Piaolin Hu, Jupeng Li*, Ruohan Ma, Kai Zhang, Yong Guo and Gang Li

Multi-stage image pyramid network for coarse-to-fine segmentation of temporomandibular joint in CBCT images

https://doi.org/10.1515/bmt-2024-0112
Received March 14, 2024; accepted June 27, 2025;
published online July 21, 2025

Abstract

Objectives: Accurate segmentation of the temporomandibular joint (TMJ) in cone beam computed tomography (CBCT) images is crucial for the diagnosis of temporomandibular joint osteoarthritis (TMJ-OA), particularly of the condyle. The aim of this study was to use a multi-stage image pyramid network for precise segmentation of the TMJ in CBCT images.
Methods: A multi-stage image pyramid network (MIP-Net) was proposed in this study for accurate TMJ segmentation in CBCT images. This network utilizes stepwise down-sampling to form an image pyramid, and feature fusion is employed to integrate features of different scales, resulting in accurate segmentation of the TMJ in CBCT images.
Results: The network was trained and evaluated using a clinical image dataset, and the dice similarity coefficient (DSC) was 98.02 ± 0.65 %, the average surface distance (ASD) was 0.0605 ± 0.0229 mm, and the Hausdorff distance (HD) was 0.1907 ± 0.0627 mm.
Conclusions: The proposed MIP-Net achieved accurate segmentation of the TMJ in CBCT images, which may help to improve the clinical diagnosis and treatment of TMJ-OA.

Keywords: medical image segmentation; TMJ-OA; muti-stage image pyramid; feature fusion; temporomandibular joint

Introduction

Temporomandibular joint osteoarthritis (TMJ-OA) is a common clinical manifestation of temporomandibular joint disorders with a high prevalence. Once diagnosed, it will seriously affect patients' daily functions such as chewing and speaking, which in turn significantly reduces the quality of life [1, 2]. In clinical practice, the diagnosis of TMJ-OA is usually based on a comprehensive evaluation, including history taking, clinical signs (e.g., joint murmur, limitation of opening, joint tenderness, etc.), assessment of dysfunction, and imaging. Among them, cone beam computed tomography (CBCT), as a high-resolution, low-radiation-dose, three-dimensional imaging technique, is widely used to evaluate structural changes in the TMJ, especially in the early stages of bone remodeling and articular surface degeneration.

Accurate segmentation of mandibular structures in CBCT images is crucial for the diagnosis and treatment of TMJ-OA. Accurate structural segmentation not only helps to identify lesion areas, such as bone destruction of the condyle, narrowing of the articular space, and articular fossa remodeling, but also assists physicians in quantitative analysis, such as calculating condylar volume, curvature, symmetry, and other key indexes, which provides a more objective and reproducible basis for diagnosis [3].

In addition, accurate mandibular segmentation can also assist physicians in formulating individualized treatment plans, such as the need for physical therapy, occlusal adjustment, pharmacological intervention or even surgical treatment. Meanwhile, it also plays an important role in evaluating the efficacy of the treatment process, which can be done by comparing the segmentation results with multiple imaging to observe whether the joint structure is improved, thus optimizing the treatment strategy, improving the efficacy and improving the patient's prognosis.

In recent years, researchers have proposed various deep learning-based image segmentation network models for bone structure segmentation in the temporomandibular joint region. Shelhamer et al. introduced fully convolutional networks (FCN), which utilizes convolutional encoders to extract features and convolutional decoders to restore the feature map size [4]. This network replaces the traditional fully connected layers with convolutional layers, enhancing its capability to process larger images and serving as a fundamental advancement in the field of semantic segmentation research. Ronneberger et al. proposed U-Net medical image segmentation network, which has outstanding performance and has received widespread attention [5]. Çiçek et al. extended the

