# Enhancing 3D Lesion Segmentation in PET-CT Scans Using MIP-Segmentation-Reconstruction Techniques

Wenyuan Chen
Stanford University
wenyuanc@stanford.edu

Thodsawit Tiyarattanachai
Stanford University
ttiya@stanford.edu

Jirayu Burapacheep
Stanford University
jirayu@stanford.edu

## Abstract

*Accurate segmentation of lesions in Positron Emission Tomography (PET) and Computed Tomography (CT) scans benefits cancer diagnosis and treatment planning. However, this task faces significant challenges due to the limited availability of annotated data. In this work, we propose a novel framework that leverages Maximum Intensity Projection (MIP) images to improve the segmentation performance of 3D PET-CT scans. Our approach combines a 2D MIP segmentation model, trained on synthetic data generated by a diffusion model, with a 3D multi-modal refinement model that integrates information from PET, CT, and the reconstructed MIP segmentation mask. Through extensive experiments on a dataset of whole-body PET-CT scans, we demonstrate that our method outperforms the baseline by achieving higher DICE scores and lower false positive volumes, indicating improved overall segmentation performance. We provide insights into the key components of our framework and discuss strategies to mitigate failure cases. Our work contributes to the development of more accurate and reliable lesion segmentation techniques, ultimately enhancing cancer diagnosis and treatment planning.*

## 1. Introduction

Positron Emission Tomography (PET) is a medical imaging modality that enables the detection of abnormal metabolic activities in cancer lesions. In clinical practice, PET scans are performed in conjunction with Computed Tomography (CT) scans, referred to as PET-CT, to align PET signals with anatomical locations on CT images. This alignment allows for precise localization of metabolic abnormalities and differentiation of true lesions from false positives, such as organs with naturally high metabolic activity.

Machine learning approaches have been introduced to address the challenges of lesion segmentation in PET scans [8, 10], where the goal is to predict the presence or absence of a lesion at every location in the 3D image. However, the segmentation of abnormal lesions in PET scans poses significant challenges due to the inherent noise in the PET images and the limited availability of PET-CT data. Furthermore, the limited study of enhancement approaches, including a synthetic generation of 3D medical images, further compounds the difficulties in developing accurate and reliable lesion segmentation techniques.

To address the limitations in the 3D segmentation of PET-CT scans, we propose a novel framework that leverages the nnUNet model architecture, designed for 3D medical image segmentation tasks, as the backbone. Our approach extends the vanilla nnUNet by introducing a Maximum Intensity Projection (MIP) module to improve the learning capability of the model by incorporating 2D segmentation results. The MIP module generates 2D maximum intensity projections along the axial, coronal, and sagittal axes from the 3D PET scan, which are then segmented using a 2D segmentation model trained on synthetic data generated by diffusion models. The segmentation results from the 2D projections are combined back into a 3D mask, which is then utilized as an additional input modality, along with the PET and CT scans, to the nnUNet model for 3D segmentation. Through this framework, we aim to overcome the challenges associated with limited PET-CT data, improve the accuracy and reliability of lesion segmentation, and remove organs that naturally exhibit high metabolic activity on PET scans but are not indicative of cancer lesions, thereby enhancing the overall reliability of the segmentation process.

## 2. Related Works

**Lesion segmentation in PET and PET-CT scans.** In the field of medical imaging, PET/CT segmentation tasks play a crucial role in diagnosis and treatment planning. However, as Prevedello et al. [8] highlight, the availability of high-quality, annotated whole-body PET/CT data is often limited, posing significant challenges for developing robust diagnostic and segmentation algorithms. Furthering the development of segmentation models, our approach modifies the 2D and 3D U-Net frameworks, which Ronneberger et al. [10] established as the current de facto standard for precise biomedical segmentation. We adapt these models to focus on the 2D MIP space, which, despite its potential, has been under-explored in the field as noted by Toosi et al. [1] and Girum et al. [4]. These sources point out that nuclear medicine physicians often use 2D PET MIP views for visual interpretation due to the noisy nature of 3D volumetric PET data. Leveraging this insight, our project utilizes MIP views not only for visual interpretation but also for synthetic image generation and segmentation tasks, addressing both computational efficiency and data scarcity.

**Synthetic data generation for medical imaging.** To overcome the scarcity of high-quality annotated PET/CT data, our project draws inspiration from the synthetic image generation techniques discussed by Chambon et al. [2] and Tanenbaum et al. [11]. These methods have successfully augmented data in various modalities such as chest X-ray and brain MRI, enhancing the robustness of machine learning models. Similarly, we aim to use synthetic 2D MIP images to augment our PET/CT data, navigating around the computational intensity and resource demands that Man et al. [6] indicate as barriers in generating synthetic 3D images. By incorporating synthetic data generation into our approach, we aim to improve the performance and robustness of our segmentation models, ultimately contributing to more accurate and efficient medical imaging diagnostics.

