1.** Internal test performance on the held-out institutional cohort. Dice and macro lesion-wise F1 are reported as mean  $\pm$  standard deviation (SD) across test scans.

| Method                                                    | Test Dice $\pm$ SD                    | Macro Lesion-wise F1 $\pm$ SD       | Global Lesion-wise F1 |
|-----------------------------------------------------------|---------------------------------------|-------------------------------------|-----------------------|
| <b>CE+Dice baseline</b>                                   | 0.7182 $\pm$ 0.2408                   | 0.624 $\pm$ 0.325                   | 0.617                 |
| <b>PET+CT ICI Net</b>                                     | 0.1096 $\pm$ 0.1892                   | 0.152 $\pm$ 0.166                   | 0.115                 |
| <b>Tversky</b>                                            | 0.6017 $\pm$ 0.2845                   | 0.536 $\pm$ 0.257                   | 0.498                 |
| <b>Hausdorff</b>                                          | 0.5921 $\pm$ 0.2964                   | 0.545 $\pm$ 0.273                   | 0.502                 |
| <b>DiceFocal</b>                                          | 0.6548 $\pm$ 0.2852                   | 0.573 $\pm$ 0.253                   | 0.532                 |
| <b>GDFL</b>                                               | 0.5930 $\pm$ 0.2975                   | 0.577 $\pm$ 0.251                   | 0.530                 |
| <b>GDL</b>                                                | 0.6251 $\pm$ 0.2787                   | 0.599 $\pm$ 0.246                   | 0.554                 |
| <b>LfalseInstance</b>                                     | 0.6985 $\pm$ 0.2455                   | 0.681 $\pm$ 0.268                   | 0.639                 |
| <b>BlobLoss</b>                                           | 0.7144 $\pm$ 0.2328                   | 0.695 $\pm$ 0.270                   | 0.654                 |
| <b>ICI</b>                                                | 0.7194 $\pm$ 0.2356                   | 0.715 $\pm$ 0.262                   | 0.654                 |
| <b>PWEG-FNFP loss</b>                                     | 0.6932 $\pm$ 0.2482                   | 0.693 $\pm$ 0.259                   | 0.659                 |
| <b>PWEG-FNFP loss + <math>\mathcal{L}_{Center}</math></b> | <b>0.7262 <math>\pm</math> 0.2326</b> | <b>0.720 <math>\pm</math> 0.240</b> | <b>0.693</b>          |

Paired Wilcoxon analysis showed that PWEG-FNFP loss with  $\mathcal{L}_{Center}$  significantly improved per-scan macro lesion-wise F1 over ICI after Holm correction ( $\Delta F1 = +0.044$ ,  $p_{Holm} = 0.016$ ), but not over BlobLoss ( $\Delta F1 = +0.017$ ,  $p_{Holm} = 0.428$ ). Since ICI already includes center-of-instance supervision, this suggests that the FN/FP terms provide complementary error-gated information beyond center localization.

Size-stratified analysis was limited to CE+Dice, PET+CT ICI Net (as the prior best lesion-wise model), and the strongest lesion-aware methods from Table 1, as conventional voxel-wise losses had lower lesion-wise F1.

### 3.6 Small-Lesion Analysis

Because the PWEG-FNFP loss was designed to improve lesion-level detection, we further analysed performance as a function of lesion size. Sensitivity was stratified by GT lesion volume, whereas precision was stratified by predicted component volume. Table 2 reports the size-stratified lesion-wise sensitivity and precision.

**Table 2.** Size-stratified lesion-wise performance on the held-out institutional cohort. Sensitivity is reported as TP/GT by GT lesion size, whereas precision is reported as TP/PC by predicted component size.