***Corresponding author: Jupeng Li**, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China, E-mail: lijupeng@bjtu.edu.cn. https://orcid.org/0000-0002-2414-7440
Piaolin Hu, **Kai Zhang and Yong Guo**, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China
Ruohan Ma and Gang Li, Department of Oral and Maxillofacial Radiology, Peking University School and Hospital of Stomatology, Beijing, China

Open Access. © 2025 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

U-Net network with three-dimension convolution to design 3D U-Net, which has a simple coder-decoder structure and widely applied in medical image segmentation tasks [6]. Zhang et al. used a U-Net to achieve multi-task segmentation of the craniomaxillofacial region and landmark detection, resulting in better image processing results than single-task segmentation [3]. This method uses a sliding window approach to extract smaller image blocks and then completes the overall segmentation through multiple image segmentations, which greatly limits the receptive field of network. Egger et al. proposed one CT image segmentation method for the mandible based on the FCN network, and the performance of image segmentation networks is constrained by the absence of inter-slice information of the volume CT images [7]. Torosdagli et al. proposed a CBCT image mandible segmentation method based on the U-Net network [8], which expands the range of segmentation objects by embedding a Dense-Block [9] in the U-Net network and further refines the segmentation results. Qiu et al. used three 2D sectional images (axial, coronal, and sagittal) as inputs to the U-Net network to obtain more semantic information [10]. However, this method is essentially a 2D slice image segmentation and cannot utilize inter-slice context information. To address this issue, Qiu et al. proposed a recurrent convolutional mandible segmentation network that employs RNNs to guide the model's attention towards the inter-slice information between 2D images [11]. Belikova et al. proposed a mandible structure segmentation algorithm based on the background area around the mandibular condyle [12]. He et al. developed a statistical deformation model to enhance medical image segmentation, achieving state-of-the-art results by introducing realistic deformations in CT images [13]. Liang et al. (2021) enhanced abdominal CT segmentation by developing a Hybrid Deformable Model (HDM) for data augmentation, which significantly improved segmentation accuracy through a 3-D attention U-Net, achieving an average Dice Similarity Coefficient of 0.852 [14]. Liang et al. (2021) proposed an innovative deep unsupervised learning framework for automated prostate contour propagation from pCT to CBCT, demonstrating high accuracy with an average DICE Similarity Coefficient up to 0.85 [15].

In 2020, Milletari et al. proposed a negative volume TMJ segmentation method based on the domain around the mandibular condyle with a network based on V-Net [16]. The method can be divided into two phases; the first phase localizes the TMJ and segments the approximate region where the TMJ is located to obtain the target region; the second phase performs a fine segmentation of the small-sized region obtained. In 2021, Kim et al. proposed a model for cortical bone segmentation of the head of the mandibular condyle in CBCT images [17], and an automated method was proposed to measure the cortical thickness. The study used a large number of CBCT images as a gold standard, combined improved U-Net and convolutional neural networks (CNNs) for segmentation model construction, and evaluated the model performance. The results showed that the method achieved high accuracy in cortical bone segmentation and cortical thickness measurement. In 2022, Wu et al. proposed a coarse-to-fine multiscale structured C2Ftrans segmentation model for the most important subtypes of articular abnormalities in patients with temporomandibular joint disorders, i.e., the different forms of disk displacement and deformation, describing the medical image segmentation as a coarse-to-fine process capable of performing accurate target boundary segmentation with a much lower computational complexity for accurate target boundary segmentation [18]. In 2023, Kumar et al. proposed a three-stage deep learning model that achieves accurate segmentation of mandibular condyles in X-ray scanned images in three stages by combining the 3D U-Net and ResNet50 models, and uses a crawling search algorithm for hyperparameter optimization [19]. In 2024, Liu et al. proposed a three-branch joint feature and topology decoder for TMJ disk and condylar segmentation, which effectively preserves the topological information of the disk structure and enhances the features. The mechanism captures the spatial, channel, and cross-dimensional information of the decoded features, thus improving the segmentation performance [20].

