Evaluation of Deep Neural Network Models for Instance Segmentation of Lumbar Spine MRI

Jiasong Chen [a], *Linchen Qian* [a], *Linhai Ma* [a], *Timur Urakov* [b], *Weiyong Gu* [c], and *Liang Liang* [a]

[a] *Department of Computer Science, University of Miami, Coral Gables, FL*

[b] *Department of Neurological Surgery, University of Miami, Coral Gables, FL*

[c] *Department of Mechanical and Aerospace Engineering, University of Miami, Coral Gables, FL*

For correspondence:

Liang Liang, Ph.D.

Department of Computer Science

University of Miami

Ungar Building, Room 330K

Coral Gables, FL, 33146

Tel: (305) 284-8381; Email: liang@cs.miami.edu

Abstract

Intervertebral disc disease, a prevalent ailment, frequently leads to intermittent or persistent low back pain, and diagnosing and assessing of this disease rely on accurate measurement of vertebral bone and intervertebral disc geometries from lumbar MR images. Deep neural network (DNN) models may assist clinicians with more efficient image segmentation of individual instances (discs and vertebrae) of the lumbar spine in an automated way, which is termed as instance image segmentation. In this work, we evaluated 15 existing DNN models for lumbar spine MR image segmentation. We introduced a new data augmentation technique to create synthetic yet realistic MR image dataset, named SSMSpine, which is made publicly available. The 15 image segmentation models are evaluated on our private in-house dataset and the public SSMSpine dataset, using two metrics, Dice Similarity Coefficient and 95% Hausdorff Distance. The SSMSpine dataset are available at https://github.com/jiasongchen/SSMSpine.

Keywords: Lumbar spine MRI, Medical image instance segmentation, Data augmentation

1. Introduction

The intervertebral discs in humans can undergo a profound degenerative process as early in the adolescence (Cox et al., 2014; Kos et al., 2019), which can be accompanied by facet arthropathy and hypertrophy. This degeneration can manifest as various conditions, including discogenic low back pain, disc herniation, spinal stenosis, and spondylolisthesis, which may necessitate the implementation of surgical or non-surgical interventions aimed at alleviating pain and restoring normal functionality. Magnetic resonance imaging (MRI) is the most widely used technique for specifically quantifying intervertebral discs degeneration (IDD) by assessing changes in disc geometry deformation and signal strength degradation (Mallio et al., 2022; Roberts et al., 2021; Tamagawa et al., 2022). The information derived from imaging data is of utmost importance for medical professionals in terms of both diagnosing medical conditions and planning appropriate treatments. Furthermore, this information serves as a critical foundation for developing patient-specific computational models, which hold the potential to mature over time and eventually enable accurate predictions of treatment outcomes within clinical settings. Presently, the process of geometry reconstruction, signal measurements, and grading from magnetic resonance (MR) images heavily relies on manual annotation. However, this process is not only time-consuming but also vulnerable to human bias. Consequently, there is an urgent need for automated MR image analysis methods to address these challenges.

In medical imaging, semantic/instance image segmentation, which divides the images into distinct sections at the pixel level so that each pixel belongs to a specific region, has the potential to be carried out through automated techniques (Galbusera et al., 2019). The traditional methods, such as watershed and level set, have demonstrated satisfactory performance in medical image segmentation tasks. The watershed method treats an image as a topological map where intensity represents the altitude of the pixels. The watershed segmentation is determined by the watershed lines on a topographic surface (Chevrefils et al., 2007; Huang and Chen, 2004). The level set method performs image segmentation by utilizing dynamic variational boundaries (Huang et al., 2013). However, the traditional method suffers from the clinical variation of different patients and the noise effect of different medical imaging equipment, problems like the over-segmentation and time-consuming consist (Li et al., 2007).

Since the increasingly vast amount of medical imaging data and computational resources have become available, machine learning (ML) methods, especially deep neural network techniques, show superior performance than traditional methods. Convolutional neural network (CNN) has a significant edge over its predecessors in that it possesses the capability to recognize essential components/features without requiring any human intervention (Suganyadevi et al., 2022). CNNs are specifically designed to effectively utilize spatial and configural information by accepting 2D or 3D images as input. This approach helps to prevent the loss or disruption of structural and configural information in medical images (Shen et al., 2017).Various deep CNNs, including UNet++ (Zhou et al., 2018), Attention U-Net (Oktay et al., 2018), MultiResUNet (Ibtehaz and Rahman, 2020) and UNeXt (Valanarasu and Patel, 2022) have been proposed for image segmentation for different medical imaging modalities and different organs (e.g. heart (Cao et al., 2023; Gao et al., 2021; Huang et al., 2023), lung (Zhou et al., 2018), brain (Hatamizadeh et al., 2022, 2021; Hu et al., 2022; Valanarasu et al., 2021), pancreas (Oktay et al., 2018), gland (Valanarasu et al., 2021; Wang et al., 2022), spine (Sekuboyina et al., 2018; Wang et al., 2023), retina blood vessels (Moccia et al., 2018; Soomro et al., 2019), aorta (Berhane et al., 2020; Noothout et al., 2018; Pepe et al., 2020), etc). Although these methods have achieved promising performance, there are still some limitations in a more complex context coping with long-range dependency explicitly due to the intrinsic locality of convolutions.

Recently, Transformer, an ML technique, has shown exceptional performance not only on natural language processing (NLP) challenges like machine translation (Vaswani et al., 2017), but also image analysis tasks including image classification (Shamshad et al., 2023) and segmentation (Chen et al., 2021; Hatamizadeh et al., 2022, 2021; Liu et al., 2021; Wang et al., 2022). Various variations of Transformer models have demonstrated that the global information perceived by the self-attention operations is beneficial in medical imaging tasks. TransUNet was the first Transformer-based network specifically for medical image segmentation on the synapse multi-organ segmentation dataset (Chen et al., 2021). Wang et al. (2022) substituted the original skip connection scheme of U-Net with the proposed UCTransNet that includes a multi-scale Channel Cross fusion Transformer and a Channel-

wise Cross-Attention and tested the network on the gland segmentation dataset (Sirinukunwattana et al., 2017) and synapse multi-organ segmentation dataset (Landman et al., 2015). Hatamizadeh et al., (2021, 2022) proposed both UNETR and Swin UNETR for 3D medical imaging segmentation. UNETR utilizes a U-shape network with a vision Transformer as the encoder and a CNN-based decoder. Swin UNETR is constructed by replacing the vision transformer encoder in UNETR architecture with the Swin Transformer encoder. Feng et al. (2022) proposed SLT-Net to utilize CSwin Transformer (Dong et al., 2022) as the encoder for feature extraction and the multi-scale context Transformer as the skip connection for skin lesion segmentation. Swin-Unet adopted Swin Transformer (Liu et al., 2021) with shifted windows as encoder and a symmetric Swin Transformer-based decoder with patch expanding layer as decoder for multi-organ segmentation task (Cao et al., 2023). Pu et al., (2023) proposed a semi-supervised learning framework with Inception-SwinUnet adopting convolution and sliding window attention in different channels for vessel segmentation on small amount of labeled data. Besides the self-attention mechanism, position embeddings are another crucial component of Transformer models. Regarding changing the order of the input, a Transformer model is invariant (Vaswani et al., 2017) without position embeddings. However, since text data inherently has a sequential structure, the absence of position information results in the ambiguous or undefined meaning of a sentence (Dufter et al., 2022). For image segmentation, usually, an image patch is treated as a token, and Transformers process the entire input sequence of tokens in parallel. With position embeddings, a Transformer would be able to differentiate between image patches with similar content that appear in different positions in the input image, which is beneficial for image segmentation applications. A variety of different methods may be used to incorporate the position information into Transformer models. Absolute position encoding and relative position encoding are two main categories to encode a token's position information. Vaswani firstly introduced absolute and relative position embedding in the vanilla Transformer model (Vaswani et al., 2017). Shaw extended the self-attention mechanism with the capacity of effectively incorporating the representation of relative position (Shaw et al., 2018). Valanarasu et al. (2021) proposed a gated position-sensitive axial attention mechanism to cope the difficulty in learning position encoding for the images.

Specifically for lumbar spine research, instance segmentation of MR images is preferred, which not only determines whether or not a pixel belongs to a disc, but also labeling the precise instance to which it belongs (Galbusera et al., 2019). In recent years, most instance segmentation methods for spine image segmentation are based on CNNs-only networks, and only a few Transformer-based networks are employed. For example, Kuang et al. (2020) built an unsupervised segmentation network for spine image segmentation using the rule-based region of interest (ROI) detection, a voting mechanism accompanied by a CNN network. Sekuboyina et al. (2018) proposed a dual branch fully convolutional network that take advantages of both low-resolution attention information on two-dimensional sagittal slices and high-resolution segmentation context on three-dimensional patches for effective segmentation of the vertebrae. MLKCA-Unet incorporates multi-scale large-kernel convolution and convolutional block attention into the U-net architecture for efficient feature extraction in spine MRI segmentation (Wang et al., 2023). Pang et al. (2022) introduced a mixed-supervised segmentation network and it was trained on a strongly supervised dataset with full segmentation labels and a weakly-supervised dataset with only key points. BianqueNet combined new modules with a modified deeplabv3+ network (Chen et al., 2018), which includes a Swin Transformer-skip connection module, for segmentation of lumbar intervertebral disc degeneration related regions (Zheng et al., 2022).

It was shown that the Transformer-based models only perform effectively when trained on large-scale datasets since the lack of inductive bias (Dosovitskiy et al., 2021). The utilization of Transformer-based networks for medical imaging tasks poses a challenge due to the limited availability of labeled images in medical datasets. Obtaining well-annotated medical imaging datasets presents significantly greater challenges compared to curating traditional computer vision datasets. Dealing with expensive imaging equipment, complex image acquisition pipelines, expert annotation requirements, and privacy concerns are all part of the problematic issues (Litjens et al., 2017). This scarcity hampers the effective application of Transformer-based models in the medical domain. In such scenarios, the adoption of suitable and feasible data augmentation techniques becomes crucial prior to model training. These techniques can help to increase the effective size of the medical image dataset and improve the performance of the Transformer-based model.

In this study, we evaluated 15 DNN modes for lumbar spinal MRI instance segmentation. For this purpose, we developed a novel data synthesis method based on statistical shape model (SSM) and biomechanics. This SSM-biomechanics-based data synthesis method generates lumbar spine images with large and plausible deformations, which can be used for model training and evaluation.

