IMPROVING VERTEBRAE SEGMENTATION USING A CENTROID DETECTION-GUIDED TRANSFORMER-BASED NETWORK

Sevde Aydodgu[⋆] *Ka-Wai Yung*[⋆] *Danail Stoyanov*[⋆] *Deepak Kalaskar*[†] *Evangelos Mazomenos*[⋆]

[⋆] Wellcome/EPSRC Centre for Interventional and Surgical Sciences,
Department of Medical Physics and Biomedical Engineering,
[†] Division of Surgery and Interventional Science,
University College London

ABSTRACT

Segmentation and identification of vertebrae are crucial tasks for diagnosing spinal deformities and treatment planning. However, past methods have often treated these tasks separately, neglecting their inherent relationship. This paper proposes a single-stage 2D centroid-detection guidance segmentation network (CD-VerTransUNet) that utilizes global information between vertebrae and the relationship between the two tasks. Moreover, a resampler module enhances the segmentation of rare (e.g. T13/L6) vertebrae. The proposed model demonstrates state-of-the-art segmentation performance for 2D models on the VerSe'20 dataset, achieving a dice-coefficient (DSC) of 75.15% for sagittal and 71.16% for coronal plane. Our novel multitasking approach even shows comparable performance to 3D architectures, yielding a DSC of 77.02% on the VerSe'20 and 71.75% on a scoliotic dataset.

***Index Terms*—** Vertebrae segmentation, Vision transformer, Centroid detection

1. INTRODUCTION

Scoliosis, vertebral deformities, and stenosis are common spinal conditions that significantly impact mobility and overall health [1]. Accurate diagnosis is critical for treatment, especially in cases where surgery is required. Computed tomography (CT) is extensively used for pre-operative planning and post-operative monitoring. It enables surgeons to localize the spinal pathology and develop patient-specific treatment. Automated identification and segmentation of the vertebrae column offers many benefits, accelerating surgical planning and eliminating manual subjectivity, albeit being a challenging computational task. Challenges include the high morphological similarity of neighboring vertebrae (inter-class), occasionally significant intra-class variation, where the same vertebrae may differ considerably in patients with different pathologies, and even variations in the total number (cases of an additional vertebra (T13/L6) or lack of one (T12)) among individuals [2]. Furthermore, the absence of a reference vertebra, top (C1) or bottom (L5), vertebra in varying or limited field of view CT imaging, where only a portion of the thoracic spine is imaged, further complicates automated identification.

Initial efforts focused on model-fitting approaches [3], while the current state-of-the-art concentrates on learning-based methods to perform vertebrae segmentation and identification [4]. Most methods are structured as multi-stage, cascaded networks that first perform localization, followed by segmentation of detected vertebrae. Payer et al. introduced a 3-stages framework involving spine recognition, localization of vertebrae centroids, and vertebrae segmentation using three separate U-Net models [5]. Cheng et al. proposed two cascaded U-Nets for vertebrae detection and segmentation [6]. Despite promising results, cascaded architectures demand independent training and optimisation for each stage, thus increasing computation and susceptibility to error accumulation, ultimately producing models that can not generalize effectively. Alternatively, vertebrae identification and segmentation can be considered an instance segmentation problem and directly addressed with single-stage frameworks. Lessman et al. proposed an iterative 3D Convolutional Neural Network (CNN) where segmented vertebrae are kept in an instance memory, and subsequent vertebrae are progressively detected and segmented in a sliding window fashion [7]. For this, repetitive searching is required for identifying the first vertebrae. Recently, a multi-task, single-stage approach is proposed for lumbar vertebrae segmentation in magnetic resonance (MR) imaging [8]. The detection branch generates a probability heatmap of each vertebra's location, assigning the highest value as the centroid of the predicted vertebra. However, calculating maximum heatmap values is non-differentiable and takes place offline, thus not contributing to model optimization and precluding end-to-end training. Typically, 3D single-stage models utilize cropped

Corresponding authors: Sevde Aydogdu (sevde.aydogdu.21@ucl.ac.uk), Evangelos Mazomenos (e.mazomenos@ucl.ac.uk). This research was partly supported by the Turkish Ministry of National Education, Republic of Turkiye, a Wellcome Innovator Award [223793/21/Z] and the Wellcome/ EPSRC Centre for Interventional and Surgical Sciences [203145/16/Z].