| Method                                     | Sens. <0.1<br>cm <sup>3</sup><br>(TP/GT) | Prec. <0.1<br>cm <sup>3</sup><br>(TP/PC) | Sens. 0.1–<br>1 cm <sup>3</sup><br>(TP/GT) | Prec. 0.1–<br>1 cm <sup>3</sup><br>(TP/PC) | Sens. 1–5<br>cm <sup>3</sup><br>(TP/GT) | Prec. 1–5<br>cm <sup>3</sup><br>(TP/PC) | Sens. >5<br>cm <sup>3</sup><br>(TP/GT) | Prec. >5<br>cm <sup>3</sup><br>(TP/PC) |
|--------------------------------------------|------------------------------------------|------------------------------------------|--------------------------------------------|--------------------------------------------|-----------------------------------------|-----------------------------------------|----------------------------------------|----------------------------------------|
| CE+Dice<br>baseline                        | 0.11<br>(37/331)                         | 0.33<br>(87/259)                         | 0.55<br>(259/467)                          | 0.72<br>(268/373)                          | 0.79<br>(203/255)                       | 0.77<br>(153/198)                       | 0.91<br>(133/146)                      | 0.85<br>(118/139)                      |
| PET+CT ICI<br>Net                          | 0.00<br>(0/331)                          | 0.21<br>(20/95)                          | 0.02<br>(8/467)                            | 0.63<br>(28/44)                            | 0.12<br>(33/255)                        | 0.68<br>(24/35)                         | 0.26<br>(39/146)                       | 0.53<br>(8/15)                         |
| BlobLoss                                   | 0.24<br>(81/331)                         | <b>0.39</b><br>(158/401)                 | 0.73<br>(339/467)                          | 0.75<br>(338/449)                          | 0.89<br>(228/255)                       | 0.81<br>(176/218)                       | 0.95<br>(138/146)                      | 0.84<br>(114/135)                      |
| ICI                                        | 0.35<br>(117/331)                        | 0.32<br>(172/537)                        | 0.79<br>(371/467)                          | 0.71<br>(371/522)                          | 0.93<br>(237/255)                       | 0.81<br>(195/241)                       | <b>0.97</b><br>(141/146)               | 0.86<br>(128/149)                      |
| PWEG-FNFP<br>loss                          | 0.27<br>(90/331)                         | 0.37<br>(125/340)                        | 0.72<br>(338/467)                          | 0.69<br>(328/474)                          | 0.88<br>(225/255)                       | 0.85<br>(196/231)                       | 0.88<br>(128/146)                      | <b>0.89</b><br>(132/149)               |
| PWEG-FNFP<br>loss + $\mathcal{L}_{Center}$ | <b>0.40</b><br>(132/331)                 | 0.34<br>(124/367)                        | <b>0.81</b><br>(377/467)                   | <b>0.75</b><br>(389/514)                   | <b>0.89</b><br>(229/255)                | <b>0.85</b><br>(222/260)                | 0.89<br>(130/146)                      | 0.84<br>(135/160)                      |

PWEG-FNFP loss +  $\mathcal{L}_{Center}$  achieved the highest sensitivity below 0.1 cm<sup>3</sup> and between 0.1 and 1 cm<sup>3</sup>, suggesting improved recovery of small and medium-sized lesions. Precision for predicted components smaller than 0.1 cm<sup>3</sup> remained limited across all methods, indicating a persistent FP burden among very small components. Overall, the results suggest that FN/FP error correction and center localization are complementary.

### 3.7 External Evaluation on AutoPET

For external evaluation, the strongest lesion-aware methods were tested on AutoPET, which was not used during training. Tables 3 and 4 report external Dice, lesion-wise F1, and size-stratified sensitivity/precision. The PWEG-FNFP loss achieved the highest Dice, while very small lesions and FP control remained challenging across methods.

**Table 3.** External evaluation on AutoPET

| Method                        | CE+Dice base-line   | PET+CT ICI Net      | BlobLoss            | ICI                 | PWEG-FNFP loss                        | PWEG-FNFP loss + $\mathcal{L}_{center}$ |
|-------------------------------|---------------------|---------------------|---------------------|---------------------|---------------------------------------|-----------------------------------------|
| Test Dice $\pm$ SD            | 0.3642 $\pm$ 0.2362 | 0.0454 $\pm$ 0.0989 | 0.3938 $\pm$ 0.2282 | 0.4068 $\pm$ 0.2338 | <b>0.4664 <math>\pm</math> 0.2290</b> | 0.3941 $\pm$ 0.2410                     |
| Macro Lesion-wise F1 $\pm$ SD | 0.378 $\pm$ 0.274   | 0.067 $\pm$ 0.097   | 0.391 $\pm$ 0.273   | 0.409 $\pm$ 0.282   | <b>0.4199 <math>\pm</math> 0.252</b>  | 0.4190 $\pm$ 0.254                      |
| Global Lesion-wise F1         | 0.404               | 0.070               | 0.388               | 0.394               | 0.397                                 | <b>0.407</b>                            |

**Table 4.** External AutoPET size-stratified lesion-wise sensitivity and precision. TP: True Positive, GT: Ground Truth, PC: Predicted Components

| Method                                  | Sens. <0.1 cm <sup>3</sup> (TP/GT) | Prec. <0.1 cm <sup>3</sup> (TP/PC) | Sens. 0.1–1 cm <sup>3</sup> (TP/GT) | Prec. 0.1–1 cm <sup>3</sup> (TP/PC) | Sens. 1–5 cm <sup>3</sup> (TP/GT) | Prec. 1–5 cm <sup>3</sup> (TP/PC) | Sens. >5 cm <sup>3</sup> (TP/GT) | Prec. >5 cm <sup>3</sup> (TP/PC) |
|-----------------------------------------|------------------------------------|------------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|-----------------------------------|----------------------------------|----------------------------------|
| CE+Dice baseline                        | 0.00<br>(0/197)                    | <b>0.24</b><br>(117/483)           | 0.30<br>(106/357)                   | <b>0.45</b><br>(205/451)            | 0.60<br>(178/299)                 | 0.49<br>(81/164)                  | 0.81<br>(180/223)                | <b>0.50</b><br>(61/123)          |
| PET+CT ICI Net                          | 0.00<br>(0/197)                    | 0.05<br>(19/356)                   | 0.00<br>(0/357)                     | 0.16<br>(18/113)                    | 0.04<br>(12/299)                  | 0.16<br>(11/69)                   | 0.21<br>(46/223)                 | 0.20<br>(10/49)                  |
| BlobLoss                                | 0.00<br>(0/197)                    | 0.19<br>(140/733)                  | 0.38<br>(137/357)                   | 0.44<br>(234/535)                   | 0.64<br>(190/299)                 | <b>0.49</b><br>(82/169)           | 0.82<br>(183/223)                | 0.48<br>(54/113)                 |
| ICI                                     | 0.00<br>(0/197)                    | 0.20<br>(183/936)                  | 0.48<br>(170/357)                   | 0.42<br>(244/581)                   | 0.71<br>(213/299)                 | 0.47<br>(81/171)                  | 0.84<br>(187/223)                | 0.49<br>(62/126)                 |
| PWEG-FNFP loss                          | 0.03<br>(5/197)                    | 0.18<br>(201/1131)                 | <b>0.62</b><br>(223/357)            | 0.42<br>(285/680)                   | <b>0.76</b><br>(226/299)          | 0.44<br>(88/200)                  | <b>0.86</b><br>(191/223)         | 0.44<br>(71/163)                 |
| PWEG-FNFP loss + $\mathcal{L}_{center}$ | <b>0.03</b><br>(5/197)             | 0.22<br>(241/1090)                 | 0.59<br>(212/357)                   | 0.41<br>(250/615)                   | 0.75<br>(224/299)                 | 0.42<br>(77/183)                  | 0.84<br>(188/223)                | 0.48<br>(61/128)                 |

