

# VerSe Data Augmentation Enables per Vertebral Body Instance Segmentation in HSCT Patient Scans

Lucas J. Powers<sup>1</sup>, Reza Babaei<sup>1</sup>, Elnaz Aghdaei<sup>1</sup>, Joseph P. Havlicek<sup>1</sup>, Samuel Cheng<sup>1</sup>, Shangqing Zhao<sup>2</sup>, Christopher G. Kanakry<sup>3</sup>, Peter Choyke<sup>3</sup>, Sara Vesely<sup>4</sup>, Jennifer Holter Chakrabarty<sup>5</sup>, and Kirsten M. Williams<sup>6</sup>

**Abstract**—Accurate and non-invasive monitoring of hematopoietic stem cell transplantation (HSCT) patients is crucial but challenging due to factors including the limitations of traditional biopsies and the intensive manual analysis required for emerging comprehensive imaging techniques like FLT PET/CT. Furthermore, low-dose CT resolution in this vulnerable patient population often hinders precise vertebral body segmentation, a critical step for comprehensive marrow compartment assessment. This paper presents a novel approach for per-vertebral body instance segmentation in HSCT FLT PET/CT scans, addressing the inherent difficulties of low-resolution data and limited annotated training cases. Our method leverages an attention-gated U-Net architecture, significantly enhanced by a novel data augmentation strategy involving downsampled high-resolution VerSe dataset images. We demonstrate, for the first time, accurate vertebral body segmentation on this challenging low-resolution dataset. Our approach integrates an attention-based U-Net model and is compared against TotalSegmentator as a baseline, showing superior segmentation performance, particularly in the anatomically complex upper spine where TotalSegmentator exhibits suboptimal results. To the best of our knowledge, this work reports fully automated high-quality instance segmentation results for individual vertebral bodies in CT volumes of HSCT FLT PET/CT patients for the first time, promising to facilitate automation for critical quantitative assessments like SUV measurement and ultimately improve long-term patient management and outcomes.

**Index Terms**—vertebrae segmentation, HSCT, FLT, PET/CT.

## I. INTRODUCTION

Hematopoietic stem cell transplantation (HSCT) serves as a crucial and often last-resort intervention for patients diagnosed with severe hematologic disorders. This procedure involves the infusion of healthy stem cells to restore normal blood cell production, offering a potential cure for disorders including leukemia, lymphoma, and aplastic anemia. HSCT is a com-

plex, multi-stage process involving marrow ablation, donor cell infusion, and subsequent hematopoietic recovery [1].

Traditionally, post-transplant monitoring has relied on invasive single aspirate bone marrow biopsies, which pose risks such as infection and patient discomfort. Importantly, single aspirate biopsies often suffer from sampling bias and are inherently incapable of providing a comprehensive characterization of the entire full-body marrow compartment. For instance, a single aspirate biopsy drawn from the pelvis (typical) may fail to reveal a recurrence of cancer cells in a spinal vertebra or in the sternum leading to relapse that could prove fatal to the patient. To address this significant inherent shortcoming of single aspirate biopsies, emerging advances in noninvasive imaging techniques and molecular biomarkers are transforming post-transplant assessment by offering dramatically more comprehensive insight into hematopoietic recovery and disease status, with the promise of ultimately improving patient management and long-term prognosis in the clinical setting. Among these emerging techniques, full-body FLT PET/CT imaging has shown strong potential for playing a critical role in assessing disease progression and treatment response in oncology by providing comprehensive, high-sensitivity metabolic insights without the need for invasive biopsies [1]–[5].

However, while full-body FLT PET/CT imaging offers important advantages including comprehensive, high-sensitivity, and high-specificity assessment of the entire full-body marrow cavity, its widespread clinical translation is hindered by some significant challenges that still remain.

First, interpretation of these images generally requires intensive labor by an expert physician to perform manual Region of Interest (ROI) designation and analysis across numerous anatomical structures (e.g., all vertebral bodies, sternum, pelvis, certain organs). Second, radiation dosing considerations in this inherently vulnerable patient population necessitate low-dose CT protocols. This results in poor CT resolution, making it extremely difficult to perform accurate identification of individual bones and their respective marrow cavities. Accurate identification of intervertebral boundaries along the spinal column in particular remains a significant and persistent challenge, both for physicians attempting to perform manual segmentation and for machine algorithms, at least because of indistinct anatomical margins resulting from dosing considerations which typically lead to poor CT resolution in full-body PET/CT volumes acquired from HSCT patients [6]. The development of new accurate and fully automated machine algorithms for vertebral body segmentation in the CT low dose

This work was supported by the National Cancer Institute, NIH, under contract number HHSN261200800001E and grant number 5R01HL146668.

<sup>1</sup>L.J. Powers, R. Babaei, E. Aghdaei, S. Cheng and J.P. Havlicek are with the University of Oklahoma, School of Electrical & Computer Engineering, Norman, OK 73019, USA.

<sup>2</sup>S. Zhao is with the University of Oklahoma, School of Computer Science, Norman, OK 73019, USA.

<sup>3</sup>C.G. Kanakry and P. Choyke are with the National Cancer Institute and National Institutes of Health, Bethesda, MD 20892, USA.

<sup>4</sup>S. Vesely is with the Dept. of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center

<sup>5</sup>J. Holter Chakrabarty is with the Division of Marrow Transplantation and Cell Therapy, Stephenson Cancer Center, Oklahoma City, OK 73104, USA.

<sup>6</sup>K.M. Williams is with the Dept. of Pediatrics, Emory University, Children's Healthcare of Atlanta, and the AFLAC Cancer and Blood Disorder Center, Atlanta, GA 30322, USA.

regime thus remains an important open problem, the solution to which is urgently needed in order to enable clinically practical robust quantitative analysis including fully-automated per-bone standardized uptake value (SUV) measurements in the PET modality which are critical for comprehensive monitoring of disease progression and treatment response in HSCT patients.