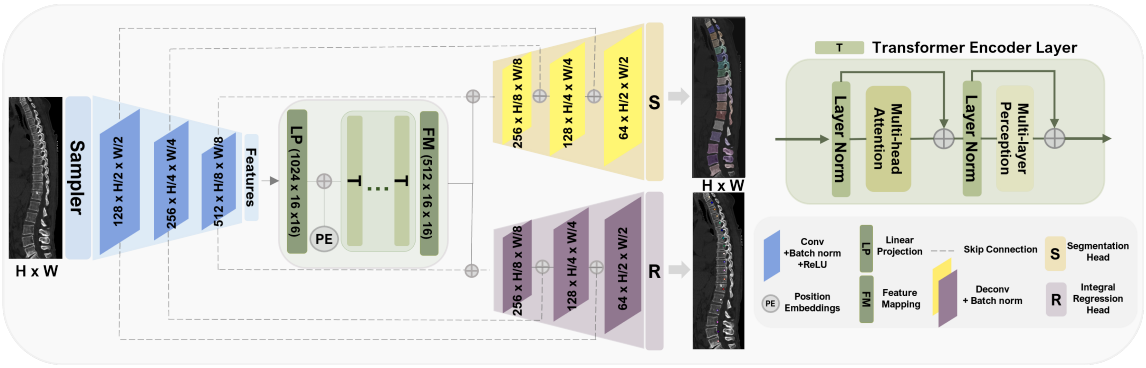

Fig. 1. CD-VerTransUNet network architecture with modified-ResNet-50-based feature extractor, transformer encoder layers, and parallel decoder branches with skip connections.

patches, thus operate on a specific region of the CT volume [9]. This limits feature extraction with global context, which affects vertebrae identification.

This study presents a detection-guided, single-stage, end-to-end trainable model (CD-VerTransUNet) for the segmentation and localization of vertebrae centroids in CT. Our work represents the first attempt at developing a single-stage, multi-task model that utilizes centroid detection to optimize vertebrae segmentation in CT imaging. We incorporate a Vision-Transformer (ViT) in a UNet base model to exploit the sequential relationships between vertebrae. A shared encoder module facilitates feature generation by considering the inherent relationship between the segmentation and localization tasks, emphasizing on learning the spatial relationships of the vertebrae throughout the spinal column. **To enhance the efficiency of ViT, we utilize 2D longitudinal slices of the CT volume instead of 3D cropped patches**, which differentiates our work from 3D approaches [10, 11]. Different to these methods, our model's ViT encoder is deployed in a multi-tasking architecture, utilizing the correlation between the two tasks to accurately identify the number of vertebrae in each CT volume. **To improve the efficiency of predicting the longitudinal structure of the spinal column, our goal is to include a higher number of vertebrae within the receptive field of our model in the spatial domain**. In the detection path, we employ integral regression to estimate the vertebrae centroids and guide the segmentation task. In contrast to direct regression, which can potentially ignore spatial information due to the fully connected layer pathway, we utilize the estimated heatmaps to predict the centroids' coordinates. This ensures that spatial information is preserved, leading to accurate centroid estimation while the shared features are used to optimize segmentation. Our contributions are summarised as:

- We present CD-VerTransUNet, a detection-guided segmentation model for vertebrae segmentation and identification in CT images. Through multi-tasking, the model leverages the detection of vertebrae centroid to enhance the segmentation task.
- Integrating ViT to UNet, our model captures contextual information along the spine's length, even between vertebrae at different sections (cervical-lumbar).
- As a training augmentation step, we introduce a resampling module to address rare vertebrae cases (e.g. T13/L6) and tackle the dataset imbalance. This enables the model to generalize more effectively, improving its ability to segment and identify rare vertebrae.
- We comprehensively validate the proposed model on the publicly available VerSe'20 dataset [12] and a scoliotic cadaver dataset. CD-VerTransUNet exhibits comparable performance to 3D approaches, achieving a Dice Score Coefficient (DSC) of 77.02% on the VerSe'20 and 71.75% on a scoliotic cadaver dataset.

2. METHODOLOGY

We propose a TransUNet-style [13] architecture, illustrated in Figure 1, with a shared encoder, combining ViT and CNN blocks to leverage the relationship between the two semantic tasks to extract high-quality features for both. A modified ResNet-50 encoder, excluding max pooling layers to preserve high spatial resolution, is the backbone for mapping high-level features. The features from the last CNN layer are then collapsed, splitting the feature maps into 16×16 sequential patches, used as input to the ViT. These patches are mapped into an embedding space, where each patch is represented numerically, capturing unique features. Transformers inherently lack an understanding of patch sequence order, requiring positional encoding to retain localization details. A learnable positional encoder is used to encode the patch's positional information within a sequence. The multiple attention heads process the information to capture different aspects of the patch relationships within the sequence, facilitating the model to learn diverse patterns and representations.