## 3. Dataset

We used a dataset of 1,014 whole-body FDG-PET/CT scans collected at the University Hospital Tübingen [3] between 2014 and 2018. The dataset comprises 1,014 imaging studies, consisting of 501 scans from patients with confirmed malignant lesions, including malignant lymphoma (145), melanoma (188), and non-small cell lung cancer (NSCLC) (168), as well as 513 negative control scans from patients without PET-positive malignant lesions. Each study contains CT, PET, and segmentation files, totaling 916,957 individual slices. The files are provided in the processed NIfTI format [7], commonly used to store brain imaging data. Each file has dimensions of $400 \times 400 \times z$, where $z$ represents the number of slices, varying based on the patient and the machine's configuration. We randomly split the dataset into train/test sets with an 80/20 ratio, ensuring a similar distribution of disease types and negative studies in each split. The splitting process was done at the patient level, such that studies from the same patient were always in the same set. The inclusion of both positive and negative cases provides a representative dataset for developing and evaluating lesion segmentation models. See Figure 1 for an example of the imaging study.

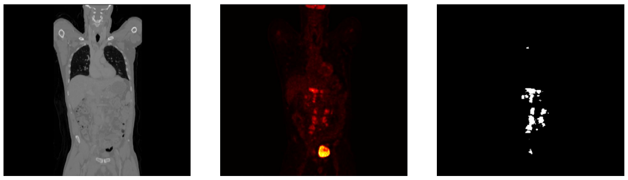


Figure 1: An example of coronal slice from CT (left), PET (middle), and segmentation (right).

For data normalization, we mainly employ two methods from nnUNet: CT normalization for CT modality and Zscore normalization for the others. CT normalization collects intensity values from the foreground class from all training cases and computes the mean and standard deviation. It then clips every intensity value of every training case with the 0.5 and 99.5 percentiles of all values. Each clipped value is followed by a subtraction of the mean and division with the standard deviation. Zscore normalization is applied to each training case separately, where each value is subtracted by its mean and divided by its standard deviation. The train/test splits we used for all training steps are the same 80/20 splits at a patient level, ensuring no patient leakage and similar disease distribution between splits.

## 4. Methodology

In this section, we present an overview of our framework for improving lesion segmentation in PET-CT scans through MIPs. The proposed methodology consists of two main components: MIP Segmentation (Section 4.1) and Multimodal Refinement model (Section 4.2). The MIP Segmentation focuses on generating a 3D segmentation mask by projecting the PET scan into 2D MIP images, performing segmentation on each MIP, and reconstructing the results back into a 3D mask. The Segmentation Model refines the 3D segmentation mask by incorporating information from the corresponding PET and CT scans.

### 4.1. MIP Segmentation

The MIP Segmentation component aims to generate a 3D segmentation mask by leveraging the information from 2D MIP images. The key insight behind this approach is that with an adequate number of 2D MIP projections, reconstruction of the 3D segmentation mask can be done accurately.

**MIP generation.** The PET scan is transformed into multiple 2D MIP images from different angles, capturing the maximum intensity values along various projection planes (Figure 3). To create the MIP images, the 3D PET scan is rotated by different degrees around a rotation axis and projected onto a perpendicular axis, selecting the maximum