2. Review of Transformer-based DNN models for Lumbar Image Segmentation

2.1. Transformer-based Networks for the segmentation of non-spine medical images

Transformer-based networks have shown promising performance on medical image segmentation because of their ability to capture long-range dependencies. Comparably, the inductive bias of CNN networks benefits from its local connectivity and parameter sharing property. Therefore, many networks, combining CNN and Transformers and leveraging both benefits, have been proposed in the past few years. TransUNet (Chen et al., 2021) firstly combines CNNs and Transformer in a cascaded manner in its encoder for medical imaging tasks, in which the low-level features are collected from the CNNs and then fed to the Transformer to capture global interactions. Other designs of CNN and Transformer networks, including UNETR (Hatamizadeh et al., 2021), UTNet (Gao et al., 2021), UCTransNet (Wang et al., 2022), Swin UNETR (Hatamizadeh et al., 2022) and MedT (Valanarasu et al., 2021) have shown better segmentation performances in different medical image modalities, compared to CNN-only networks. UNETR replaces the encoder of a UNet with Transformer layers with 1D learnable positional embedding, and its decoder only has convolution layers (Hatamizadeh et al., 2021). UTNet inserts Transformer layers into a Unet in a sequential manner: a convolution layer followed by a Transformer layer with learnable relative position embedding (Gao et al., 2021). UCTransNet embeds Transformer layers into the skip-connections of a Unet with fully learnable absolute position embedding (Wang et al., 2022). Swin UNETR replaces the Transformer in UNETR with Swin-Transformer (Hatamizadeh et al., 2022). Medical Transformer (MedT) uses gated-axial Transformer layers in the encoder of a Unet (Valanarasu et al., 2021). HSNet used PVTv2 (Wang et al., 2022) as encoder and a dual-branch structure which Transformer branch and CNN branch fused by element-wise product as decoder for polyp segmentation (Zhang et al., 2022).

2.2. Transformer-based Networks for the segmentation of spine images

To the best of our knowledge, there are only a few Transformer-based networks (You et al., 2022; Tao et al., 2022) specifically designed for lumbar spine image segmentation, including EG-Trans3DUNet (You et al., 2022), Spine-transformers (Tao et al., 2022), APSegmenter (Zhang et al., 2022), and BianqueNet (Zheng et al., 2022). However, it is worth noting that most of these networks were developed for the segmentation of vertebral bodies using CT modality, which may not be directly applicable to the segmentation of the intervertebral discs in order to study disc degeneration. Amony those networks, only the code of BianqueNet is publicly available.

EG-Trans3DUNet combines two vision transformer branches to handle both local patches and resized global spinal CT images (You et al., 2022), and it merges edge characteristics and semantic features generated by a CNN-based edge detection block. Spine-transformers was designed for spinal CT image segmentation with a two-stage pipeline to handle arbitrary Field-Of-View input images (Tao et al., 2022). In its first stage, a Transformer with a CNN backbone is utilized as a 3D object detector to locate individual vertebrae, and then the input image is cropped into regions of individual vertebrae. In its second stage, a multi-task encoder-decoder CNN network is applied to each cropped region to segment the vertebra. The source code of the second stage is not publicly available. APSegmenter (Zhang et al., 2022) combines a ViT-style Transformer with a mask Transformer to segment spine X-ray images, and an adaptive postprocessing is applied to further refine the result. BianqueNet (Zheng et al., 2022) employed a resnet101 network to perform feature extraction, followed by upsampling using the Swin Transformer-skip connection module and a double upsampling operations. It also used a multi-scale feature fusion module to generate the segmentation of regions associated with intervertebral disc degeneration.

In addition to the segmentation of vertebral bodies, the segmentation of intervertebral discs (IVDs) is vital for lumbar spinal disease diagnosis and treatment. Since the water content in IVDs cannot be revealed on CT images, currently, MRI is the gold standard imaging modality for the evaluation of IVD pathologies (Kirnaz et al., 2022). Our study aims to explore the benefit of combining CNN and transformer for instance segmentation of lumbar spine MR images.

2.3. Self-Attention in Transformer

As opposed to convolutional operations, the self-attention mechanism within a Transformer network has the fundamental advantage of effectively capturing global features and long-range contextual dependency. It uses the Key, Query and Value vectors to better describe the features' connections. Nonetheless, because of the inherited properties of self-attention, it doesn't retrieve the position information on its own, which is important for instance segmentation. One of the best ways to tackle this problem is to use a well-designed position embedding mechanism to inject the position relationships into the self-attention calculation.

2.3.1. The classic self-attention mechanism with additive position embedding

The plain Transformer is constructed with the multi-head self-attention modules (MHSA), which enable Transformer to capture and utilize more accurate and detailed spatial information (Vaswani et al., 2017). Given an input token set (e.g., image patches) X, three individual linear transformations (W_Q, W_K, W_V) are applied to X to generate query embedding (Q), key embedding (K), and value embedding (V). Then, the self-attention score, $Attn$, is calculated as a scale-product of these three embedding as following:

$$Attn = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \tag{1}$$

$$Out = Attn \times V \tag{2}$$

In the above equations, $Q = (X + P)W_Q$, $K = (X + P)W_K$, $V = (X + P)W_V$. P is the encoded position, and d is the dimension of embedding in each head. Out is the final output of the self-attention module.

2.3.2. Position Embedding

Without using any position embedding, the self-attention mechanism in Eq.(1) is permutation-invariant and cannot distinguish tokens (e.g., image patches) at different spatial locations. Therefore, it is essential to design efficient position embedding cooperating with the self-attention mechanism.

The attention matrix in the classic self-attention Eq.(1) can be decomposed into three terms:

$$Attn \sim QK^T = \underbrace{XW_Q(XW_K)^T}_{\text{Content-Content}} + \underbrace{P\,W_Q(XW_K)^T + XW_Q(PW_K)^T}_{\text{Content-Position}} + \underbrace{PW_Q(PW_K)^T}_{\text{Position-Position}} \tag{3}$$

Therefore, the attention considers three interactions/correlations among tokens: $XW_Q(XW_K)^T$ for content to content interaction, $P\,W_Q(XW_K)^T + XW_Q(PW_K)^T$ for interaction between content and position, and $PW_Q(PW_K)^T$ for position to position interaction. As the instance segmentation task is location-specific, a well-designed interaction between content and position could improve self-attention ability to utilize both content and position information to accomplish the instance segmentation task.

Generally, there are mainly two steps to define the position embedding. The first step is defining the position function or distance function, which is used for encoding the position information of input tokens. There are plenty of position functions such as index function, Euclidean distance, and sinusoidal functions etc., The second step is defining methods to incorporate the encoded position information into self-attention.

Absolute position embedding and relative position embedding are two main position representation methods to incorporate the position information into input tokens. Absolute position embedding encodes the absolute positions of each input tokens as individual encoding vectors, and relative position embedding focus on the relative positional relationships of pairwise input tokens (Lin et al., 2022; Wu et al., 2021). In the vanilla Transformer designed for NLP (Vaswani et al., 2017), it used a combination of absolute and relative position embedding to add position information to the tokens. It is inconclusive that if relative position embedding is better or worse than absolute position embedding, and the answer seems to be dependent on specific applications (Dufter et al., 2022; Huang et al., 2020; Shaw et al., 2018; Wu et al., 2021). Relative position encoding benefits from capturing the details of relative distance/direction and is invariant to tokens' shifting. The intuition is that, in the self-attention mechanism, the pairwise positional relationship (both in terms of direction and distance) between input elements might be more advantageous than absolute position of individual elements (Lin et al., 2022). In such a case, position information in Transformer is an extensive research area, and various relative position encodings have been proposed for medical imaging segmentation (Dosovitskiy et al., 2021). For example, UTNet proposed the 2-dimensional relative position encoding by adding relative height and width information (Gao et al., 2021). MedT updated self-attention mechanism with position encoding along the width axis with the inspiration of axial attention (Valanarasu et al., 2021; Wang et al., 2020; Zhang and Zhang, 2022). The Parameter-Efficient Transformer added a trainable position vector to the input to encode relative distances (Hu et al., 2022). In this work, we propose a novel relative position embedding method for segmentation performance improvement.

2.3.3. Image Self-Attention in the Existing Image Segmentation Models

We compared instance segmentation performances of 15 representative image segmentation models, including 11 Transformer-based models and 4 CNN-only models.

Table 1 summarizes the image self-attention mechanisms in the existing 11 Transformer-based models. TransUnet and UNETR add the position embedding directly into the input patches (Chen et al., 2021; Hatamizadeh et al., 2021). Swin-Unet, Swin UNETR, and Inception-SwinUnet incorporate the position embedding into attention score instead of input tokens. (Cao et al., 2023; Hatamizadeh et al., 2022; Pu et al., 2023). BianqueNet utilized the position embedding within Swin-Transformer (Zheng et al., 2022). According to the source code, SLT-Net introduces a lepe distance to represent the position bias (embedding), which is directly added into the output matrix (Feng et al., 2022). MedT proposed a gated position-sensitive axial attention mechanism where four learnable gates (G) control the amount of position embedding contained in key (K), query (Q) and value (V) embeddings (Valanarasu et al., 2021). UTNet introduced the 2-dimensional relative position encoding by adding relative position logits along height and width dimensions(R_{height}, R_{width}) into the key embedding (Gao et al., 2021). HSNet and UCTransNet do not include position embedding in their models (Zhang et al., 2022; Wang et al., 2022)

Table 1. Image Self-Attention in the attention/Transformer-based segmentation models

Model	Self-Attention
Swin UNETR	$softmax\left(\frac{XW_Q(XW_K)^T}{\sqrt{d}}+R\right)(XW_V)$ Note: R is called relative position bias in the reference
SLT-Net	$softmax\left(\frac{XW_Q(XW_K)^T}{\sqrt{d}}\right)(XW_V)+lepe(XW_V)$ Note: $lepe$ is a convolution kernel
UNETR	$softmax\left(\frac{(X+P)W_Q\left((X+P)W_K\right)^T}{\sqrt{d}}\right)(X+P)W_V$
Inception-SwinUnet	$softmax\left(\frac{XW_Q(XW_K)^T}{\sqrt{d}}+R\right)(XW_V)$ Note: R is called relative position bias in the reference
HSNet	$softmax\left(\frac{XW_Q(XW_K)^T}{\sqrt{d}}\right)(XW_V)$
Swin-Unet	$softmax\left(\frac{XW_Q(XW_K)^T}{\sqrt{d}}+R\right)(XW_V)$ Note: R is called relative position bias in the reference
TransUNet	$softmax\left(\frac{(X+P)W_Q\left((X+P)W_K\right)^T}{\sqrt{d}}\right)(X+P)W_V$
MedT	$softmax\left(XW_Q(XW_K)^T+G_QXW_Q\left(R_Q\right)^T+G_KXW_K(R_K)^T\right)\left(G_{V_1}XW_V+G_{V_2}R_V\right)$ Note: $G_Q, G_K, G_{V_1}, G_{V_2}$ are learnable parameters for gating mechanism. R_Q, R_K, R_V are the position bias for query, key and value
UTNet	$softmax\left(\frac{XW_Q\left(XW_K+R_{width}+R_{height}\right)^T}{\sqrt{d}}\right)XW_V$ Note: R_{width} and R_{height} are the relative height and width information
UCTransNet	$softmax\left(\frac{X_cW_Q(X_cW_K)^T}{\sqrt{d}}\right)(X_cW_V)$ Note: X_c is composed of image channels instead of image patches.
BianqueNet	$softmax\left(\frac{XW_Q(XW_K)^T}{\sqrt{d}}+R\right)(XW_V)$ Note: R is called relative position bias in the reference

