Response letter for the manuscript titled '**3D Lymphoma Segmentation on PET/CT Images via Multi-Scale Information Fusion with Cross-Attention**'

Dear Editor and Reviewers,

We are grateful for the comments and suggestions from the editors and the reviewers, which are crucial for improving our work. We have revised the manuscript to address the reviewers' comments fully. Our point-by-point reply to the review comments is summarized below. In this document, the original reviewers' comments are in **black**; our responses are in **blue**; the quotations in the revised manuscript are in **red**.

Reviewer #1:

General Comments:

1. This research paper proposes a new deep learning-based method for segmenting DLBCL lesions in PET/CT images. The method uses a dual-branch encoder with shifted window transformers and a MSIF module to integrate features from both PET and CT modalities. The MSIF module uses cross-attention mechanisms to enhance the interaction between features at different scales, improving segmentation accuracy. The study also examines the impact of various network parameters on segmentation performance and assesses the method's ability to calculate the TMTV. The authors conclude that their proposed method outperforms existing methods in terms of segmentation accuracy and TMTV calculation, offering a promising tool for lymphoma diagnosis and treatment.

Reply: Thank you for your detailed and positive feedback. We greatly appreciate your recognition of our work and the potential clinical value it offers for lymphoma diagnosis and treatment. Your insightful comments are highly encouraging and affirm the significance of our study. Thank you again for your unique perspective and thoughtful review.

2. The study used a dataset of 165 PET/CT scans from patients clinically diagnosed with DLBCL, provided by Peking University People's Hospital, using the Discovery VCT PET/CT scanner (GE Healthcare, Milwaukee, Wisconsin, USA). These datasets were used for both training and testing, employing a 5-fold cross-validation approach. Based on this, the study is a single-center study, which affects the generalizability of the segmentation model. It is highly recommended to include publicly available datasets, such as autoPET (which contains more than 100 lymphoma cases with manual segmentation), to assess the model's performance. Otherwise, the metrics provided by the authors may not be valid for general application of the proposed segmentation model for lymphoma. Further validation on a larger, more diverse dataset is needed to demonstrate the robustness and clinical applicability of the model.

Reply: Thank you for your insightful suggestion. We fully agree with your point regarding the need to validate the model's generalizability on publicly available datasets. To address this, we conducted additional experiments using the lymphoma cases in the autoPET dataset. Specifically, we trained and tested our method on the autoPET dataset following the same 5-fold cross-validation protocol as used for our private dataset. The results from these experiments confirm the robustness and generalizability of our method across diverse datasets. These findings have been incorporated into the revised manuscript, specifically in the [Abstract, Section 2.1, Section 2.2, Section 2.5, Section 3.1, and Section 3.2]. We sincerely appreciate your suggestion, which has significantly enhanced the rigor and applicability of our study.

(Abstract)

Results: The model was trained and validated on a private dataset of 165 DLBCL patients and a publicly available dataset (autoPET) containing 145 PET/CT scans of lymphoma patients. Both datasets were analyzed using 5-fold cross-validation. On the private dataset, our model achieved a DSC of 0.7512, sensitivity of 0.7548, precision of 0.7611, an Average Surface Distance (ASD) of 3.61 mm, and a Hausdorff Distance at the 95th percentile (HD95) of 15.25 mm. On the autoPET dataset, the model achieved a DSC of 0.7441, sensitivity of 0.7573, precision of 0.7427, ASD of 5.83 mm, and HD95 of 21.27 mm, outperforming state-of-the-art methods ($p < 0.05$, t-test). For TMTV quantification, Pearson correlation coefficients of 0.91 (private dataset) and 0.86 (autoPET) were observed, with R^2 values of 0.89 and 0.75, respectively. Extensive ablation studies demonstrated the MSIF module's contribution to enhanced segmentation accuracy.

(Section 2.1 Dataset)

This study utilized two datasets: (1) our private dataset comprising 165 PET/CT scan datasets from patients clinically diagnosed with DLBCL, provided by Peking University People's Hospital, and (2) the FDG-PET/CT dataset (autoPET), comprising 145 PET/CT scans of lymphoma patients with manually annotated tumor lesions, obtained from The Cancer Imaging Archive (TCIA). The use of the autoPET dataset in this study complies with TCIA's data usage policy and is authorized for research purposes[23].

(Section 2.2 Data preprocessing)

For the autoPET dataset, we followed the official preprocessing steps provided by TCIA. These steps included rigid-body registration, resampling consistent voxel spacing, and intensity normalization. To ensure compatibility with our network, we further applied center cropping to obtain slices of 224×224 pixels.

(Section 2.5 Implementation and experiments)

We evaluated the effectiveness of our method on both the private dataset and the autoPET dataset, comparing its performance with various state-of-the-art (SOTA) methods, including UnetR[32], Swin-UnetR[21], Att-Unet[33], Unet++[34], SegResNet[35], and SwinCross[22]. Consistent data splitting was ensured for all methods by employing the same 5-fold cross-validation approach. In each fold, the training set consisted of 60% of the data, while the validation and test set each accounted for 20%.
To ensure fairness, all experiments were conducted within the same computational environment, using identical hardware and software configurations. A sliding window technique was employed to reduce GPU memory consumption, extracting 32 consecutive slices per batch to form a 3D volume. Hyperparameter optimization, including adjustments to learning rates, batch sizes, and optimizer settings, was performed for each method based on validation set performance.

(Section 3.1 Results of segmentation)

Table 1: Results of different methods on the private dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.7107±0.0178 **	0.7608±0.0128	0.6686±0.0298 **	4.10±0.20 **	18.05±2.36
SegResNet	0.7223±0.0146 *	0.7175±0.0466	0.7289±0.0125 **	4.61±0.26 **	21.01±0.69 **
Swin-UnetR	0.7271±0.0163 *	**0.7659±0.0123**	0.7041±0.0246 **	3.92±0.22 *	15.74±0.98
SwinCross	0.7414±0.0209	0.7405±0.0213	0.7432±0.0176	4.04±0.22 **	16.82±1.51
Unet++	0.7446±0.0129	0.7322±0.0072 **	0.7577±0.0137	4.21±0.09 **	18.05±1.51 **
Att-Unet	0.7463±0.0113	0.7622±0.0075	0.7314±0.0179 *	4.75±0.04 **	17.16±2.26
Ours	**0.7512±0.0078**	0.7548±0.0063	**0.7611±0.0078**	**3.61±0.11**	**15.20±0.78**

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates $p < 0.05$, and ** indicates $p < 0.01$ when compared to our method.

Table 1 presents the segmentation performance of various methods on our private dataset. Our model achieves the highest DSC (0.7512) and precision (0.7611), demonstrating significant advantages in overall performance and false positive reduction. Although our sensitivity score (0.7548) was slightly lower than that of Swin-UnetR (0.7659), it remains competitive, indicating that our approach effectively balances multiple metrics for accurate and reliable tumor segmentation. Furthermore, our method achieved the lowest ASD (3.61 mm) and HD95 (15.20 mm), confirming its effectiveness in capturing accurate tumor boundaries.

Table 2: Results of different methods on the autoPET dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.6865±0.0478 **	0. 6924±0.0812	0.6851±0.0404 *	6.65±0.73	22.38±1.95 *
SegResNet	0.6740±0.0412 *	0.6951±0.0627	0.6483±0.0505 *	6.12±1.02	21.26±1.29
Swin-UnetR	0.7282±0.0605	0.7311±0.0833	0.7274±0.0450	5.40±0.92	**19.08±2.63**
SwinCross	0.7267±0.0146 **	0.7382±0.0717 *	0.7233±0.0525	6.40±1.48	23.37±2.95
Unet++	0.7302±0.0192	0.7424±0.0818	0.7277±0.0523	**5.11±0.92**	19.92±1.59
Att-Unet	0.6941±0.0261 **	0.7016±0.0657	0.6917±0.0401 **	6.17±1.04	21.29±1.25
Ours	**0.7441±0.0241**	**0.7573±0.0874**	**0.7427±0.0647**	5.83±1.18	21.27±1.44

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates $p < 0.05$, and ** indicates $p < 0.01$ when compared to our method.

To validate the generalizability of our method, we further evaluated its performance on the autoPET dataset. Table 2 presents the results of various methods on this publicly available dataset. Our method achieved the highest DSC (0.7441), sensitivity (0.7573), and precision (0.7427), demonstrating robust segmentation accuracy and reliability across datasets. While the ASD (5.83 mm) and HD95 (21.27 mm) of our method were comparable to other approaches, they did not exhibit a significant advantage. This indicates that while our model exceled in capturing lesion characteristics and reducing false positives, further refinement may be required to enhance boundary delineation accuracy, particularly in datasets with greater variability.

To evaluate the stability of our model under different conditions, we employed the box plots to display the distribution of DSC, sensitivity, precision, ASD and HD95 across five-fold cross-validation on both the private and autoPET datasets. Fig. 4 (a) shows the results on the private dataset and autoPET dataset. In both cases, our method demonstrates a more concentrated distribution with less variability compared to other methods, indicating higher stability across experiments.

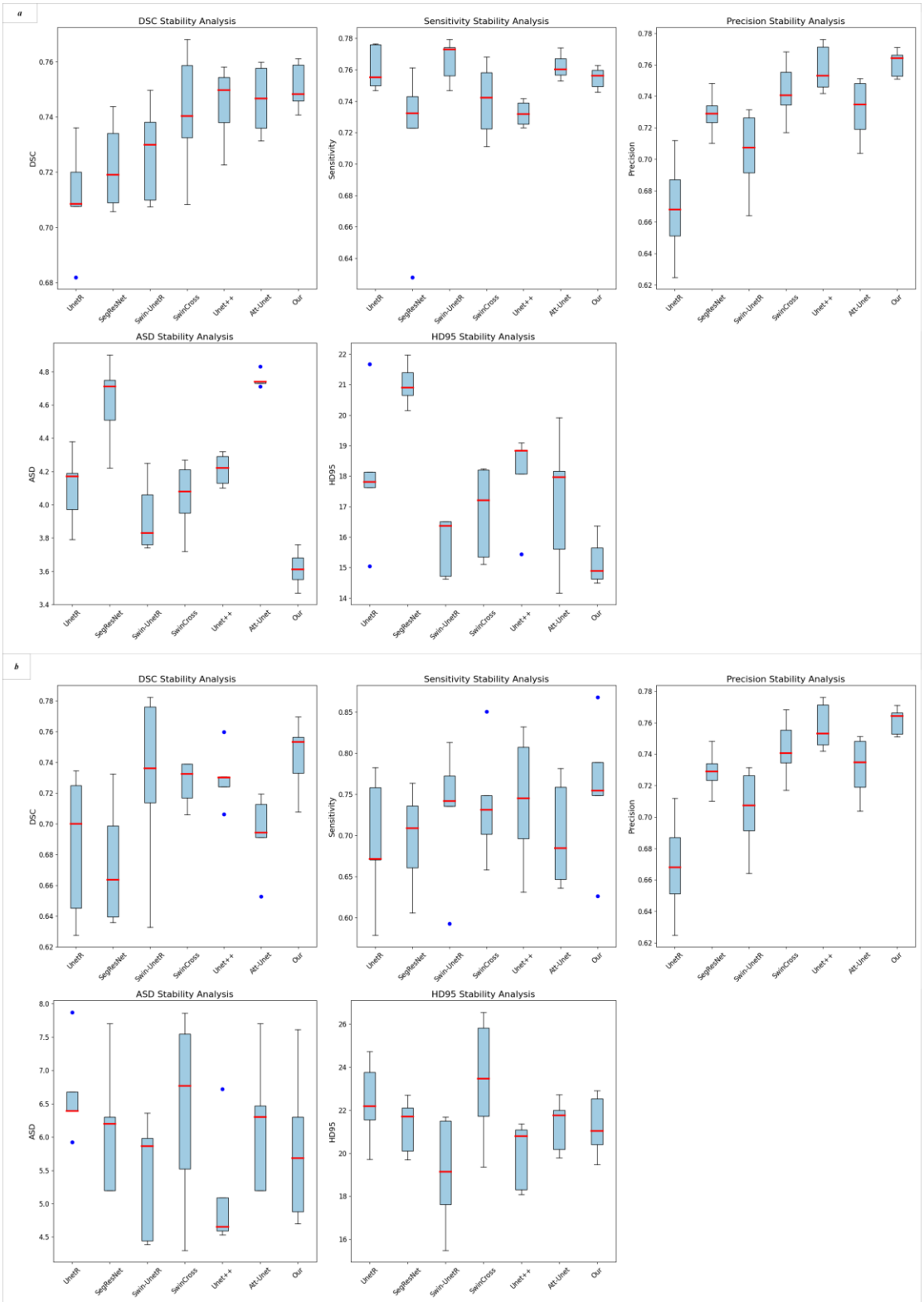


Fig. 4. Stability analysis using box plots for the private and autoPET datasets: This Fig. presents the ranges of DSC, sensitivity, precision, ASD and HD95 across five cross-validation folds for different models. Subfigure (a) displays the results on the private dataset, while subfigure (b) shows the corresponding results on the autoPET dataset. The box shows the 1st quartile (lower boundary), median (red line), and 3rd quartile (upper boundary). The whiskers represent the range of data, excluding outliers, which are marked as blue dots.

To provide a comprehensive comparison of different methods, we visualized the segmentation results on both the private dataset and the autoPET dataset. Fig. 5 and Fig. 6 show the difference maps generated by each method, highlighting true positive (green), false negative (red), and false positive (blue) regions. For the private dataset (shown in Fig. 5), our method demonstrated superior accuracy, particularly in smaller lesion regions and areas with complex shapes or blurred edges. Similarly, on the autoPET dataset (shown in Fig. 6), our method consistently reproduced the ground truth with higher precision, confirming its robustness and effectiveness on a public dataset.

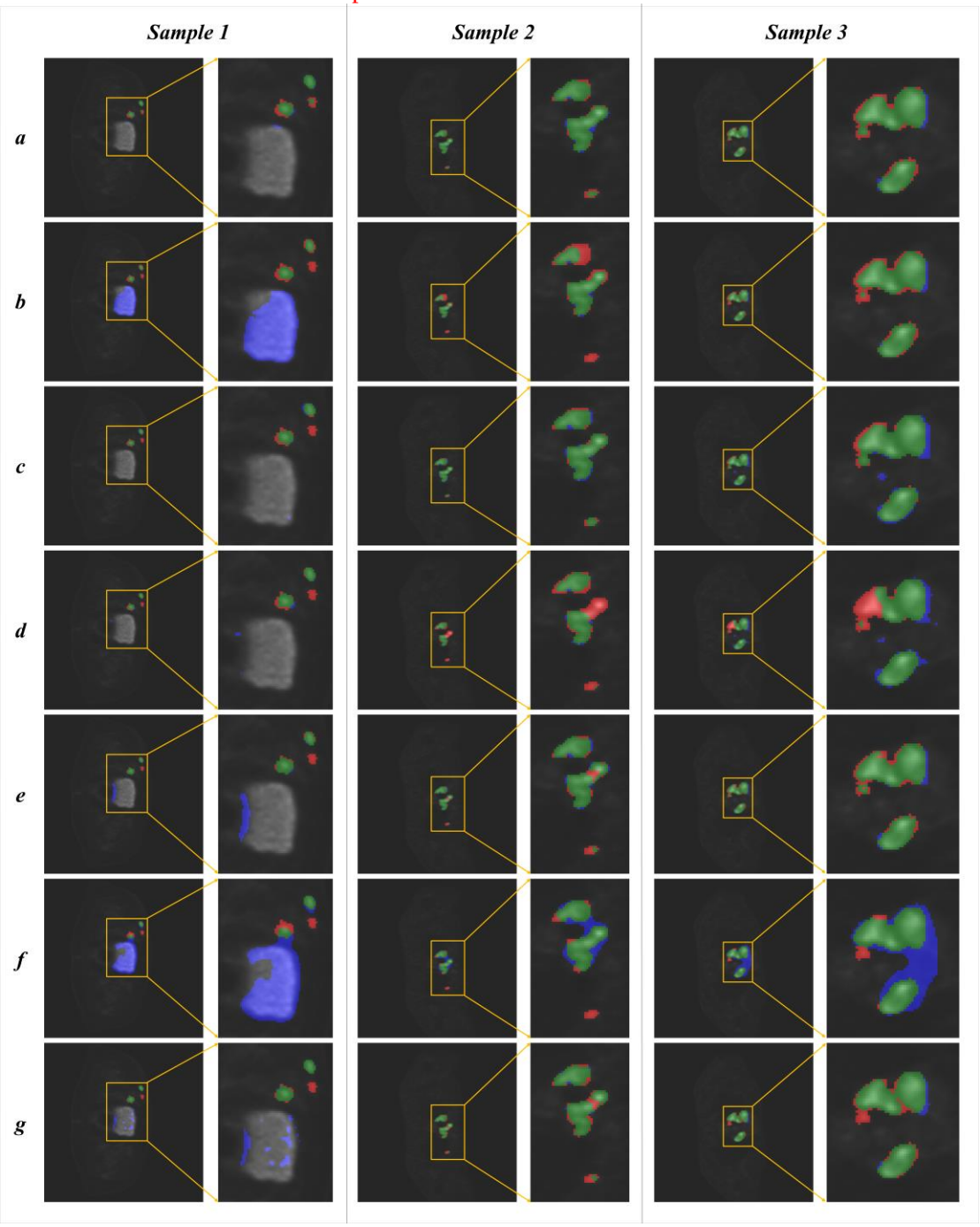


Fig. 5. Difference maps of segmentation results compared with ground truth for private datasets. The green, red, and blue regions represent true positive, false negative, and false positive pixels, respectively. Subfigures (***a***)–(***g***) show results generated by our method, Att-Unet, Unet++, SwinCross, Swin-UnetR, SegResNet, and UnetR, respectively.