Independent decoder pathways are utilized. For recovering the full-size segmentation mask ($H \times W$), up-sampling convolutions are applied on the feature maps. Skip connec-

tions between corresponding encoder/decoder layers enable the propagation of gradients during training and help incorporate contextual information, enhancing the model's ability to understand and represent fine details for the segmentation task. For centroid regression, we follow a similar approach to [14, 15] and use integral regression to predict the center coordinates of the vertebrae directly. Similar to the segmentation decoder, the regression pathway produces heatmaps in the original size. We then apply softmax for normalization, with each pixel value in the normalized maps representing the probability of the vertebrae centroid position. The normalized probability maps are weighted using a trainable parameter and then summed with their corresponding coordinates. This generates a coordinate prediction for each dimension (x and y) in the 2D domain, as:

$$f_{\text{ni}}^{\text{reg}}(v) = \sum_{j}^{H \times W} W_{ij} \cdot P_n(v) \tag{1}$$

where H, W is the height and width of the CT slices. $P_n(v)$ is the n^{th} vertebra normalized probability map, while W_{ij} is a trainable parameter that represents the pixel coordinates into the normalized probability map with size $2 \times H \times W$.

2.1. Training strategy

To optimize CD-VerTransUNet, a weighted combination of cross-entropy and dice loss is used for segmentation, and the mean square error (MSE) is employed for centroid regression. The overall loss is defined as: $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{mse}}$ where hyper-parameters, α and β, are set to 0.5, weighting each segmentation loss equally. Additionally, a vertebrae resampling module is integrated into the framework to address the class imbalance issue and contribute to vertebrae identification. Extreme vertebrae cases, such as lumbarization vertebrae (L6), occur infrequently, which raises challenges due to the imbalanced class distribution. The module oversamples rare vertebrae (e.g. L6, T13) by a factor of 10 and undersamples input slices without any vertebrae by a factor of 0.1.

2.2. Implementation details

CT datasets are preprocessed via resampling into an isotropic voxel space of $1 \times 1 \times 1\ mm^3$ and reorienting to the same anatomic orientation (PIR; Posterior, Inferior, and Right) to mitigate variations introduced during image acquisition. Images and corresponding masks are cropped using the smallest bounding box, including the full spine, to reduce input size. Voxel values are normalized to [0,1] with z-score normalization, facilitating faster network convergence. Random rotation and cropping are applied for data augmentation. Networks were initialized using pre-trained ImageNet-21k weights and trained for 200 epochs. The AdamW optimizer was employed with a base learning rate of 5×10^{-4} and a weight decay rate of 1×10^{-4} ($\beta_1 = 0.9$, $\beta_1 = 0.999$).

Table 1. CD-VerTransUNet results and comparison with 2D/3D single-stage methods on the VerSe'20 benchmark (DSC(%), HD95(mm), id-Rate(%), and d_{mean}(mm))

	Segmentation		Identification	
Methods	DSC	HD95	id-Rate	d_{mean}
3D Single-stage Approaches				
Mask Ret-Net [16]	65.2	20.4	**84.9**	**10.2**
A^2Unet [16]	81.7	15.7	3.4	204.8
VerteFormer [11]	**86.5**	**3.6**	*	*
2D Single-stage Approaches				
U-Net [16]	25.5	240.6	*	*
Mask R-CNN [16]	58.2	99.8	9.2	191.0
CD-VerTransUNet-S	**75.2**	**11.1**	63.4	17.6
CD-VerTransUNet-C	71.2	12.9	**65.2**	**16.4**
Multi-Plane Fusion Single-stage Approach				
CD-VerTransUNet-MP	**77.0**	**10.3**	**63.8**	**17.0**

CD-VerTransUNet-S: Sagittal, -C: Coronal, and -MP: Multi-Planes

All models and experiments were implemented in PyTorch. After obtaining segmentation probability maps and centroid predictions of all vertebrae in the sagittal and coronal planes, we fuse them to generate final predictions. This leverages the 3D spatial context of the volumetric CT data. We then map the 2D slices back into 3D space to obtain a volume-based output.