3. Methods

3.1. Novel data augmentation/synthesis method based on SSM and biomechanics

For medical image data augmentation, elastic deformation is often used for nonlinear deformation of the images to increase diversity of training data (Ronneberger et al., 2015). Briefly, the input space is discretized by a grid, and a random displacement field on the grid is generated by sampling from a normal distribution with standard deviation equal to $\sigma \times$ grid resolution (i.e., the size of a grid cell). The parameter σ determines deformation magnitude. To ensure a large deformation with diffeomorphism, the grid needs to be coarser than the input size (i.e., 512 ×512). In this study, we applied two successive elastic deformations to each training image, with grid sizes of 9 × 9 and 17 × 17. As shown in Figure 1, when the deformation parameter σ is larger than 0.5, the generated images and spine shapes are highly unrealistic.

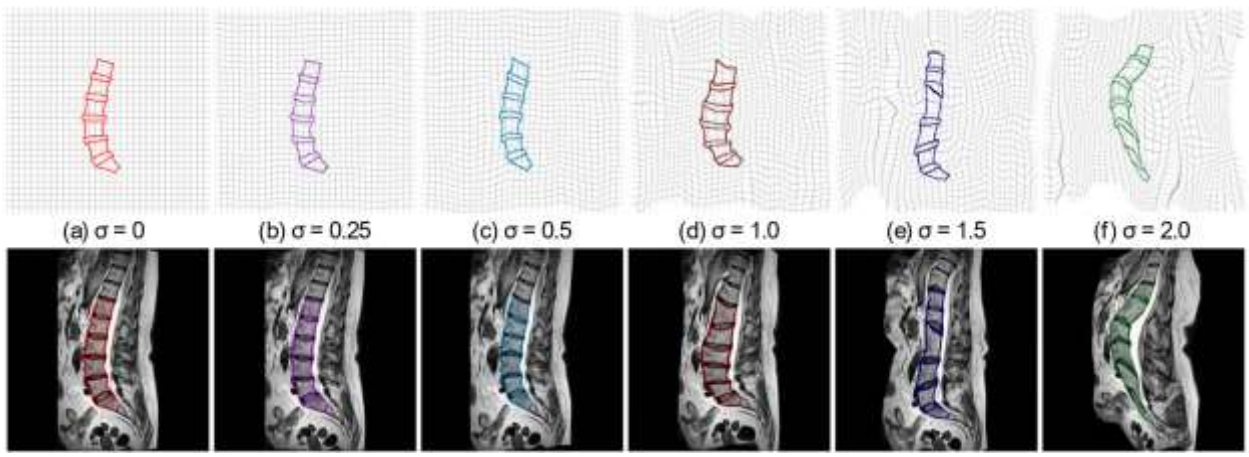

Fig. 1. Data augmentation/synthesis examples using elastic deformation with sigma from σ to 2.0.

In this work, we developed a new method to synthesize lumbar spine MR images suitable for model training and evaluation. First, we built a statistical shape model (SSM) of lumbar spine shapes (i.e., contours of discs and vertebrae) in a dataset set, and the SSM represents the probability distribution of lumbar spine shapes. We refer the reader to the reference papers (Ambellan et al., 2019; Sarkalkan et al., 2014; Hufnagel et al., 2007; Davies et al., 2003; Cootes et al., 1995) for the details of constructing an SSM. By sampling from the SSM, different lumbar spine shapes can be generated, and each generated lumbar spine shape could be considered from a virtual patient. We note that the SSM technique has been used to generate virtual but realistic patient geometries in many applications, such as generating aortic aneurysm geometries (Liang et al., 2017; van Veldhuizen et al., 2022; Wiputra et al., 2023). Given a lumbar spine shape, if a lumbar spine MR image can be generated and consistent with the shape, then we will have a new sample with ground-truth. For this purpose, we developed a biomechanics-based method to generate a lumbar spine MR image $\tilde{I}$ from a lumbar spine shape $\tilde{S}$ by using a reference image I with its ground-truth shape S. Intuitively speaking, a nonlinear spatial transform from the shape S to the shape $\tilde{S}$ is determined by using biomechanics principles, and then $\tilde{I}$ is obtained by applying the spatial transform to I. The generated images are visually plausible, as shown in Figure 2.

In the implementation, we obtain the spatial transform T from $\tilde{S}$ to S, and apply the spatial transform T to a regular mesh grid around the shape $\tilde{S}$ to obtain a deformed grid in the space of the reference image I, and then $\tilde{I}$ is obtained by interpolating pixel values of I at each node of the deformed grid, i.e., $\tilde{I}(x, y) = I(T(x, y))$ where (x, y) denotes a 2D spatial point and $T(x, y)$ is the transformed point. By using biomechanics and finite element analysis (FEA), the spatial transform, i.e., the deformation field on the mesh grid, is obtained by minimizing the following energy/loss function Π:

$$\Pi = \int_V \Psi \, dV + \lambda \cdot avg_i \left\| T(\tilde{S}(i)) - S(i) \right\|^4 \tag{1}$$

In the above Eq.(1), V represents the undeformed mesh grid of the image $\tilde{I}$ to be generated, $\tilde{S}(i)$ represents the i-th point location of the shape $\tilde{S}$, $T(\tilde{S}(i))$ is the transformed point location that needs to be equal to $S(i)$, and $\|\blacksquare\|$ denotes vector L2 norm. avg is the average operator. λ is a weight constant (set to 16 in experiments). Ψ is the strain energy density function that is determined by deformation and mechanical property of soft biological tissues around the lumbar spine. From the perspective of FEA and biomechanics, the Eq.(1) simulates the scenario that under the external "force" proportional to $T(\tilde{S}(i)) - S(i)$ at each point of the lumbar spine shape, the soft biological tissues of human body will deform and reach to an equilibrium state of minimum energy. To speed up the optimization process, we use a deep neural network with sine activation functions to parameterize the transform T, i.e., $T(x, y) = DNN(x, y)$, and then the energy function in Eq.(1) becomes a function of the DNN internal parameters. The energy optimization problem is resolved by adjusting/optimizing the parameters of the DNN. Once the optimization is done, the deformation field is obtained and then the image $\tilde{I}$ is generated. Since our goal is to generate plausible images for model training and evaluation in the image segmentation tasks, not for patient-specific FEA simulation of human body deformation, we made an assumption about the strain energy density function to reduce computation cost: tissue mechanical behavior follows the Ogden hyperelastic model with homogeneous tissue properties (Ogden and Hill, 1997; Dwivedi et al., 2022). The whole procedure is implemented by using our newly developed PyTorch-FEA library for large deformation biomechanics (Liang et al., 2023).

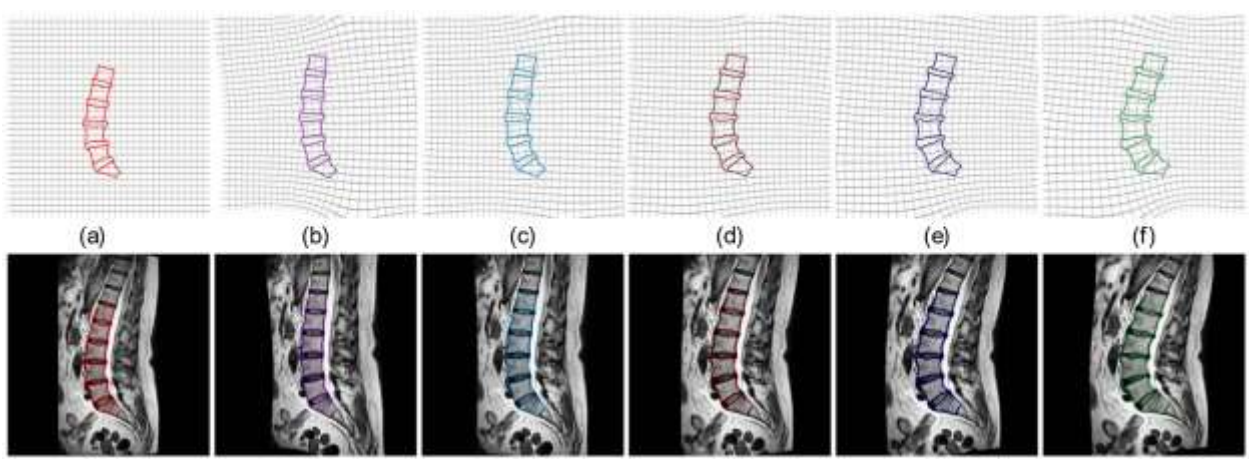

Fig. 2. data augmentation/synthesis examples (a-f) using our method. Please zoom in for better visualization.

3.2. Loss function

For each model, we use the original loss function if it applicable to our application. If the original loss function is not suitable (e.g., it is for binary classification only), then we use the loss function $\mathcal{L}$ that combines a Dice loss $\mathcal{L}_{Dice}$ and an area-weighted cross entropy loss $\mathcal{L}_{aw_ce}$.

$$\mathcal{L} = 0.5 L_{Dice} + 0.5 L_{aw_ce} \tag{2}$$

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_{i,j} \left(y(i,j) \hat{p}_m(i,j) \right) + \epsilon}{\sum_{i,j} \left(y(i,j) + \hat{p}_m(i,j) \right) + \epsilon} \tag{3}$$

$$\mathcal{L}_{aw_ce} = -\sum_{i,j,m} w_m y_m(i,j) log\left(\hat{p}_m(i,j) \right) \tag{4}$$

$\hat{p}_m(i,j)$ is the m-th element in the output tensor from the softmax layer at the pixel location (i,j), which corresponds to the m-th object (a disc or a vertebra) at the location (i,j). $y(i,j)$ is the true label of the pixel at location (i,j).

w_m is a nonnegative weight inversely proportional to the area of the m-th object, and $\sum_m w_m = 1$. ϵ is a small constant (1e-4) to prevent the case of 0/0 in Eq.(3). In a lumbar spine image, where the background area is substantially larger than the combined area of the discs and vertebrae, employing area-weighted cross entropy loss effectively reduces the influence of the background in the loss function.