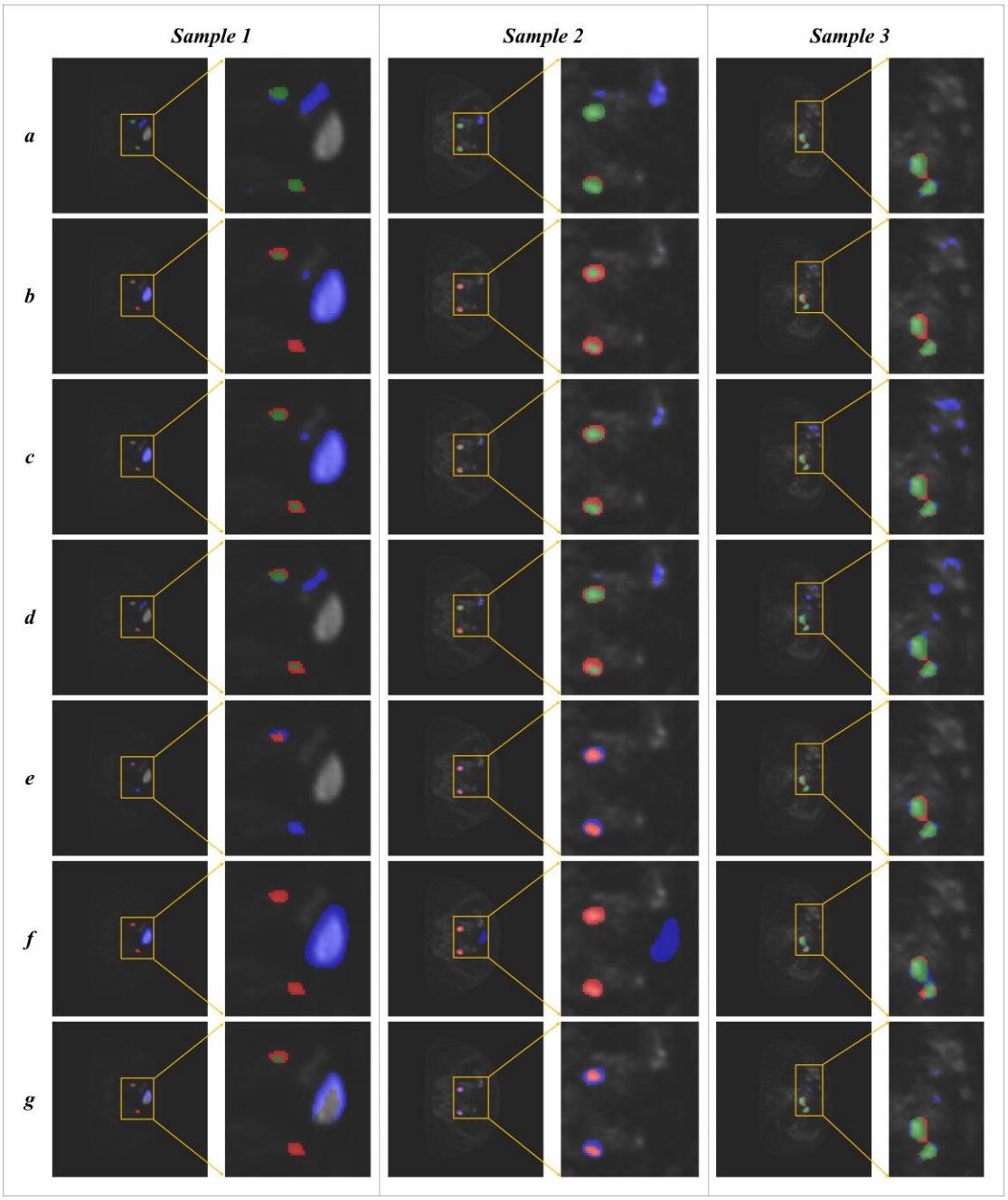
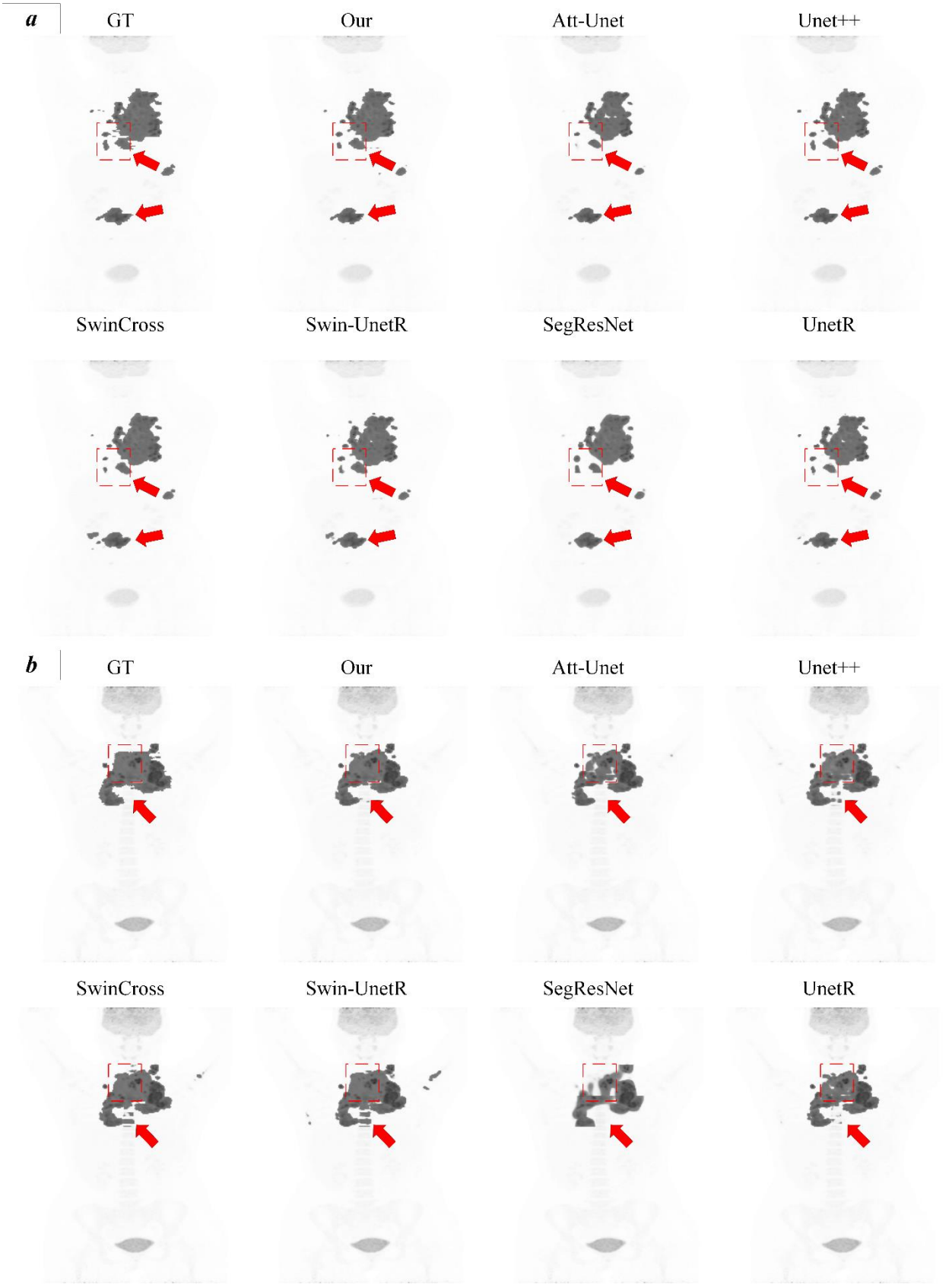


Fig. 6. Difference maps of segmentation results compared with ground truth for autoPET datasets. The green, red, and blue regions represent true positive, false negative, and false positive pixels, respectively. Subfigures (***a***)–(***g***) show results generated by our method, Att-Unet, Unet++, SwinCross, Swin-UnetR, SegResNet, and UnetR, respectively.

Fig. 7 presents the segmentation results visualized on the maximum intensity projection of the PET images for whole-body lymphoma cases. On both the private dataset (Fig. 7 (a)) and the autoPET dataset (Fig. 7 (b)), our method achieved superior lesion delineation compared to other methods. The improved performance was particularly noticeable in regions with irregular boundaries, emphasizing the generalizability of our approach across datasets.


a
GT
Our
Att-Unet
Unet++
SwinCross
Swin-UnetR
SegResNet
UnetR
b
GT
Our
Att-Unet
Unet++
SwinCross
Swin-UnetR
SegResNet
UnetR

140

Fig. 7. Segmentation results visualized on the maximum intensity projection of PET images for the private (a) and the autoPET (b) datasets. Ground truth (GT) masks and predicted segmentation masks are overlaid on the maximum intensity projection of PET images.

(Section 3.2 Results of TMTV)

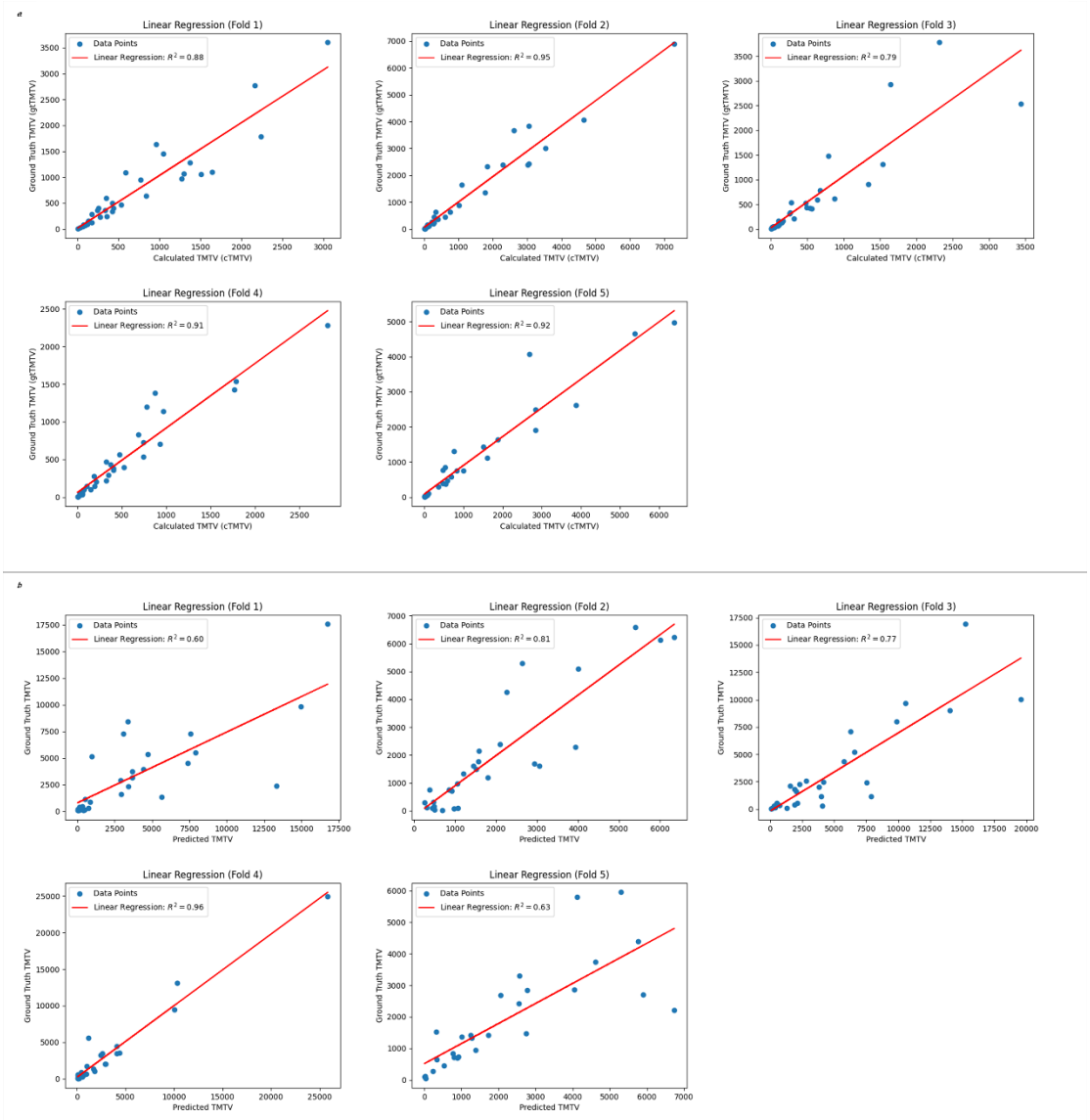


Fig. 8. Linear regression results of the predicted TMTV as cTMTV vs. the ground truth of TMTV as gtTMTV on: (a) private dataset and (b) autoPET dataset. The red line represents the linear regression fit, with the R^2 indicating the goodness of fit.

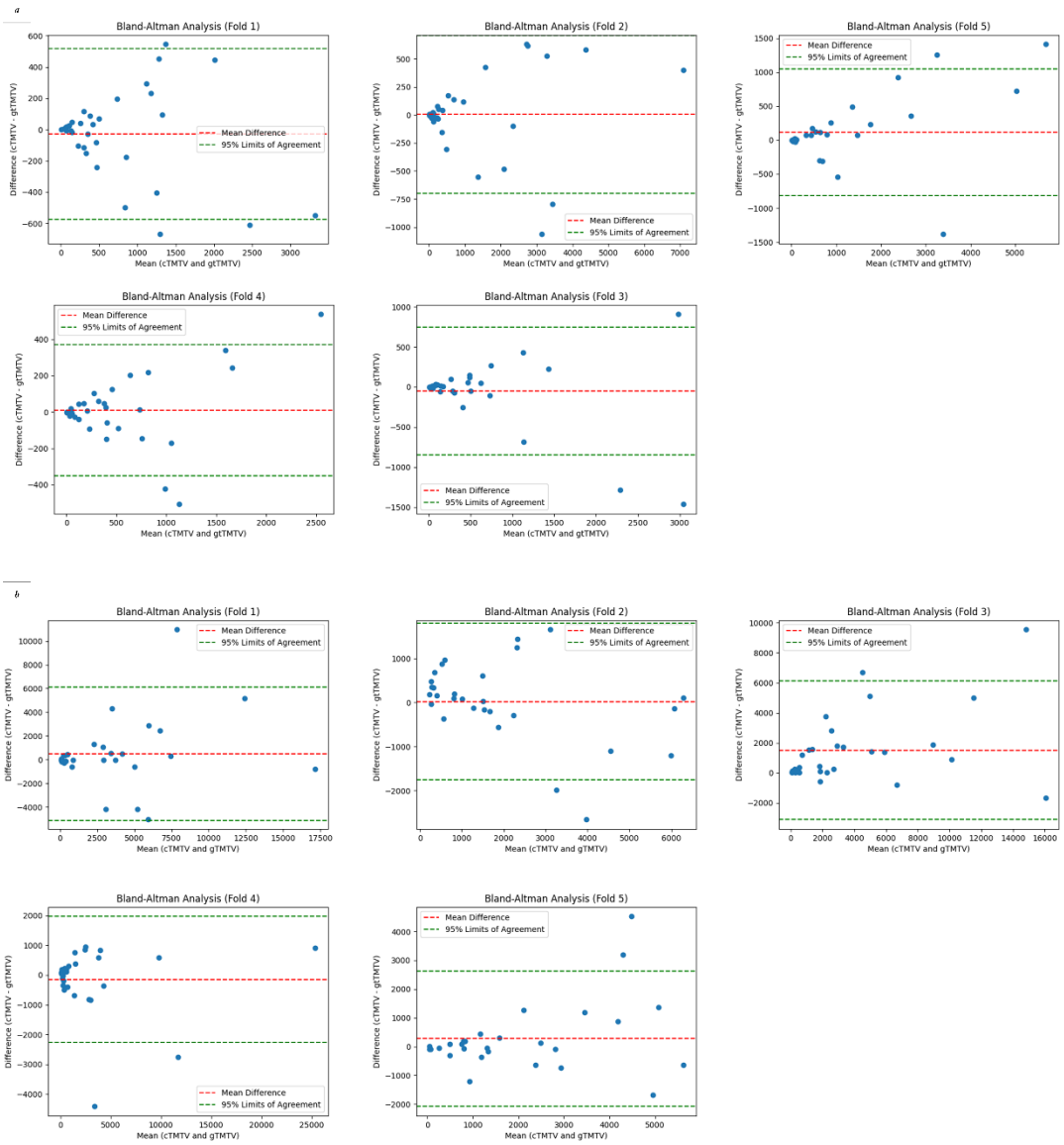


Fig. 9. Bland-Altman analysis for cTMTV vs. gtTMTV on (a) our private dataset and (b) autoPET dataset. The horizontal axis represents the mean of cTMTV and gtTMTV, while the vertical axis represents their difference. The red dashed line shows the mean difference, and the green dashed lines represent the 95% limits of agreement, calculated as the mean difference ± 1.96 standard deviations of the differences.

We evaluated TMTV predictions on both the private dataset and the autoPET dataset using five-fold cross-validation. For the private dataset, the mean cTMTV was 802.69 ± 1192.95 mL, and the mean gtTMTV was 792.36 ± 1133.78 mL, with a mean difference of 10.33 ± 360.03 mL. Similarly, on the autoPET dataset, the mean cTMTV was 3074.83 ± 4080.80 mL, and the mean gtTMTV was 2637.70 ± 3636.79 mL, resulting in a mean difference of 437.12 ± 1961.38 mL. To further quantify prediction accuracy, we calculated the MAE and MRE. For the private dataset, the MAE was 123.42 ± 61.84 mL, with an MRE of 15.75% ± 2.00%. Meanwhile, on the autoPET dataset, the MAE was 1069.17 ± 1699.52 mL, and the MRE reached 173.91 % ± 741.42 %. Although the MRE on the autoPET dataset appeared high, this was driven by a few extreme cases involving small ground-truth volumes. Overall, these results indicated that our model achieved robust performance across both datasets, maintaining a relatively low absolute error compared to the total volume in most cases.