In addition to segmentation of individual targets, there are also papers that propose simultaneous segmentation of multiple targets within a region, such as simultaneous segmentation of both parts of the condylar head and the articular fossa. In their paper, Liu Yin et al. first focused on the simultaneous segmentation of the temporomandibular joint fossa and condyle, and developed an automatic segmentation algorithm, which uses a U-Net network for the initial segmentation, and then post-processes the boundaries of the entire CT image based on the generalized gradient vector flow snake model and tracking idea. This study shows that the new algorithm has high accuracy and segmentation efficiency [21].

Inter-slice information, which is the contextual data between neighboring slices in 3D volumetric CBCT images, is essential for accurately capturing the continuity of complex anatomical structures such as the temporomandibular joint. Although 3D convolution is theoretically able to utilize this information, its high computational complexity and large number of parameters limit its wide application in high-resolution medical images. Meanwhile, many traditional methods use a global downsampling strategy, which often loses critical inter-slice contextual information in the process of reducing spatial dimensionality, resulting in difficulty in preserving fine structures and affecting segmentation accuracy.

To overcome these problems, we propose a multi-stage image pyramid network (MIP-Net) based on image pyramids for accurate segmentation of TMJ from coarse to fine in CBCT images. The method constructs an image pyramid by stepwise downsampling, extracts multi-scale feature information, and designs a residual feature fusion module to efficiently transfer contextual information among different resolutions. In particular, we achieve more adequate inter-slice feature representation and effectively improve the segmentation accuracy while keeping the overall GPU of the network largely unchanged.

Materials and methods

Data collection

Data information

A total of 88 patients' CBCT image data from Peking University Stomatological Hospital were collected and anonymized for research purposes. The CBCT scanner used was a NewTom VG, with the following scanning parameters: field of view of 16 × 13 cm, voltage of 120 kV, current of 5 mA, and voxel resolution of 0.125 mm. All image data were automatically de-identified by the hospital's data system in order to protect patient privacy.

The data used in this study were collected retrospectively, and all included patients underwent CBCT examinations due to mandibular-related conditions. The inclusion criteria were that the mandibular structures, including the condyle, were completely and clearly visible in the CBCT images. The exclusion criteria included severe image artifacts, incomplete scanning range, or the presence of significant maxillofacial malformations or structural abnormalities. The data collection process was approved by the Ethics Committee of Peking University Stomatological Hospital and strictly adhered to the ethical standards of medical research.

To determine the required sample size of 88 CBCT cases, we referred to representative studies [22, 23] on deep learning-based segmentation of mandibular lesions, in which the average sensitivity and specificity were reported around 0.923–0.975. Based on this, we used a simplified sample size estimation formula commonly applied in classification model evaluation:

$$N = \frac{Z^2 \cdot P(1-P)}{d^2} \tag{1}$$

Where Z=1.96 for a 95 % confidence level, P is the expected sensitivity, and d=0.05 is the allowable margin of error. Substituting these values, we obtained an estimated sample size of approximately 38–109. To ensure sufficient data diversity and robustness in model training, we ultimately selected 88 cases. This estimation is also consistent with dataset sizes reported in comparable CBCT segmentation studies. A slightly larger sample size was chosen to enhance the reliability and generalizability of the results.

Data annotation

The annotation work is completed by a core team of annotation experts from the Imaging Department of Peking University Stomatology Hospital, which consists of a total of five doctors: three of them are radiologists with more than three years of experience, one is an imaging expert with more than 10 years of clinical experience, and the other is a senior imaging expert with rich experience in deep learning annotation projects, who also serves as the project annotation administrator.