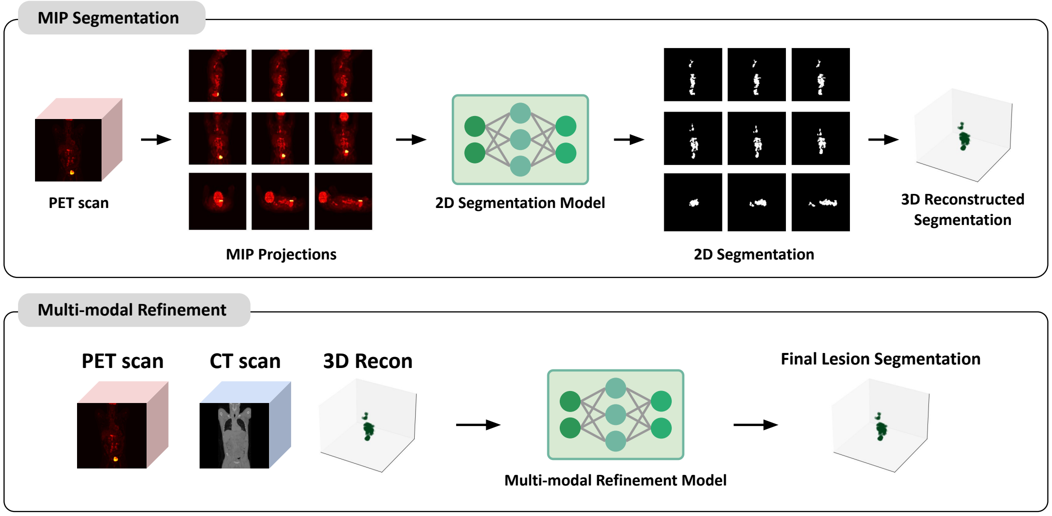


Figure 2: Proposed framework for improving 3D lesion segmentation in PET-CT scans using 2D synthetic data. The process involves projecting the 3D PET scan into 2D MIP images, segmenting each 2D image, and combining the results into a 3D mask, refined with corresponding PET and CT scans.

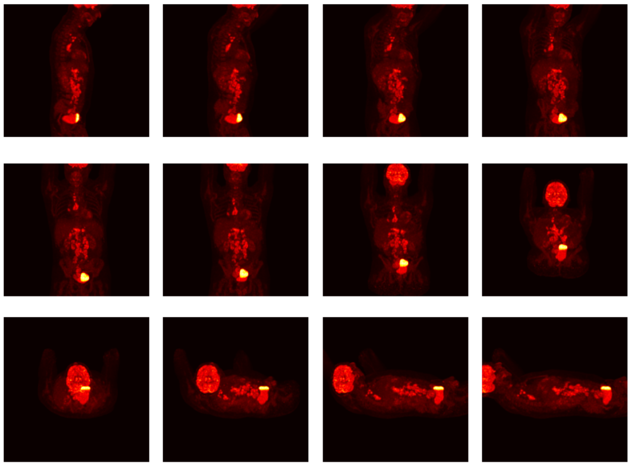


Figure 3: MIPs along different projection axes (different rows). For each projection axis, the 3D PET scan is rotated by different degrees (different columns) before applying maximum intensity projection.

intensity value. This process is repeated for the $x$, $y$, and $z$ projection axes. A hyperparameter, degree increment, determines the number of 2D MIPs; smaller increments yield more MIPs, resulting in a more accurate 3D reconstruction of the segmentation mask. Let $R$ be the number of rotations per each projection axis. The total number of MIPs for each 3D PET scan is $3R$.

**2D segmentation.** A 2D segmentation model is used to segment the lesions in each MIP image. The model is designed to accurately identify lesion boundaries within the MIP images. The input of the model is the extracted MIP, and the prediction ground truth is the projected lesion mask at the same angle and rotation axis. We employ the nnUNet framework [5] to train this 2D segmentation model. nnUNet is an automatic framework for UNet-based segmentation, renowned for its robustness and flexibility in medical imaging tasks.

The 2D U-Net architecture [10] with 7 stages for both downsampling and upsampling. The number of features per stage is defined as [32, 64, 128, 256, 512, 512, 512], with the number of features increasing in the downsampling path and decreasing in the upsampling path. Each stage consists of two convolutional layers with a $3 \times 3 \times 3$ kernel size. The strides for the convolutional layers use a stride of 1 for the first stage and a stride of 2 for the remaining stages to perform spatial downsampling. The model is optimized using the Adam optimizer with an initial learning rate of 0.01 and a linear learning rate decay schedule. The training is performed for 1,000 epochs with a batch size of 2 and leaky ReLU activation. We adopt the same training loss as described in [5], where the network is trained with a combination of DICE and cross-entropy loss:

$$L_{\text{total}} = L_{\text{DICE}} + L_{\text{CE}} \tag{1}$$

$$L_{\text{DICE}} = -2 \cdot \frac{\sum_{i \in I} u_i v_i}{\sum_{i \in I} u_i + \sum_{i \in I} v_i} \tag{2}$$