Fig. 8 illustrates the linear regression analysis for both datasets. On the private dataset, the R^2 ranged from 0.79 to 0.95 across folds, with lower values potentially influenced by outliers in specific folds. In contrast, the autoPET dataset showed a broader span of R^2 (0.60-0.96), indicating that while some folds achieved strong linear correlations, others exhibited weaker fits—likely reflecting greater variability or differences in data collection and labeling. Nonetheless, these findings suggest that, overall, our model captures an appreciable linear relationship between cTMTV and gtTMTV in both datasets.

Fig. 9 shows the Bland-Altman analysis for both datasets. Most differences fell within the acceptable range, indicating general alignment with true TMTV values, though a few outliers were noted. These discrepancies stemmed from patient-specific variations, such as tumor irregularities or low contrast in certain PET regions, which challenged the boundary detection. Additionally, noise in PET images, particularly in low-activity regions, influenced the measurement accuracy. The analysis revealed a trend of increasing discrepancies with higher TMTV values, which was attributed to the limited number of cases, reducing the generalizability for larger tumor volumes and increases variability in predictions.

3. The discussion section lacks a comparison of ease of use with other existing techniques for TMTV segmentation and quantification. Addressing this would provide a more comprehensive assessment of the proposed method's practicality.

Reply: Thank you for your insightful comment. In response, we have revised [Section 4.2] in the discussion to include a detailed comparison of the ease of use and practicality of our method relative to other state-of-the-art techniques for TMTV segmentation and quantification.

(Section 4.2 Comparison with Existing TMTV Calculation Methods)

Compared to existing techniques for TMTV segmentation and quantification, our method offers clear advantages in practicality, automation, and accuracy. While Yousefirizi et al.[37,38] and Blanc-Durand et al.[14] evaluated their approaches on single-center or multi-center datasets, the validation on publicly available datasets was not conducted, limiting the generalizability of their methods. In contrast, our approach was rigorously evaluated on both a private dataset and the autoPET dataset, demonstrating consistent performance across varied imaging protocols. This dual validation underscores the robustness of our method in diverse clinical settings.

In terms of accuracy, Yousefirizi et al.[38] achieved a Pearson correlation coefficient of R^2=0.83 for TMTV quantification but reported lower segmentation accuracy (DSC=0.68) due to single-modality constraints. Similarly, Blanc-Durand et al.[14] achieved a DSC of 0.73 but faced significant TMTV underestimation on the external validation dataset. By leveraging multi-scale and cross-modal feature fusion, our method effectively integrates PET and CT information, achieving superior segmentation accuracy with DSC values of 0.7512 on the private dataset and 0.7441 on the autoPET dataset.

Furthermore, the semi-automated workflows proposed by Burggraaff et al.[39] required manual threshold adjustments, leading to time-consuming processes prone to variability. In contrast, our fully automated pipeline eliminates the need for manual intervention, ensuring consistent, reproducible results while significantly reducing analysis time. These features make our method particularly suitable for routine clinical applications.

In conclusion, our method combines automation, accuracy, and generalizability to provide a practical and efficient solution for TMTV segmentation and quantification, supporting both research and clinical workflows.

Specific Comments:

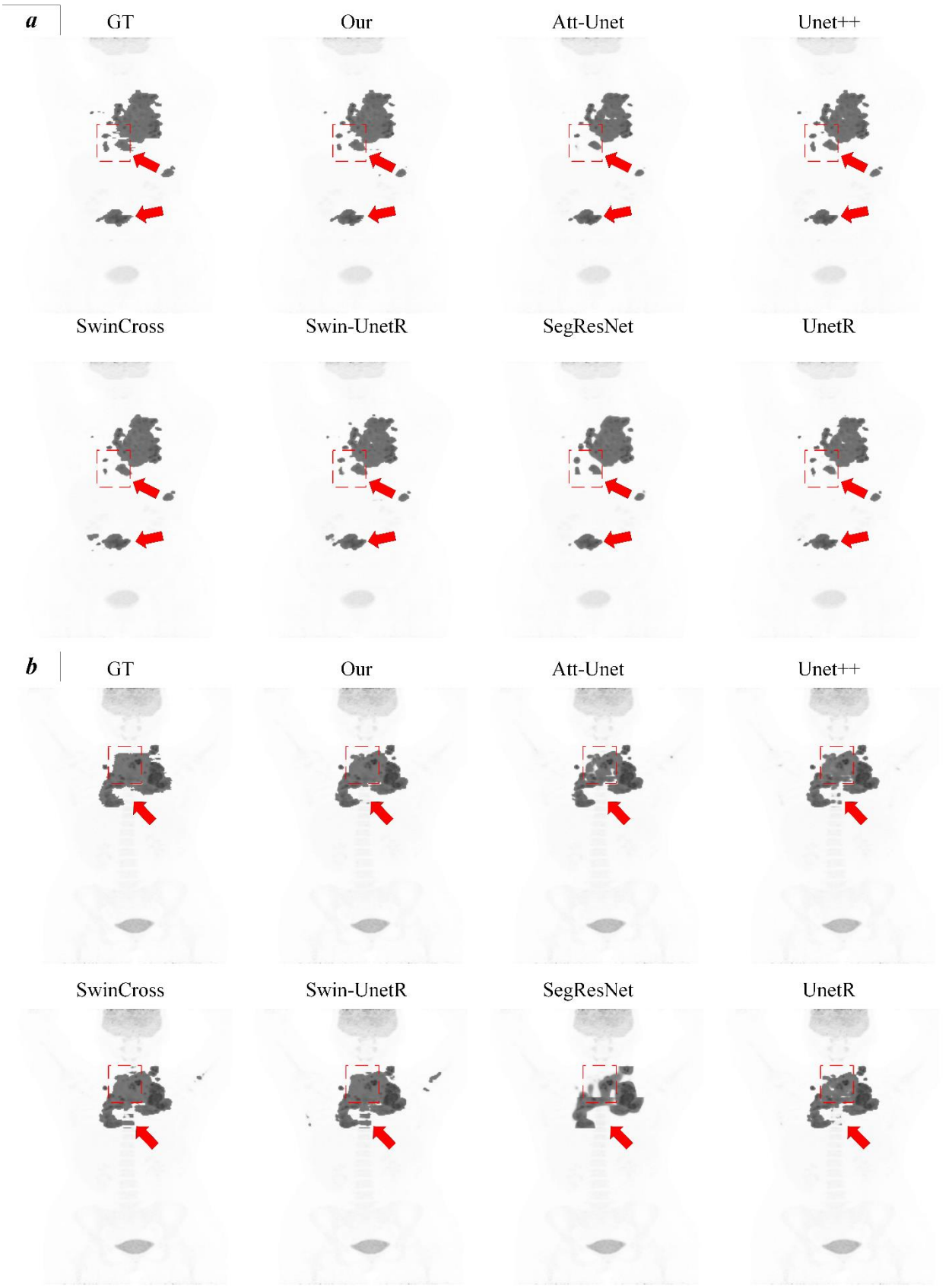
1. The paper could benefit from a Fig. in the results section showing the segmentation ground truth mask and the predicted mask by the model, overlaid on maximum intensity images, to provide a visual inspection

measure. Since lymphoma lesions can appear anywhere in the body, a whole-body Fig. comparison would be beneficial.

Reply: Thank you for your valuable suggestion. We completely agree that visualizing segmentation results is essential for providing a clear inspection measure. In response, we have added a new Fig. in [Section 3.1], showing the ground truth masks and the predicted masks overlaid on maximum intensity projections (MIPs) of PET images.

(Section 3.1 Results of segmentation)

Fig. 7 presents the segmentation results visualized on the maximum intensity projection of the PET images for whole-body lymphoma cases. On both the private dataset (Fig. 7 (a)) and the autoPET dataset (Fig. 7 (b)), our method achieved superior lesion delineation compared to other methods. The improved performance was particularly noticeable in regions with irregular boundaries, emphasizing the generalizability of our approach across datasets.


a
GT
Our
Att-Unet
Unet++
SwinCross
Swin-UnetR
SegResNet
UnetR
b
GT
Our
Att-Unet
Unet++
SwinCross
Swin-UnetR
SegResNet
UnetR

227

Fig. 7. Segmentation results visualized on the maximum intensity projection of PET images for the private (a) and the autoPET (b) datasets. Ground truth (GT) masks and predicted segmentation masks are overlaid on the maximum intensity projection of PET images.

2. The authors mentioned: "Rigid-body registration was used to align the PET and CT volumes into the same coordinate space, a standard practice in PET/CT segmentation. The PET images were upsampled to a target size of 256×256 using bicubic interpolation, while the CT images were downsampled to the same size." Upsampling and downsampling can introduce errors. It is unclear why the authors did not downsample the CT images to match the PET size, instead of upsampling PET and downsampling CT. Clarifying this choice would strengthen the methodology.

Reply: To address your concern, we have revised [Section 2.2] to clarify our choice of upsampling PET images to 256×256 and downsampling CT images to the same resolution. This approach was selected to preserve the anatomical details critical for segmentation accuracy, which would otherwise be significantly degraded if CT images were directly downsampled to the PET resolution of 128×128.

Additionally, we conducted experiments to compare segmentation performance under different resolutions for PET and CT images. As shown in table RL1, our method achieved significantly better results at a resolution of 256×256 for both PET and CT images. These results confirmed the importance of preserving anatomical details through our chosen preprocessing strategy.

Table RL1: Segmentation performance under different resolutions of PET and CT images

CT Resolution	PET Resolution	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
256×256	256×256	**0.7512±0.0078**	**0.7548±0.0063**	**0.7611±0.0078**	**3.61±0.11**	**15.20±0.78**
128×128	128×128	0.7135±0.0225	0.7316±0.0562	0.7219±0.0211	4.71±0.42	17.79±1.12

Our method achieved significantly better results (e.g., higher DSC and lower ASD) at a resolution of 256×256 for both PET and CT images. These results confirm the importance of preserving detail through our chosen preprocessing strategy.

(Section 2.2 Data preprocessing)

To ensure optimal detail preservation, we upsampled the PET images to a target resolution of 256×256 using bicubic interpolation, while downsampling the CT images to the same resolution. This approach was chosen to avoid the substantial loss of anatomical detail essential for segmentation accuracy which would occur if the CT images were downsampled directly to 128×128, the original PET resolution. Subsequently, all slices were cropped to 224×224 pixels, removing peripheral regions irrelevant to segmentation.

3. In addition to Dice, sensitivity, and specificity, the paper would benefit from using additional performance metrics, such as the Hausdorff Distance at the 95th percentile or the Normalized Surface Distance, to better evaluate segmentation quality.

Reply: Thank you for your valuable suggestion. We completely agree that incorporating additional performance metrics can provide a more comprehensive evaluation of segmentation quality. In response, we have added the Hausdorff Distance at the 95th percentile (HD95) and the Average Surface Distance (ASD) as supplementary evaluation metrics in the results section. The corresponding results and discussions have been included in [Section 2.4, Section 3.1 and Section 4.1] of the revised manuscript.

(Section 2.4 Segmentation evaluation criteria)

Average Surface Distance (ASD): ASD calculates the average distance between the predicted and ground truth boundaries, providing a measure of segmentation accuracy in terms of surface alignment. The normalized ASD is defined in Eq. 18:

$$ASD(A,B) = \frac{1}{|A| + |B|}\left(\sum_{a \in A} min_{b \in B} d(a,b) + \sum_{b \in B} min_{a \in A} d(b,a)\right) \quad (18)$$

where $d(a,b)$ represents the distance from point $a \in A$ to the nearest point $b \in B$, and $|A|$ and $|B|$ are the number of points in sets A and B, respectively.

Hausdorff Distance at the 95th percentile (HD95): HD95 measures the maximum distance between the predicted and ground truth segmentation boundaries, taking the 95th percentile of all such distances to avoid extreme outliers, as shown in Eq. 19:

$$HD95(A,B) = max\big(percentile(d(A,B), 95), percentile(d(B,A), 95)\big) \quad (19)$$

where $d(A,B)$ represents the distance from each point in set A to its nearest point in set B, and $percentile(d(A,B), 95)$ denotes the 95th percentile of these distances.

(Section 3.1 Results of segmentation)

Table 1: Results of different methods on the private dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.7107±0.0178 **	0.7608±0.0128	0.6686±0.0298 **	4.10±0.20 **	18.05±2.36
SegResNet	0.7223±0.0146 *	0.7175±0.0466	0.7289±0.0125 **	4.61±0.26 **	21.01±0.69 **
Swin-UnetR	0.7271±0.0163 *	**0.7659±0.0123**	0.7041±0.0246 **	3.92±0.22 *	15.74±0.98
SwinCross	0.7414±0.0209	0.7405±0.0213	0.7432±0.0176	4.04±0.22 **	16.82±1.51
Unet++	0.7446±0.0129	0.7322±0.0072 **	0.7577±0.0137	4.21±0.09 **	18.05±1.51 **
Att-Unet	0.7463±0.0113	0.7622±0.0075	0.7314±0.0179 *	4.75±0.04 **	17.16±2.26
Ours	**0.7512±0.0078**	0.7548±0.0063	**0.7611±0.0078**	**3.61±0.11**	**15.20±0.78**

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates p < 0.05, and ** indicates p < 0.01 when compared to our method.

Table 1 presents the segmentation performance of various methods on our private dataset. Our model achieves the highest DSC (0.7512) and precision (0.7611), demonstrating significant advantages in overall performance and false positive reduction. Although our sensitivity score (0.7548) was slightly lower than that of Swin-UnetR (0.7659), it remains competitive, indicating that our approach effectively balances multiple metrics for accurate and reliable tumor segmentation. Furthermore, our method achieved the lowest ASD (3.61 mm) and HD95 (15.20 mm), confirming its effectiveness in capturing accurate tumor boundaries.

Table 2: Results of different methods on the autoPET dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.6865±0.0478 **	0. 6924±0.0812	0.6851±0.0404 *	6.65±0.73	22.38±1.95 *
SegResNet	0.6740±0.0412 *	0.6951±0.0627	0.6483±0.0505 *	6.12±1.02	21.26±1.29
Swin-UnetR	0.7282±0.0605	0.7311±0.0833	0.7274±0.0450	5.40±0.92	**19.08±2.63**
SwinCross	0.7267±0.0146 **	0.7382±0.0717 *	0.7233±0.0525	6.40±1.48	23.37±2.95
Unet++	0.7302±0.0192	0.7424±0.0818	0.7277±0.0523	**5.11±0.92**	19.92±1.59
Att-Unet	0.6941±0.0261 **	0.7016±0.0657	0.6917±0.0401 **	6.17±1.04	21.29±1.25
Ours	**0.7441±0.0241**	**0.7573±0.0874**	**0.7427±0.0647**	5.83±1.18	21.27±1.44

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates p < 0.05, and ** indicates p < 0.01 when compared to our method.

To validate the generalizability of our method, we further evaluated its performance on the autoPET dataset. Table 2 presents the results of various methods on this publicly available dataset. Our method achieved the highest DSC (0.7441), sensitivity (0.7573), and precision (0.7427), demonstrating robust segmentation accuracy and reliability across datasets. While the ASD (5.83 mm) and HD95 (21.27 mm) of our method were comparable to other approaches, they did not exhibit a significant advantage. This indicates that while our model exceled in capturing lesion characteristics and reducing false positives, further refinement may be required to enhance boundary delineation accuracy, particularly in datasets with greater variability.

To evaluate the stability of our model under different conditions, we employed the box plots to display the distribution of DSC, sensitivity, precision, ASD and HD95 across five-fold cross-validation on both the private and autoPET datasets. Fig. 4 (a) shows the results on the private dataset and autoPET dataset. In

both cases, our method demonstrates a more concentrated distribution with less variability compared to other methods, indicating higher stability across experiments.

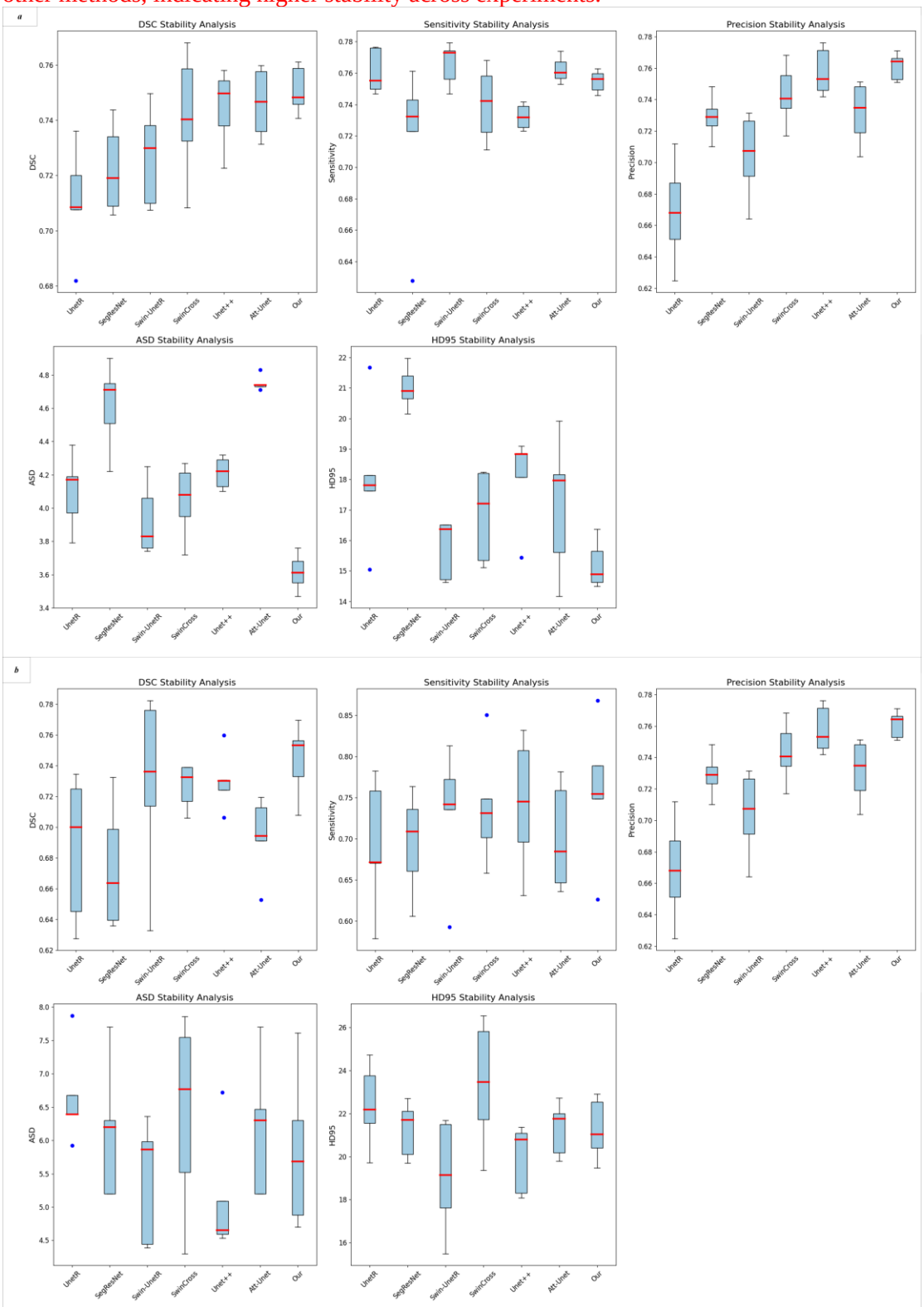


Fig. 4. Stability analysis using box plots for the private and autoPET datasets: This figure presents the ranges of DSC, sensitivity, precision, ASD and HD95 (shown in mm) across five cross-validation folds for different models. Subfigure (a) displays the results on the private dataset, while subfigure (b) shows the corresponding results on the autoPET dataset. The box shows the 1st quartile

(lower boundary), median (red line), and 3rd quartile (upper boundary). The whiskers represent the range of data, excluding outliers, which are marked as blue dots.