To ensure consistency of standards among the team, a calibration session was conducted before the formal annotation: all annotation specialists jointly annotated, compared and discussed representative CBCT data, and clarified the annotation boundary standards and specifications. The calibration session significantly improved the consistency of the annotation. The specific labeling process was as follows: first, two experienced radiologists collaborated to manually label the CBCT images layer by layer using ITK-SNAP software, mainly outlining the axial images and completing the region labeling with the software's auto-filling function. In order to maintain the structural continuity between slices, the doctors simultaneously referred to the sagittal and coronal images for optimization. Afterwards, the third doctor strictly reviewed and corrected all the preliminary data to ensure the accuracy and objectivity of the results. The fourth senior expert then randomly selected some samples for secondary validation to monitor and improve the overall annotation quality, and finally, the project quality director organized a review and audit to confirm the final Ground Truth data.

After the annotation was completed, inter-operator and intra-operator consistency were evaluated, and Dice coefficients and consistency indexes were calculated by repeating the annotation on a random sample of 20 images. The results show that the average inter-operator Dice coefficient is 0.94 and the average intra-operator Dice coefficient is 0.96, indicating that the labeling results have high consistency and reliability.

Data preprocessing

Uniformization of image resolution

Given that CBCT images are collected from various devices, it is essential to standardize all images to an isotropic resolution format of $0.3 \times 0.3 \times 0.3$ mm^3 before feeding the jawbone ROI data into the target segmentation network. This standardization is crucial to strike a balance between computational efficiency and segmentation accuracy.

Voxel intensity value normalization

The differences in voxel intensity value distribution among CBCT images from different devices are a critical factor affecting the repeatability and generalizability of disease intelligent diagnostic algorithms. To achieve normalization of voxel intensity value distribution, our algorithm design encompasses two aspects:

(a) Intra-image intensity value normalization: We employ linear functions such as Min-Max Normalization and Z-Score Standardization to normalize voxel intensities within CBCT images to the range [0, 1]. Furthermore, to mitigate the impact of extreme values within CBCT images (especially in regions affected by metal artifacts), we clip the intensity values of each CBCT scan to the interval [0, 2,500] before performing voxel intensity value normalization.

(b) Inter-image intensity value normalization within the dataset: We select a single CBCT image from the dataset as a reference template and employ the Histogram Matching regularization algorithm. This process aligns the histograms of all CBCT images in the dataset with the predefined template image, ensuring that they share the same voxel intensity value distribution. This further enhances the training accuracy of subsequent network models for target segmentation and lesion detection in CBCT images.

Multi-stage image pyramid network

The image pyramid is a common method for multi-scale representation of images, providing an effective means of representing images at various resolutions, and finding widespread application in tasks such as image segmentation [24, 25]. In this paper, the image pyramid consists of three levels: low-resolution, medium-resolution, and high-resolution. The image size decreases from high to low, with image pyramids of $481 \times 481 \times 481$, $240 \times 240 \times 240$ and $120 \times 120 \times 120$ containing image data of different scales. The high-resolution image contains rich image details but has the largest data volume and is only suitable for networks with fewer parameters. The low-resolution image contains the global information of the image, and the amount of data is small, making it suitable for designing networks with more parameters.

Figure 1 shows the overall network architecture of MIP-Net. Our approach is inspired by the concept of image pyramids and is combined with the U-Net model to enhance feature extraction at multiple scales. Each U-Net network is specifically designed for different image resolutions, enabling effective learning of image features at various scales. (a) The first stage processes the low-resolution image from the image pyramid. This low-resolution image provides global information and context, allowing the network to capture broad, high-level features of the TMJ. (b) The second stage uses the medium-resolution image from the pyramid, which contains more detailed information compared to the low-resolution image. The coarse segmentation results from the first stage are fused with the medium-resolution image to refine the segmentation performance. This stage benefits from both global context and enhanced detail. (c) The third stage operates on the high-resolution image, which is the original CBCT image with rich details. Here, the segmentation results from both the first and second stages are integrated with the high-resolution image. This fusion assists the network in achieving precise segmentation of the TMJ, leveraging detailed features for accurate results.

By employing this multi-stage approach, MIP-Net effectively captures and integrates features at different scales, leading to improved segmentation accuracy. This method distinguishes itself from existing models like 3D U-Net and V-Net by its explicit stage-wise processing and feature fusion, providing a novel contribution to the field.