3. RESULTS

3.1. Dataset and evaluation criteria

The VerSe'20 benchmark with its standard dataset splits is used [16]. To further evaluate the model, four CT scans of scoliotic cadavers from the Royal National Orthopaedic Hospital are included. Our study follows the VerSe'20 challenge metrics, using the DSC and Hausdorff Distance (HD95) for segmentation and the Identification Rate (id-Rate) along with the distance error (d_{mean}) for vertebrae identification.

3.2. Comparison against 2D/3D single-stage methods

CD-VerTransUNet results from both coronal and sagittal planes alongside a comparison to state-of-the-art 2D and 3D single-stage models are summarized in Table 1. Our method clearly outperforms previous 2D approaches, which typically report only sagittal results, achieving a 17.0% higher DSC and 88.7 mm shorter HD95 distance compared to Mask R-CNN [16]. CD-VerTransUNet has similar performance in both planes, with the coronal output showing an improvement in labeling performance but a small decrease in segmentation compared to the sagittal.

Against 3D single-staged models, CD-VerTransUNet using the multi-plane approach exhibits comparable performance with VerteFormer reporting superior segmenta-

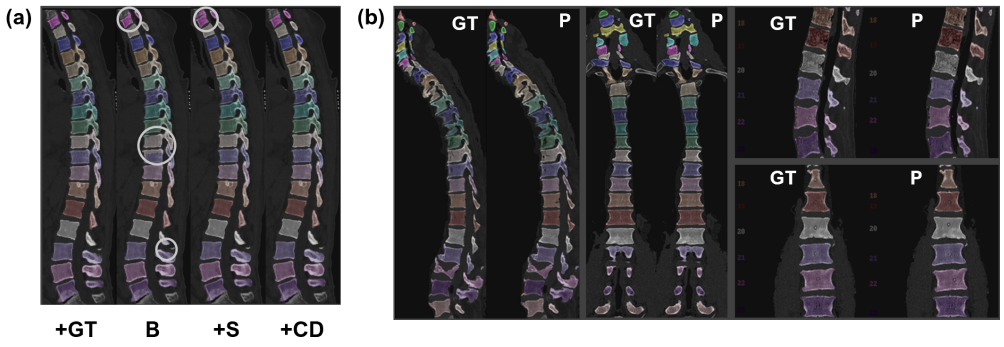

Fig. 2. Visualization of the predicted segmentation masks (P) and their corresponding ground truth masks (GT); (a) for the ablation study (baseline, + resampling, + centroid-detection) of CD-VerTransUNet on VerSe'20, and (b) for the scoliotic dataset

Table 2. Ablation study results on the entire spine, specific regions, and the lumbarization L6 vertebra (DSC(%)/HD95(mm))

	B	***B + S***	***B + S + CD***
Spine	68.3 / 14.5	75.6 / 11.0	77.0 / 10.3
Cervical	74.9 / 6.6	82.5 / 4.1	82.8 / 4.3
Thoracic	62.2 / 26.1	69.7 / 20.4	70.9 / 18.7
Lumbar	61.3 / 19.7	69.2 / 14.2	75.1 / 12.4
L6	21.6 / 29.5	39.8 / 14.3	71.4 / 7.9

tion [11], but without considering the identification task. Compared against A^{2}Unet, a 3D multi-task framework, CD-VerTransUNet obtains an increase of 60.4% in the id-Rate performance and a decrease of 187.8 mm in d_{mean}, thanks to our integral regression-based approach [16]. This highlights the effectiveness of integral regression in detecting vertebrae centroids.

3.3. Ablation study

An ablation study of CD-VerTransUNet on the VerSe20 dataset is performed to highlight each module's impact (see Table 2). Starting with the baseline model without flipping augmentation (B) for multi-label vertebrae segmentation, we explored the module performance, the vertebrae resampling module (S), and the regression task for centroid detection (CD). Relative to the baseline model, our model shows significant improvements across all spinal regions: the averages of DSCs increased by 7.9% (cervical), 8.7% (thoracic), and 13.8% (lumbar), with a remarkable 49.8% increase for the L6. Correspondingly, it decreased the HD95 distances by 2.3 mm, 7.4 mm, and 7.3 mm, notably by 21.6 mm for L6, resulting in an overall 8.7% improvement in segmentation DSC and a 4.2 mm reduction in HD95 distance. Our model efficiently extracts selective feature information along the vertebral column, proving effective for vertebrae segmentation. Figure 2a visualizes examples of improved vertebrae labeling and refined segmentation due to the resampling (+S) and centroid detection (+CD) modules. The addition of these components addresses class imbalance, due to rare vertebrae, and significantly boosts segmentation by providing extra supervision and more informative global features.