(Section 4.1 Optimization of Network Design and Ablation Study)

Table 4: Impact of each module on overall model performance.

Model	MSM	CMA	GFM	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
Baseline	☐	☐	☐	0.7291±0.0112	0.7405±0.0123	0.7152±0.0184	4.35±0.25	22.98±2.64
MSM	☑	☐	☐	0.7386±0.0121	0.7549±0.0115	0.7228±0.0173	4.01±0.18	21.77±1.71
CMA	☐	☑	☐	0.7405±0.0101	0.7538±0.0099	0.7284±0.0112	3.79±0.26	18.23±1.64
GFM	☑	☐	☑	0.7369±0.0132	0.7443±0.0109	0.7415±0.0118	4.22±0.21	19.47±2.13
MSM+CMA	☑	☑	☐	0.7458±0.0137	0.7544±0.0172	0.7329±0.0201	3.77±0.19	17.43±1.65
Full Model	☑	☑	☑	0.7466±0.0118	0.7648±0.0145	0.7591±0.0178	3.62±0.17	16.35±1.55

Table 4 highlights the contributions of each module to overall model performance, offering insight into their roles in improving segmentation accuracy and robustness. Below, we provide a detailed analysis of each configuration:

- **Baseline**: The Baseline model excludes multi-scale features and advanced cross-modal fusion mechanisms. Instead, it employs feature map concatenation and convolution layers for feature fusion across PET and CT modalities. This design limits the model's ability to leverage complementary information from the two modalities, resulting in suboptimal performance (DSC: 0.7291, ASD: 4.35 mm). The achieved HD95 (22.98 mm) indicates difficulties in accurately delineating complex tumor boundaries using the baseline model, especially in regions with blurred edges or low contrast.
- **MSM**: Based on the baseline model, we added the MSM to extract the multi-scale features through convolutional layers with varying kernel sizes. However, the feature fusion across modalities and scales still relies on feature map concatenation and convolution layers. This addition improves DSC and sensitivity (0.7386 and 0.7549, respectively), highlighting the importance of multi-scale feature aggregation in capturing fine-grained tumor details. Nevertheless, the modest improvement in precision (0.7228) suggests that the increased feature complexity may amplify noise in certain regions, resulting in false positives.
- **CMA**: Based on the baseline model, the CMA model employs the cross-modal attention mechanism for feature fusion between PET and CT modalities. This mechanism aligns spatial features from CT with metabolic features from PET, improving boundary delineation and enhancing the integration of complementary information. As a result, the CMA model significantly reduced the ASD (3.79 mm) and HD95 (18.23 mm) compared to baseline and MSM models, while the improved the DSC to 0.7405. However, the limited gain in precision (0.7284) indicates that challenges remain in handling low-contrast or heterogeneous regions.
- **GFM**: By adding the GFM individually to baseline model, we replaced the concatenation and convolution operation for multi-scale feature fusion with a gated fusion mechanism. This allows the model to dynamically adjust feature contributions from different scales, selectively emphasizing high-confidence features and suppressing noise. While GFM showed limited improvement in DSC (0.7369) and sensitivity (0.7443), the precision was increased notably to 0.7415, reflecting the module's effectiveness in mitigating false positives. However, the slightly higher ASD (4.22 mm) and HD95 (19.47 mm) suggested a focus on local optimization at the expense of global consistency.
- **MSM+CMA**: Combining MSM and CMA introduces multi-scale features with cross-modal attention for modality fusion but retains concatenation and convolution for multi-scale feature integration. This configuration achieved the highest DSC (0.7458) among other baseline models, reflecting the complementary roles of multi-scale and cross-modal mechanisms. The sensitivity (0.7544) was also improved significantly, but the achieved precision remained moderate (0.7329) due to the lack of dynamic scale weighting. The reduced HD95 (17.43 mm) indicated better handling of outliers and boundary refinement.
- **Full Model**: The Full Model integrates all three components, combining multi-scale features, cross-

modal attention, and gated fusion. This configuration achieved the best overall performance (DSC: 0.7466, ASD: 3.62 mm, HD95: 16.35 mm). The integration of GFM with MSM and CMA allows for robust feature selection and fusion, balancing local and global segmentation challenges. The significant improvement in precision (0.7591) confirmed the effectiveness of GFM in addressing false positives, while MSF and CMA enhance sensitivity and boundary delineation. These results demonstrate the comprehensive capabilities of the Full Model in extracting lymphoma lesions with high accuracy and robustness using CT and PET images.

4. Although correlation analysis and Bland-Altman are useful for evaluating TMTV quantification performance, the authors could also report absolute and relative errors to better demonstrate quantification accuracy.

Reply: Thank you for your insightful suggestion. We fully agree that absolute and relative errors are essential for a more comprehensive evaluation of TMTV quantification accuracy. In response, we have added the mean absolute error (MAE) and mean relative error (MRE) as additional metrics in [Section 2.6 and Section 3.2].

(Section 2.6 TMTV calculation)

Furthermore, we calculate the Mean Absolute Error (MAE) and Mean Relative Error (MRE) to quantify the accuracy of cTMTV. The formulas are defined as shown in Eqs. 22 and 23:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|cTMTV_i - gtTMTV_i| \quad (22)$$

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{cTMTV_i - gtTMTV_i}{gtTMTV_I} \times 100 \quad (23)$$

where N represents the total number of samples, $cTMTV_i$ is the calculated TMTV for the i-th sample, $gtTMTV_i$ is the ground truth TMTV for the i-th sample.

(Section 3.2 Results of TMTV)

We evaluated TMTV predictions on both the private dataset and the autoPET dataset using five-fold cross-validation. For the private dataset, the mean cTMTV was 802.69 ± 1192.95 mL, and the mean gtTMTV was 792.36 ± 1133.78 mL, with a mean difference of 10.33 ± 360.03 mL. Similarly, on the autoPET dataset, the mean cTMTV was 3074.83 ± 4080.80 mL, and the mean gtTMTV was 2637.70 ± 3636.79 mL, resulting in a mean difference of 437.12 ± 1961.38 mL. To further quantify prediction accuracy, we calculated the MAE and MRE. For the private dataset, the MAE was 123.42 ± 61.84 mL, with an MRE of 15.75% ± 2.00%. Meanwhile, on the autoPET dataset, the MAE was 1069.17 ± 1699.52 mL, and the MRE reached 173.91 % ± 741.42 %. Although the MRE on the autoPET dataset appeared high, this was driven by a few extreme cases involving small ground-truth volumes. Overall, these results indicated that our model achieved robust performance across both datasets, maintaining a relatively low absolute error compared to the total volume in most cases.

5. The Bland-Altman analysis indicates some inconsistencies in TMTV predictions, with a few outliers. Discussing potential reasons for these discrepancies would be valuable.

Reply: Thank you for your thoughtful comment. We have revised [Section 3.2] to discuss potential reasons for the observed outliers in the Bland-Altman analysis. Specifically, we attribute these discrepancies to

factors such as patient-specific variations (e.g., tumor irregularities or low contrast in certain PET regions) and noise in PET images, which may impact boundary detection and volume measurement accuracy. This additional discussion provides a deeper understanding of the challenges and limitations of our method.

(Section 3.2 Results of TMTV)

Fig. 9 shows the Bland-Altman analysis for both datasets. Most differences fell within the acceptable range, indicating general alignment with true TMTV values, though a few outliers were noted. These discrepancies stemmed from patient-specific variations, such as tumor irregularities or low contrast in certain PET regions, which challenged the boundary detection. Additionally, noise in PET images, particularly in low-activity regions, influenced the measurement accuracy. The analysis revealed a trend of increasing discrepancies with higher TMTV values, which was attributed to the limited number of cases, reducing the generalizability for larger tumor volumes and increases variability in predictions.

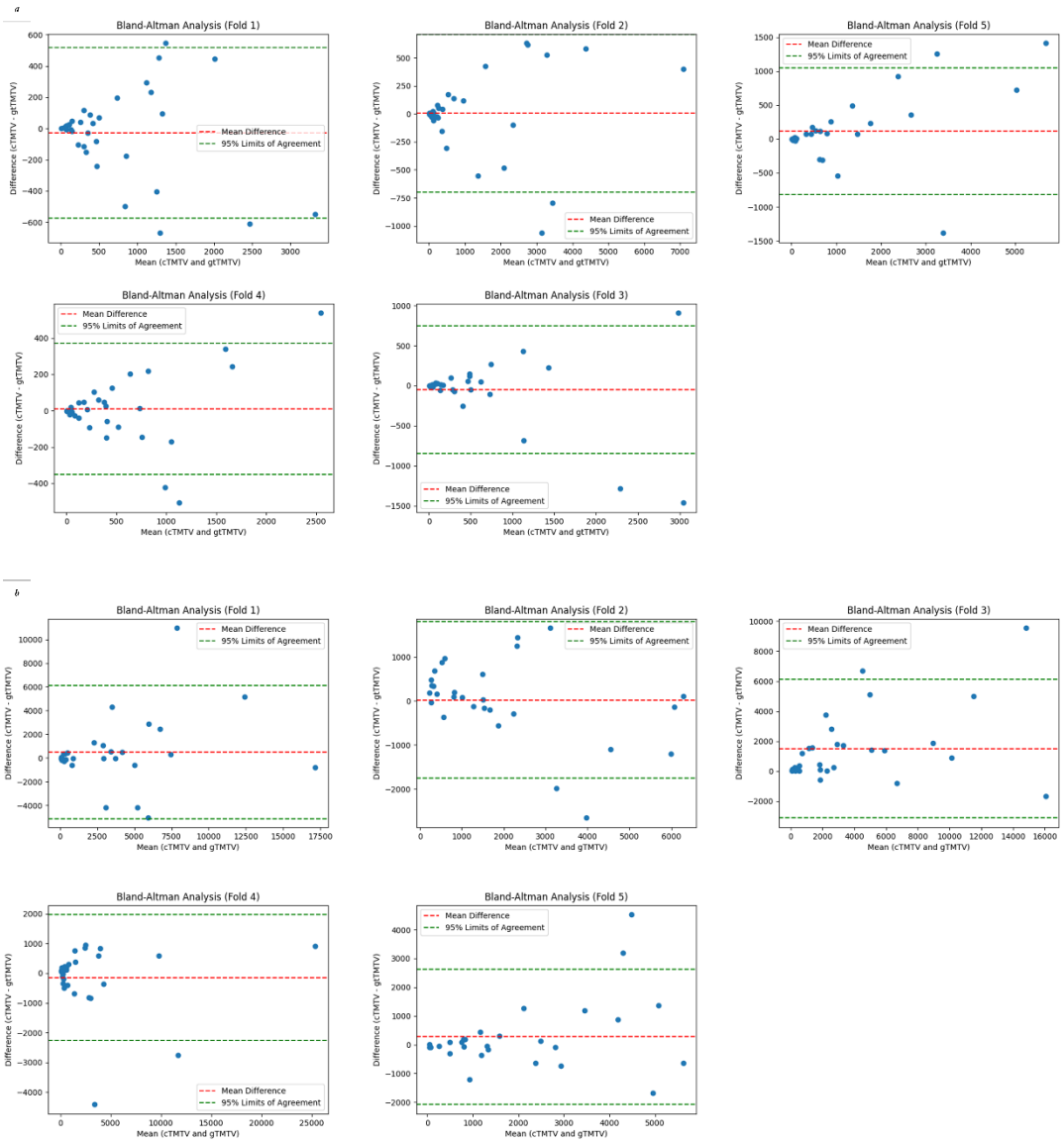


Fig. 9. Bland-Altman analysis for cTMTV vs. gtTMTV on (a) private dataset and (b) autoPET dataset. The horizontal axis represents the mean of cTMTV and gtTMTV, while the vertical axis represents their difference. The red dashed line shows the mean difference, and the green dashed lines represent the 95% limits of agreement, calculated as the mean difference ± 1.96 standard deviations of the differences.

6. It would be helpful to clearly mention the statistical tests used in Table 1 and how the authors ensured consistent data splitting across all the state-of-the-art (SOTA) techniques compared. The significant digits could also be reduced to two for metric evaluations, as four digits may be excessive in the context of segmentation metrics.

Thank you for your valuable suggestion. In response, we have clarified in [Section 2.5] that consistent data splitting across all SOTA methods was ensured by employing identical five-fold cross-validation sets. Additionally, we have explicitly stated the statistical test used (paired t-tests) in Table 1 and Table 2 under [Section 3.1]. Regarding the significant digits, while we appreciate your suggestion to reduce them to two decimal places, we chose to retain four decimal places for the following reasons:

1) Precision in Statistical Comparison: Small differences in segmentation metrics (e.g., DSC, Sensitivity, Precision) can be statistically significant, especially when values are close. Retaining four decimal places ensures these differences are not lost or misinterpreted.
2) Consistency with Related Literature: Reporting four decimal places aligns with practices in recent studies on medical image segmentation, facilitating direct comparison with existing methods.
3) High Sensitivity Metrics: Segmentation metrics like DSC are highly sensitive, and truncating to two decimal places could obscure the nuanced differences between methods.

We believe this approach maintains clarity while ensuring scientific rigor. Thank you again for your valuable feedback, which has significantly improved the clarity and comprehensiveness of our results.

(Section 2.5 Implementation and experiments)
We evaluated the effectiveness of our method on both the private dataset and the autoPET dataset, comparing its performance with various state-of-the-art (SOTA) methods, including UnetR[32], Swin-UnetR[21], Att-Unet[33], Unet++[34], SegResNet[35], and SwinCross[22]. Consistent data splitting was ensured for all methods by employing the same 5-fold cross-validation approach. In each fold, the training set consisted of 60% of the data, while the validation and test set each accounted for 20%.
To ensure fairness, all experiments were conducted within the same computational environment, using identical hardware and software configurations. A sliding window technique was employed to reduce GPU memory consumption, extracting 32 consecutive slices per batch to form a 3D volume. Hyperparameter optimization, including adjustments to learning rates, batch sizes, and optimizer settings, was performed for each method based on validation set performance.

(Section 3.1 Results of segmentation)

Table 1: Results of different methods on the private dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.7107±0.0178 **	0.7608±0.0128	0.6686±0.0298 **	4.10±0.20 **	18.05±2.36
SegResNet	0.7223±0.0146 *	0.7175±0.0466	0.7289±0.0125 **	4.61±0.26 **	21.01±0.69 **
Swin-UnetR	0.7271±0.0163 *	**0.7659±0.0123**	0.7041±0.0246 **	3.92±0.22 *	15.74±0.98
SwinCross	0.7414±0.0209	0.7405±0.0213	0.7432±0.0176	4.04±0.22 **	16.82±1.51
Unet++	0.7446±0.0129	0.7322±0.0072 **	0.7577±0.0137	4.21±0.09 **	18.05±1.51 **
Att-Unet	0.7463±0.0113	0.7622±0.0075	0.7314±0.0179 *	4.75±0.04 **	17.16±2.26
Ours	**0.7512±0.0078**	0.7548±0.0063	**0.7611±0.0078**	**3.61±0.11**	**15.20±0.78**

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates $p < 0.05$, and ** indicates $p < 0.01$ when compared to our method.

Table 2: Results of different methods on the autoPET dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.6865±0.0478 **	0. 6924±0.0812	0.6851±0.0404 *	6.65±0.73	22.38±1.95 *
SegResNet	0.6740±0.0412 *	0.6951±0.0627	0.6483±0.0505 *	6.12±1.02	21.26±1.29
Swin-UnetR	0.7282±0.0605	0.7311±0.0833	0.7274±0.0450	5.40±0.92	**19.08±2.63**
SwinCross	0.7267±0.0146 **	0.7382±0.0717 *	0.7233±0.0525	6.40±1.48	23.37±2.95

Unet++	0.7302±0.0192	0.7424±0.0818	0.7277±0.0523	**5.11±0.92**	19.92±1.59
Att-Unet	0.6941±0.0261 **	0.7016±0.0657	0.6917±0.0401 **	6.17±1.04	21.29±1.25
Ours	**0.7441±0.0241**	**0.7573±0.0874**	**0.7427±0.0647**	5.83±1.18	21.27±1.44

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates $p < 0.05$, and ** indicates $p < 0.01$ when compared to our method.