$$L_{\text{CE}} = -\sum_{i \in I} v_i \log(u_i) \tag{3}$$

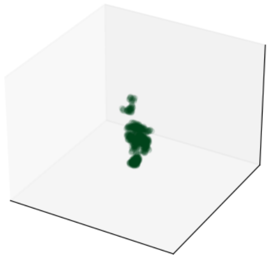


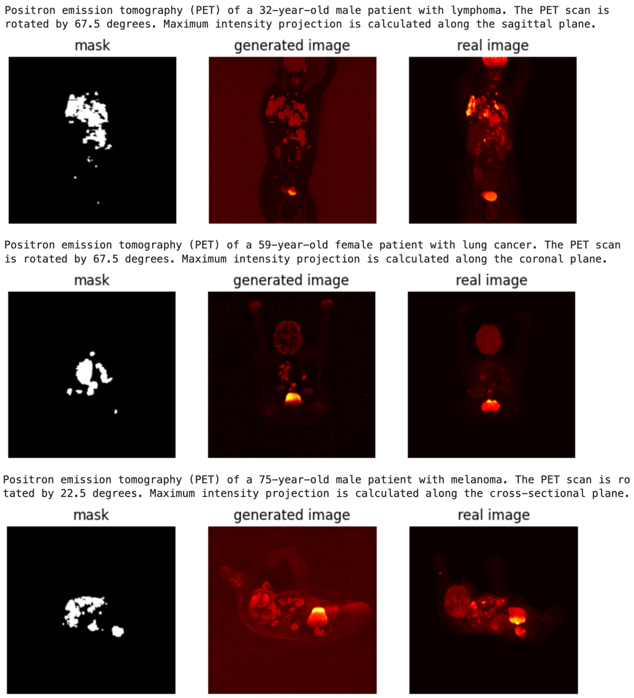
Figure 4: Ground truth (left) and reconstructed (right) 3D segmentation masks. To validate the performance of the reconstruction method, we generate MIPs from the 3D ground truth segmentation mask and then reconstruct the 3D segmentation mask from the MIPs. With 16 rotations per projection axis, the reconstruction method achieves a DICE score of 0.86 for this case, which has complex-shape segmentation.

where $u$ is the softmax output of the network, and $v$ is a one-hot encoding of the ground truth segmentation map.

**3D reconstruction.** The segmented MIP images are then reconstructed into a 3D segmentation mask. This involves aligning the segmented 2D masks back into their original 3D space. The final 3D segmentation mask is then obtained by taking the intersection of the projected masks, ensuring that only the regions consistently identified as lesions across all MIP views are included. See an example of reconstructed MIP of the ground truth in Figure 4. We discuss the effectiveness of 3D reconstruction in Section 5.2. We report the performance of 3D segmentation masks reconstructed from outputs of the 2D MIP Segmentation model (before going through the Refinement Model) in Table 1, denoted as MIP Seg.

**Improving 2D segmentation model using synthetic MIP images.** To further improve the segmentation performance, we introduce synthetic training data in the MIP space. By generating synthetic MIP images with corresponding segmentation masks, we can augment the training dataset and aim to enhance the performance of the segmentation model. We employed ControlNet [12], a model based on latent diffusion, to generate synthetic MIP images. The model was trained to generate MIP images conditioned on 1) a text description of patient age, sex, cancer diagnosis, MIP rotation angle and projection axis, and 2) a 2D lesion mask that was projected by the same method as MIP. Using these two inputs, the model was conditioned to generate MIP images that align with the text description and contain lesion(s) on the locations indicated by the mask.

Briefly, ControlNet adopts layers and pretrained weights from Stable Diffusion v1.5 [9], and makes a copy of these layers for receiving an additional Control image. The layers in the original copy are locked and used for receiving text prompts, while the additional copy (which receives the control image) is trainable. In our case, the control image is the lesion mask. Features from both branches are added to generate output images conditioned by the text prompt and control image.



Positron emission tomography (PET) of a 32-year-old male patient with lymphoma. The PET scan is rotated by 67.5 degrees. Maximum intensity projection is calculated along the sagittal plane.

Positron emission tomography (PET) of a 59-year-old female patient with lung cancer. The PET scan is rotated by 67.5 degrees. Maximum intensity projection is calculated along the coronal plane.

Positron emission tomography (PET) of a 75-year-old male patient with melanoma. The PET scan is rotated by 22.5 degrees. Maximum intensity projection is calculated along the cross-sectional plane.

Figure 5: Examples of generated synthetic MIP images (middle), along with input conditions including text prompts (above each row) and lesion masks (left). Real images are shown on the right.

We trained ControlNet using 9744 MIP images in the training set (812 studies × 3 projection axes × 4 rotations per axis). To facilitate training on a GPU with limited VRAM (Nvidia Tesla T4 GPU, 16 GB VRAM), we applied gradient accumulation by feeding 1 training instance into the model at each iteration, accumulating gradients over 4 iterations, and updating the model's parameters every 4 iterations. The model was trained for 10 epochs (total of $10 \times 9744/4 = 24360$ rounds of parameter updates) with a learning rate of $10^{-5}$. The MIP images and lesion masks were zero-padded to match ControlNet's mask and output image dimension of $512 \times 512$ pixels. During inference, the following settings were used: sampling steps $= 100$, control strength $= 1.0$ (range [0.0, 2.0]), guidance scale $=$ 9.0 (range [0.1, 30.0]). Examples of synthetic MIP images, along with input conditions, are shown in Figure 5.