4. Experiments

4.1. Original Dataset and Augmented Datasets

Our dataset consists of a total of 100 patients' lumbar spine MR images from the University of Miami medical school, with personal identification information removed. Following the protocol in (Hu et al., 2018), five lumbar discs (D1, D2, D3, D4, D5) and six vertebral bones (L1, L2, L3, L4, L5, S1) in each patient's mid-sagittal MR image was manually annotated by three trained operators to identify and mark the boundaries and landmarks of the lumbar discs and vertebrae. To ensure accuracy and consistency, the three operators engaged in discussion to reach a consensus on the best annotation (i.e., ground-truth), for each mid-sagittal MR image. In the literature, lumbar disc D1 is also called L1/L2, similarly D2 for L2/L3, D3 for L3/L4, D4 for L4/L5, and D5 for L5/S1. The MR images are of various resolutions, and each of the images are resized to 512 × 512. Each image is also pre-processed independently by normalizing the intensities into range [0,1]. The average pixel spacing is 0.7004mm. The dataset of 100 patients was divided into 70 training samples, 10 validation samples, and 20 test samples, and those samples are referred to as the original training/validation/test samples in this paper.

Subsequently, the augmented dataset, named SSMSpine, is generated by using our method in Section 3.1. The SSMSpine dataset is divided into three sets: an augmented training set with 7000 samples, an augmented validation set with 250 samples, and an augmented test set with 2500 samples. To generate the augmented test set, an SSM was constructed using the 20 original test samples, and then 125 virtual shapes were generated from the SSM. Using each of the 20 original test samples as a reference image and each of the 125 virtual shapes, 20 × 125 (=2500) new MR images were generated using the method in Section 3.1. The augmented training and validation sets were generated in a similar way using another SSM built on the original training and validation samples.

4.2. Model evaluation and comparison

In our study, we trained and compared a total of 15 models on our lumbar spine MRI dataset. These models are Attention U-Net (Oktay et al., 2018), HSNet (Zhang et al., 2022), Inception-SwinUnet (Pu et al., 2023), MedT (Valanarasu et al., 2021), MultiResUNet (Ibtehaz and Rahman, 2020), SLT-Net (Feng et al., 2022), Swin-Unet (Cao et al., 2023), UNETR (Hatamizadeh et al., 2021), Swin UNETR (Hatamizadeh et al., 2022), TransUNet (Chen et al., 2021), UCTransNet (Wang et al., 2022), UNet++ (Zhou et al., 2018), UNeXt (Valanarasu and Patel, 2022), UTNet (Gao et al., 2021), and BianqueNet (Zheng et al., 2022). We were unable to test all models mentioned in Section 2 due to either unavailability (e.g., no source code) or incompatibility (e.g., size not matching). By training and evaluating these models, we aimed to compare their performance and determine the most effective approach for spine MRI instance segmentation on our dataset. To ensure compatibility with our dataset, we made minor adjustments to the original codes of some models if necessary.

We conducted two experiments: experiment-A and experiment-B. In experiment-A, each model is trained using the original training set with elastic deformation (Section 3.1). The top 4 models in experiment-A are selected for training using the augmented training set in experiment-B. In both experiments, we applied random translations to the input images within 16 pixels during training, which is a common data augmentation method. In both experiments, the augmented validation set is used for hyper-parameter tuning. The evaluation process contains instance segmentation evaluation.

In the instance segmentation evaluation, we assess model segmentation performance for individual lumbar spine instances/objects in the MR images. The instance segmentation task is formulated as a task of labeling 12 distinct objects (5 lumbar discs, 6 vertebrae, and a background). The input to each model is a single-channel mid-sagittal lumbar spine MR image with size of 512 × 512 pixels. In the segmentation output, each class is represented by a distinct channel as a binary segmentation map. We employ both the Dice Similarity Coefficient and the 95% Hausdorff Distance (HD95) as evaluation metrics. In both experiments, the original test set with 20 samples and the augmented test set with 2500 samples are used separately for model performance assessment on unseen data.

Each model was trained on a Nvidia A6000 GPU with 48GB VRAM. During the training process, a batch size of 6 was used for most of the models, except for training MedT, where a batch size of 2 was utilized due to its large model size. The Adam optimizer with an initial learning rate of 0.0001 was employed for model optimization. A low learning rate is generally preferred to ensure stable convergence during training. Although a low learning rate might slow down the convergence process, it helps avoid convergence failures. Gradient clipping is applied during training to prevent potentially large gradients from causing instability in the learning process. We performed model selection based on the performance on the validation set.

4.3. Results of Experiment-A with the original training set

In experiment-A, the 15 models were trained on the original training set with elastic deformation and random-shift, and then the models were evaluated on both the original test set and the augmented test set to measure instance segmentation accuracy. Figure 3 displays the performance of the top 4 models.

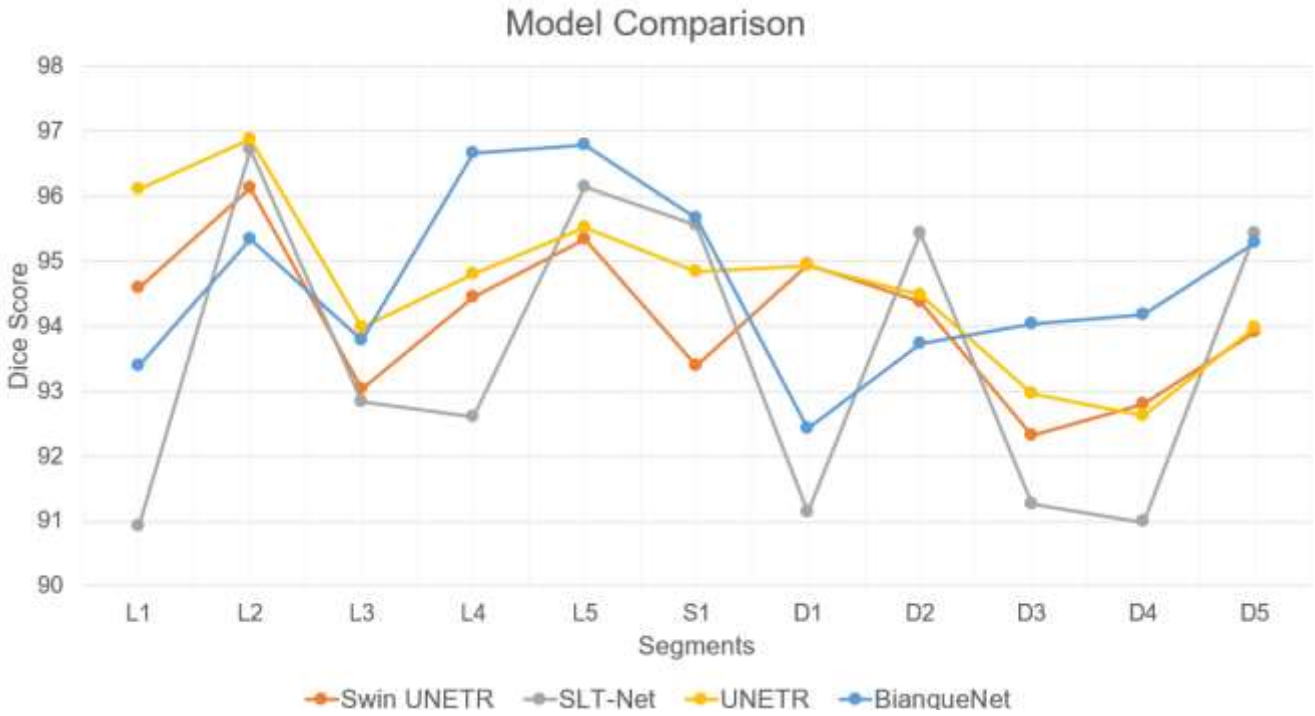

Fig. 3. Top 4 Model Comparison Results (Dice) on the augmented test set

4.3.1. Instance Segmentation Evaluation

Table 2 summarizes the instance segmentation results for vertebrae bodies (VB) and intervertebral discs (IVD) in terms of the Dice Similarity Coefficient on the original test set consisting of 20 samples. For better clarity and ease of understanding, we have converted Dice Similarity Coefficient into percentage ratios between 0 and 100%. The results show that Transformer-based models surpasses the CNN-only models.