7. The model (s) and codes have to be publicly shared for evaluation.

Reply: Thank you for your valuable suggestion. We agree that sharing the model and code is crucial for ensuring reproducibility and facilitating further research. In response, we have made our implementation publicly available on GitHub at [https://github.com/chenzhao2023/lymphoma_seg]. This information has been added to the revised manuscript in the abstract section.

(Abstract)

The code for the proposed method is available at https://github.com/chenzhao2023/lymphoma_seg.

Minor comments:

1. Figure 5 does not seem effective for showing visual performance comparison through inspection. A more informative visual representation should be considered.

Reply: Thank you for your valuable suggestion. To address this issue, we have replaced the original Fig. 5 with two new figures (Fig. 5 and Fig. 6), which provide a more detailed and informative visual representation.

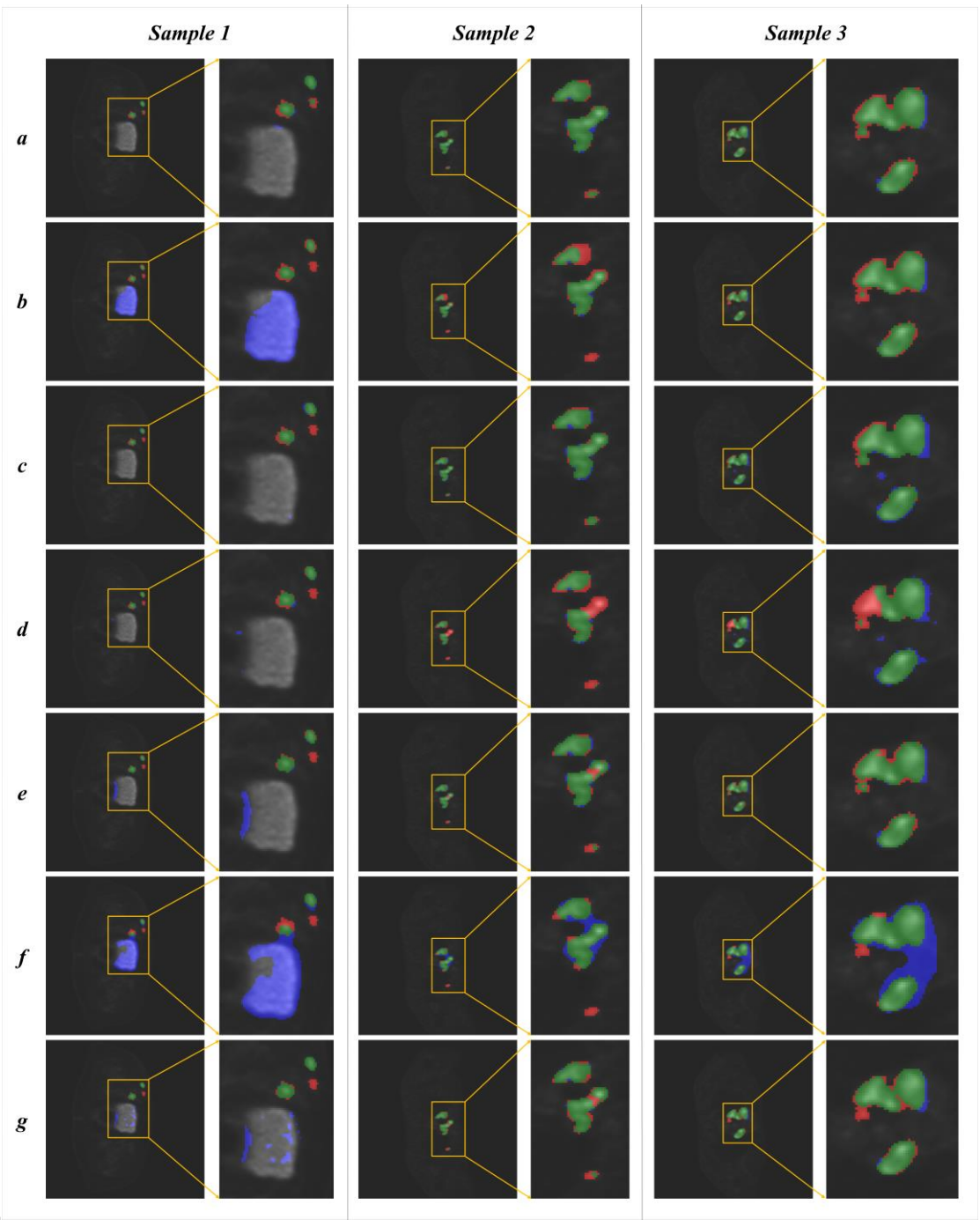


Fig. 5. Difference maps of segmentation results compared with ground truth for private datasets. The green, red, and blue regions represent true positive, false negative, and false positive pixels, respectively. Subfigures (***a***)–(***g***) show results generated by our method, Att-Unet, Unet++, SwinCross, Swin-UnetR, SegResNet, and UnetR, respectively.

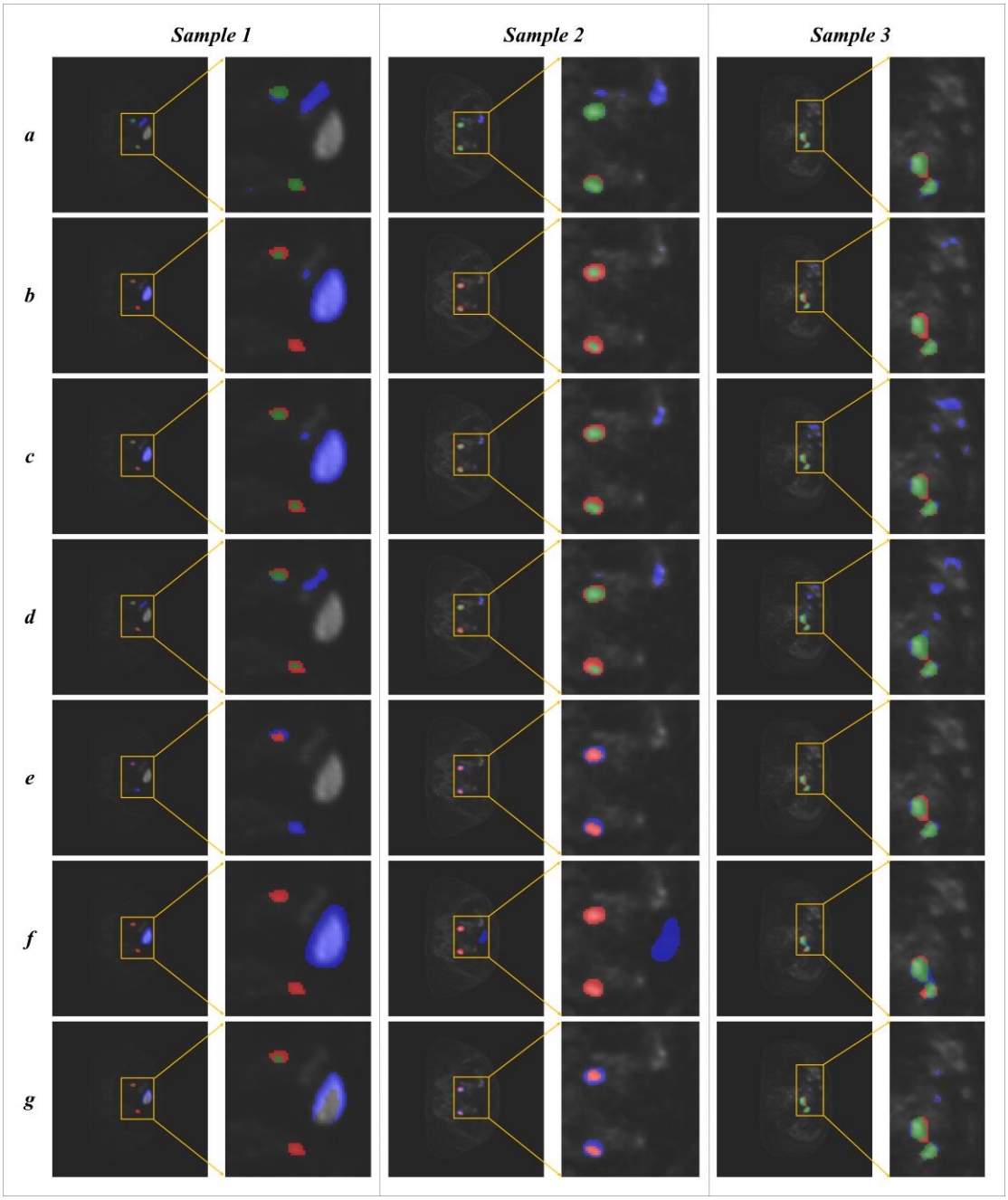


Fig. 6. Difference maps of segmentation results compared with ground truth for autoPET datasets. The green, red, and blue regions represent true positive, false negative, and false positive pixels, respectively. Subfigures (***a***)–(***g***) show results generated by our method, Att-Unet, Unet++, SwinCross, Swin-UnetR, SegResNet, and UnetR, respectively.

Reviewer #2:

General Comments (Required)

1. This paper presents a study on lymphoma segmentation using an enhanced network designed with PET and CT images. The primary originality of this research lies in the MSIF method, which effectively utilizes the features from both imaging modalities within the network.

Reply: Thank you for your positive feedback on the originality of our research. We appreciate your recognition of the MSIF method as a key innovation in leveraging features from both PET and CT modalities. Your encouraging comments affirm the significance of our work and inspire us to continue improving its impact.

2. While the accompanying figures are high-quality and excellent for illustrating the concept, the written explanation lacks sufficient detail for readers to fully understand the method. Furthermore, in terms of discussion on the results, the study only includes an ablation study, which simply verifies the functionality of the network.

Reply: Thank you for your constructive feedback. We have carefully revised the manuscript to include more detailed explanations in [Section 2.3.2], ensuring that readers can fully understand our method. Additionally, we have expanded the discussion section to provide a broader analysis of the results beyond the ablation study. Specifically, in [Section 4.2], we compared our approach with other state-of-the-art methods, highlighting its advantages in segmentation accuracy and its practical strengths in real-world clinical applications.

(Section 2.3.2 Multi-Scale Information Fusion)

As shown in Fig. 3 (***i***), we compute $Q\ (query)$, $K\ (key)$, and $V(value)$for each patch as show in Eqs. 2 to 4:

$$Q^l_{modal_n} = F^l_{modal_n} W_Q \tag{2}$$

$$K^l_{modal_n} = F^l_{modal_n} W_K \tag{3}$$

$$V^l_{modal_n} = F^l_{modal_n} W_V \tag{4}$$

where l is the layer of Swin Transformer. $W_Q, W_K, W_V \in R^{D_f \times D_q}$ are the weight matrices, where D_f is the feature dimension, and D_q is the dimension of the queries and keys. Here, Q captures the specific features within a modality's patch that should attend to features in the other modality. K represents the features from the complementary modality that the query seeks alignment with, while V provides the actual information from the complementary modality that will contribute to the fused representation.

As shown in Fig. 3 (***ii***), cross-modal attention is then computed as shown in Eqs. 5 and 6:

$$Att^l_1 = softmax\left(\frac{Q^l_1 \left(K^l_2\right)^T}{\sqrt{D_k}}\right) V^l_1 \tag{5}$$

$$Att^l_2 = softmax\left(\frac{Q^l_2 \left(K^l_1\right)^T}{\sqrt{D_k}}\right) V^l_2 \tag{6}$$

where D_k is the dimension of the keys and queries, T denotes the matrix transpose. Normalizing the dot product of queries and keys by $\sqrt{D_k}$ ensures stable gradient flow by preventing excessively large values during the Softmax operation. This mechanism allows the PET query (Q^l_1) to selectively attend to relevant CT features (K^l_2) and vice versa$\left(Q^l_2\ and\ K^l_1\right)$. This bidirectional interaction enables spatial features from CT to provide anatomical context for PET's metabolic activity, while PET's metabolic features enhance CT's structural understanding. To maintain consistency and stability in feature alignment, we empirically set $D_q = D_k$.

At each layer l, the image is partitioned into windows of size $M \times M \times M$. In the subsequent layer $l+1$, these windows are shifted by $\left[\frac{M}{2}, \frac{M}{2}, \frac{M}{2}\right]$ voxels, allowing interaction between adjacent windows and reducing redundant calculations. This shifting strategy eliminates the repeated processing of overlapping regions, which is common in fixed-window attention mechanisms, thus optimizing computational efficiency. Moreover, by enabling neighboring regions to interact across layers, the shifted window multi-head self-attention (SW_MSA) facilitates seamless information flow, addressing the issue of isolated window processing.

(Section 4.1 Optimization of Network Design and Ablation Study)

Table 3: Optimal Network Configuration.

Configuration	Attention Heads	Swin Transformer Layers	Patch Embedding Dimension	Windows Size
Best Setting	[3,6,12,24)	[2, 2, 2, 2]	24	[3,3,3]

To identify the optimal network configuration, we conducted ablation experiments by varying key parameters, including the number of attention heads, the number of layers in each stage of the Swin Transformer, and the patch embedding dimension. Based on these experiments, we determined the optimal configuration that balances segmentation performance and computational efficiency. Table 3 summarizes the best settings for each parameter.

To evaluate the contributions of the MSM, CMA, and GFM, we conducted an ablation study with the following experimental setups in Table 4. The Baseline model was designed with single-scale fusion and dual encoders, without attention mechanisms or gated networks, serving as the benchmark for comparison.

Table 4: Impact of each module on overall model performance.

Model	MSM	CMA	GFM	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
Baseline	☐	☐	☐	0.7291±0.0112	0.7405±0.0123	0.7152±0.0184	4.35±0.25	22.98±2.64
MSM	☑	☐	☐	0.7386±0.0121	0.7549±0.0115	0.7228±0.0173	4.01±0.18	21.77±1.71
CMA	☐	☑	☐	0.7405±0.0101	0.7538±0.0099	0.7284±0.0112	3.79±0.26	18.23±1.64
GFM	☑	☐	☑	0.7369±0.0132	0.7443±0.0109	0.7415±0.0118	4.22±0.21	19.47±2.13
MSM+CMA	☑	☑	☐	0.7458±0.0137	0.7544±0.0172	0.7329±0.0201	3.77±0.19	17.43±1.65
Full Model	☑	☑	☑	0.7466±0.0118	0.7648±0.0145	0.7591±0.0178	3.62±0.17	16.35±1.55

Table 4 highlights the contributions of each module to overall model performance, offering insight into their roles in improving segmentation accuracy and robustness. Below, we provide a detailed analysis of each configuration:

- **Baseline**: The Baseline model excludes multi-scale features and advanced cross-modal fusion mechanisms. Instead, it employs feature map concatenation and convolution layers for feature fusion across PET and CT modalities. This design limits the model's ability to leverage complementary information from the two modalities, resulting in suboptimal performance (DSC: 0.7291, ASD: 4.35 mm). The achieved HD95 (22.98 mm) indicates difficulties in accurately delineating complex tumor boundaries using the baseline model, especially in regions with blurred edges or low contrast.
- **MSM**: Based on the baseline model, we added the MSM to extract the multi-scale features through convolutional layers with varying kernel sizes. However, the feature fusion across modalities and scales still relies on feature map concatenation and convolution layers. This addition improves DSC and sensitivity (0.7386 and 0.7549, respectively), highlighting the importance of multi-scale feature aggregation in capturing fine-grained tumor details. Nevertheless, the modest improvement in precision (0.7228) suggests that the increased feature complexity may amplify noise in certain regions, resulting in false positives.

- **CMA**: Based on the baseline model, the CMA model employs the cross-modal attention mechanism for feature fusion between PET and CT modalities. This mechanism aligns spatial features from CT with metabolic features from PET, improving boundary delineation and enhancing the integration of complementary information. As a result, the CMA model significantly reduced the ASD (3.79 mm) and HD95 (18.23 mm) compared to baseline and MSM models, while the improved the DSC to 0.7405. However, the limited gain in precision (0.7284) indicates that challenges remain in handling low-contrast or heterogeneous regions.
- **GFM**: By adding the GFM individually to baseline model, we replaced the concatenation and convolution operation for multi-scale feature fusion with a gated fusion mechanism. This allows the model to dynamically adjust feature contributions from different scales, selectively emphasizing high-confidence features and suppressing noise. While GFM showed limited improvement in DSC (0.7369) and sensitivity (0.7443), the precision was increased notably to 0.7415, reflecting the module's effectiveness in mitigating false positives. However, the slightly higher ASD (4.22 mm) and HD95 (19.47 mm) suggested a focus on local optimization at the expense of global consistency.
- **MSM+CMA**: Combining MSM and CMA introduces multi-scale features with cross-modal attention for modality fusion but retains concatenation and convolution for multi-scale feature integration. This configuration achieved the highest DSC (0.7458) among other baseline models, reflecting the complementary roles of multi-scale and cross-modal mechanisms. The sensitivity (0.7544) was also improved significantly, but the achieved precision remained moderate (0.7329) due to the lack of dynamic scale weighting. The reduced HD95 (17.43 mm) indicated better handling of outliers and boundary refinement.
- **Full Model**: The Full Model integrates all three components, combining multi-scale features, cross-modal attention, and gated fusion. This configuration achieved the best overall performance (DSC: 0.7466, ASD: 3.62 mm, HD95: 16.35 mm). The integration of GFM with MSM and CMA allows for robust feature selection and fusion, balancing local and global segmentation challenges. The significant improvement in precision (0.7591) confirmed the effectiveness of GFM in addressing false positives, while MSF and CMA enhance sensitivity and boundary delineation. These results demonstrate the comprehensive capabilities of the Full Model in extracting lymphoma lesions with high accuracy and robustness using CT and PET images.

(Section 4.2 Comparison with Existing TMTV Calculation Methods)

Compared to existing techniques for TMTV segmentation and quantification, our method offers clear advantages in practicality, automation, and accuracy. While Yousefirizi et al.[37,38] and Blanc-Durand et al.[14] evaluated their approaches on single-center or multi-center datasets, the validation on publicly available datasets was not conducted, limiting the generalizability of their methods. In contrast, our approach was rigorously evaluated on both a private dataset and the autoPET dataset, demonstrating consistent performance across varied imaging protocols. This dual validation underscores the robustness of our method in diverse clinical settings.