Recent advances in deep learning have introduced novel approaches for analyzing medical images, particularly through pattern recognition models that facilitate automated interpretation. In this context, segmentation models play an important role by identifying and delineating regions of interest in complex medical images [7]. Among deep learning architectures, U-Net has gained prominence in medical image analysis due to its ability to preserve important structural details while operating effectively even in cases of severely limited training data and/or poor image quality [8]. A widely recognized extension of this architecture, nnU-Net, has been successfully employed as a benchmark model for medical image segmentation. One of its notable applications is TotalSegmentator, a tool designed for general medical image segmentation tasks and for the CT modality in particular [9]–[12]. However, in our specific use case of low-resolution HSCT FLT PET/CT scans, TotalSegmentator demonstrated suboptimal performance against the CT modality and even complete failure on some vertebrae, particularly in the segmentation of the upper spine where anatomical complexities pose additional challenges not only for machine algorithms but even for expert physicians performing manual segmentations. While TotalSegmentator was utilized for PET/CT in [10], it was applied only for organs and not for vertebrae, underscoring the unique and unsolved challenges posed by vertebral segmentation in this specific clinical context and data modality.

In this paper, we address these critical challenges and the identified segmentation gap by introducing a novel and highly specialized deep learning framework for accurate per-vertebral body instance segmentation in low-resolution FLT PET/CT scans of HSCT patients. As is typical in many practical medical imaging scenarios, we faced the difficult problem of a severe paucity of training data – which represents a significant challenge impeding the development of effective practical deep learning algorithms. A cornerstone of our approach is a new and effective data augmentation strategy designed to overcome this severe paucity of annotated clinical training samples. We leverage cases from the well-established VerSe dataset [13], [14] to create a more robust and relevant training corpus by resampling these high-resolution images and processing the corresponding ground truth labels to precisely match the 5mm slice thickness of our clinical cases as well as satisfy the need to extract accurate SUV measurements from the marrow cavities of the vertebral bodies **only**. We propose an attention-based U-Net model specifically tailored for upper spine segmentation, which demonstrates superior segmentation performance on low-resolution PET/CT images of HSCT patients compared to the small number of existing benchmarks such as [6], [15], as well as to

TotalSegmentator in this low-resolution problem domain. As shown by our results given in Table I below, our proposed U-Net architecture provides a significant performance advantage relative to TotalSegmentator on some regions of the spine; the inherent difficulty of this problem is further demonstrated by TotalSegmentator’s failure on some vertebrae including most notably C2-C6 and T5-T6. To the best of our knowledge, this paper represents the first study to successfully perform accurate automated segmentation of the individual vertebral bodies in low-resolution PET/CT scans of HSCT patients while achieving high precision as quantified by Dice scores. Moreover our proposed method accomplishes this without relying on cues from the PET modality, making it applicable at early observation points where the PET signature is expected to be low. Our findings highlight the profound potential of deep learning in overcoming image quality constraints and illustrate how creative data augmentation strategies can be used to compensate for severe paucity of training data. In a larger context, our proposed segmentation methodology provides a tangible step towards development and clinical translation of fully automated quantitative assessments such as per-bone SUV measurements with potential to enhance patient outcomes and long-term prognosis.

## II. RELATED WORK

Automated vertebrae segmentation from CT scans is a critical medical imaging task supporting diagnosis, treatment planning, and surgical navigation. Research in this domain has evolved from traditional image processing to advanced deep learning, consistently aiming for enhanced accuracy and robustness. This section provides an overview of recent advances, highlighting methodologies and the specific challenges our current work addresses.

Early segmentation relied on classical methods like intensity-based thresholding, active shape models (ASMs), and deformable models. While foundational, these techniques often struggled with spinal anatomical variability, noise, artifacts, and required extensive manual initialization, limiting their generalizability across diverse patient populations and imaging protocols [16]–[18].

Deep learning has revolutionized medical image analysis. For example, recent Transformer models capable of capturing global contextual information include VerteFormer [19] (Vision Transformer with ED and GIE blocks) and LumVert-CancNet [20] (Swin Transformer for centroid localization and hybrid encoder-decoder). While these models represent the current state-of-the-art on high-resolution general CT data, their generalizability to low-dose, low-resolution clinical data from specific cohorts like HSCT patients and their inherent computational demands require further investigation for practical clinical deployment.

Traditional deep learning approaches often employ convolutional neural networks (CNNs) and fully convolutional networks (FCNs). Chen et al. [21] proposed an FCN-based framework integrating a hidden Markov model, achieving high identification rates. Klein et al. [22] introduced VertDetect, a

3D vertebral segmentation model leveraging a shared CNN backbone and graph convolutional network (GCN), achieving a Dice similarity coefficient (DSC) of 0.883. Other works combine CNNs with recurrent networks; Liao et al. [23] used a multi-task 3D FCN with a bidirectional recurrent neural network, while Lessmann et al. [24] proposed an iterative FCN. Despite their robust performance, these models often demand extensive annotated datasets which are scarce in specialized clinical contexts like low-dose HSCT imaging. These models may also struggle with the fine-grained details and indistinct boundaries prevalent in low-resolution acquisitions.

U-Net [8] remains a foundational architecture in biomedical image segmentation due at least in part to its effective preservation of structural details. Falk, et al. [25], for example, applied U-Net broadly for cell detection and analysis. The Attention U-Net [26] in particular is an important U-Net variation that uses attention gates for improved feature focus. The general effectiveness of U-Net including U-Net variants in medical image segmentation has also been well demonstrated in the VerSe Challenge [7], [13].

However, while U-Net performance is generally robust in a wide variety of medical image segmentation problems importantly including many where limited training data is a concern, obtaining good performance against *extremely low-resolution and low-contrast* clinical data such as the HSCT patients scans we consider in this paper remains challenging.