On AutoPET, the PWEG-FNFP loss achieved the highest Dice among lesion-aware methods, while the PWEG-FNFP loss +  $\mathcal{L}_{center}$  achieved the highest global lesion-wise F1-score. Performance remained lower than on the institutional cohort, indicating sensitivity to domain shift across datasets with different lesion distributions, acquisition protocols, and annotation characteristics.

## 4 Discussion and Conclusion

This study shows that the PWEG-FNFP loss is not simply another lesion-aware re-weighting term, but an error-gated component objective that changes where optimization pressure is applied. The FN term increases foreground support inside missed GT components, while the FP term suppresses unsupported predicted components. Because probabilities are averaged within each selected component, the added correction is component-normalized rather than dominated by lesion volume. This explains why the standalone loss improves over CE+Dice, while the best result with  $\mathcal{L}_{center}$  indicates complementarity with lesion-center localization.

The main limitation is that component selection remains threshold-based. Thus, the loss is differentiable only within the currently selected FN/FP sets and can change when components appear, disappear, split, or merge. This likely contributes to the remaining

sensitivity–precision trade-off, especially for very small predicted components. External AutoPET evaluation further shows that lesion-level optimization does not remove domain shift. Future work should investigate probabilistic component selection, threshold robustness, and stronger FP suppression.

**Acknowledgments.** This work was supported by Université Côte d’Azur and Centre Antoine Lacassagne. The authors thank the clinical teams at all participating sites for data collection and annotation.

## References

1. Boellaard, R., et al.: FDG PET/CT: EANM procedure guidelines for tumor imaging: Version 2.0. *Eur. J. Nucl. Med. Mol. Imaging* 42(2), 328–354 (2015). <https://doi.org/10.1007/s00259-014-2961-x>
2. Gatidis, S., Früh, M., Fabritius, M.P., et al.: Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging. *Nature Machine Intelligence* 6, 1396–1405 (2024).
3. Guo, W., et al.: Volumetric parameters of the primary tumor and whole-body tumor burden derived from baseline  $^{18}\text{F}$ -FDG PET/CT can predict overall survival in non-small cell lung cancer patients: initial results from a single institution. *EJNMMI Rep.* 6, 28 (2022). <https://doi.org/10.1186/s41824-022-00158-x>
4. Andrearczyk, V., Oreiller, V., Abobakr, M., et al.: Overview of the HECKTOR Challenge at MICCAI 2022: Automatic head and neck tumor segmentation and outcome prediction in PET/CT. In: *Head and Neck Tumor Segmentation and Outcome Prediction, LNCS*. Springer (2023).
5. 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, 203–211 (2021).
6. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: *International Conference on 3D Vision*, pp. 565–571 (2016).
7. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *DLMIA/ML-CDS, LNCS*, pp. 240–248. Springer (2017).
8. Shirokikh, B., Shevtsov, A., Kurmukov, A., et al.: Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation. In: *MICCAI, LNCS*, pp. 523–532. Springer (2020).
9. Kofler, F., Shit, S., Ezhov, I., et al.: Blob loss: Instance imbalance aware loss functions for semantic segmentation. In: *IPMI, LNCS*, pp. 755–767. Springer (2023).
10. Rachmadi, F., Poon, C., Skibbe, H.: Improving segmentation of objects with varying sizes in biomedical images using instance-wise and center-of-instance segmentation loss function. In: *Medical Imaging with Deep Learning, PMLR* 227, pp. 286–300 (2024).
11. Rachmadi, M.F., Byra, M., Skibbe, H.: A new family of instance-level loss functions for improving instance-level segmentation and detection of white matter hyperintensities in routine

clinical brain MRI. *Comput. Biol. Med.* 174, 108414 (2024). <https://doi.org/10.1016/j.compbio-med.2024.108414>