Table 2. Dice (the higher, the better) of each model on the original test set

	L1	L2	L3	L4	L5	S1	D1	D2	D3	D4	D5	Average
Swin UNETR	93.126 ±7.356	96.813 ±1.641	93.874 ±11.389	**95.805** ±3.189	94.380 ±4.816	92.194 ±6.449	92.757 ±4.568	**94.299** ±3.519	92.821 ±6.563	89.777 ±7.645	91.077 ±10.179	**93.357** ±4.098
SLT-Net	91.967 ±6.991	97.052 ±0.902	**94.328** ±10.261	93.619 ±13.504	95.763 ±2.553	93.805 ±1.815	92.990 ±4.893	92.938 ±6.914	90.282 ±16.896	91.481 ±5.639	92.589 ±6.048	93.347 ±5.886
UNETR	**94.355** ±3.131	96.795 ±1.709	91.900 ±18.125	95.546 ±3.196	94.099 ±3.814	92.988 ±4.539	93.261 ±3.878	93.554 ±4.254	90.931 ±9.003	90.435 ±5.678	91.446 ±4.198	93.210 ±4.333
BianqueNet	89.330 ±21.689	93.332 ±16.860	91.819 ±18.667	95.245 ±6.709	**96.011** ±2.155	**94.193** ±5.037	90.069 ±21.017	92.566 ±9.930	**93.635** ±4.852	**93.539** ±3.373	**94.326** ±4.325	93.097 ±2.840
Inception-SwinUnet	91.397 ±19.415	**97.335** ±0.812	91.926 ±20.366	93.837 ±12.637	91.156 ±18.678	91.592 ±8.410	**93.533** ±4.614	93.281 ±7.055	89.739 ±20.753	87.727 ±14.249	91.233 ±11.216	92.069 ±9.514
HSNet	94.240 ±12.384	94.116 ±10.544	90.123 ±22.355	91.649 ±21.243	93.456 ±11.688	92.662 ±9.901	92.948 ±7.903	88.941 ±21.43	89.008 ±21.266	89.158 ±17.099	91.986 ±12.085	91.662 ±13.085
Swin-Unet	90.839 ±20.912	93.514 ±16.261	91.781 ±21.085	91.781 ±21.08	90.076 ±21.037	92.197 ±6.612	91.638 ±11.014	90.000 ±20.632	89.755 ±20.720	87.864 ±20.629	88.554 ±16.934	90.727 ±16.984
UNeXt	92.354 ±15.512	95.835 ±4.808	91.220 ±21.036	89.282 ±21.93	88.818 ±18.828	87.325 ±14.428	92.082 ±9.072	91.271 ±16.774	87.280 ±21.834	85.790 ±19.574	89.065 ±17.579	90.029 ±12.849
TransUNet	92.245 ±9.199	87.526 ±25.59	88.178 ±24.371	91.605 ±21.122	91.219 ±21.025	89.705 ±20.496	81.444 ±32.928	86.309 ±24.176	88.812 ±21.076	88.060 ±20.596	90.294 ±20.782	88.672 ±20.293
MedT	88.707 ±20.984	94.461 ±8.57	91.011 ±20.989	90.999 ±20.981	87.121 ±25.111	87.183 ±15.317	87.639 ±20.944	89.689 ±19.384	87.318 ±20.279	82.663 ±23.855	85.180 ±22.131	88.361 ±18.641
UTNet	81.322 ±33.18	82.886 ±31.442	85.159 ±29.06	88.802 ±24.539	90.776 ±20.988	87.080 ±14.823	77.511 ±32.691	78.711 ±32.650	83.856 ±28.478	85.608 ±21.674	89.312 ±18.307	84.639 ±23.716
UCTransNet	75.346 ±31.945	78.997 ±28.28	81.862 ±26.04	80.907 ±29.371	86.107 ±22.629	86.968 ±20.202	80.058 ±25.635	80.407 ±22.833	76.787 ±30.266	81.144 ±25.322	88.063 ±19.596	81.513 ±20.303
Attention U-Net	74.333 ±35.642	74.288 ±33.164	68.732 ±36.905	76.499 ±29.188	86.964 ±22.583	85.474 ±25.111	73.700 ±37.113	68.943 ±37.476	68.092 ±35.719	80.147 ±25.685	87.694 ±21.912	76.806 ±25.335
UNet++	68.734 ±40.68	66.998 ±35.694	74.033 ±27.911	71.468 ±40.229	81.297 ±31.709	83.980 ±27.527	65.450 ±39.937	71.473 ±28.932	67.208 ±40.117	79.834 ±28.708	84.529 ±26.387	74.091 ±26.461
MultiResUNet	77.421 ±31.596	74.309 ±29.915	66.317 ±34.739	66.828 ±35.945	76.870 ±35.698	84.644 ±22.624	75.710 ±33.508	70.025 ±27.873	58.563 ±39.241	72.553 ±33.342	75.217 ±31.98	72.587 ±26.547

Table 3 shows the instance segmentation results measured by 95% Hausdorff distance (HD95) on the original test set. The results show that Transformer-based models surpasses the CNN-only models.

Table 3. HD95 (the lower, the better) of each model on the original test set.

	L1	L2	L3	L4	L5	S1	D1	D2	D3	D4	D5	Average
Swin UNETR	8.594 ±12.047	11.024 ±33.718	2.992 ±3.483	3.147 ±6.521	3.759 ±5.905	9.454 ±13.059	19.057 ±41.678	1.894 ±2.472	3.466 ±5.801	3.814 ±3.828	8.128 ±12.917	6.848 ±10.137
SLT-Net	13.716 ±16.122	1.332 ±0.64	2.371 ±3.562	2.778 ±5.493	2.363 ±2.252	8.11 ±17.93	2.348 ±3.912	2.004 ±2.338	3.118 ±5.032	2.865 ±2.174	4.762 ±12.584	4.161 ±4.089
UNETR	3.744 ±6.201	1.512 ±1.261	3.679 ±7.65	2.237 ±1.722	5.927 ±6.302	**4.724** ±5.909	2.442 ±3.25	**1.704** ±1.3	3.589 ±4.982	3.277 ±2.968	7.517 ±8.093	3.668 ±2.34
BianqueNet	5.056 ±10.105	7.158 ±15.206	3.624 ±7.993	3.375 ±7.988	**2.290** ±2.285	5.308 ±9.319	4.278 ±9.886	5.106 ±10.580	3.138 ±7.382	5.857 ±12.764	**3.713** ±6.602	4.446 ±9.739
Inception-SwinUnet	3.378 ±7.275	**1.197** ±0.555	3.5 ±8.067	3.033 ±6.334	3.362 ±3.809	6.39 ±10.21	7.53 ±25.967	2.051 ±3.129	**1.59** ±0.876	3.957 ±4.714	5.715 ±9.957	3.791 ±4.134
HSNet	3.006 ±6.714	3.634 ±8.238	4.121 ±8.874	3.94 ±9.34	5.547 ±11.36	5.196 ±10.482	2.869 ±7.442	5.378 ±11.527	4.145 ±9.551	5.966 ±11.862	5.045 ±11.035	4.441 ±7.376
Swin-Unet	**2.131** ±3.459	2.226 ±4.686	1.753 ±1.591	**1.971** ±2.484	2.767 ±2.596	4.732 ±7.47	**2.221** ±3.907	2.937 ±6.313	1.679 ±1.159	**2.08** ±1.308	6.339 ±10.19	**2.803** ±2.24
UNeXt	2.805 ±4.959	2.016 ±2.768	**1.933** ±1.542	7.259 ±19.018	13.485 ±17.479	10.8 ±15.286	5.98 ±14.131	2.536 ±4.028	5.69 ±15.364	7.292 ±13.081	11.01 ±15.926	6.437 ±8.162
TransUNet	7.727 ±15.874	8.556 ±20.98	9.261 ±19.824	3.836 ±9.163	4.372 ±9.063	7.595 ±12.805	8.968 ±22.198	8.662 ±19.145	3.876 ±9.056	5.176 ±10.127	4.15 ±9.494	6.562 ±11.646
MedT	8.665 ±14.016	2.557 ±2.564	6.168 ±11.39	2.386 ±2.751	7.056 ±10.149	10.406 ±10.254	5.684 ±10.784	7.024 ±11.272	63.762 ±83.727	5.359 ±8.133	18.524 ±33.337	12.508 ±7.994
UTNet	6.897 ±12.365	7.209 ±12.726	10.986 ±16.485	5.092 ±10.427	8.523 ±13.468	20.203 ±35.303	7.421 ±13.484	13.133 ±18.251	11.486 ±17.583	10.463 ±17.173	8.903 ±13.608	10.029 ±12.359
UCTransNet	11.554 ±13.053	21.176 ±20.644	30.833 ±45.698	12.752 ±15.773	11.034 ±15.015	9.536 ±15.319	20.424 ±32.647	23.576 ±27.95	20.578 ±22.595	14.814 ±16.837	7.444 ±13.146	16.702 ±12.286
Attention U-Net	9.394 ±12.866	13.821 ±17.43	16.946 ±18.222	21.225 ±32.217	10.944 ±15.876	7.474 ±13.279	13.878 ±18.345	15.805 ±19.732	28.632 ±35.708	10.027 ±15.161	7.124 ±13.263	14.116 ±13.022
UNet++	6.733 ±11.95	16.872 ±17.754	22.11 ±20.232	15.901 ±19.52	12.515 ±16.298	9.518 ±14.037	9.862 ±14.455	21.599 ±20.421	19.863 ±26.463	10.369 ±16.036	9.929 ±15.288	14.116 ±11.826
MultiResUNet	8.841 ±11.614	17.292 ±16.037	19.035 ±17.013	15.814 ±15.802	9.93 ±13.24	9.582 ±13.03	9.63 ±14.817	18.331 ±18.619	22.223 ±19.32	10.918 ±15.737	10.632 ±14.034	13.839 ±11.524

We also assessed the segmentation performance of all models on the augmented test set consisting of 2500 samples. Table 4 (Dice) and Table 5 (HD95) summarize the instance segmentation performance of each model evaluated on the augmented test set.

Table 4. Dice (the higher, the better) of each model on the augmented test set

	L1	L2	L3	L4	L5	S1	D1	D2	D3	D4	D5	Average
Swin UNETR	93.108 ±8.399	96.337 ±3.289	93.286 ±14.544	95.270 ±5.054	94.695 ±4.905	93.072 ±5.52	93.127 ±5.447	**94.191** ±4.227	91.970 ±9.192	90.438 ±6.799	91.308 ±6.215	93.346 ±4.983
SLT-Net	91.925 ±11.161	96.736 ±2.401	93.889 ±12.742	92.538 ±18.37	94.557 ±8.863	92.437 ±8.045	92.712 ±7.199	93.833 ±4.888	89.858 ±18.676	90.419 ±11.831	91.149 ±11.399	92.732 ±9.07
UNETR	93.839 ±6.152	96.260 ±3.454	91.773 ±18.475	94.984 ±5.13	94.052 ±5.003	89.817 ±10.051	93.039 ±3.998	93.865 ±4.332	90.486 ±9.81	90.317 ±5.791	89.976 ±6.801	92.583 ±5.384
BianqueNet	92.103 ±16.133	95.329 ±10.126	**94.670** ±10.097	**95.486** ±7.397	**96.003** ±2.653	95.191 ±2.992	92.275 ±13.814	93.724 ±7.138	**93.752** ±6.614	**93.384** ±5.339	94.187 ±4.917	**94.191** ±8.996
Inception-SwinUnet	91.988 ±18.095	**96.977** ±3.847	91.923 ±20.632	92.382 ±18.076	94.237 ±8.303	92.904 ±9.058	**93.660** ±6.763	93.689 ±6.904	89.551 ±20.673	88.781 ±14.933	91.991 ±7.246	92.553 ±10.665
HSNet	**94.628** ±14.373	95.868 ±10.502	92.061 ±18.001	91.512 ±21.12	95.849 ±8.781	**95.682** ±3.865	93.117 ±13.282	93.874 ±10.222	89.820 ±19.811	89.645 ±17.916	**94.570** ±4.643	93.330 ±9.196
Swin-Unet	91.328 ±18.901	93.299 ±17.946	91.703 ±21.094	91.669 ±21.131	91.198 ±18.97	91.366 ±12.268	91.283 ±14.972	90.320 ±20.105	89.189 ±20.739	87.815 ±20.844	90.297 ±13.743	90.861 ±17.347
UNeXt	91.280 ±16.796	95.036 ±8.14	91.013 ±21.094	88.953 ±23.119	91.170 ±13.263	89.812 ±12.544	91.255 ±11.141	92.186 ±12.876	87.131 ±22.224	87.645 ±15.647	90.288 ±10.6	90.525 ±13.111
TransUNet	92.319 ±16.2	93.879 ±16.636	92.586 ±17.395	94.676 ±10.604	94.741 ±7.245	94.003 ±5.979	90.806 ±18.041	91.343 ±16.735	90.749 ±16.481	91.550 ±10.169	92.894 ±6.709	92.686 ±10.357
MedT	86.861 ±23.723	92.629 ±15.296	91.421 ±19.013	89.039 ±21.809	88.147 ±20.587	84.398 ±18.84	86.124 ±22.55	89.722 ±19.247	86.538 ±20.526	83.335 ±23.549	83.075 ±21.478	87.390 ±18.852
UTNet	82.709 ±28.31	80.782 ±30.95	84.073 ±26.989	85.725 ±27.893	91.096 ±16.215	91.306 ±9.014	76.822 ±30.988	80.621 ±28.224	83.916 ±27.028	85.401 ±20.584	90.734 ±11.839	84.835 ±19.788
UCTransNet	72.343 ±34.307	76.444 ±29.91	81.393 ±24.405	79.606 ±28.661	87.287 ±18.466	92.642 ±8.999	73.083 ±32.634	81.171 ±23.118	74.872 ±29.442	79.460 ±23.231	89.673 ±13.893	80.725 ±18.567
Attention U-Net	68.915 ±36.382	69.948 ±35.755	65.592 ±34.768	75.681 ±28.502	86.721 ±22.809	93.227 ±8.204	66.819 ±38.116	63.876 ±37.865	66.104 ±32.424	75.905 ±28.8	91.537 ±11.763	74.939 ±20.816
UNet++	65.096 ±40.792	69.498 ±36.367	69.993 ±31.743	66.925 ±37.244	81.392 ±28.029	90.460 ±13.191	64.565 ±40.33	72.047 ±31.75	63.061 ±38.234	72.931 ±31.167	87.159 ±20.266	73.012 ±23.317
MultiResUNet	73.766 ±33.559	72.475 ±32.877	70.752 ±32.574	70.171 ±35.186	78.962 ±32.165	87.809 ±18.254	68.730 ±36.759	69.854 ±34.213	68.738 ±34.452	72.479 ±33.345	79.657 ±28.326	73.945 ±25.091