For example, Carson et al. employed an ensemble of U-Nets to perform approximate segmentation of individual vertebral bodies in low-dose PET/CT scans of HSCT patients in [6]. Their method involves first using the CT data to obtain a segmentation of the spinal column as a unit and then using the PET data to find a best axial plane for approximating each intervertebral boundary. This approach suffers from two main limitations. First, because each approximated intervertebral boundary is constrained to lie entirely within a single axial plane, the method is inherently incapable of accounting for the natural curvature of the spine which can introduce significant segmentation errors particularly in the smaller vertebrae. Second, since the segmentation of individual vertebral bodies relies on having a strong PET signature, their method may not be applicable at early observation points before or soon after transplant (e.g., 1 day before and/or 3 and 5-9 days post-HSCT) when early detection or prediction of events such as graft failure and relapse could be most beneficial in terms of improving patient outcomes.

The availability of large-scale annotated datasets such as Sekuboyina et al.'s 374 multi-detector spine CT scans [13] and Liebl et al.'s VerSe 2020 dataset with 4142 vertebrae annotations [14] has significantly facilitated and positively impacted vertebrae segmentation research. However, a critical scarcity still persists for datasets specific to low-resolution clinical protocols in specialized patient populations such as the low-dose FLT PET/CT HSCT patients we consider in this paper. A key component of our contribution in this paper is to demonstrate how an innovative augmentation strategy can bridge this domain gap, enabling the relatively rich availability

of labeled segmentation training samples in datasets such as VerSe to be leveraged for alleviating the profound paucity of training samples available with small cohort low-dose datasets such as the HSCT patient scans considered here.

In summary, while significant progress has been made in automated vertebrae segmentation through advanced deep learning, the specific challenges of low-resolution, low-dose clinical data in vulnerable populations such as HSCT patients coupled with a profound paucity of annotated data represent a critical unmet need that is impeding the development of practical fully automatic machine segmentation and SUV measurement algorithms that could play a role in facilitating widespread clinical translation of full-body FLT PET/CT imaging in the treatment of HSCT patients. We directly address these limitations in this paper by developing a highly specialized framework for robust and accurate vertebral body segmentation under such constrained conditions.

### III. METHODOLOGY

Developing robust machine learning models for medical image segmentation generally requires large quantities of high-quality labeled data. However, in many medical imaging applications, particularly at the research stage, large numbers of data samples are simply not available. Moreover, even when large quantities of data are available, acquiring comprehensive annotations is generally labor-intensive, sometimes constrained by ethical considerations, and may also be hampered by limited expert availability. In this study, we faced a significant data paucity issue with access to only 27 labeled low-resolution FLT PET/CT clinical volumes from 18 patients – thus rendering the problem of how to train a generalizable segmentation model for vertebral bodies exceedingly challenging. The CT voxel resolution was fixed at  $1.17 \text{ mm} \times 1.17 \text{ mm} \times 5.00 \text{ mm}$  due to dosing considerations while the PET data were acquired isotropically at a voxel resolution of  $4 \text{ mm} \times 4 \text{ mm} \times 4 \text{ mm}$ . In addition to the small cohort size and corresponding general lack of adequate training samples, the very thick 5 mm CT axial slice spacing in this low-dose data is a primary factor obscuring the intervertebral boundaries, especially in the cervical region, and making individual vertebral body segmentation difficult for machines and even for human experts. Also, because we desire a fully automatic solution that, unlike the previous methods given in [6] and [15], can be applied at early observation points when the PET signature may be very low (i.e., at -1, +3, +5-9 days relative to transplant), we sought an approach that could rely on the CT volumes alone for performing vertebral body instance segmentation without depending on cues from the PET modality.

To address the significant data paucity limitation, we placed data augmentation at the center of our strategy and sought a feasible way to leverage the publicly available VerSe dataset [13], [14] to expand our training corpus. Doing so was not entirely obvious and straightforward for two main reasons:

- First, the various images in the VerSe dataset have a variety of voxel spacings ranging from 0.6 mm to 2.5 mm

that are isotropic in some cases and anisotropic in others. Resampling the VerSe CT data to match the  $1.17 \text{ mm} \times 1.17 \text{ mm} \times 5.00 \text{ mm}$  voxel spacing of our HSCT patient scans, thus making the resampled VerSe images appear as though they had been acquired under the imaging conditions of our clinical protocol, was straightforward; we accomplished this using the `Spacingd` transform provided by MONAI with trilinear interpolation.<sup>1</sup> The slightly tricky aspect lies in realizing that the corresponding VerSe ground truth label images must also be resampled to the same resolution, but that a different interpolation scheme is required since the ground truth labels must take integer values. We called `Spacingd` with nearest neighbor interpolation to resample the labels.

- Second, our ultimate goal is to develop an accurate, fully automatic technique for extracting clinically significant SUV measurements from the marrow cavity of each vertebral body. This results in an anatomical mismatch with the VerSe ground truth data wherein the entire vertebra is labeled including the vertebral body, pedicles, lamina, and the spinous, transverse, and superior articular processes. Consequently, for both training and evaluation it was necessary for us to process the VerSe ground truth to remove the pedicles, lamina, and processes. We performed this processing at the higher native VerSe resolution prior to resampling the CT data and labels.

Our processing to remove the unwanted structures from the VerSe ground truth begins by applying a 2D morphological opening in axial planes with a dynamically sized disk-shaped structuring element. Each image of VerSe ground truth labels was first converted to the RAS orientation to ensure consistency with respect to the directionality of the labels associated with the spinous process relative to those associated with the main vertebral body. The radius of the structuring element was dynamically adjusted on a per-slice basis, guided by the pixel area of the vertebra labels in the current axial slice. The reason for dynamically adjusting the structuring element size was to account for the variation in the size of the vertebral body in traversing from the cervical region through to the lumbar region. Within each axial slice, the opening results in a binary image where the pedicles, lamina, and processes are either separated from the main vertebral body or eradicated altogether. We then apply connected component labeling with region removal to retain only the labels associated with the vertebral body. Since connected components corresponding to one or more of the vertebral processes may actually have a larger area than the component corresponding to the vertebral body, particularly in the lumbar region, it is not sufficient to simply retain the largest connected component. Instead, to identify the vertebral body while rejecting the unwanted structures in the labels image, we retain the connected component lying closest to the average spatial centroid computed across the last five slices processed (with obvious modification to account for the first few axial slices in the cervical region).