The generated synthetic MIP images along with their corresponding lesion masks serve as additional training data for the 2D MIP segmentation model. These generated synthetic MIP images are then combined with the original MIP images in different proportions $p$ to train the 2D MIP segmentation model, leveraging the augmented data to improve its performance. We discuss the effectiveness of the synthetic MIP generation approach in Section 5.2.

### 4.2. Multi-modal Refinement Model

The multi-modal refinement model aims to refine the 3D segmentation mask by incorporating information from the corresponding PET and CT scans. The original PET scan highlights areas with high metabolic activity, which can indicate the presence of lesions, while the CT scan provides anatomical context and helps to remove false positives caused by organs with naturally high metabolic activity. By combining these two complementary imaging modalities, the Refinement Model can leverage the strengths of each modality to produce a more accurate and robust 3D segmentation mask. We employ a 3D nnUNet architecture for this component where the inputs are the PET scan, the CT scan, and the 3D segmentation mask obtained from the MIP Segmentation component.

To train the multi-modal refinement model, we used the 3D reconstruction of output segmentation from the 2D MIP segmentation model, with $p = 15\%$ (our best params). Together with 3D PET and CT scans, these images form the input data for our training process. CT scans were normalized by CT normalization. PET and reconstructed MIP segmentation were normalized by Zscore normalization. The model uses a 3D U-Net architecture [10] with 6 stages for both downsampling and upsampling. The number of features per stage is defined as [32, 64, 128, 256, 320, 320], with the number of features increasing in the downsampling path and decreasing in the upsampling path. Each stage consists of two convolutional layers with a $3 \times 3 \times 3$ kernel size. The strides for the convolutional layers use a stride of 1 for the first stage and a stride of 2 for the remaining stages to perform spatial downsampling. The training loss used in the model is described in Equation 1. The model is optimized using the Adam optimizer with an initial learning rate of 0.01 and a linear learning rate decay schedule. The training is performed for 1,000 epochs with a batch size of 2 and leaky ReLU activation.

## 5. Experiments

This section presents empirical experiments to evaluate the effectiveness of our proposed method. In particular, we aim to show that the inclusion of MIP can effectively improve the overall lesion segmentation performance and provide a discussion on the success and failure modes of our framework. Our code is released publicly for reproducible research at https://github.com/Top34051/lesion-segmentation-mip.

### 5.1. Experiment Details

**Baseline.** Our baseline is nnUNet receiving two modalities as inputs: PET and CT scans. This model is based on a 3D U-Net architecture and provides a comparison with our proposed method. The inputs were 3D PET volumes converted to SUV units and CT volumes of the same resolution. The training was performed with the train split (80% of patients) with a max epoch set to 1,000 and an initial learning rate of $10^{-4}$.

**Evaluation metrics.** We consider three metrics to evaluate the performance of the methods: DICE score, False Positive Volumes (FPV), and False Negative Volumes (FNV). While the DICE score is the default metric for the segmentation task, the latter two are also important in biomedical domain.

- **DICE score**: This metric is commonly adopted for evaluating segmentation models. The DICE score measures the overlap between the predicted segmentation and the ground truth segmentation. It is defined as:

$$\text{DICE} = \frac{2|A \cap B|}{|A| + |B|}$$

  where $A$ is the set of predicted lesion voxels and $B$ is the set of ground truth lesion voxels. DICE score is defined as 1 when $|A| + |B| = 0$. A higher DICE score indicates better overlap and, thus, better segmentation performance.

- **False Positive Volume (FPV)**: FPV measures the volume of non-lesion tissue that is incorrectly classified as lesion by the segmentation model. Specifically, FPV is defined as the volume of false positive connected components in the predicted segmentation mask that do not overlap with tumor regions in the ground truth segmentation mask. This can include areas of physiological FDG uptake (e.g., brain, heart, kidneys) that are erroneously classified as tumors.

- **False Negative Volume (FNV)**: FNV measures the volume of actual lesion tissue that is incorrectly classified as non-lesion by the segmentation model. Specifically, FNV is defined as the volume of connected components in the ground truth segmentation mask (i.e., tumor lesions) that do not overlap with the predicted segmentation mask. These are tumor lesions that are entirely missed by the segmentation model.