In terms of accuracy, Yousefirizi et al.[38] achieved a Pearson correlation coefficient of R^2=0.83 for TMTV quantification but reported lower segmentation accuracy (DSC=0.68) due to single-modality constraints. Similarly, Blanc-Durand et al.[14] achieved a DSC of 0.73 but faced significant TMTV underestimation on the external validation dataset. By leveraging multi-scale and cross-modal feature fusion, our method effectively integrates PET and CT information, achieving superior segmentation accuracy with DSC values of 0.7512 on the private dataset and 0.7441 on the autoPET dataset.

Furthermore, the semi-automated workflows proposed by Burggraaff et al.[39] required manual threshold adjustments, leading to time-consuming processes prone to variability. In contrast, our fully automated

pipeline eliminates the need for manual intervention, ensuring consistent, reproducible results while significantly reducing analysis time. These features make our method particularly suitable for routine clinical applications.

In conclusion, our method combines automation, accuracy, and generalizability to provide a practical and efficient solution for TMTV segmentation and quantification, supporting both research and clinical workflows.

3. However, the significance of the research and its results are sufficiently meaningful, and thus the following revision comments are provided.

Reply: Thank you for recognizing the significance of our research and its results. We greatly appreciate your constructive comments, which have provided valuable insights for improving the manuscript.

Specific Comments:

1. The study appears to be a retrospective clinical data analysis, but there is a lack of explicit mention regarding research ethics.

Reply: Thank you for your valuable comment. To address this, we have revised [Section 2.1] to explicitly state that ethical approval was obtained from the Institutional Review Board of Peking University People's Hospital for the private dataset. Additionally, we ensured that all data was anonymized to protect patient privacy and adhered to ethical guidelines for retrospective clinical data analysis.

(Section 2.1 Dataset)

This study utilized two datasets: (1) our private dataset comprising 165 PET/CT scan datasets from patients clinically diagnosed with DLBCL, provided by Peking University People's Hospital, and (2) the FDG-PET/CT dataset (autoPET), comprising 145 PET/CT scans of lymphoma patients with manually annotated tumor lesions, obtained from The Cancer Imaging Archive (TCIA). The use of the autoPET dataset in this study complies with TCIA's data usage policy and is authorized for research purposes[23].

For the private dataset, ethical approval was obtained from the hospital's Institutional Review Board, and all data was de-identified to protect patient privacy, adhering to ethical guidelines for retrospective clinical data analysis.

2. If multiple nuclear medicine experts were involved in manual segmentation, there needs to be a discussion on how differences in their segmentation abilities might have impacted the validation of the algorithm's performance.

Reply: Thank you for your insightful comment. To address this, we have revised [Section 2.1] to clarify how the ground truth segmentations were validated.

(Section 2.1 Dataset)

Ground truth volumes of interest (VOI) were manually extracted on PET images by two experienced nuclear medicine experts. The extracted VOIs were cross-reviewed and confirmed by both experts to ensure consistency and accuracy.

3. While the introduction briefly mentions the Swin Transformer, it is necessary to provide further explanation on why it was used for encoding in this study's network and why it is superior compared to other networks.

Reply: Thank you for your insightful comment. In response, we have revised the introduction to provide a detailed explanation of why the Swin Transformer was chosen as the backbone encoder for this study.

(Section 1. Introduction)

However, existing convolutional neural networks (CNNs) face challenges in fully leveraging multimodal PET/CT data, particularly due to their limited receptive field, which restricts their ability to capture global and local contextual information[16-18].

Recent approaches, such as the Vision Transformer (ViT)[19], have introduced global self-attention mechanisms to capture long-range dependencies. However, the high computational cost and absence of hierarchical feature representation in ViT limit its practicality for high-resolution medical image processing. The Swin Transformer addresses these limitations by combining a hierarchical structure with a shifted window attention mechanism[20-22]. This design enables the Swin Transformer to effectively capture global and local context while reducing computational complexity, making it particularly suited for multimodal PET/CT segmentation.

4. The originality of the paper should be discussed in more detail in the methods or discussion sections.

Reply: Thank you for your valuable suggestion. We have revised the manuscript to include a more detailed discussion of the originality of our work. In [Section 2.3], we elaborated on the innovations introduced by our MSIF module, highlighting its dynamic modality contribution adjustment, enhanced multi-scale feature fusion, and computational efficiency, which address the limitations of existing methods. In [Section 4.2], we emphasized the unique contributions of our approach compared to state-of-the-art techniques, particularly its ability to leverage multi-modal feature fusion effectively and its clinical applicability in lymphoma diagnosis.

(Section 2.3 Network architecture)

The MSIF module draws inspiration from existing multi-modal and multi-scale feature fusion methods in medical imaging, such as MDRANet[27], CA-Net[28], MFCPNet[29], and SwinCross[22]. These methods have made significant progress in leveraging multi-scale information and cross-modal interactions; however, they exhibit certain limitations. For instance, MDRANet and CA-Net focus on single-modality feature enhancement and lack mechanisms for cross-modal interaction, while MFCPNet employs fixed fusion strategies that hinder its adaptability in multi-modal tasks. SwinCross captures complementary PET and CT information through shifted window-based attention but does not dynamically adjust modality contributions and does not fully exploit multi-scale feature integration. Building upon these approaches, the MSIF module addresses these limitations through dynamic modality contribution adjustment, enhanced multi-scale feature fusion, and computational efficiency. These innovations enable robust and flexible performance across a wide range of medical imaging tasks, including scenarios with limited data availability or complex modality interactions.

(Section 4.2 Comparison with Existing TMTV Calculation Methods)

Compared to existing techniques for TMTV segmentation and quantification, our method offers clear advantages in practicality, automation, and accuracy. While Yousefirizi et al.[37,38] and Blanc-Durand et al.[14] tested their approaches on single-center or multi-center datasets, they lacked validation on publicly available datasets, limiting the generalizability of their methods. In contrast, our approach was rigorously evaluated on both a private dataset and the autoPET dataset, demonstrating consistent performance across varied imaging protocols. This dual validation underscores the robustness of our method in diverse clinical settings.
In terms of accuracy, Yousefirizi et al.[38] achieved a Pearson correlation coefficient of R^2=0.83 for TMTV quantification but reported lower segmentation accuracy (DSC=0.68) due to single-modality constraints. Similarly, Blanc-Durand et al.[14] achieved a DSC of 0.73 but faced significant TMTV underestimation on the external validation dataset. By leveraging multi-scale and cross-modal feature fusion, our method

effectively integrates PET and CT information, achieving superior segmentation accuracy with DSC values of 0.7512 on the private dataset and 0.7441 on the autoPET dataset.
Furthermore, the semi-automated workflows proposed by Burggraaff et al.[39] required manual threshold adjustments, leading to time-consuming processes prone to variability. In contrast, our fully automated pipeline eliminates the need for manual intervention, ensuring consistent, reproducible results while significantly reducing analysis time. These features make our method particularly suitable for routine clinical applications.

In conclusion, our method combines automation, accuracy, and generalizability to provide a practical and efficient solution for TMTV segmentation and quantification, supporting both research and clinical workflows.

5. There seem to be several multi-scale information fusion methods similar to MSIF; are there any references to such methods?

Reply: Thank you for your insightful comment. In response, we have revised [Section 2.3] to include a discussion of the relationship between the MSIF module and existing multi-modal and multi-scale feature fusion methods, such as MDRANet[17], CA-Net[18], MFCPNet[19], and SwinCross[7]. These methods represent significant advancements in medical imaging but exhibit certain limitations, such as the lack of cross-modal interaction mechanisms or dynamic contribution adjustment. Building upon these approaches, we highlighted the unique contributions of the MSIF module, including dynamic modality contribution adjustment, enhanced multi-scale feature fusion, and computational efficiency. These additions clarify the relationship between MSIF and existing methods while emphasizing its innovations.

(Section 2.3 Network architecture)

The MSIF module draws inspiration from existing multi-modal and multi-scale feature fusion methods in medical imaging, such as MDRANet[27], CA-Net[28], MFCPNet[29], and SwinCross[22]. These methods have made significant progress in leveraging multi-scale information and cross-modal interactions; however, they exhibit certain limitations. For instance, MDRANet and CA-Net focus on single-modality feature enhancement and lack mechanisms for cross-modal interaction, while MFCPNet employs fixed fusion strategies that hinder its adaptability in multi-modal tasks. SwinCross captures complementary PET and CT information through shifted window-based attention but does not dynamically adjust modality contributions and does not fully exploit multi-scale feature integration. Building upon these approaches, the MSIF module addresses these limitations through dynamic modality contribution adjustment, enhanced multi-scale feature fusion, and computational efficiency. These innovations enable robust and flexible performance across a wide range of medical imaging tasks, including scenarios with limited data availability or complex modality interactions.

6. Are the terms query, key, and value adopted from databases, and how exactly are the three defined and distinguished in PET and CT data?

Reply: Thank you for your thoughtful comment. To address your concern, we have expanded [Section 2.3.2] to clarify the roles and definitions of query (Q), key (K), and value (V) in the context of PET and CT data fusion. These terms, originally adopted from attention mechanisms in natural language processing (NLP), share conceptual similarities with database terms but are specifically tailored for feature alignment in multi-modal data fusion.

In our framework, Q represents the features within one modality (e.g., PET) that attend to complementary features in the other modality (e.g., CT), defined as K, while V provides the actual information from the complementary modality that contributes to the fused representation.

(Section 2.3.2 Multi-Scale Information Fusion)

As shown in Fig. 3 (***i***), we compute Q $(query)$, K (key), and $V(value)$for each patch as show in Eqs. 2 to 4:

$$Q^l_{modal_n} = F^l_{modal_n} W_Q \quad (2)$$

$$K^l_{modal_n} = F^l_{modal_n} W_K \quad (3)$$

$$V^l_{modal_n} = F^l_{modal_n} W_V \quad (4)$$

where l is the layer of Swin Transformer. $W_Q, W_K, W_V \in R^{D_f \times D_q}$ are the weight matrices, where D_f is the feature dimension, and D_q is the dimension of the queries and keys. Here, Q captures the specific features within a modality's patch that should attend to features in the other modality. K represents the features from the complementary modality that the query seeks alignment with, while V provides the actual information from the complementary modality that will contribute to the fused representation.

7. An explanation of how the equation (Eq. 5) was derived and why this equation facilitates better fusion of PET and CT.

Reply: Thank you for your insightful comment. In response, we have revised [Section 2.3.2] to provide a more detailed explanation of how Eq. (5) was derived and its role in facilitating the fusion of PET and CT data. Specifically, Eq. (5) is derived from the scaled dot-product attention mechanism, which normalizes the dot product of queries and keys by the square root of their dimensionality $\sqrt{D_k}$. This normalization reduces excessively large values in the Softmax operation, ensuring stable gradient flow during training. By enabling selective attention between PET and CT features, this mechanism effectively captures spatial and metabolic complementarities, leading to more precise feature fusion.

(Section 2.3.2 Multi-Scale Information Fusion)

As shown in Fig. 3 (***ii***), cross-modal attention is then computed as shown in Eqs. 5 and 6:

$$Att^l_1 = softmax\left(\frac{Q^l_1 (K^l_2)^T}{\sqrt{D_k}}\right) V^l_1 \quad (5)$$

$$Att^l_2 = softmax\left(\frac{Q^l_2 (K^l_1)^T}{\sqrt{D_k}}\right) V^l_2 \quad (6)$$

where D_k is the dimension of the keys and queries, T denotes the matrix transpose. Normalizing the dot product of queries and keys by $\sqrt{D_k}$ ensures stable gradient flow by preventing excessively large values during the Softmax operation. This mechanism allows the PET query (Q^l_1) to selectively attend to relevant CT features (K^l_2) and vice versa$\left(Q^l_2 \; and \; K^l_1\right)$. This bidirectional interaction enables spatial features from CT to provide anatomical context for PET's metabolic activity, while PET's metabolic features enhance CT's structural understanding. To maintain consistency and stability in feature alignment, we empirically set $D_q = D_k$.

8. The rationale behind setting Dq=Dk, and how changing this value affects network performance.

Reply: Thank you for your insightful comment regarding the rationale behind setting $D_q = D_k$. This choice is a standard practice in attention-based models, including Swin Transformer, due to the mathematical and functional consistency it ensures in the attention mechanism. To clarify, this setting is derived from the core formula of the scaled dot-product attention:

$$AttentionScores = \frac{Q \cdot K^T}{\sqrt{D_k}}$$

here, Q $(Query)$ and K (Key) represent feature embeddings of dimension D_q and D_k, respectively. For the dot-product operation $Q \cdot K^T$ to be valid, it is necessary that $D_q = D_k$. This ensures alignment in feature space, allowing meaningful computation of similarity scores. Additionally, the normalization term $\sqrt{D_k}$ is specifically designed to stabilize the gradient flow during training, and altering the relationship between D_q and D_k could disrupt this stability.

Maintaining $D_q = D_k$ also ensures semantic consistency in comparing query and key vectors, as both are projected into the same latent space, which is critical for effective attention computation. This design principle has been extensively validated in foundational works, including Vaswani et al.[20], $Attention\ is\ All\ You\ Need$ $(NeurIPS\ 2017)$, and has been widely adopted in subsequent Transformer-based models[3,14-16]. Given the theoretical basis and its established effectiveness, we followed this standard practice in our implementation. We hope this explanation sufficiently clarifies our rationale, and we appreciate the opportunity to provide additional context.

9. By shifting the layers l and l+1, why does this reduce redundant calculations, and how does it enhance feature extraction capabilities for multi-modal imaging?

Reply: Thank you for your thoughtful comment. The shifted window mechanism in the Swin Transformer reduces redundant calculations and enhances feature extraction capabilities, particularly in multi-modal imaging tasks, through a combination of local attention and dynamic window shifting. By restricting self-attention computation to fixed-size windows (e.g. 7×7), the computational complexity decreases from $O(n^2)$ to $O(n)$, where n is the number of input pixels. This ensures that the complexity scales linearly with the input size, making it computationally efficient. The shift in windows across layers facilitates information exchange between adjacent windows, overcoming the limitations of isolated computations and enabling better integration of global and local features[16]. For multi-modal imaging, such as PET and CT, this mechanism enhances feature fusion by allowing complementary information, such as PET's metabolic activity and CT's anatomical structure, to be effectively integrated across window boundaries. The progressive patch merging in the hierarchical structure further enlarges the receptive field, capturing both fine details and contextual information critical for understanding multi-modal relationships. This design reduces redundant computations while maximizing the extraction of meaningful features, thus enabling the Swin Transformer to deliver both computational efficiency and superior performance in complex imaging scenarios.

(Section 2.3.2 Multi-Scale Information Fusion)

At each layer l, the image is partitioned into windows of size $M \times M \times M$. In the subsequent layer $l + 1$, these windows are shifted by $\left[\frac{M}{2}, \frac{M}{2}, \frac{M}{2}\right]$ voxels, allowing interaction between adjacent windows and reducing redundant calculations. This shifting strategy eliminates the repeated processing of overlapping regions, which is common in fixed-window attention mechanisms, thus optimizing computational efficiency. Moreover, by enabling neighboring regions to interact across layers, the shifted window multi-head self-attention (SW_MSA) facilitates seamless information flow, addressing the issue of isolated window processing.

10. Although various networks were selected for comparison, was the optimization of parameters for each comparison network conducted? Since the performance of a network can vary depending on the input data, verification of whether network performance was optimized accordingly is necessary.

Reply: Thank you for your insightful comment. In response, we have revised [Section 2.5] to clarify that we optimized the hyperparameters for each state-of-the-art method based on validation set performance. Detailed hyperparameter settings for each method are provided in the table RL2.

Table RL2. Detailed hyperparameter settings for the peer models for performance comparison.

Method	Learning rate	Batch Size	Optimizer	Weight Decay	Additional Settings
UnetR	0.0001	2	Adam	0.0001	feature size=16, hidden size =768, num heads = 12, dropout rate = 0.2
SegResNet	0.0001	8	Adam	0.0001	Spatial dims = 3, dropout prob = 0.2, norm = batch
Swin-UnetR	0.0001	2	Adam	0.0001	Feature size =12, depths = (2,2,2,2), num heads = (3,6,12,24), drop rate = 0.2, dropout rate = 0.1
SwinCross	0.0001	4	Adam	0.0001	patch size = 2, embed dim = 48, depths = (2,4,2,2), num heads = (3,6,12,24), window size = (3,3,3), drop rate =0
Unet++	0.0001	4	Adam	0.0001	Features = [32,32,64,128,256,32], deep supervision = False dropout = 0
Att-Unet	0.0001	4	Adam	0.0001	channels = (16,32,64,128,256), strides = (2,2,2,2), dropout = 0.2

(Section 2.5 Implementation and experiments)

We evaluated the effectiveness of our method on both the private dataset and the autoPET dataset, comparing its performance with various state-of-the-art (SOTA) methods, including UnetR[32], Swin-UnetR[21], Att-Unet[33], Unet++[34], SegResNet[35], and SwinCross[22]. Consistent data splitting was ensured for all methods by employing the same 5-fold cross-validation approach. In each fold, the training set consisted of 60% of the data, while the validation and test set each accounted for 20%.

To ensure fairness, all experiments were conducted within the same computational environment, using identical hardware and software configurations. A sliding window technique was employed to reduce GPU memory consumption, extracting 32 consecutive slices per batch to form a 3D volume. Hyperparameter optimization, including adjustments to learning rates, batch sizes, and optimizer settings, was performed for each method based on validation set performance.

11. Sections 4.1 and 4.2 in the discussion mainly focus on network structure optimization. It is recommended to summarize this content into a single paragraph by presenting a condensed table and mentioning only the key features. Rather than merely describing the results of ablation studies, it would be beneficial to discuss the reasons behind the observed results.