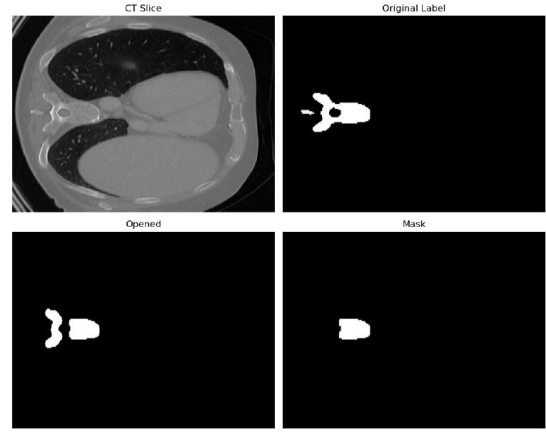


Fig. 1. Processing to remove unwanted structures from the VerSe ground truth labels. Upper left: example axial plane from the original VerSe CT volume. Upper right: axial binary label image associated with the vertebra. Lower left: result of the morphological opening. Lower right: binary mask corresponding to the vertebral body only, obtained by connected components analysis.

An example of this processing is shown in Fig. 1 where the obtained binary mask corresponding to the labels associated with the vertebral body only is shown in the lower right panel. Once the unwanted structures are removed from the VerSe ground truth labels associated with all vertebrae, both the CT and ground truth volumes are resampled to a voxel spacing of  $1.17 \times 1.17 \times 5 \text{ mm}$  using MONAI as described above. An example is given in Fig. 2, where the upper panel shows the original VerSe CT data with ground truth labels overlayed while the lower panel shows the downsampled CT data with processed and downsampled ground truth overlayed.

By making augmentation the cornerstone of our pipeline, we significantly enhanced the diversity and relevance of the training data, directly improving the robustness and clinical applicability of our segmentation model under severe data constraints.

#### A. Network Architecture

We employed the attention gated U-Net architecture, following the design proposed by Oktay et al. [26]. This architecture was specifically chosen for its ability to integrate attention mechanisms into the standard U-Net framework, which selectively emphasizes relevant image regions and features. This capability is particularly important for improving segmentation accuracy in our challenging low-resolution, low-contrast medical images, where indistinct boundaries and subtle features might otherwise be overlooked. The architecture is a fully 3D U-Net, making it inherently suitable for volumetric medical image segmentation.

The encoder branch comprises five layers of residual units, where each unit consists of convolutional blocks incorporating skip connections. These residual connections are vital for mitigating the vanishing gradient problem, enabling the training of deeper networks and promoting more effective gradient flow. All convolutions throughout the network are 3D convolutions,

<sup>1</sup><https://docs.monai.io/en/stable/transforms.html>

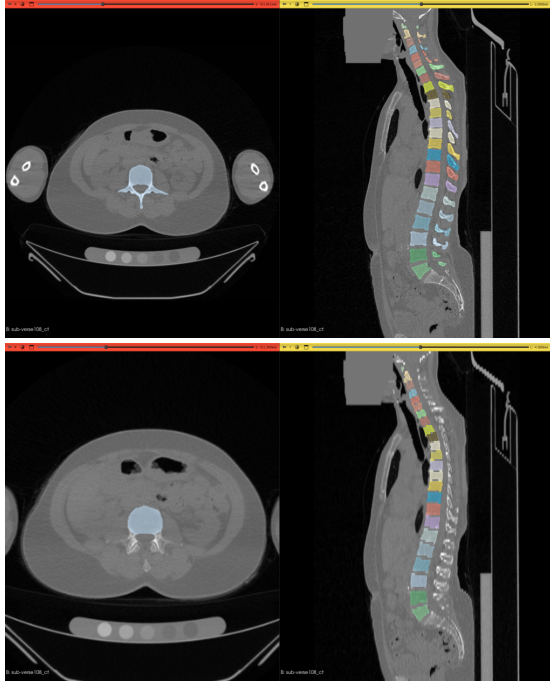


Fig. 2. Processing to remove unwanted structures from the VerSe ground truth and downsample both ground truth and CT data. Upper: original VerSe CT data with ground truth overlaid; both shown at native VerSe voxel resolution. Lower: result after processing to remove pedicles, lamina, and processes and downsampling to match the poor CT resolution of our HSCT patient CT volumes. The red bar indicates axial slice while yellow indicates sagittal.

preserving the spatial context in all three dimensions. The channel configuration systematically scales, starting with 16 channels in the initial layer and expanding to 256 channels in the final encoder layer.

To prevent overfitting, dropout regularization is applied selectively within the encoder: the first two layers have no dropout, while the third, fourth, and final layers employ dropout rates of 0.1, 0.2, and 0.3, respectively. This selective application allows for robust feature learning in earlier layers while introducing regularization in deeper layers more prone to overfitting. The decoder branch incorporates attention gates before each up-convolution operation, directing the network’s focus to the most salient spatial features of the input images. Batch normalization is applied throughout the network to stabilize training and accelerate convergence, and parametric ReLU (PReLU) is utilized as the activation function, allowing for learning the slope of the negative part of the rectifier, thereby preventing the “dying ReLU” problem. The overall network architecture is depicted in Fig. 3.