### 5.2. Results and Discussion

**Inclusion of MIP segmentation helps improve overall segmentation performance.** Table 1 shows that our proposed method, which incorporates MIP segmentation, significantly improves overall segmentation performance compared to the baseline methods. The inclusion of MIP segmentation helps the model to better understand the lesion boundaries by leveraging the maximum intensity projections from different angles. This multi-view information enhances the accuracy of the 3D segmentation, resulting in higher average DICE scores and lower false positive volume. See Figure 6 for a breakdown of segmentation performance.

Table 1: Comparison of segmentation performance with and without MIP segmentation. **Bold** values indicate improved performance from the baseline.

| **Method** | **Modalities** | | | **Metrics** | | |
|---|---|---|---|---|---|---|
| | PET | CT | MIP | DICE | FPV | FNV |
| Baseline | ✓ | ✓ | | 0.42 | 21.11 | 6.66 |
| MIP Seg. | ✓ | | ✓ | **0.64** | **1.04** | 39.62 |
| Ours | ✓ | ✓ | ✓ | **0.66** | **4.38** | 34.83 |

**Ablation study on a number of rotations per axis $R$.** We conduct an ablation study to investigate the effect of the number of rotations per axis on the quality of the 3D reconstructed segmentation. Table 2 indicates that increasing the number of rotations enhances the module's ability to capture diverse views of the lesion, thereby improving segmentation accuracy. However, beyond a certain point, the performance gains diminish, suggesting an optimal number of rotations for achieving the best trade-off between computational cost and reconstruction quality. According to these results, we select $R = 4$ for our main experiments.

Table 2: Ablation study results on the number of rotations per axis ($R$). The table shows the DICE score, FPV, FNV, and the time to compute MIPs per one 3D image, averaged over the test set. The compute time grows linearly with $R$ as expected.

| $R$ | **DICE** ↑ | **FPV** ↓ | **FNV** ↑ | **Compute time** ($s$) |
|---|---|---|---|---|
| 1 | 0.917 | 3.905 | 0.000 | 15.83 |
| 2 | 0.931 | 0.620 | 0.157 | 31.66 |
| 4 | 0.931 | 0.230 | 1.313 | 63.31 |
| 8 | 0.934 | 0.127 | 1.355 | 126.62 |
| 16 | 0.933 | 0.066 | 1.425 | 233.24 |

**Ablation study on synthetic MIP images proportions $p$ for the 2D segmentation model.** We perform an ablation study to assess the impact of varying the proportion of synthetic MIP images ($p$) used for training the 2D segmentation model. The parameter $p$ indicates that the number of synthetic MIP images is $p$ percent of the real MIP images. As shown in Table 3, incorporating synthetic MIP images into the training process improves the segmentation model's robustness and generalization capabilities. A moderate proportion of synthetic images provides the best results, balancing the benefits of data augmentation with the risk of overfitting to synthetic data characteristics. We select the best performing $p$ in our experiments based on these results.

Table 3: Ablation study results on the proportion of synthetic MIP images ($p$).

| $p$ | **DICE Score** | **FPV** | **FNV** |
|---|---|---|---|
| 0% | 0.63 | 0.61 | 41.45 |
| 15% | 0.64 | 1.04 | 39.62 |
| 50% | 0.62 | 1.03 | 38.24 |

**Discussion on the failure case.** In the case of positive samples (i.e., samples containing lesion(s)), both MIP segmentation alone and our combined method exhibit lower DICE scores compared to the baseline, as shown in Table 4. These results suggest that while the integration of MIP segmentation improves overall performance, there are challenges in training the model to effectively leverage the information from PET and CT scans, which are included in the baseline method. Theoretically, the model should be able to utilize the information from both PET and CT modalities effectively. This may be attributed to two potential factors: 1) over-reliance on the output from the MIP segmentation component, as we can see that the model is improving beyond just taking from them but still relying heavily on it, as observed from their similar performance; and 2) the generalized loss function provided by the nnUNet framework, which may not be optimal for combining the multi-modal information in this specific task. In figure 7, high dice score example is provided in panel a. Panel b is positive case completely missed by the model (FNV_only) where panel c is negative case predicted positive by model (FPV_only).