Reply: Thank you for your valuable suggestion. We have carefully revised the manuscript to consolidate [Section 4.1] and [Section 4.2] into a single section, [Section 4.1]. This revision includes a summary of ablation experiments in Table 3 (network configuration performance) and Table 3 (module contributions) and a streamlined narrative that highlights key findings. Additionally, we have provided detailed explanations for the observed results, emphasizing the contributions of different network configurations and modules to overall segmentation performance.

(Section 4.1 Optimization of Network Design and Ablation Study)

Table 3: Optimal Network Configuration.

Configuration	Attention Heads	Swin Transformer Layers	Patch Embedding Dimension	Windows Size
Best Setting	[3,6,12,24)	[2, 2, 2, 2]	24	[3,3,3]

To identify the optimal network configuration, we conducted ablation experiments by varying key parameters, including the number of attention heads, the number of layers in each stage of the Swin Transformer, and the patch embedding dimension. Based on these experiments, we determined the optimal

configuration that balances segmentation performance and computational efficiency. Table 3 summarizes the best settings for each parameter.

To evaluate the contributions of the MSM, CMA, and GFM, we conducted an ablation study with the following experimental setups in Table 4. The Baseline model was designed with single-scale fusion and dual encoders, without attention mechanisms or gated networks, serving as the benchmark for comparison.

Table 4: Impact of each module on overall model performance.

Model	MSM	CMA	GFM	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
Baseline	☐	☐	☐	0.7291±0.0112	0.7405±0.0123	0.7152±0.0184	4.35±0.25	22.98±2.64
MSM	☑	☐	☐	0.7386±0.0121	0.7549±0.0115	0.7228±0.0173	4.01±0.18	21.77±1.71
CMA	☐	☑	☐	0.7405±0.0101	0.7538±0.0099	0.7284±0.0112	3.79±0.26	18.23±1.64
GFM	☑	☐	☑	0.7369±0.0132	0.7443±0.0109	0.7415±0.0118	4.22±0.21	19.47±2.13
MSM+CMA	☑	☑	☐	0.7458±0.0137	0.7544±0.0172	0.7329±0.0201	3.77±0.19	17.43±1.65
Full Model	☑	☑	☑	0.7466±0.0118	0.7648±0.0145	0.7591±0.0178	3.62±0.17	16.35±1.55

Table 4 highlights the contributions of each module to overall model performance, offering insight into their roles in improving segmentation accuracy and robustness. Below, we provide a detailed analysis of each configuration:

- **Baseline**: The Baseline model excludes multi-scale features and advanced cross-modal fusion mechanisms. Instead, it employs feature map concatenation and convolution layers for feature fusion across PET and CT modalities. This design limits the model's ability to leverage complementary information from the two modalities, resulting in suboptimal performance (DSC: 0.7291, ASD: 4.35 mm). The achieved HD95 (22.98 mm) indicates difficulties in accurately delineating complex tumor boundaries using the baseline model, especially in regions with blurred edges or low contrast.
- **MSM**: Based on the baseline model, we added the MSM to extract the multi-scale features through convolutional layers with varying kernel sizes. However, the feature fusion across modalities and scales still relies on feature map concatenation and convolution layers. This addition improves DSC and sensitivity (0.7386 and 0.7549, respectively), highlighting the importance of multi-scale feature aggregation in capturing fine-grained tumor details. Nevertheless, the modest improvement in precision (0.7228) suggests that the increased feature complexity may amplify noise in certain regions, resulting in false positives.
- **CMA**: Based on the baseline model, the CMA model employs the cross-modal attention mechanism for feature fusion between PET and CT modalities. This mechanism aligns spatial features from CT with metabolic features from PET, improving boundary delineation and enhancing the integration of complementary information. As a result, the CMA model significantly reduced the ASD (3.79 mm) and HD95 (18.23 mm) compared to baseline and MSM models, while the improved the DSC to 0.7405. However, the limited gain in precision (0.7284) indicates that challenges remain in handling low-contrast or heterogeneous regions.
- **GFM**: By adding the GFM individually to baseline model, we replaced the concatenation and convolution operation for multi-scale feature fusion with a gated fusion mechanism. This allows the model to dynamically adjust feature contributions from different scales, selectively emphasizing high-confidence features and suppressing noise. While GFM showed limited improvement in DSC (0.7369) and sensitivity (0.7443), the precision was increased notably to 0.7415, reflecting the module's effectiveness in mitigating false positives. However, the slightly higher ASD (4.22 mm) and HD95 (19.47 mm) suggested a focus on local optimization at the expense of global consistency.

- **MSM+CMA**: Combining MSM and CMA introduces multi-scale features with cross-modal attention for modality fusion but retains concatenation and convolution for multi-scale feature integration. This configuration achieved the highest DSC (0.7458) among other baseline models, reflecting the complementary roles of multi-scale and cross-modal mechanisms. The sensitivity (0.7544) was also improved significantly, but the achieved precision remained moderate (0.7329) due to the lack of dynamic scale weighting. The reduced HD95 (17.43 mm) indicated better handling of outliers and boundary refinement.

- **Full Model**: The Full Model integrates all three components, combining multi-scale features, cross-modal attention, and gated fusion. This configuration achieved the best overall performance (DSC: 0.7466, ASD: 3.62 mm, HD95: 16.35 mm). The integration of GFM with MSM and CMA allows for robust feature selection and fusion, balancing local and global segmentation challenges. The significant improvement in precision (0.7591) confirmed the effectiveness of GFM in addressing false positives, while MSF and CMA enhance sensitivity and boundary delineation. These results demonstrate the comprehensive capabilities of the Full Model in extracting lymphoma lesions with high accuracy and robustness using CT and PET images.

12. Additionally, discussing other studies related to the comments made above, as well as peculiarities in the network or parameters, and elaborating on the topics mentioned in Section 4.3 limitations, such as generalization and plans for multi-center studies, would enhance the quality of the paper.

Reply: Thank you for your valuable suggestion. We have added [Section 4.2] to provide a detailed discussion of related studies, highlighting the advantages and limitations of current methods in comparison to ours. Additionally, we expanded [Section 4.3] to address generalization challenges and interpretability, proposing federated learning and explainable AI as potential future directions.

(Section 4.2 Comparison with Existing TMTV Calculation Methods)

Compared to existing techniques for TMTV segmentation and quantification, our method offers clear advantages in practicality, automation, and accuracy. While Yousefirizi et al.[37,38] and Blanc-Durand et al.[14] tested their approaches on single-center or multi-center datasets, they lacked validation on publicly available datasets, limiting the generalizability of their methods. In contrast, our approach was rigorously evaluated on both a private dataset and the autoPET dataset, demonstrating consistent performance across varied imaging protocols. This dual validation underscores the robustness of our method in diverse clinical settings.

In terms of accuracy, Yousefirizi et al.[38] achieved a Pearson correlation coefficient of R^2=0.83 for TMTV quantification but reported lower segmentation accuracy (DSC=0.68) due to single-modality constraints. Similarly, Blanc-Durand et al.[14] achieved a DSC of 0.73 but faced significant TMTV underestimation on the external validation dataset. By leveraging multi-scale and cross-modal feature fusion, our method effectively integrates PET and CT information, achieving superior segmentation accuracy with DSC values of 0.7512 on the private dataset and 0.7441 on the autoPET dataset.

Furthermore, the semi-automated workflows proposed by Burggraaff et al.[39] required manual threshold adjustments, leading to time-consuming processes prone to variability. In contrast, our fully automated pipeline eliminates the need for manual intervention, ensuring consistent, reproducible results while significantly reducing analysis time. These features make our method particularly suitable for routine clinical applications.

In conclusion, our method combines automation, accuracy, and generalizability to provide a practical and efficient solution for TMTV segmentation and quantification, supporting both research and clinical workflows.

(Section 4.3 Limitations)

Although our model demonstrates strong performance in segmenting DLBCL lesions in PET/CT images, its effectiveness is highly dependent on the quality and consistency of the training data. Variations in imaging protocols and equipment across different centers or institutions may introduce significant data variability, which could limit the model's generalizability. Federated learning offers a promising solution by enabling model training on distributed data sources without the need to directly share sensitive patient information. This approach helps mitigate the impact of inter-center variability and enhances the model's robustness across heterogeneous datasets.

Another limitation of the current model is its interpretability, particularly in understanding how it integrates multimodal information (e.g., PET and CT images) for segmentation decisions. Understanding the model's decision-making process is crucial for its clinical adoption. In this context, explainable AI techniques, such as uncertainty quantification, can play an essential role. By quantifying the uncertainty in model predictions, clinicians can receive more reliable guidance, enabling more informed clinical decision-making.

Minor comments:

1. Please revise the reference list to accurately match the required format.

Reply: Thank you for pointing this out. We have carefully revised the reference list to ensure it aligns with the required format.

2. In the second paragraph of the introduction, the last sentence is a redundant expression of the same content as the last sentence of the third paragraph. Kindly revise this to avoid repetition.

Reply: Thank you for highlighting this redundancy. We have revised the introduction to eliminate repetition and improve clarity.

(Section 1 Introduction)

Moreover, total metabolic tumor volume (TMTV), which quantifies the metabolic activity of tumors, is a key prognostic biomarker for DLBCL[8]. Accurate lymphoma segmentation is essential for determining TMTV, but manual delineation is both time-consuming and subjective. Traditional lymphoma segmentation methods, such as thresholding and region growing, have inherent limitations. Thresholding, while straightforward, lacks adaptability[9], especially when image conditions cause lymphoma and normal tissue to appear with similar gray values. Region growing is highly dependent on initial seed points[10], which require careful selection to handle the diverse shapes and sizes of lymphoma. Recent advances in deep learning have led to the development of automated segmentation methods, providing greater consistency and accuracy[11].

Li et al.[12] proposed an end-to-end network for semi-supervised lymphoma segmentation, achieving a Dice similarity coefficient (DSC) of 0.72 using PET/CT data from 80 lymphoma cases. Yuan et al.[13] introduced a dual-branch encoder network for lymphoma segmentation, achieving a DSC of 0.73 on 45 DLBCL patients. Blanc-Durand et al.[14] achieved a DSC of 0.73 using their 3D U-Net, trained and validated on PET/CT data from 639 DLBCL patients, with 94 cases reserved for testing. Yousefirizi et al.[15] proposed a cascaded approach for automated tumor delineation in lymphoma involving 1418 PET/CT scans from multiple centers. This approach combined multi-resolution 3D U-Nets and model ensembling, achieving an average DSC of 0.68 on internal test data and 0.66 on external multi-site data. However, existing convolutional neural networks (CNNs) face challenges in fully leveraging multimodal PET/CT data, particularly due to their limited receptive field, which restricts their ability to capture global and local contextual information[16-18].

3. For Figure 4, improve readability by aligning the order of the plots with the table.

Reply: Thank you for the suggestion. We have revised Fig. 4 in [Section 3.1] to ensure the order of the plots aligns with the corresponding table.

(Section 3.1 Results of segmentation)

Table 1: Results of different methods on the private dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.7107±0.0178 **	0.7608±0.0128	0.6686±0.0298 **	4.10±0.20 **	18.05±2.36
SegResNet	0.7223±0.0146 *	0.7175±0.0466	0.7289±0.0125 **	4.61±0.26 **	21.01±0.69 **
Swin-UnetR	0.7271±0.0163 *	**0.7659±0.0123**	0.7041±0.0246 **	3.92±0.22 *	15.74±0.98
SwinCross	0.7414±0.0209	0.7405±0.0213	0.7432±0.0176	4.04±0.22 **	16.82±1.51
Unet++	0.7446±0.0129	0.7322±0.0072 **	0.7577±0.0137	4.21±0.09 **	18.05±1.51 **
Att-Unet	0.7463±0.0113	0.7622±0.0075	0.7314±0.0179 *	4.75±0.04 **	17.16±2.26
Ours	**0.7512±0.0078**	0.7548±0.0063	**0.7611±0.0078**	**3.61±0.11**	**15.20±0.78**

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates $p < 0.05$, and ** indicates $p < 0.01$ when compared to our method.

Table 2: Results of different methods on the autoPET dataset for lymphoma segmentation.

Method	DSC ↑	Sensitivity ↑	Precision ↑	ASD (mm) ↓	HD95 (mm) ↓
UnetR	0.6865±0.0478 **	0. 6924±0.0812	0.6851±0.0404 *	6.65±0.73	22.38±1.95 *
SegResNet	0.6740±0.0412 *	0.6951±0.0627	0.6483±0.0505 *	6.12±1.02	21.26±1.29
Swin-UnetR	0.7282±0.0605	0.7311±0.0833	0.7274±0.0450	5.40±0.92	**19.08±2.63**
SwinCross	0.7267±0.0146 **	0.7382±0.0717 *	0.7233±0.0525	6.40±1.48	23.37±2.95
Unet++	0.7302±0.0192	0.7424±0.0818	0.7277±0.0523	**5.11±0.92**	19.92±1.59
Att-Unet	0.6941±0.0261 **	0.7016±0.0657	0.6917±0.0401 **	6.17±1.04	21.29±1.25
Ours	**0.7441±0.0241**	**0.7573±0.0874**	**0.7427±0.0647**	5.83±1.18	21.27±1.44

The best metric is shown in bold. Statistical significance was assessed using paired t-tests across all metrics, where * indicates $p < 0.05$, and ** indicates $p < 0.01$ when compared to our method.

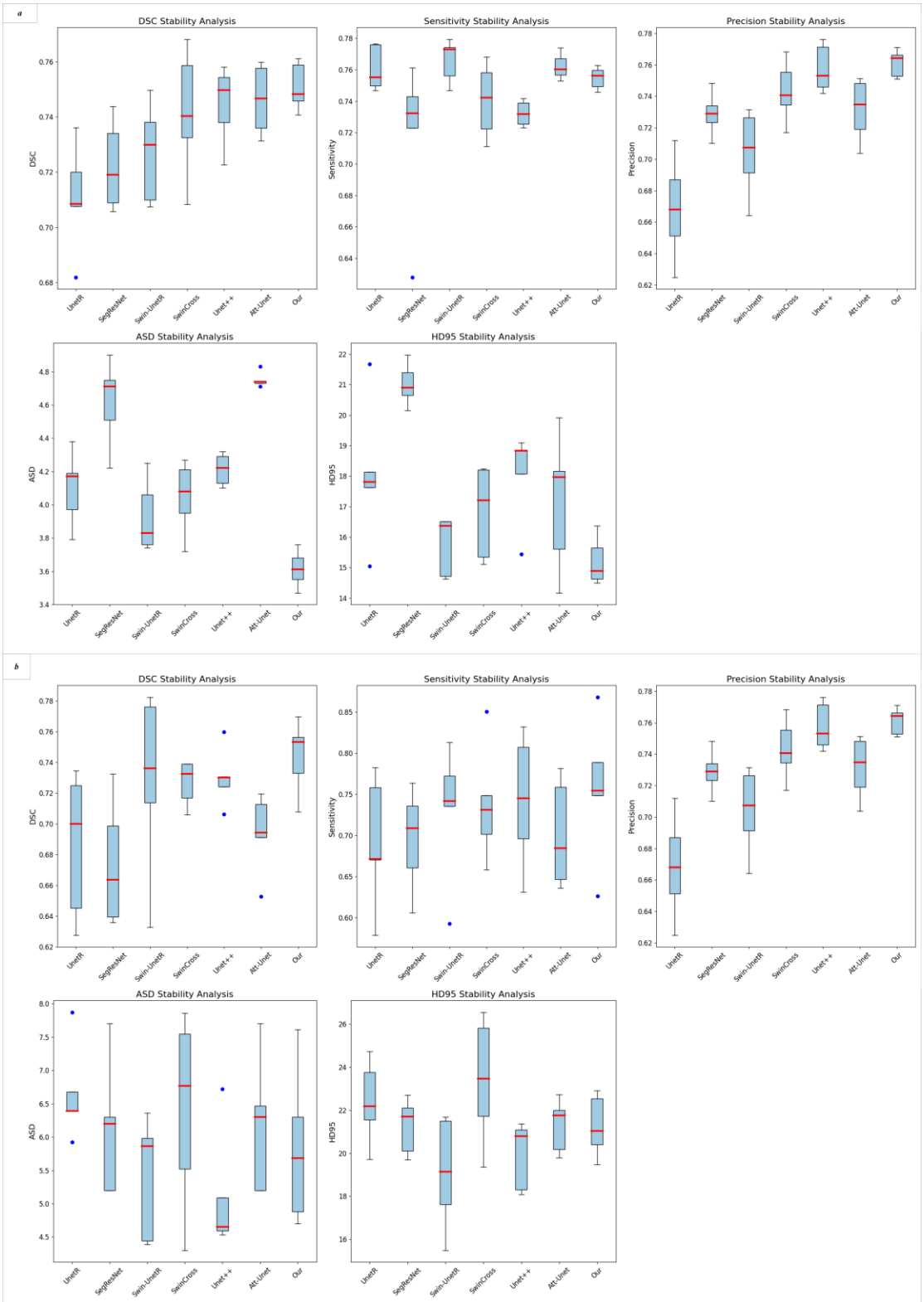
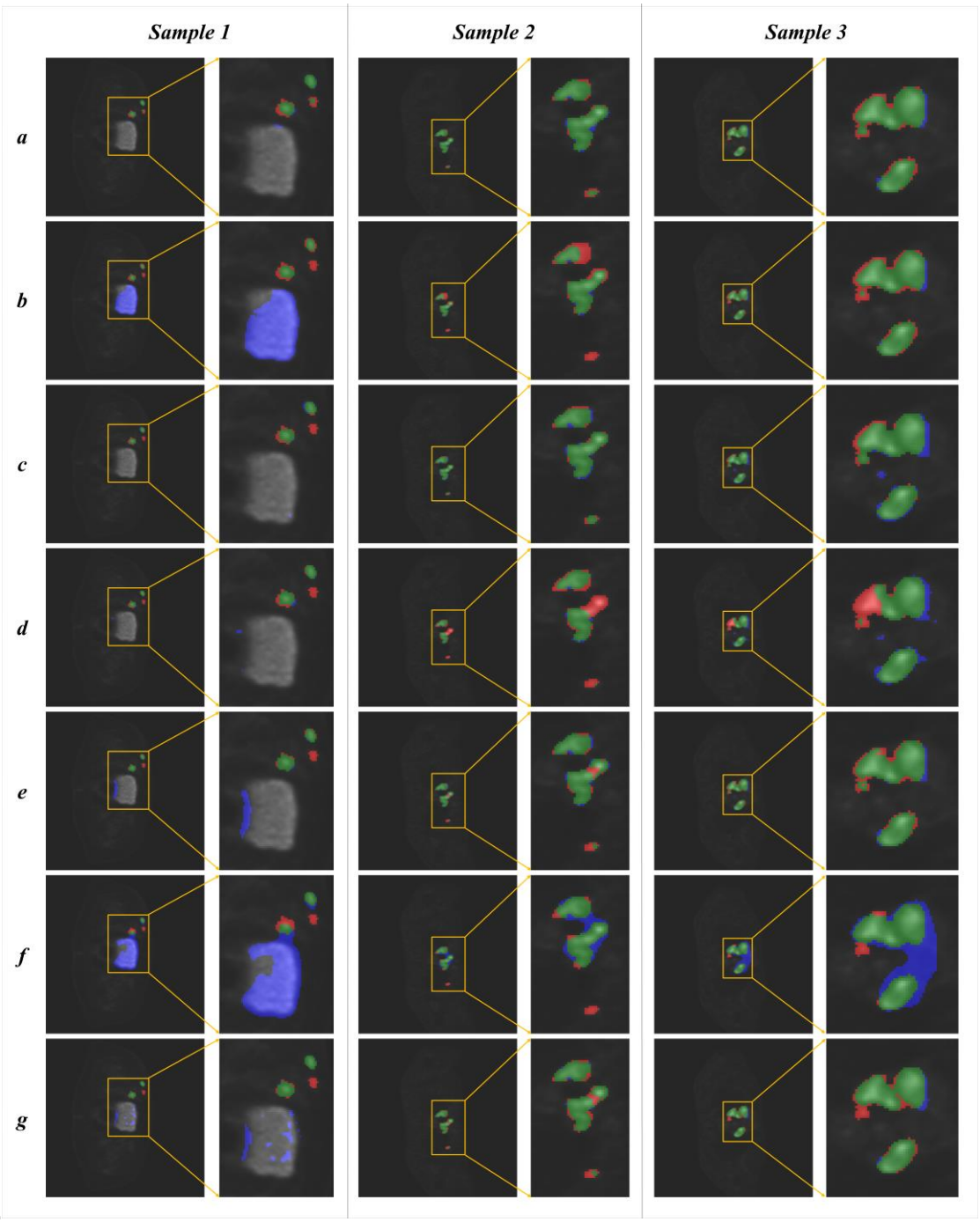