Despite its comprehensive capabilities, the model consists of a relatively compact 4,787,484 trainable parameters, making it lightweight and efficient for volumetric medical imaging applications. This compact size, combined with the inclusion of attention gates and residual connections, facilitates effective feature extraction and robust segmentation performance while still being amenable to clinical deployment.

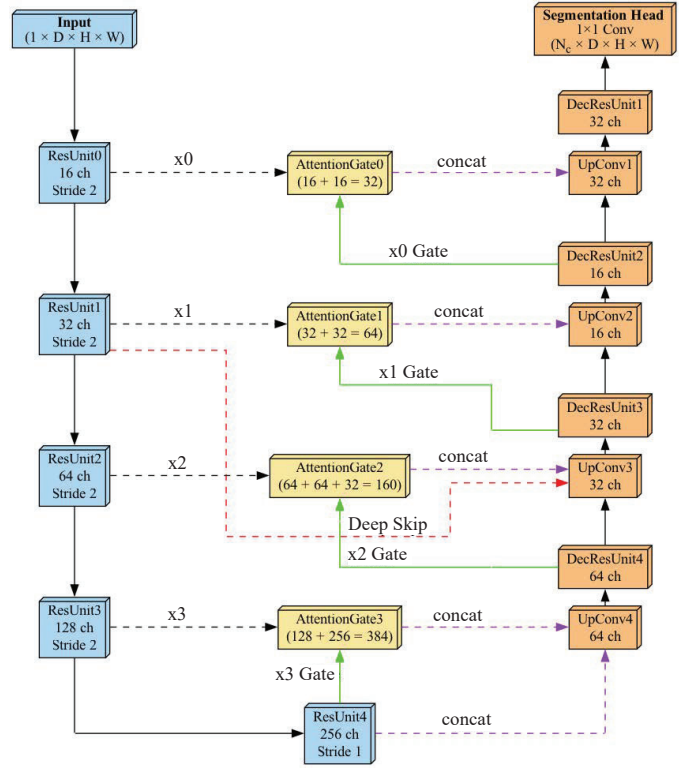


Fig. 3. Proposed attention gated U-Net architecture. The encoder (left) consists of residual units with increasing channels and dropout, while the decoder (right) incorporates attention gates before each up-convolution to focus on salient features.

## B. Data and Preprocessing

This study utilized  $^{18}\text{F}$ -fluorothymidine (FLT) PET/CT scan data from hematopoietic stem cell transplantation (HSCT) patients. The clinical dataset comprised 27 labeled scans from 18 patients, with all manual annotations meticulously performed by a registered trainee to ensure high accuracy for model training and evaluation. As previously noted, the limited sample size of these clinical cases underscored the necessity for robust data augmentation strategies. The CT scans in our clinical dataset had a coarse axial resolution with a slice thickness of 5 mm. This low resolution significantly complicated vertebral body segmentation due to reduced boundary detail and inherent blur. Additionally, the dataset incorporated temporal variations in FLT uptake, with scans acquired at one day before transplantation, 5-9 days post-transplant, and 28 days post-transplant.

For training, our combined dataset consisted of 93 downsampled VerSe volumes and 27 clinical HSCT volumes. Model validation was conducted using four independent clinical HSCT volumes from four patients. The segmentation task involved 26 distinct anatomical classes, covering individual vertebral bodies. Class 0 (background) was explicitly excluded from both the training loss calculation and Dice validation to focus performance metrics solely on the anatomical structures of interest. Due to its unique and often variable morphology

and lesser clinical significance in this context, the C1 vertebra was omitted from training.

Preprocessing was applied to ensure data consistency and quality. All images were standardized to the LPS (Left, Posterior, Superior) coordinate system to maintain a consistent anatomical orientation. CT intensities were normalized to a standardized range of -100 to 250 Hounsfield Units (HU), effectively normalizing tissue densities across different scans. Images smaller than the designated patch size (96, 96, 64) were padded accordingly to ensure uniform input dimensions for the 3D convolutional network. Training samples were generated through random patch extraction, a common strategy to manage memory constraints with volumetric data and to encourage the model to learn local features.

To further enhance robustness and mitigate overfitting given the limited clinical data, conventional data augmentation strategies were employed on the already VerSe-augmented dataset, including small random elastic transformations to simulate realistic anatomical deformations, random affine transformations (rotations/scaling/shearing) to enhance viewpoint invariance, and random intensity shifts with an offset of 15% relative to the original distribution, mimicking variations in scanner parameters and patient attenuation.

### C. Training Setup

The attention gated U-Net model was trained with a batch size of 1, primarily due to memory constraints associated with processing large 3D volumetric medical images. Training was conducted over 750 epochs. The AdamW optimizer was utilized for optimization, chosen for its strong performance in deep learning applications, particularly with its decoupled weight decay. A fixed learning rate of  $1 \times 10^{-4}$  and a weight decay (L2 regularization) of  $1 \times 10^{-3}$  were applied.

The loss function was a composite of Dice loss and cross-entropy loss with one-hot encoding. This combination is standard practice in medical image segmentation; Dice loss effectively handles class imbalance (e.g., small foreground objects like vertebrae against a large background) by maximizing the overlap between predicted and ground truth segmentations, while cross-entropy loss provides robust pixel-wise classification.

To mitigate overfitting, which is a significant concern with limited clinical data, data augmentation played a pivotal role, leveraging both the spatial transformations and intensity-based augmentations already described in Section III-B. Although experiments with learning rate schedulers were conducted, they were found to exacerbate overfitting rather than improve generalization in this specific setup, suggesting that the fixed learning rate combined with aggressive data augmentation provided high-quality and effective regularization. The primary evaluation metric for model performance during training and validation was the Dice similarity coefficient, a widely accepted metric for segmentation accuracy.

### D. Experimental Validation

To quantitatively evaluate our framework, we conducted a comprehensive experimental validation, assessing both our

novel data augmentation strategy and comparative performance against a leading state-of-the-art model.