Table 3. Performance evaluation on the scoliotic dataset (DSC(%), HD95(mm), id-Rate(%), and d_{mean}(mm)

	Segmentation		Identification	
Methods	DSC	HD95	id-Rate	d_{mean}
CD-VerTransUNet-S	72.4	10.4	42.8	27.5
CD-VerTransUNet-C	51.5	22.31	41.6	25.1
CD-VerTransUNet-MP	71.75	9.81	42.3	26.3

3.4. Scoliotic dataset evaluation

Table 3 presents CD-VerTransUNet performance in the scoliotic cadaver dataset. Despite trained exclusively on non-scoliotic cases, our approach exhibits good generalization as demonstrated by the sagittal and multi-plane results. The decrease in id-Rate and d_{mean} is due to the high pathological variability and degree of deformation among the scoliotic samples, also illustrated in Figure 2b.

4. CONCLUSION

This paper introduces CD-VerTransUNet, a single-stage, centroid detection guided, and end-to-end trainable model integrating vertebrae segmentation and identification. Leveraging multi-head global attention and co-optimizing features both for segmentation and centroid regression, significantly boost segmentation performance. Our model outperforms 2D architectures, with a 17% higher DSC and 88.7mm lower HD95 in VerSe'20, and compares favourably with state-of-the-art 3D models. Prelimary validation on a scoliotic dataset shows encouraging generalization outcomes despite the model not being optimized in these type of spinal deformation.

5. COMPLIANCE WITH ETHICAL STANDARDS

This article contains no studies with human participants performed by any authors.

6. REFERENCES

[1] Jack C Cheng et al., "Adolescent idiopathic scoliosis," *Nature reviews disease primers*, vol. 1, no. 1, pp. 1–21, 2015.

[2] Shumao Pang et al., "SpineParseNet: Spine Parsing for Volumetric MR Image by a Two-Stage Segmentation Framework with Semantic Image Representation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 262–273, 1 2021.

[3] Robert Korez et al., "A Framework for Automated Spine and Vertebrae Interpolation-Based Detection and Model-Based Segmentation," *IEEE Trans Med Imaging*, vol. 34, no. 8, pp. 1649–1662, 8 2015.

[4] Cheng-Hung Chuang et al., "Efficient triple output network for vertebral segmentation and identification," *IEEE Access*, vol. 7, pp. 117978–117985, 2019.

[5] Christian Payer et al., "Coarse to fine vertebrae localization and segmentation with spatialconfiguration-Net and U-Net," *VISIGRAPP (5: VISAPP)*, pp. 124–133, 2020.

[6] Pengfei et al. Cheng, "Automatic vertebrae localization and segmentation in CT with a two-stage Dense-U-Net," *Sci. Rep.*, vol. 11, no. 1, 2021.

[7] Nikolas Lessmann et al., "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Med. Image Anal.*, vol. 53, pp. 142–155, 4 2019.

[8] Shumao Pang et al., "Dgmsnet: Spine segmentation for mr image by a detection-guided mixed-supervised segmentation network," *Medical Image Analysis*, vol. 75, pp. 102261, 2022.

[9] Xin You et al., "Learning with explicit shape priors for medical image segmentation," *arXiv preprint arXiv:2303.17967*, 2023.

[10] Rong Tao et al., "Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers," *Med. Image Anal.*, vol. 75, 1 2022.

[11] Xin You et al., "Verteformer: A single-staged transformer network for vertebrae segmentation from ct images with arbitrary field of views," *Medical Physics*, 2023.

[12] Hans Liebl et al., "A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data," *Sci Data*, vol. 8, no. 1, pp. 1–7, 2021.

[13] Jieneng Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *preprint arxiv:2102.04306*, 2 2021.

[14] Xiao Sun et al., "Integral human pose regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 529–545.

[15] Chunli Qin et al., "Vertebrae labeling via end-to-end integral regression localization and multi-label classification network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2726–2736, 2021.

[16] Anjany Sekuboyina et al., "VERSE: A Vertebrae labelling and segmentation benchmark for multi-detector CT images," *Med. Image Anal.*, vol. 73, pp. 102–166, 10 2021.