Considering the resolution of the TMJ CBCT images ($481 \times 481 \times 481$) and the experimental test results, we set L=2. The top layer of the network is a low-resolution layer with an image size of $120 \times 120 \times 120$, associated with the network Net-2 as shown in Table 1. Net-2 comprises 10 units, including an initial convolution unit (F_0), four down-sampling units (F_1~F_4), four up-sampling units (F_5~F_8), and a convolutional output unit (F_9). The initial convolution unit (F_0) contains two $3 \times 3 \times 3$ convolutional layers designed to extract underlying image features, with each layer followed by a ReLU activation function to introduce non-linearity. The down-sampling unit (F_1~F_4) includes three convolutional layers: the first is a $3 \times 3 \times 3$ convolutional subsampling layer with a stride of 2, which reduces the image size to one-fourth of the original dimensions; the subsequent two layers are standard convolutional layers used to capture more abstract features from the down-sampled image. The up-sampling unit (F_5~F_8) also consists of three convolutional layers, with the first being a

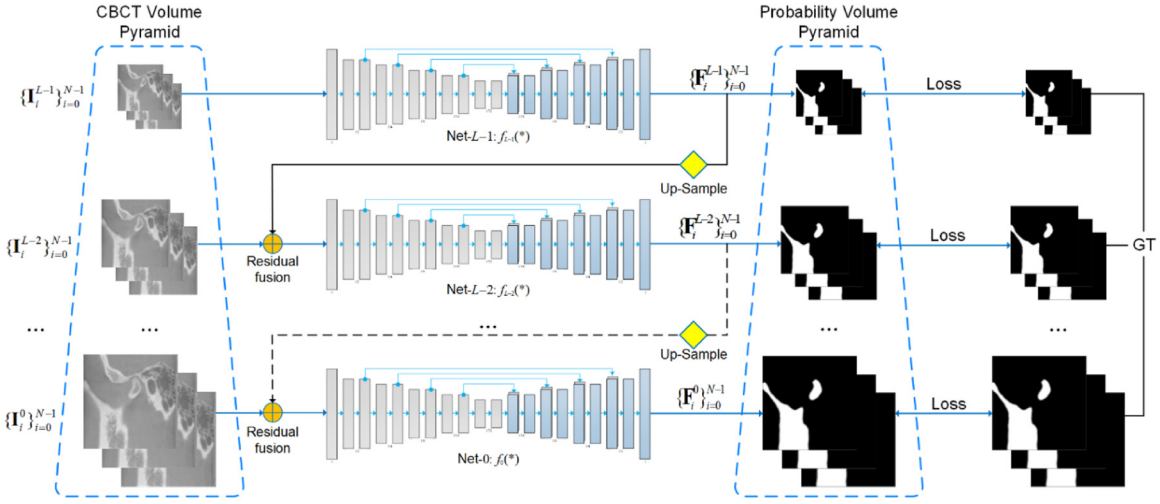


Figure 1: Multi-stage image pyramid network for coarse-to-fine segmentation of TMJ in CBCT image. We highlight that all segmentation was done in 3D throughout this paper. For clarity and simplicity, we depict the 2D formulation of our method in this figure.

$3 \times 3 \times 3$ deconvolutional layer (with a stride of 2) that doubles the feature map size, followed by two $3 \times 3 \times 3$ convolutional layers for further feature extraction and refinement. The final convolutional output unit (F_9) utilizes a $1 \times 1 \times 1$ convolutional layer to adjust the number of channels in the feature map to one, providing the final segmentation output.

The middle layer of the image pyramid, with an image size of $240 \times 240 \times 240$, is handled by the network model Net-1.

Table 1: Network structure of layers for TMJ CBCT images segmentation in our MIP-Net, considering each stage input image/feature size.