First, we analyzed the impact of incorporating the downsampled VerSe dataset. Table I presents the results of a three-fold cross-validation, summarizing the average Dice Similarity Coefficient (DSC) across all vertebral bodies as well as per-vertebra segmentation. Our findings reveal a substantial improvement when the downsampled VerSe dataset was introduced, with average Dice scores increasing by approximately 39% (e.g., from 0.537 to **0.751** for C4). This demonstrates the significant role and viability of our data augmentation in overcoming acute scarcity of annotated clinical data, underscoring its potency for boosting segmentation model performance in challenging clinical settings.

Second, we performed a comparative analysis against TotalSegmentator [9], a state-of-the-art nnU-Net based model. Both models were evaluated on four independent, previously unseen clinical test cases, focusing particularly on the anatomically complex and challenging upper spine regions where segmentation is difficult due to intricate structures and the lower image quality of low-dose CT. Detailed comparative performance metrics are presented in Table I. We used the publicly released TotalSegmentator model without retraining or fine-tuning. As a general-purpose segmentation framework trained on high-resolution CT data, TotalSegmentator remains one of the most widely used tools in medical image segmentation and provides a practical off-the-shelf baseline. Retraining the TotalSegmentator model with our domain-specific data in this study was not feasible due to several factors, including: (1) even with VerSe augmentation, our limited dataset size is insufficient to support effective re-training of a model of this size; (2) its broad multi-class output space is misaligned with our focus on vertebral bodies only; and (3) adapting the training pipeline, which is not fully documented for task-specific retraining, would require substantial engineering effort beyond the scope of this study. Thus, we do not present TotalSegmentator as a competing method but rather as a representative general-purpose benchmark. Its underperformance in this context helps to highlight the unique challenges posed by our clinical imaging domain and motivates the need for tailored, domain-aware solutions like the one we propose.

As shown in Table I, our proposed model consistently outperforms TotalSegmentator on all but three vertebrae. For instance, our model achieves a Dice score of **0.781  $\pm$  0.069** for C5 (vs.  $0.633 \pm 0.047$  for TotalSegmentator) and a remarkable **0.904  $\pm$  0.038** for T10, (vs.  $0.778 \pm 0.273$  for TotalSegmentator). In addition, TotalSegmentator generally exhibits a high sample standard deviation in the thoracic region and even failure to segment the vertebral body in some instances in the cervical and thoracic regions. These data support our hypothesis that an attention-enhanced U-Net coupled with domain-aware data augmentation can achieve superior performance even under low-resolution constraints. While TotalSegmentator delivers the best average Dice scores for certain vertebrae (T2, T12, and L1), the performance gain provided by TotalSegmentator on these vertebrae relative



TABLE I  
AVERAGE DICE SCORES  $\pm$  STANDARD DEVIATION PER VERTEBRA FROM  
3-FOLD CROSS-VALIDATION FOR NO VERSe, WITH VERSe, AND TOTAL  
SEGMENTATOR SETTINGS

Vertebra	Proposed		Total Segmentator
	w/out VerSe Augmentation	w/ VerSe Augmentation	
C2	0.574 $\pm$ 0.126	<b>0.730 <math>\pm</math> 0.082</b>	0.679 $\pm$ 0.053
C3	0.562 $\pm$ 0.154	<b>0.755 <math>\pm</math> 0.048</b>	0.679 $\pm$ 0.051
C4	0.537 $\pm$ 0.160	<b>0.751 <math>\pm</math> 0.055</b>	0.658 $\pm$ 0.043
C5	0.600 $\pm$ 0.129	<b>0.781 <math>\pm</math> 0.069</b>	0.633 $\pm$ 0.047
C6	0.647 $\pm$ 0.113	<b>0.796 <math>\pm</math> 0.051</b>	0.695 $\pm$ 0.045
C7	0.661 $\pm$ 0.105	<b>0.777 <math>\pm</math> 0.059</b>	0.735 $\pm$ 0.056
T1	0.645 $\pm$ 0.085	<b>0.797 <math>\pm</math> 0.046</b>	0.742 $\pm$ 0.093
T2	0.679 $\pm$ 0.086	0.824 $\pm$ 0.047	<b>0.825 <math>\pm</math> 0.054</b>
T3	0.628 $\pm$ 0.095	<b>0.848 <math>\pm</math> 0.030</b>	0.803 $\pm$ 0.130
T4	0.574 $\pm$ 0.134	<b>0.839 <math>\pm</math> 0.036</b>	0.804 $\pm$ 0.119
T5	0.551 $\pm$ 0.153	<b>0.841 <math>\pm</math> 0.037</b>	0.684 $\pm$ 0.309
T6	0.513 $\pm$ 0.182	<b>0.835 <math>\pm</math> 0.021</b>	0.603 $\pm$ 0.301
T7	0.554 $\pm$ 0.258	<b>0.868 <math>\pm</math> 0.050</b>	0.710 $\pm$ 0.285
T8	0.582 $\pm$ 0.214	<b>0.879 <math>\pm</math> 0.040</b>	0.710 $\pm$ 0.270
T9	0.548 $\pm$ 0.229	<b>0.885 <math>\pm</math> 0.036</b>	0.738 $\pm$ 0.277
T10	0.591 $\pm$ 0.169	<b>0.904 <math>\pm</math> 0.038</b>	0.778 $\pm$ 0.273
T11	0.635 $\pm$ 0.172	<b>0.902 <math>\pm</math> 0.032</b>	0.838 $\pm$ 0.101
T12	0.591 $\pm$ 0.268	0.907 $\pm$ 0.029	<b>0.925 <math>\pm</math> 0.030</b>
L1	0.590 $\pm$ 0.299	0.907 $\pm$ 0.019	<b>0.913 <math>\pm</math> 0.027</b>
L2	0.738 $\pm$ 0.147	<b>0.928 <math>\pm</math> 0.017</b>	0.899 $\pm$ 0.065
L3	0.725 $\pm$ 0.178	<b>0.923 <math>\pm</math> 0.017</b>	0.897 $\pm$ 0.091
L4	0.686 $\pm$ 0.213	<b>0.926 <math>\pm</math> 0.023</b>	0.833 $\pm$ 0.259
L5	0.667 $\pm$ 0.151	<b>0.844 <math>\pm</math> 0.248</b>	0.790 $\pm$ 0.327