Fig. 4. Stability analysis using box plots for the private and autoPET datasets: This Fig. presents the ranges of DSC, sensitivity, precision, ASD and HD95 across five cross-validation folds for different models. Subfigure (a) displays the results on the private dataset, while subfigure (b) shows the corresponding results on the autoPET dataset. The box shows the 1st quartile (lower boundary), median (red line), and 3rd quartile (upper boundary). The whiskers represent the range of data, excluding outliers, which are marked as blue dots.

4. In Figure 5, please use arrows or other pointing elements to highlight the emphasized regions.

Reply: Thank you for your valuable suggestion. To address this issue, we have replaced the original Fig. 5 with two new figures (Fig. 5 and Fig. 6), which provide a more detailed and informative visual representation.



.

Fig. 5. Difference maps of segmentation results compared with ground truth for private datasets. The green, red, and blue regions represent true positive, false negative, and false positive pixels, respectively. Subfigures (***a***)–(***g***) show results generated by our method, Att-Unet, Unet++, SwinCross, Swin-UnetR, SegResNet, and UnetR, respectively.

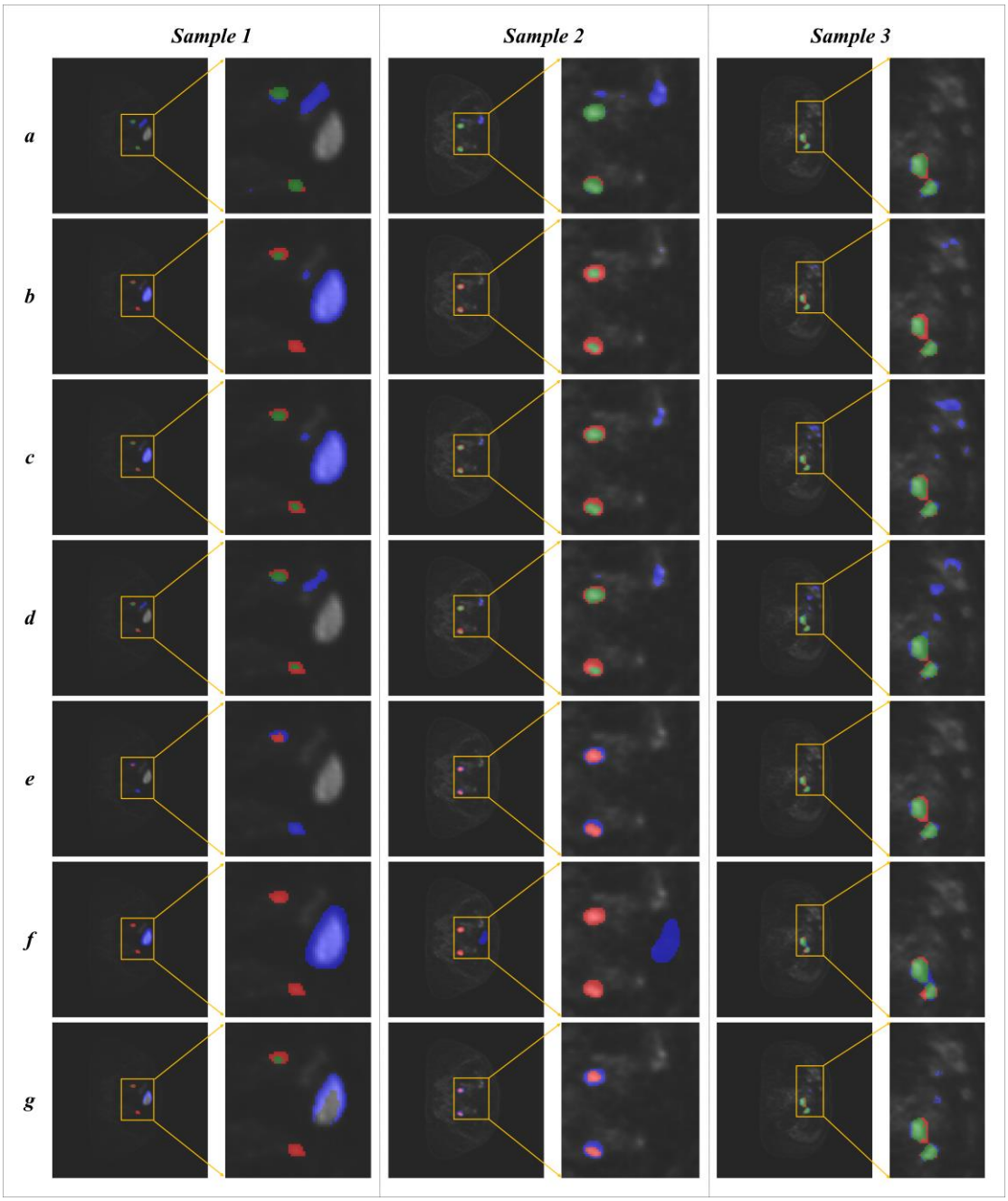


Fig. 6. Difference maps of segmentation results compared with ground truth for autoPET datasets. The green, red, and blue regions represent true positive, false negative, and false positive pixels, respectively. Subfigures (***a***)–(***g***) show results generated by our method, Att-Unet, Unet++, SwinCross, Swin-UnetR, SegResNet, and UnetR, respectively.

5. Correct the typo "gtTMTV" in the caption of Fig. 7.

Reply: Thank you for pointing this out. This Fig. has been corrected in the revised manuscript.

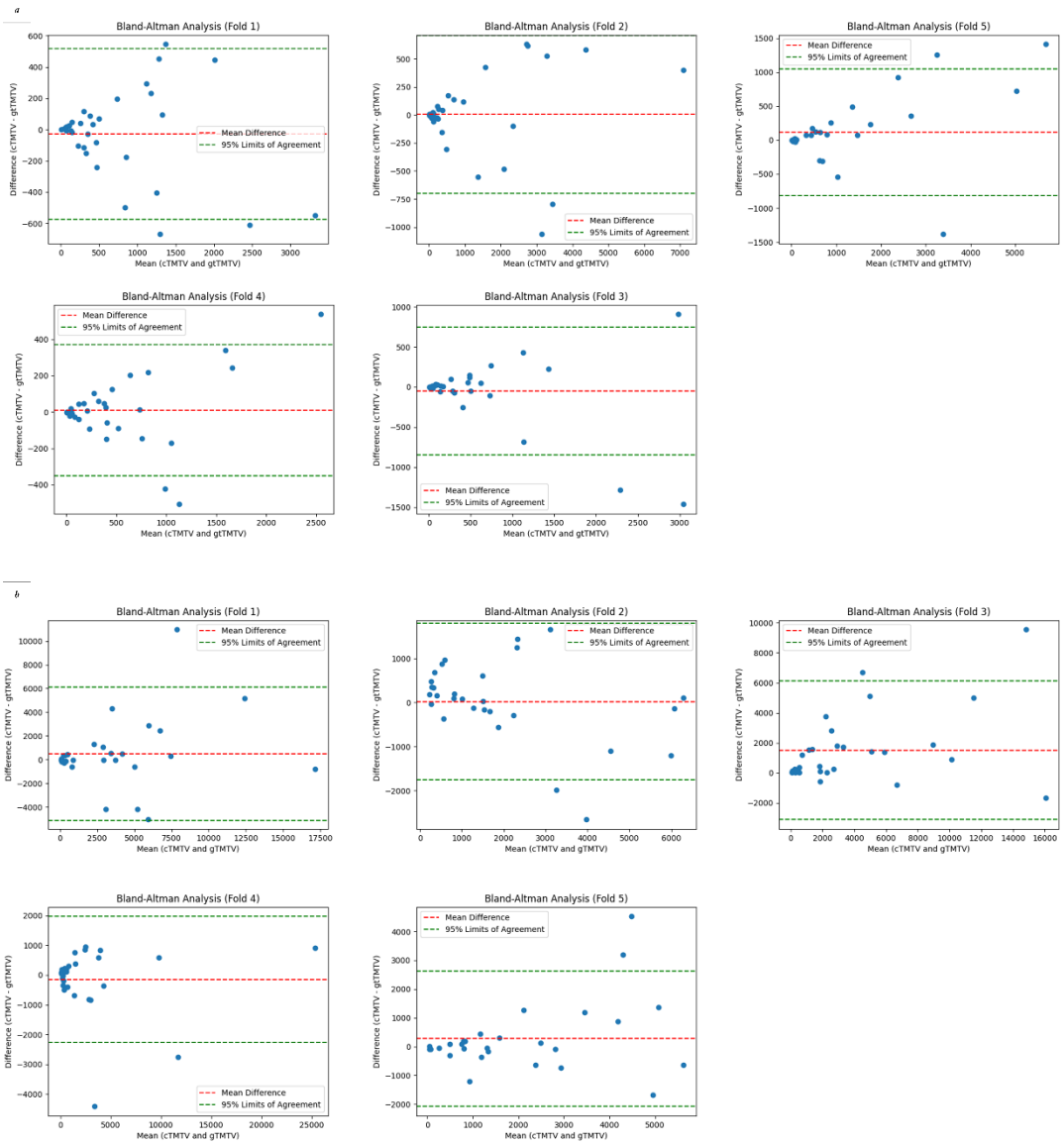


Fig. 9. Bland-Altman analysis for cTMTV vs. gtTMTV on (a) private dataset and (b) autoPET dataset. The horizontal axis represents the mean of cTMTV and gtTMTV, while the vertical axis represents their difference. The red dashed line shows the mean difference, and the green dashed lines represent the 95% limits of agreement, calculated as the mean difference ± 1.96 standard deviations of the differences.

6. In Eq. 6 and Eq. 8, "(W-MSA)" and "(SW-MSA)" seem to be used as operators, but they may be misunderstood as a minus sign. Consider connecting them with an underscore or assigning a different designation to avoid confusion.

Reply: Thank you for your observation. We have addressed this issue in the revised manuscript by modifying the notation for "(W-MSA)" and "(SW-MSA)" to avoid any potential confusion with a minus sign.

(Section 2.3.2 Multi-Scale Information Fusion)

The outputs for layers l and $l+1$ are computed using Eqs. 7 to 10:

$$\hat{A}^{l}_{modal\_n} = W\_MSA\left(LN\left(A^{l-1}_{modal\_n}\right)\right) + \hat{A}^{l-1}_{modal\_n} \quad (7)$$

$$A^{l}_{modal\_n} = MLP\left(LN\left(\hat{A}^{l}_{modal\_n}\right)\right) + \hat{A}^{l}_{modal\_n} \quad (8)$$

$$\hat{A}^{l+1}_{modal\_n} = SW\_MSA\left(LN\left(A^{l}_{modal\_n}\right)\right) + A^{l}_{modal\_n} \quad (9)$$

$$A^{l+1}_{modal\_n} = MLP\left(LN\left(\hat{A}^{l+1}_{modal\_n}\right)\right) + \hat{A}^{l+1}_{modal\_n} \quad (10)$$

In these equations, W_MSA and SW_MSA stand for regular and shifted window multi-head self-attention modules, respectively.

Reference:

1. Gatidis S, Kuestner T. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions) [Dataset]. *The Cancer Imaging Archive*. 2022. doi: 10.7937/gkr0-xv29.
2. Hatamizadeh A, Tang Y, Nath V, et al. Unetr: Transformers for 3d medical image segmentation. Paper presented at: Proceedings of the IEEE/CVF winter conference on applications of computer vision2022.
3. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. Paper presented at: International MICCAI brainlesion workshop2021.
4. Wang S, Li L, Zhuang X. AttU-Net: attention U-Net for brain tumor segmentation. Paper presented at: International MICCAI brainlesion workshop2021.
5. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. Paper presented at: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 42018.
6. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. Paper presented at: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 42019.
7. Li GY, Chen J, Jang SI, Gong K, Li Q. SwinCross: Cross-modal Swin transformer for head-and-neck tumor segmentation in PET/CT images. *Medical physics.* 2024;51(3):2096-2107.
8. Yousefirizi F, Ahamed S, Bloise I, Saboury B, Rahmim A. Semi-supervised and unsupervised convolutional neural networks for automated lesion segmentation in PET imaging of lymphoma. In.: Soc Nuclear Med; 2022.
9. Yousefirizi F, Jha A, Ahamed S, et al. A novel loss function for improved deep learning-based segmentation: implications for TMTV computation. In.: Soc Nuclear Med; 2022.
10. Blanc-Durand P, Simon.Jégou, Kanoun S, et al. Fully automatic segmentation of diffuse large B cell lymphoma lesions on 3D FDG-PET/CT for total metabolic tumour volume prediction using a convolutional neural network. *European journal of nuclear medicine and molecular imaging.* 2021;48(5):1362-1370.
11. Burggraaff CN, Rahman F, Kaßner I, et al. Optimizing workflows for fast and reliable metabolic tumor volume measurements in diffuse large B cell lymphoma. *Molecular Imaging and Biology.* 2020;22:1102-1110.
12. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems.* 2016;29.
13. Jacobsen J-H, Van Gemert J, Lou Z, Smeulders AW. Structured receptive fields in cnns. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2016.
14. Chen J, Lu Y, Yu Q, Luo X, Zhou Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. 2021.
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020.
16. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Paper presented at: Proceedings of the IEEE/CVF international conference on computer vision2021.
17. Fu J, Li W, Peng X, et al. MDRANet: A multiscale dense residual attention network for magnetic resonance and nuclear medicine image fusion. *Biomedical Signal Processing and Control*. 2023;80:104382.
18. Wang X, Li Z, Huang Y, Jiao Y. Multimodal medical image segmentation using multi-scale context-aware network. *Neurocomputing.* 2022;486:135-146.

19. Hou L, Yan Z, Desrosiers C, Liu H. MFCPNet: Real time medical image segmentation network via multi-scale feature fusion and channel pruning. *Biomedical Signal Processing and Control.* 2025;100:107074.
20. Vaswani A. Attention is all you need. *Advances in Neural Information Processing Systems.* 2017.
21. Hasani N, Paravastu SS, Farhadi F, et al. Artificial intelligence in lymphoma PET imaging: a scoping review (current trends and future directions). *PET clinics.* 2022;17(1):145-174.
22. Hellwig D, Graeter TP, Ukena D, et al. 18F-FDG PET for mediastinal staging of lung cancer: which SUV threshold makes sense? *Journal of Nuclear Medicine.* 2007;48(11):1761-1766.
23. Hatt M, Lee JA, Schmidtlein CR, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. *Medical physics.* 2017;44(6):e1-e42.
24. Huang L, Denoeux T, Tonnelet D, Decazes P, Ruan S. Deep PET/CT fusion with Dempster-Shafer theory for lymphoma segmentation. 2021.
25. Li H, Jiang H, Li S, et al. DenseX-net: an end-to-end model for lymphoma segmentation in whole-body PET/CT images. *Ieee Access.* 2019;8:8004-8018.
26. Yuan C, Zhang M, Huang X, Xie W, Qian D. Diffuse Large B 丶ell Lymphoma Segmentation in PET Images via Hybrid Learning for Feature Fusion. *Medical Physics.* 2021;48(7).
27. Yousefirizi F, Klyuzhin IS, O JH, et al. TMTV-Net: fully automated total metabolic tumor volume segmentation in lymphoma PET/CT images—a multi-center generalizability analysis. *European Journal of Nuclear Medicine and Molecular Imaging.* 2024.1-18.