Network layer	Kernel size/(input output) channels (Net-2)	Kernel size/(input output) channels (Net-1)	Kernel size/(input output) channels (Net-0)
F_0	$3 \times 3 \times 3$, (1, 16)	$3 \times 3 \times 3$, (2, 12)	$3 \times 3 \times 3$, (2, 3, s=2)
	$3 \times 3 \times 3$, (16, 24)		$3 \times 3 \times 3$, (4, 6)
F_1	$3 \times 3 \times 3$, (24, 24, s=2)	$3 \times 3 \times 3$, (12, 12, s=2)	MaxPool3D, (6, 6)
	$3 \times 3 \times 3$, (24, 24)	$3 \times 3 \times 3$, (13, 24)	$3 \times 3 \times 3$, (7, 16)
	$3 \times 3 \times 3$, (24, 64)		$3 \times 3 \times 3$, (16, 16)
F_2	$3 \times 3 \times 3$, (64, 64, s=2)	$3 \times 3 \times 3$, (24, 24, s=2)	MaxPool3D, (16, 16)
	$3 \times 3 \times 3$, (64, 64)	$3 \times 3 \times 3$, (24, 24)	$3 \times 3 \times 3$, (16, 64)
	$3 \times 3 \times 3$, (64, 128)	$3 \times 3 \times 3$, (24, 48)	$3 \times 3 \times 3$, (64, 64)
F_3	$3 \times 3 \times 3$, (128, 128, s=2)	$3 \times 3 \times 3$, (48, 48, s=2)	MaxPool3D, (64, 64)
	$3 \times 3 \times 3$, (128, 128)	$3 \times 3 \times 3$, (48, 48)	$3 \times 3 \times 3$, (64, 128)
	$3 \times 3 \times 3$, (128, 256)	$3 \times 3 \times 3$, (48, 128)	$3 \times 3 \times 3$, (128, 128)
F_4	$3 \times 3 \times 3$, (256, 256, s=2)	$3 \times 3 \times 3$, (128, 128, s=2)	MaxPool3D, (128, 128)
	$3 \times 3 \times 3$, (256, 256)	$3 \times 3 \times 3$, (128, 128)	$3 \times 3 \times 3$, (128, 128)
	$3 \times 3 \times 3$, (256, 512)	$3 \times 3 \times 3$, (128, 256)	$3 \times 3 \times 3$, (128, 192)
F_5	$3 \times 3 \times 3$, (512, 512, s=2)	$3 \times 3 \times 3$, (256, 256, s=2)	Trilinear, (192, 192)
	$3 \times 3 \times 3$, (768, 768)	$3 \times 3 \times 3$, (384, 384)	$3 \times 3 \times 3$, (320, 128)
	$3 \times 3 \times 3$, (768, 256)	$3 \times 3 \times 3$, (384, 128)	$3 \times 3 \times 3$, (128, 128)
F_6	$3 \times 3 \times 3$, (256, 256, s=2)	$3 \times 3 \times 3$, (128, 128, s=2)	Trilinear, (128, 128)
	$3 \times 3 \times 3$, (384, 384)	$3 \times 3 \times 3$, (176, 176)	$3 \times 3 \times 3$, (192, 64)
	$3 \times 3 \times 3$, (384, 128)	$3 \times 3 \times 3$, (176, 48)	$3 \times 3 \times 3$, (64, 64)
F_7	$3 \times 3 \times 3$, (128, 128, s=2)	$3 \times 3 \times 3$, (48, 48, s=2)	Trilinear, (64, 64)
	$3 \times 3 \times 3$, (192, 192)	$3 \times 3 \times 3$, (72, 72)	$3 \times 3 \times 3$, (80, 16)
	$3 \times 3 \times 3$, (192, 64)	$3 \times 3 \times 3$, (72, 24)	$3 \times 3 \times 3$, (16, 16)
F_8	$3 \times 3 \times 3$, (64, 64, *s*=2)	$3 \times 3 \times 3$, (24, 24, s=2)	Trilinear, (16, 16)
	$3 \times 3 \times 3$, (88, 88)	$3 \times 3 \times 3$, (36, 12)	$3 \times 3 \times 3$, (22, 8)
	$3 \times 3 \times 3$, (88, 24)		
F_9	$1 \times 1 \times 1$, (24, 1)	$1 \times 1 \times 1$, (12, 1)	$3 \times 3 \times 3$, (8, 1, s=2)