to our model with VerSe augmentation is relatively small; moreover, our model maintains highly competitive scores across the entire spine, exhibiting exceptional robustness in the most challenging anatomical areas where TotalSegmentator has significant difficulty and high variability. These results demonstrate our proposed model’s ability to accurately segment vertebral structures despite severe low-resolution constraints, making it a more robust and suitable choice for critical clinical applications where high-quality imaging may not always be available, e.g. due to dosing considerations associated with particularly vulnerable patient populations. Qualitative segmentation results providing visual evidence of the proposed model’s precision are given in Fig. 4. Further discussion, analysis, and implications for future research are given in Section IV.

#### IV. CONCLUSION

In this paper we have illustrated how creative data augmentation can in at least some instances be used to overcome seemingly insurmountable data quality and data paucity problems. On the face of it, the initial appearance was that it would be infeasible to consider augmenting our very limited HSCT patient low-dose PET/CT dataset with high-resolution, ground truth labeled VerSe data at least because of the stark resolution differences and because each entire vertebra is labeled in the VerSe ground truth whereas the vertebral body marrow cavity must be isolated for SUV measurement in our application. One

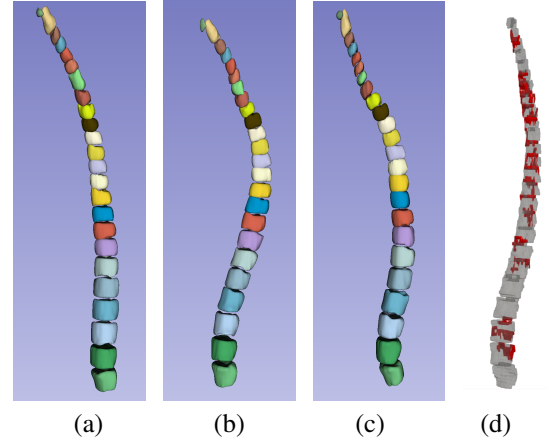


Fig. 4. 3D examples. (a)-(c): U-Net vertebral body instance segmentations obtained by applying the proposed model with VerSe augmentation to low-dose HSCT patient CT data. (d): typical volumetric difference map between the delivered segmentation and ground truth for one validation sample; red indicates misclassified voxels that disagree with ground truth.

of the key things that we have shown is that these challenges can be successfully traversed by creatively processing the VerSe ground truth at native (high) resolution to adapt it by removing the unwanted structures, and then downsampling both the VerSe CT data and the adapted ground truth labels to generate meaningful additional labeled training samples as though they had been originally acquired with the low-dose imaging configuration required for our vulnerable HSCT patient population.

This approach had a significant impact, making it possible to train an attention-gated U-Net model to deliver fully automated, high-quality segmentations of the individual vertebral bodies of the spinal column in HSCT patient low-dose FLT PET/CT scans without relying on cues from the PET modality for the first time.

This result is important for at least two reasons. First, the ability to segment the individual vertebral bodies in low-resolution CT data without relying on the PET modality means that our technique can be applied at early observation points when the PET signal may be weak. Combined with per-bone SUV measurements, this could facilitate machine assisted or even fully automatic early detection and prediction of graft failure and relapse, potentially enabling life saving intervention that would be impossible in current clinical practice. Second, compared to current clinical practice which is largely based on single aspirate biopsies for assessment of the post-HSCT marrow compartment, this result opens the door to machine assisted or even fully automatic assessment on a per-bone basis, potentially alleviating the high physician time and labor burden that currently impedes clinical translation of emerging comprehensive imaging technologies such as FLT PET/CT.

The results in Table I show that our strategy for adapting VerSe data to meaningfully augment the limited dataset of HSCT patient scans delivered a substantial performance improvement of approximately 39% in average Dice score

over the entire spinal column compared to the same U-Net architecture without VerSe augmentation. Broken out by region, the average Dice score performance gain due to VerSe augmentation was 28.69% for the cervical region, 46.60% for the thoracic region, and 33.66% for the lumbar region.

Table I also shows comparative performance of the proposed method against TotalSegmentator, a leading state-of-the-art medical image segmentation solution. While TotalSegmentator excels in generalized CT segmentation problems, on the specialized low-dose low-resolution HSCT patient scans considered here our VerSe augmented method provided a consistent and significant performance advantage, particularly in the upper cervical (C2-C7) and mid-thoracic (T4-T11) spinal regions. In particular, compared to TotalSegmentator our model achieved a Dice score improvement of 11.59% averaged over the entire spine, broken out as gains of 12.74% for the cervical region, 13.89% for the thoracic region, and 4.69% for the lumbar region. In addition, as also shown in Table I, our model achieved a lower variance in Dice score compared to TotalSegmentator on 18 of the 23 vertebrae tested.

Our ongoing work is focused in a few key areas. First, we are continuing to enroll new HSCT patients and acquire additional FLT PET/CT scans at a variety of observation points including 3, 5-9, and 28 days post-transplant. Second, we are continuing the arduous and time consuming task of manually generating ground truth annotations for our current backlog of already-acquired but still unlabeled HSCT patient low-dose FLT PET/CT scans as well as for new patient scans as they are acquired going forward. We believe that continuing to generate as large as possible a corpus of labeled HSCT patient scans to enrich the training set and enable additional cross validation folds at testing is probably the most promising avenue for obtaining further performance enhancements with our current U-Net architecture described in this paper. Third, our approach described here needs to be expanded to include additional bones, most immediately the sternum and pelvis. Doing so is impeded primarily by the current lack of sufficient labeled training data to support the additional segmentation classes. Finally, this fully automatic segmentation method needs to be leveraged to realize fully automated extraction of SUV measurements, validate them against physician measurements, and then use them to investigate and develop machine assisted and fully automatic methods for early prediction of graft failure and relapse.