Table 5. HD95 value (the lower, the better) of each model on the augmented test set

	L1	L2	L3	L4	L5	S1	D1	D2	D3	D4	D5	Average
Swin UNETR	11.353 ±33.189	9.468 ±31.478	7.005 ±23.16	5.783 ±17.828	5.375 ±11.093	16.07 ±35.564	13.829 ±40.509	3.708 ±15.61	6.191 ±18.392	4.775 ±11.302	11.15 ±17.265	8.609 ±15.102
SLT-Net	8.464 ±17.817	2.537 ±9.812	3.463 ±9.614	3.078 ±8.708	3.766 ±9.767	5.409 ±15.909	3.96 ±15.752	2.376 ±5.825	3.835 ±13.919	4.023 ±10.257	5.751 ±11.967	4.242 ±7.772
UNETR	5.873 ±10.615	2.63 ±6.155	4.111 ±7.566	2.506 ±3.004	5.014 ±7.516	10.871 ±19.237	2.772 ±4.803	**2.103** ±3.727	4.111 ±6.555	3.735 ±4.708	6.404 ±8.943	4.557 ±4.305
BianqueNet	5.555 ±10.365	5.767 ±13.823	3.398 ±8.156	3.173 ±6.849	3.318 ±7.518	3.540 ±11.313	5.303 ±11.274	5.818 ±15.604	2.231 ±4.499	2.868 ±6.468	3.384 ±7.982	4.032 ±10.045
Inception-SwinUnet	3.41 ±10.003	**1.529** ±5.754	3.685 ±10.956	3.125 ±7.22	4.276 ±10.702	5.534 ±13.91	**1.677** ±3.326	2.424 ±7.185	3.671 ±9.82	3.923 ±6.581	6.093 ±11.203	3.577 ±5.158
HSNet	**2.575** ±6.116	2.72 ±9.091	3.826 ±9.386	2.955 ±7.523	**1.958** ±5.101	**2.37** ±4.584	2.165 ±6.245	2.778 ±7.9	3.13 ±8.164	2.894 ±6.34	**2.258** ±6.682	**2.693** ±4.533
Swin-Unet	3.718 ±12.477	2.281 ±9.182	**2.092** ±5.183	**2.044** ±5.193	4.841 ±11.639	5.539 ±9.335	2.601 ±10.854	2.392 ±9.554	**2.103** ±4.736	**2.735** ±5.601	6.01 ±10.775	3.305 ±4.993
UNeXt	7.212 ±18.727	4.233 ±12.846	4.913 ±16.54	7.77 ±17.801	11.646 ±17.788	9.462 ±21.051	8.457 ±20.29	3.848 ±10.381	7.327 ±19.721	7.806 ±12.668	11.651 ±16.133	7.666 ±9.932
TransUNet	6.245 ±18.622	6.008 ±18.599	6.191 ±17.579	5.246 ±18.663	6.385 ±20.503	7.016 ±19.462	5.417 ±18.977	8.688 ±23.985	7.87 ±21.105	5.241 ±14.506	9.589 ±24.032	6.718 ±13.318
MedT	8.115 ±12.876	3.878 ±6.953	4.385 ±7.671	5.283 ±9.288	10.009 ±13.817	13.935 ±16.408	6.139 ±11.146	4.428 ±9.506	72.075 ±82.33	8.635 ±16.218	12.997 ±17.864	13.625 ±9.645
UTNet	8.667 ±17.117	10.592 ±20.449	15.484 ±20.453	9.202 ±17.838	14.237 ±27.825	14.724 ±22.491	10.674 ±16.985	15.102 ±21.452	11.572 ±19.287	10.649 ±18.77	8.581 ±16.592	11.771± 11.935
UCTransNet	14.267 ±20.828	21.247 ±25.68	23.542 ±31.268	20.769 ±31.321	12.504 ±16.412	9.226 ±22.252	21.576 ±32.676	27.888 ±31.415	28.971 ±34.852	17.921 ±25.073	10.608 ±19.215	18.956 ±13.438
Attention U-Net	21.033 ±29.85	18.984 ±20.704	22.082 ±22.065	31.271 ±36.969	11.244 ±20.511	6.783 ±16.671	18.986 ±25.984	19 ±21.259	32.27 ±31.194	19.275 ±26.051	11.469 ±28.816	19.309 ±14.67
UNet++	18.011 ±27.449	16.192 ±19.325	23.953 ±20.546	18.749 ±21.889	12.423 ±16.826	6.939 ±12.271	13.378 ±20.277	20.232 ±21.096	25.36 ±30.082	17.679 ±20.881	8.728 ±15.29	16.513 ±12.019
MultiResUNet	13.291 ±19.988	16.623 ±17.923	20.534 ±20.18	16.424 ±18.717	8.357 ±12.915	7.503 ±11.828	13.491 ±20.581	20.504 ±23.767	19.608 ±22.555	11.27 ±16.494	8.239 ±13.843	14.168 ±11.371

Figure 4 shows segmentation examples of the 15 models. All of the models produce misclassifications or fragmentation errors. For example, BianqueNet exhibits hollow holes in both D3 and D4. TransUnet incorrectly identifies D3 as D2 and exhibits a significant segmentation error in the lower left corner in the MR scan. Also seen in Figure 4, many models produce incorrect predictions around pixels in close proximity to the boundary of two adjacent objects.

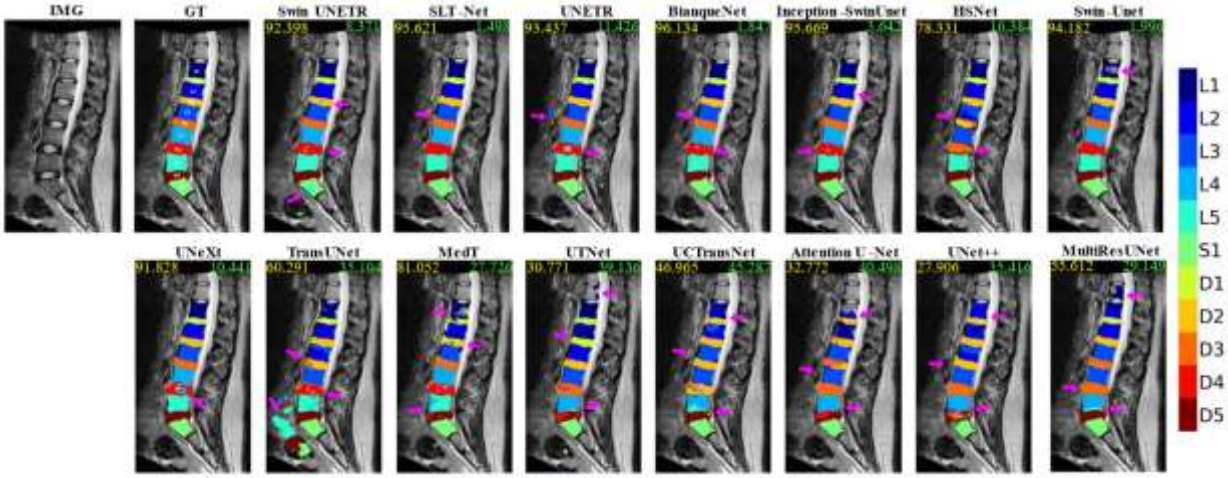

Fig. 4. Segmentation examples of the 15 models. IMG is the input image. GT indicates the Ground-Truth annotation. Dice (green color) and HD95 (yellow color) are shown on top of each image. The 11 lumbar objects are shown in different colors. Segmenttion errors are indicated by pink arrows.

4.4. Results of Experiment-B with the augmented training set

In this section, we show the advantages of our data augmentation method in Section 3.1. The top four models in Table 2 are Swin UNETR, SLT-Net, UNETR, BianqueNet. In the experiment-B, each of the four models was trained from scratch using the augmented training set consisting of 7000 samples. Model evaluations were conducted on the augmented test set (see Table 8 and Table 9). The results show that the segmentation performance of most models is improved, suggesting that the augmented data can enhance the generalization and overall performance of a model.

It is noteworthy that training Transformer-based models with limited data is challenging. Nevertheless, the new data augmentation method effectively tackles the challenge of data scarcity with the support of SSM and biomechanics. Our data augmentation can generate synthetic images that closely resemble real data, effectively enhancing the model training process. This approach assists in augmenting the available data, enabling Transformer models to learn from a more diverse and representative dataset. The augmented/synthesized datasets can be made publicly available without any concerns related to medical data privacy.

By using the instance segmentation evaluation, a comprehensive evaluation of each model's performance is achieved. This ensures that models excel not only instance segmentation ability but also maintain their accuracy under realistic and varied conditions.