This network accepts two channels as input: the current training image and the coarse prediction result from the initial stage. Net-1's architecture, detailed in the third column of Table 1, includes one initial convolution unit, four down-sampling units, four up-sampling units, and one convolutional output unit, making a total of 10 units. Unlike the low-resolution network, Net-1 features a reduced number of convolutional layers and channels due to GPU memory constraints. Specifically, the initial convolution unit (F_0) has been simplified from two convolutional layers to a single convolutional layer, with the feature map channels reduced from 48 to 12. The down-sampling units (F_1~F_4) and up-sampling units (F_5~F_8) are also streamlined to manage memory usage efficiently. Notably, the first down-sampling unit (F_1) includes a feature fusion step where the feature map is combined with the coarse prediction results from the first stage. The final convolutional output unit (F_9) employs a $1 \times 1 \times 1$ convolutional layer to adjust the number of output channels and ensure the output feature map is properly normalized.

The lowest layer of the image pyramid, denoted as the high-resolution layer, has an image size of $481 \times 481 \times 481$ and is associated with the network model Net-0. The network takes a three-channel input, consisting of the training image for this stage and the coarse segmentation results from the previous two stages. The detailed structure of Net-0 is presented in the last column of Table 1. To minimize GPU memory usage, the first layer of the network is set as a convolutional down-sampling layer. Down-sampling units (F_1~F_4) utilize 3D max-pooling, while up-sampling units (F_5~F_8) use trilinear interpolation instead of deconvolution. The network performs a total of 5 down-samplings, and deconvolution (F_9) is used in the convolutional output unit to increase the feature map size and normalize the number of feature map channels to maintain consistent input and output sizes.

Feature fusion

Incorporating features from different scales is a commonly used method to enhance the performance of network models, and several fusion techniques have been proposed [26–31]. To combine the coarse prediction results from the low-resolution layer with the intermediate feature map of the high-resolution layer network, we employed a residual fusion approach, shown in Figure 2. The feature map F_i^{l-1} is the result of coarse segmentation, which can be regarded as a soft label that is not binarized. Similar to spatial attention in the attention mechanism, convolution and concatenation of I_i^l with the feature map $\mathrm{Up}\left(F_i^{l-1}\right)$ can make the network pay more attention to the TMJ bone structure region. The final fused feature map:

$$Fusion\ \mathbf{F}_i^l = \mathrm{LN}\left(f\left(\mathbf{I}_i^l, \mathrm{Up}\left(\mathbf{F}_i^{l-1}\right)\right) + \mathbf{I}_i^l\right) \tag{2}$$

where $f(\cdot, \cdot)$ is $3 \times 3 \times 3$ convolution and concatenation of two input features, LN(·) represents layer normalization, and Up(·) means up-sampling processing.

Loss functions

Here, DICE loss function will be added on the basis of BCE loss function. DICE loss function first appeared in the V-Net model and is widely used in medical image segmentation because of its direct optimization of DICE coefficient evaluation index. In this paper, these two loss functions will be used to supervise the training of each stage of the network.