## REFERENCES

- [1] K. M. Williams and J. H. Chakrabarty, "Imaging haemopoietic stem cells and microenvironment dynamics through transplantation," *The Lancet Haematology*, vol. 7, no. 3, pp. e259–e269, 2020.
- [2] A. Agool, B.W. Schot, P.L. Jager, and E. Vellenga, "18F-FLT PET in hematologic disorders: a novel technique to analyze the bone marrow compartment," *J. Nuclear Med.*, vol. 47, no. 10, pp. 1592–1598, 2006.
- [3] A.K. Buck et al., "First demonstration of leukemia imaging with the proliferation marker 18F-fluorodeoxythymidine," *J. Nuclear Medicine*, vol. 49, no. 11, pp. 1756–1762, 2008.
- [4] M. Vanderhoek, M.B. Juckett, S.B. Perlman, R.J. Nickles, and R. Jeraj, "Early assessment of treatment response in patients with AML using [18F] FLT PET imaging," *Leukemia Res.*, vol. 35, no. 3, pp. 310–316, 2011.
- [5] S. Schelhaas et al., "Preclinical applications of 3'-deoxy-3'-[18F] Fluorothymidine in oncology—a systematic review," *Theranostics*, vol. 7, no. 1, pp. 40–50, 2017.
- [6] B.D. Carson et al., "Approximate vertebral body instance segmentation by PET-CT fusion for assessment after hematopoietic stem cell transplantation," in *2023 IEEE 23rd Int'l. Conf. Bioinformatics, Bioengr. (BIBE)*, 2023, pp. 62–69.
- [7] N. Altini et al., "Segmentation and identification of vertebrae in CT scans using CNN, k-means clustering and k-NN," *Informatics*, vol. 8, no. 2:40, 2021.
- [8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015*. Springer, 2015, pp. 234–241.
- [9] J. Wasserthal et al., "TotalSegmentator: robust segmentation of 104 anatomic structures in CT images," *Radiol.: Artificial Intell.*, vol. 5, no. 5, 2023.
- [10] S. Mostafapour et al., "Evaluation of automatic segmentation tools on low-dose and ultra-low-dose CT images in PET/CT scans," in *2024 12th European Workshop Visual Infor. Process. (EUVIP)*, 2024, pp. 1–6.
- [11] Y.-L. Chen, I.-F. Chung, C.-T. Cheng, and H.-S. Lin, "A 2-step deep learning approach to splenic injury detection," in *2023 Int'l Conf. Fuzzy Theory, Applications (iFUZZY)*, 2023, pp. 1–5.
- [12] X. Jia et al., "Enhancing single-source domain generalization from CT to MR by data augmentation based on grayscale distribution remapping," in *2024 IEEE Int'l. Conf. Med. Artif. Intell. (MedAI)*, 2024, pp. 105–113.
- [13] A. Sekuboyina et al., "VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images," *Medical Image Anal.*, vol. 73:102166, 32 pp., 2021.
- [14] H. Liebl et al., "A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data," *Sci. Data*, vol. 8:284, 2021.
- [15] C. Nguyen et al., "An automatic 3D CT/PET segmentation framework for bone marrow proliferation assessment," in *2016 IEEE Int'l. Conf. Image Process. (ICIP)*, 2016, pp. 4126–4130.
- [16] B. Glocker, D. Zikic, E. Konukoglu, D.R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine CT via dense classification from sparse annotations," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. Springer, 2013, pp. 262–270.
- [17] A. Rasouli, R. Rohling, and P. Abolmaesumi, "Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+ pose model," *IEEE Trans. Medical Imag.*, vol. 32, no. 10, pp. 1890–1900, 2013.
- [18] S.-H. Huang, Y.-H. Chu, S.-H. Lai, and C.L. Novak, "Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI," *IEEE Trans. Medical Imag.*, vol. 28, no. 10, pp. 1595–1605, 2009.
- [19] X. You, Y. Gu, Y. Liu, S. Lu, X. Tang, and J. Yang, "Verteformer: A single-staged transformer network for vertebrae segmentation from CT images with arbitrary field of views," *Med. Phys.*, vol. 50, no. 10, pp. 6296–6318, 2023.
- [20] Y. Zhang et al., "LumVertCancNet: A novel 3D lumbar vertebral body cancellous bone location and segmentation method based on hybrid swin-transformer," *Comput. in Biol., Medicine*, vol. 171:108237, 2024.
- [21] Y. Chen, Y. Gao, K. Li, L. Zhao, and J. Zhao, "Vertebrae identification and localization utilizing fully convolutional networks and a hidden Markov model," *IEEE Trans. Medical Imag.*, vol. 39, no. 2, pp. 387–399, 2019.
- [22] G. Klein, M. Hardisty, C. Whyne, and A.L. Martel, "VertDetect: Fully end-to-end 3D vertebral instance segmentation model," *arXiv preprint arXiv:2311.09958*, 2023.
- [23] H. Liao, A. Mesfin, and J. Luo, "Joint vertebrae identification and localization in spinal CT images by combining short-and long-range contextual information," *IEEE Trans. Medical Imag.*, vol. 37, no. 5, pp. 1266–1275, 2018.
- [24] N. Lessmann, B. Van Ginneken, P.A. De Jong, and I. Išgum, "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Med. Image Anal.*, vol. 53, pp. 142–155, 2019.
- [25] T. Falk et al., "U-Net: deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, 2019.
- [26] O. Oktay et al., "Attention U-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.