4.4.1. Instance Segmentation Evaluation

Table 8 (Dice) and Table 9 (HD95) present the instance segmentation evaluation results on the augmented test set. It is evident that the segmentation performances of all models have improved by using the augmented dataset, as indicated by enhancements in both evaluation metrics, except for Swin UNETR on the HD95 metric.

The augmented dataset offers a broader domain, enabling the models to acquire more diverse knowledge and achieve improved segmentation performance. This increased resilience to data variations is a notable benefit of the augmented training dataset. Increasing the complexity of the training dataset helps prevent overfitting and regulating models from memorizing the training.

Table 8. Dice (the higher, the better) of each model on the augmented test set. The "change" is the Dice difference between training on the augmented training set and training on the original training set

	L1	L2	L3	L4	L5	S1	D1	D2	D3	D4	D5	Average	Change
Swin UNETR	94.593 ±8.113	96.124 ±7.352	93.027 ±17.824	94.442 ±12.723	95.345 ±6.840	93.384 ±6.406	94.949 ±3.461	94.369 ±7.674	92.323 ±13.660	92.806 ±5.711	93.900 ±4.372	94.115 ±9.581	**+0.769**
SLT-Net	90.928 ±21.398	96.708 ±5.097	92.836 ±18.462	92.604 ±19.937	96.143 ±4.279	95.545 ±2.669	91.135 ±18.062	95.435 ±3.448	91.254 ±17.130	90.982 ±15.676	95.427 ±2.972	93.545 ±14.102	**+0.813**
UNETR	96.099 ±2.574	96.878 ±2.479	93.981 ±11.711	94.798 ±8.922	95.509 ±3.888	94.832 ±2.951	94.930 ±2.159	94.486 ±5.341	92.957 ±8.036	92.616 ±4.935	93.974 ±3.926	94.642 ±6.081	**+2.059**
BianqueNet	93.381 ±16.166	95.339 ±12.916	93.786 ±13.760	96.670 ±4.308	96.789 ±2.243	95.666 ±2.302	92.427 ±16.211	93.723 ±11.633	94.040 ±7.633	94.184 ±4.318	95.276 ±3.268	94.662 ±10.231	**+0.471**

Table 9. HD95 (the lower, the better) of each model on the augmented test set. The "change" is the HD95 difference between training on the augmented training set and training on the original training set

	L1	L2	L3	L4	L5	S1	D1	D2	D3	D4	D5	Average	Change
Swin UNETR	17.591 ±52.428	13.865 ±44.911	12.207 ±42.917	11.290 ±38.870	10.393 ±35.903	19.144 ±52.802	9.570 ±35.018	12.249 ±43.617	7.145 ±30.825	5.524 ±21.039	11.808 ±30.508	11.889 ±40.236	**+3.280**
SLT-Net	3.720 ±8.774	1.577 ±3.310	3.227 ±7.838	1.899 ±5.205	2.012 ±2.264	1.871 ±2.843	2.344 ±6.312	**1.387** ±2.514	3.100 ±8.190	2.584 ±4.752	1.602 ±2.714	2.302 ±5.549	**-1.940**
UNETR	1.923 ±2.598	1.774 ±3.164	2.679 ±4.416	3.029 ±7.859	3.056 ±5.502	3.340 ±5.462	1.329 ±1.5817	1.788 ±2.9369	2.348 ±3.860	2.640 ±3.108	2.455 ±3.079	2.396 ±4.341	**-2.161**
BianqueNet	3.815 ±9.985	2.907 ±8.092	3.849 ±10.217	2.605 ±9.389	2.268 ±7.335	2.667 ±10.589	3.383 ±9.089	4.257 ±11.508	3.149 ±10.065	2.061 ±5.126	2.860 ±13.087	3.075 ±9.734	**-0.957**

5. Conclusion

In this paper, we present an evaluation of 15 DNN models for instance segmentation of lumbar spine MR images, using the original dataset of 100 patients and the generated SSMSpine dataset of thousands of virtual patients. We developed the SSM-biomechanics based data augmentation method to further improve model performance by providing large and diverse datasets of synthetic images with ground-truth. Given that our augmented datasets consist entirely of synthetic data, we have made our augmented dataset, SSMSpine, publicly available. The results presented indicate that models trained on the augmented training set had comparably or even better performance than the same models trained on the original training set. This underscores that our data augmentation method can generate synthetic data that eliminates privacy concerns while retaining in the same image domain.

Our current study mainly focused on the mid-sagittal lumbar spine MR images for two major reasons. First, as shown in a clinical study (Hu et al., 2018), the mid-sagittal image of a patient provides the most useful information for the diagnosis of lumbar spine degeneration. Second, the slice thickness of a lumbar MR scan in the sagittal direction is often much larger than 5mm, which causes difficulties to create accurate 3D ground-truth annotation for model training. Nevertheless, the models could be directly extended to handle 3D images once the slice thickness becomes acceptably small with the advancement of imaging technology.

The 15 DNN models, including the Transformer-based models, often generate segmentation artifacts, such as: (1) extra areas not belonging to any discs or vertebrae, (2) assigning the same class label to two different discs, and (3) broken area of a disc or vertebrae. Thus, new models are needed for artifact-free geometry reconstruction of lumbar spine from MR images.

References

Ambellan, F., Lamecker, H., von Tycowicz, C., Zachow, S., 2019. Statistical Shape Models: Understanding and Mastering Variation in Anatomy, in: Rea, P.M. (Ed.), Biomedical Visualisation : Volume 3, Advances in Experimental Medicine and Biology. Springer International Publishing, Cham, pp. 67–84. https://doi.org/10.1007/978-3-030-19385-0_5

Berhane, H., Scott, M., Elbaz, M., Jarvis, K., McCarthy, P., Carr, J., Malaisrie, C., Avery, R., Barker, A.J., Robinson, J.D., Rigsby, C.K., Markl, M., 2020. Fully automated 3D aortic segmentation of 4D flow MRI for hemodynamic analysis using deep learning. Magn. Reson. Med. 84, 2204–2218. https://doi.org/10.1002/mrm.28257

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2023. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation, in: Karlinsky, L., Michaeli, T., Nishino, K. (Eds.), Computer Vision – ECCV 2022 Workshops, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 205–218. https://doi.org/10.1007/978-3-031-25066-8_9

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. https://doi.org/10.48550/arXiv.2102.04306

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Presented at the Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.

Chevrefils, C., Chériet, F., Grimard, G., Aubin, C.-E., 2007. Watershed Segmentation of Intervertebral Disk and Spinal Canal from MRI Images, in: Kamel, M., Campilho, A. (Eds.), Image Analysis and Recognition, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 1017–1027. https://doi.org/10.1007/978-3-540-74260-9_90

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active Shape Models-Their Training and Application. Comput. Vis. Image Underst. 61, 38–59. https://doi.org/10.1006/cviu.1995.1004

Cox, M., Serra, R., Shapiro, I., Risbud, M., 2014. The Intervertebral Disc: Molecular and Structural Studies of the Disc in Health and Disease.

Davies, R.H., Twining, C.J., Daniel Allen, P., Cootes, T.F., Taylor, C.J., 2003. Building optimal 2D statistical shape models. Image Vis. Comput., British Machine Vision Computing 2001 21, 1171–1182. https://doi.org/10.1016/j.imavis.2003.09.003

Dong, X., Bao, J., Chen, Dongdong, Zhang, W., Yu, N., Yuan, L., Chen, Dong, Guo, B., 2022. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. https://doi.org/10.48550/arXiv.2107.00652

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

Dufter, P., Schmitt, M., Schütze, H., 2022. Position Information in Transformers: An Overview. Comput. Linguist. 48, 733–763. https://doi.org/10.1162/coli_a_00445

Dwivedi, K.Kr., Lakhani, P., Kumar, S., Kumar, N., 2022. A hyperelastic model to capture the mechanical behaviour and histological aspects of the soft tissues. J. Mech. Behav. Biomed. Mater. 126, 105013. https://doi.org/10.1016/j.jmbbm.2021.105013

Feng, K., Ren, L., Wang, G., Wang, H., Li, Y., 2022. SLT-Net: A codec network for skin lesion segmentation. Comput. Biol. Med. 148, 105942. https://doi.org/10.1016/j.compbiomed.2022.105942

Galbusera, F., Casaroli, G., Bassani, T., 2019. Artificial intelligence and machine learning in spine research. JOR SPINE 2, e1044. https://doi.org/10.1002/jsp2.1044

Gao, Y., Zhou, M., Metaxas, D.N., 2021. UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Lecture Notes in

Computer Science. Springer International Publishing, Cham, pp. 61–71. https://doi.org/10.1007/978-3-030-87199-4_6
Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2022. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images, in: Crimi, A., Bakas, S. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 272–284. https://doi.org/10.1007/978-3-031-08999-2_22
Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D., 2021. UNETR: Transformers for 3D Medical Image Segmentation. https://doi.org/10.48550/arXiv.2103.10504
Hu, X., Chen, M., Pan, J., Liang, L., Wang, Y., 2018. Is it appropriate to measure age-related lumbar disc degeneration on the mid-sagittal MR image? A quantitative image study. Eur. Spine J. 27, 1073–1081. https://doi.org/10.1007/s00586-017-5357-3
Hu, Y., Zhang, L., Mu, N., Liu, L., 2022. Parameter-Efficient Transformer with Hybrid Axial-Attention for Medical Image Segmentation. https://doi.org/10.48550/arXiv.2211.09533
Huang, J., Jian, F., Wu, H., Li, H., 2013. An improved level set method for vertebra CT image segmentation. Biomed. Eng. OnLine 12, 48. https://doi.org/10.1186/1475-925X-12-48
Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Chi, H., Hu, X., Fan, D.-P., Dong, F., Ni, D., 2023. Segment Anything Model for Medical Images? https://doi.org/10.48550/arXiv.2304.14660
Huang, Y.L., Chen, D.R., 2004. Watershed segmentation for breast tumor in 2-D sonography. Ultrasound Med. Biol. 30, 625–632. https://doi.org/10.1016/j.ultrasmedbio.2003.12.001
Hufnagel, H., Pennec, X., Ehrhardt, J., Handels, H., Ayache, N., 2007. Shape Analysis Using a Point-Based Statistical Shape Model Built on Correspondence Probabilities, in: Ayache, N., Ourselin, S., Maeder, A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 959–967. https://doi.org/10.1007/978-3-540-75757-3_116
Ibtehaz, N., Rahman, M.S., 2020. MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. 121, 74–87. https://doi.org/10.1016/j.neunet.2019.08.025
Kirnaz, S., Capadona, C., Wong, T., Goldberg, J.L., Medary, B., Sommer, F., McGrath, L.B., Härtl, R., 2022. Fundamentals of Intervertebral Disc Degeneration. World Neurosurg. 157, 264–273. https://doi.org/10.1016/j.wneu.2021.09.066
Kos, N., Gradisnik, L., Velnar, T., 2019. A Brief Review of the Degenerative Intervertebral Disc Disease. Med. Arch. 73, 421–424. https://doi.org/10.5455/medarh.2019.73.421-424
Kuang, X., Cheung, J.P., Wu, H., Dokos, S., Zhang, T., 2020. MRI-SegFlow: a novel unsupervised deep learning pipeline enabling accurate vertebral segmentation of MRI images, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Presented at the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1633–1636. https://doi.org/10.1109/EMBC44109.2020.9175987
Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. Presented at the Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, p. 12. https://doi.org/10.7303/syn3193805
Li, N., Liu, M., Li, Y., 2007. Image Segmentation Algorithm using Watershed Transform and Level Set Method, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07. Presented at the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07, p. I-613-I–616. https://doi.org/10.1109/ICASSP.2007.365982
Liang, L., Liu, M., Elefteriades, J., Sun, W., 2023. PyTorch-FEA: Autograd-enabled finite element analysis methods with applications for biomechanical analysis of human aorta. Comput. Methods Programs Biomed. 238, 107616. https://doi.org/10.1016/j.cmpb.2023.107616
Liang, L., Liu, M., Martin, C., Elefteriades, J.A., Sun, W., 2017. A machine learning approach to investigate the relationship between shape features and numerically predicted risk of ascending aortic aneurysm. Biomech. Model. Mechanobiol. 16, 1519–1533. https://doi.org/10.1007/s10237-017-0903-9