**Replacing reconstructed Predicted MIP segmentation with reconstructed Ground truth MIP segmentation.** To validate the proposed model's effectiveness, we replaced the 3D segmentation reconstructed from the 2D segmentation model's predictions with a 3D segmentation reconstructed from 2D ground truth MIP segmentation masks. This substitution allowed us to assess the model's upper bound performance. The results indicated a significantly high performance, with a DICE score of 0.94, as shown in Table 5. These findings suggest that using the ground truth

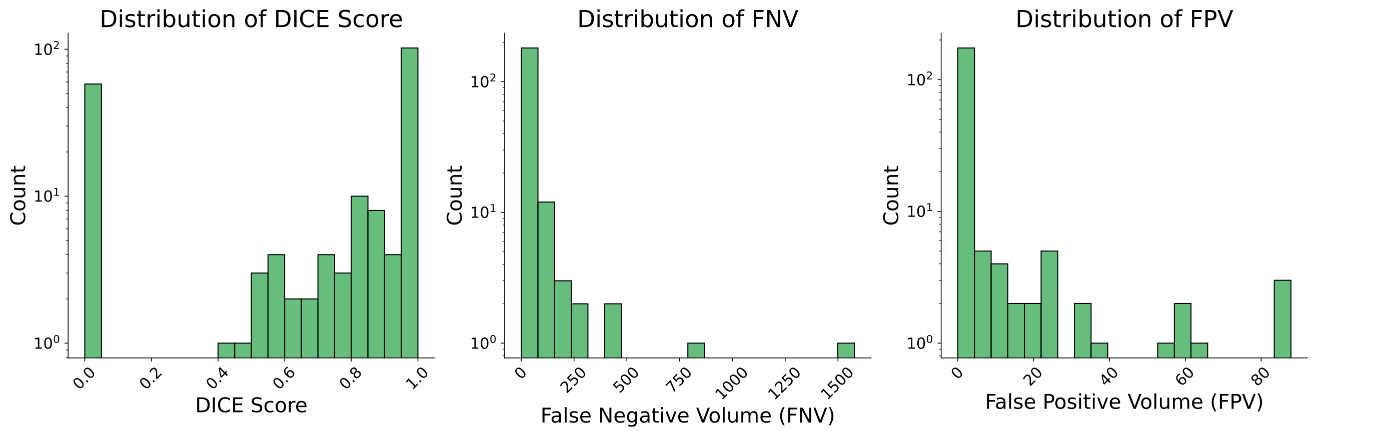


Figure 6: Distribution of segmentation performance metrics. The plots show the distributions of DICE scores, false negative volumes (FNV), and false positive volumes (FPV) across all samples, with logarithmic scaling on the $y$-axes to emphasize the spread of values.

Table 4: Performance metrics for negative and positive samples. **Bold** values indicate an improved performance compared to the baseline. Note that FNV is only defined for positive samples.

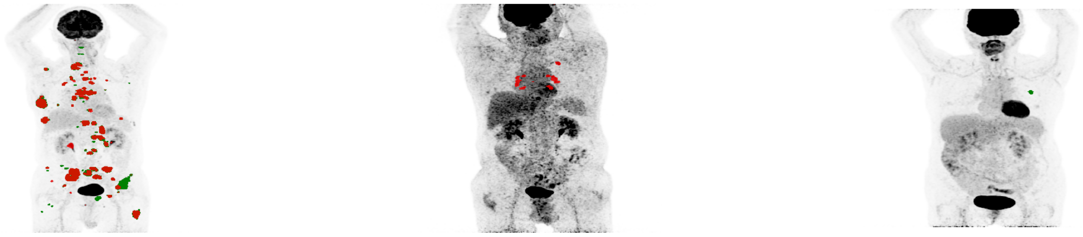
| Method | DICE Score | | | FPV | | | FNV |
|---|---|---|---|---|---|---|---|
| | Negative | Positive | Average | Negative | Positive | Average | Positive |
| Baseline | 0.16 | 0.69 | 0.42 | 29.71 | 12.17 | 21.11 | 13.58 |
| **Proposed method** | | | | | | | |
| MIP Segmentation | **0.99** | 0.27 | **0.64** | **0.02** | **2.10** | **1.04** | 80.84 |
| Ours | **0.99** | 0.32 | **0.66** | **0.25** | **8.68** | **4.38** | 71.06 |
| **Mitigation strategy** | | | | | | | |
| Baseline + MIP Segmentation | **0.90** | 0.49 | **0.68** | **23.04** | **10.24** | **16.77** | 18.85 |
| Baseline + Ours | **0.91** | 0.49 | **0.70** | **23.28** | **10.35** | **16.94** | 18.80 |

MIP segmentation sets a high benchmark for our approach, highlighting the model architecture is appropriate for learning the final segmentation from the reconstructed 3D segmentation.

Table 5: Performance metrics using 3D reconstruction of 2D ground truth MIP segmentation as input (together with CT and PET) to the Refinement Model.

| | DICE Score | FPV | FNV |
|---|---|---|---|
| **Negative** | 1.00 | 0.00 | N/A |
| **Positive** | 0.88 | 0.48 | 2.73 |
| **Total** | 0.94 | 0.24 | 2.73 |

**Threshold selection for mitigation strategy.** To mitigate the identified failure case, we propose a thresholding strategy that leverages the strengths of both methods. Specifically, we can set a threshold to dynamically decide whether to trust the outputs from our model or the baseline model. If the sum of predicted lesion volume from our model and the baseline model is lower than the threshold, we use the output from our model; otherwise, we use the output from the baseline model. This volume threshold is selected to maximize the average DICE score. This approach aims to utilize the high performance of our model on negative samples and the high performance of the baseline model on positive samples, thereby enhancing the overall segmentation performance. We select the volume threshold of 10. By doing so, we are able to achieve higher performance on negative samples and comparable performance on positive samples, compared to the baseline model, as shown in Table 4.



(a) Case with high overlap between gt and pred

(b) Case with FPV only

(c) Case with FNV only

Figure 7: Sample visualization for FPV and FNV where red represents prediction mask and green represents gt mask

## 6. Conclusion and Future Works

In conclusion, this work proposes a novel framework for improving 3D lesion segmentation in PET-CT scans by leveraging 2D maximum intensity projections (MIPs). Our approach combines a 2D MIP segmentation model, trained on synthetic data, with a 3D multi-modal refinement model that incorporates information from PET, CT, and the reconstructed MIP segmentation mask. Through extensive experiments, we demonstrate that our method outperforms the baseline by achieving higher DICE scores and lower false positive volumes, indicating improved overall segmentation accuracy.

We provide insights into the key components of our framework, such as the impact of the number of MIP rotations and the proportion of synthetic data used for training the MIP segmentation model. Furthermore, we discuss failure cases and propose mitigation strategies to leverage the strengths of different modalities effectively.

One major limitation identified is the multi-modal refinement model's underperformance on positive cases (samples with lesions present) compared to the baseline model using only PET and CT scans. While the integration of MIP segmentation improves overall metrics, the model struggles to effectively utilize the complementary information from the PET, CT and reconstructed MIP inputs for accurate lesion segmentation in positive cases.

The primary challenge in debugging and improving the multi-modal refinement model lies in the significant computational requirements and limited resources available. Training the 3D models is extremely computationally intensive, with training times exceeding 10 hours on GPUs, which severely restricts our ability to extensively explore different architectures, loss functions, and training strategies tailored to effectively combine multi-modal information. We believe that further research with access to increased computational resources holds promise for overcoming this limitation and achieving even better performance, particularly for lesions in positive cases.

## 7. Contributions and Acknowledgements

W.C., T.T., and J.B. designed the framework and evaluation. T.T. implemented MIP generation, 3D reconstruction, and trained the diffusion model for synthetic MIP generation. W.C. explored the nnUNet framework and trained the baseline model and multi-modal refinement model. J.B. trained the 2D segmentation model and multi-modal refinement model. W.C., T.T., and J.B. wrote the paper. We would like to thank all of the teaching staff of CS 231N, especially Samir Agarwala for their guidance and support, and Google Cloud for providing compute credits.

## References

[1] S. A. F. Y. F. B. C. U. A. R. Amirhosein Toosi, Sara Harsini. State-of-the-art object detection algorithms for small lesion detection in psma pet: Use of rotational maximum intensity projection (mip) images. In *Proceedings of SPIE*, volume 12464, page 124643E. SPIE, 2022.

[2] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Zambrano Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

[3] S. Gatidis, T. Hepp, M. Früh, et al. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9:601, 2022.

[4] K. B. Girum, L. Rebaud, A. S. Cottereau, et al. 18f-fdg pet maximum-intensity projections and artificial intelligence: A win-win combination to easily measure prognostic biomarkers in dlbcl patients. *Journal of Nuclear Medicine*, 63(12):1925–1932, 2022.

[5] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018.

[6] K. Man and J. Chahl. A review of synthetic image data and its use in computer vision. *Journal of Imaging*, 8(11):310, 2022.

[7] B. D. Moore C, Knipe H. Nifti (file format), 2024.

[8] L. M. Prevedello, S. S. Halabi, G. Shih, et al. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*, 1(1):e180031, 2019. Published 2019 Jan 30.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[10] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[11] L. N. Tanenbaum, A. J. Tsiouris, A. N. Johnson, et al. Synthetic mri for clinical neuroimaging: Results of the magnetic resonance image compilation (magic) prospective, multicenter, multireader trial. *AJNR American Journal of Neuroradiology*, 38(6):1103–1110, 2017.

[12] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.