Lin, T., Wang, Y., Liu, X., Qiu, X., 2022. A survey of transformers. AI Open 3, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. https://doi.org/10.48550/arXiv.2103.14030

Mallio, C.A., Vadalà, G., Russo, F., Bernetti, C., Ambrosio, L., Zobel, B.B., Quattrocchi, C.C., Papalia, R., Denaro, V., 2022. Novel Magnetic Resonance Imaging Tools for the Diagnosis of Degenerative Disc Disease: A Narrative Review. Diagnostics 12, 420. https://doi.org/10.3390/diagnostics12020420

Moccia, S., De Momi, E., El Hadji, S., Mattos, L.S., 2018. Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics. Comput. Methods Programs Biomed. 158, 71–91. https://doi.org/10.1016/j.cmpb.2018.02.001

Noothout, J.M.H., Vos, B.D. de, Wolterink, J.M., Išgum, I., 2018. Automatic segmentation of thoracic aorta segments in low-dose chest CT, in: Medical Imaging 2018: Image Processing. Presented at the Medical Imaging 2018: Image Processing, SPIE, pp. 446–451. https://doi.org/10.1117/12.2293114

Ogden, R.W., Hill, R., 1997. Large deformation isotropic elasticity – on the correlation of theory and experiment for incompressible rubberlike solids. Proc. R. Soc. Lond. Math. Phys. Sci. 326, 565–584. https://doi.org/10.1098/rspa.1972.0026

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas. https://doi.org/10.48550/arXiv.1804.03999

Pang, S., Pang, C., Su, Z., Lin, L., Zhao, L., Chen, Y., Zhou, Y., Lu, H., Feng, Q., 2022. DGMSNet: Spine segmentation for MR image by a detection-guided mixed-supervised segmentation network. Med. Image Anal. 75, 102261. https://doi.org/10.1016/j.media.2021.102261

Pepe, A., Li, J., Rolf-Pissarczyk, M., Gsaxner, C., Chen, X., Holzapfel, G.A., Egger, J., 2020. Detection, segmentation, simulation and visualization of aortic dissections: A review. Med. Image Anal. 65, 101773. https://doi.org/10.1016/j.media.2020.101773

Pu, Y., Zhang, Q., Qian, C., Zeng, Q., Li, N., Zhang, L., Zhou, S., Zhao, G., 2023. Semi-supervised segmentation of coronary DSA using mixed networks and multi-strategies. Comput. Biol. Med. 156, 106493. https://doi.org/10.1016/j.compbiomed.2022.106493

Roberts, S., Gardner, C., Jiang, Z., Abedi, A., Buser, Z., Wang, J.C., 2021. Analysis of trends in lumbar disc degeneration using kinematic MRI. Clin. Imaging 79, 136–141. https://doi.org/10.1016/j.clinimag.2021.04.028

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Sarkalkan, N., Weinans, H., Zadpoor, A.A., 2014. Statistical shape and appearance models of bones. Bone 60, 129–140. https://doi.org/10.1016/j.bone.2013.12.006

Sekuboyina, A., Kukačka, J., Kirschke, J.S., Menze, B.H., Valentinitsch, A., 2018. Attention-Driven Deep Learning for Pathological Spine Segmentation, in: Glocker, B., Yao, J., Vrtovec, T., Frangi, A., Zheng, G. (Eds.), Computational Methods and Clinical Applications in Musculoskeletal Imaging, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 108–119. https://doi.org/10.1007/978-3-319-74113-0_10

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2023. Transformers in medical imaging: A survey. Med. Image Anal. 88, 102802. https://doi.org/10.1016/j.media.2023.102802

Shaw, P., Uszkoreit, J., Vaswani, A., 2018. Self-Attention with Relative Position Representations. https://doi.org/10.48550/arXiv.1803.02155

Shen, D., Wu, G., Suk, H.-I., 2017. Deep Learning in Medical Image Analysis. Annu. Rev. Biomed. Eng. 19, 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Si, C., Yu, W., Zhou, P., Zhou, Y., Wang, X., Yan, S., 2022. Inception Transformer. https://doi.org/10.48550/arXiv.2205.12956

Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R.J., Rajpoot, N.M., 2017. Gland segmentation in colon histology images: The glas challenge contest. Med. Image Anal. 35, 489–502. https://doi.org/10.1016/j.media.2016.08.008

Soomro, T.A., Afifi, A.J., Zheng, L., Soomro, S., Gao, J., Hellwich, O., Paul, M., 2019. Deep Learning Models for Retinal Blood Vessels Segmentation: A Review. IEEE Access 7, 71696–71717. https://doi.org/10.1109/ACCESS.2019.2920616

Suganyadevi, S., Seethalakshmi, V., Balasamy, K., 2022. A review on deep learning in medical image analysis. Int. J. Multimed. Inf. Retr. 11, 19–38. https://doi.org/10.1007/s13735-021-00218-1

Tamagawa, S., Sakai, D., Nojiri, H., Sato, M., Ishijima, M., Watanabe, M., 2022. Imaging Evaluation of Intervertebral Disc Degeneration and Painful Discs—Advances and Challenges in Quantitative MRI. Diagnostics 12, 707. https://doi.org/10.3390/diagnostics12030707

Tao, R., Liu, W., Zheng, G., 2022. Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers. Med. Image Anal. 75, 102258. https://doi.org/10.1016/j.media.2021.102258

Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M., 2021. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 36–46. https://doi.org/10.1007/978-3-030-87193-2_4

Valanarasu, J.M.J., Patel, V.M., 2022. UNeXt: MLP-Based Rapid Medical Image Segmentation Network, in: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 23–33. https://doi.org/10.1007/978-3-031-16443-9_3

van Veldhuizen, W.A., Schuurmann, R.C.L., IJpma, F.F.A., Kropman, R.H.J., Antoniou, G.A., Wolterink, J.M., de Vries, J.-P.P.M., 2022. A Statistical Shape Model of the Morphological Variation of the Infrarenal Abdominal Aortic Aneurysm Neck. J. Clin. Med. 11, 1687. https://doi.org/10.3390/jcm11061687

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Wang, B., Qin, J., Lv, L., Cheng, M., Li, L., Xia, D., Wang, S., 2023. MLKCA-Unet: Multiscale large-kernel convolution and attention in Unet for spine MRI segmentation. Optik 272, 170277. https://doi.org/10.1016/j.ijleo.2022.170277

Wang, H., Cao, P., Wang, J., Zaiane, O.R., 2022. UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer. Proc. AAAI Conf. Artif. Intell. 36, 2441–2449. https://doi.org/10.1609/aaai.v36i3.20144

Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C., 2020. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 108–126. https://doi.org/10.1007/978-3-030-58548-8_7

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. PVT v2: Improved Baselines with Pyramid Vision Transformer. Comput. Vis. Media 8, 415–424. https://doi.org/10.1007/s41095-022-0274-8

Wiputra, H., Matsumoto, S., Wagenseil, J.E., Braverman, A.C., Voeller, R.K., Barocas, V.H., 2023. Statistical shape representation of the thoracic aorta: accounting for major branches of the aortic arch. Comput. Methods Biomech. Biomed. Engin. 26, 1557–1571. https://doi.org/10.1080/10255842.2022.2128672

Wu, K., Peng, H., Chen, M., Fu, J., Chao, H., 2021. Rethinking and Improving Relative Position Encoding for Vision Transformer. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10033–10041.

You, X., Gu, Y., Liu, Y., Lu, S., Tang, X., Yang, J., 2022. EG-Trans3DUNet: A Single-Staged Transformer-Based Model for Accurate Vertebrae Segmentation from Spinal Ct Images, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). Presented at the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. https://doi.org/10.1109/ISBI52829.2022.9761551

Zhang, L., Yang, J., Liu, D., Zhang, F., Nie, S., Tan, Y., Guo, T., 2022. Spine X-ray Image Segmentation Based on Transformer and Adaptive Optimized Postprocessing, in: 2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI). Presented at the 2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI), pp. 88–92. https://doi.org/10.1109/SEAI55746.2022.9832144

Zhang, W., Fu, C., Zheng, Y., Zhang, F., Zhao, Y., Sham, C.-W., 2022. HSNet: A hybrid semantic network for polyp segmentation. Comput. Biol. Med. 150, 106173. https://doi.org/10.1016/j.compbiomed.2022.106173

Zhang, Z., Zhang, W., 2022. Pyramid Medical Transformer for Medical Image Segmentation. https://doi.org/10.48550/arXiv.2104.14702

Zheng, H.-D., Sun, Y.-L., Kong, D.-W., Yin, M.-C., Chen, J., Lin, Y.-P., Ma, X.-F., Wang, H.-S., Yuan, G.-J., Yao, M., Cui, X.-J., Tian, Y.-Z., Wang, Y.-J., 2022. Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from MRI. Nat. Commun. 13, 841. https://doi.org/10.1038/s41467-022-28387-5

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation, in: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 3–11. https://doi.org/10.1007/978-3-030-00889-5_1