The specific loss function is calculated as follows: given the paired input image and the ground truth, $\left\{X^j, Y^j\right\}_{j=1,2,\cdots k} \in \mathbb{R}^{w \times h \times d}$ the width, length and depth of the image are w, h, and d, respectively, where k is the number of images in the training data. Expand the input image and the ground truth into one dimension, $X^j = \left[x_i^j\right]_{i=1,2,\cdots,m}$ and $Y^j = \left[y_i^j\right]_{i=1,2,\cdots,m}$, m=$w \times h \times d$. Set the target of the bone structure to be segmented as l_f and the background as l_b, and first calculate the BCE loss function:

$$\mathrm{L_{BCE}} = \sum_{j=1}^{N}\left(Y^j \log f\left(X^j\right) + \left(1 - Y^j\right)\log\left(1 - f\left(X^j\right)\right)\right) \tag{3}$$

DICE loss function:

$$\mathrm{L_{DICE}} = 1 - \frac{2\left|f\left(X^j\right) \cap Y^j\right|}{\left|f\left(X^j\right) \cup Y^j\right|} \tag{4}$$

The overall loss function defines as:

$$\mathrm{L} = \mathrm{L_{BCE}} + \mathrm{L_{DICE}} \tag{5}$$

where $f(\cdot)$ is the network model.

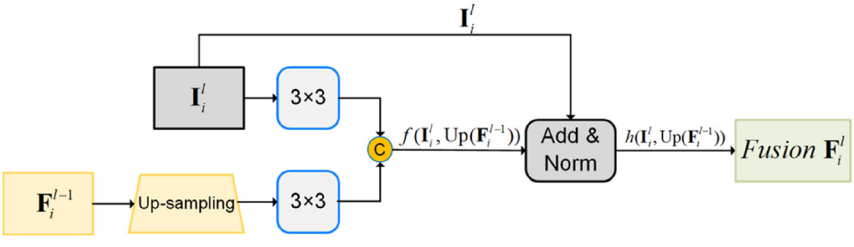


Figure 2: Residual fusion module for different scale features.

Network training

The CBCT image dataset used in this study consisted of 88 clinically acquired images, with 72 images used for training and 16 for testing. The network in each stage of the model is trained sequentially from Net-2 to Net-0, and the loss function is optimized by SGD algorithm, and the momentum value is set to 0.9. The learning rate setting adopts a segmented decay strategy, which sets the initial learning rate to 0.1, gradually decreases to 0.001 as the network iterative training times and sets the number of training cycles to 300.

Experiments and results

To assess the segmentation performance of the MIP-Net in the TMJ in CBCT images, we conducted a comparison of its performance with that of the V-Net [16] network and the U-Net [32] in the testing dataset. To facilitate a comparison of experimental results, the same dataset was used for training and testing all models. In the U-Net, the size of CBCT images varied from $481 \times 481 \times 481$ to $134 \times 134 \times 134$ after global down-sampling. V-Net, on the other hand, sets a convolutional down-sampling layer on the initial convolutional layer and can accommodate maximum input images of $240 \times 240 \times 240$ under the same GPU video memory resources.

Figure 3 shows the segmentation results of the axial layer at the condyle position. It can be observed from the comparison that the segmentation performance of the MIP-Net is significantly superior to that of other models. The U-Net model (Figure 3c) showed rough segmentation results at the condyle edge, with a small area of segmentation error appearing in the condyle position of image #81, and an obvious segmentation error appearing in the condyle edge position of image #92. The input image size of the V-Net network (Figure 3d) is relatively increased, and a smooth edge contour segmentation result can be obtained. However, it is important to note that the accuracy of edge segmentation is compromised, with image #92 displaying segmentation errors. By comparing the results of the above two network segmentation approaches, the MIP-Net proposed in this paper (Figure 3e) exhibits no error segmentation phenomenon of the above method, and is superior to other network segmentation results in the segmentation accuracy of the condylar joint edge and the integrity of internal segmentation.

A quantitative comparison of the segmentation performance of different segmentation networks is shown in Table 2. The DSC, ASD, and HD performance evaluation indicators of the segmentation results of the TMJ in 16 CBCT images in the test set were calculated. Compared with the U-Net segmentation algorithm, our proposed MIP-Net achieved higher segmentation accuracy, with a 0.53 % increase

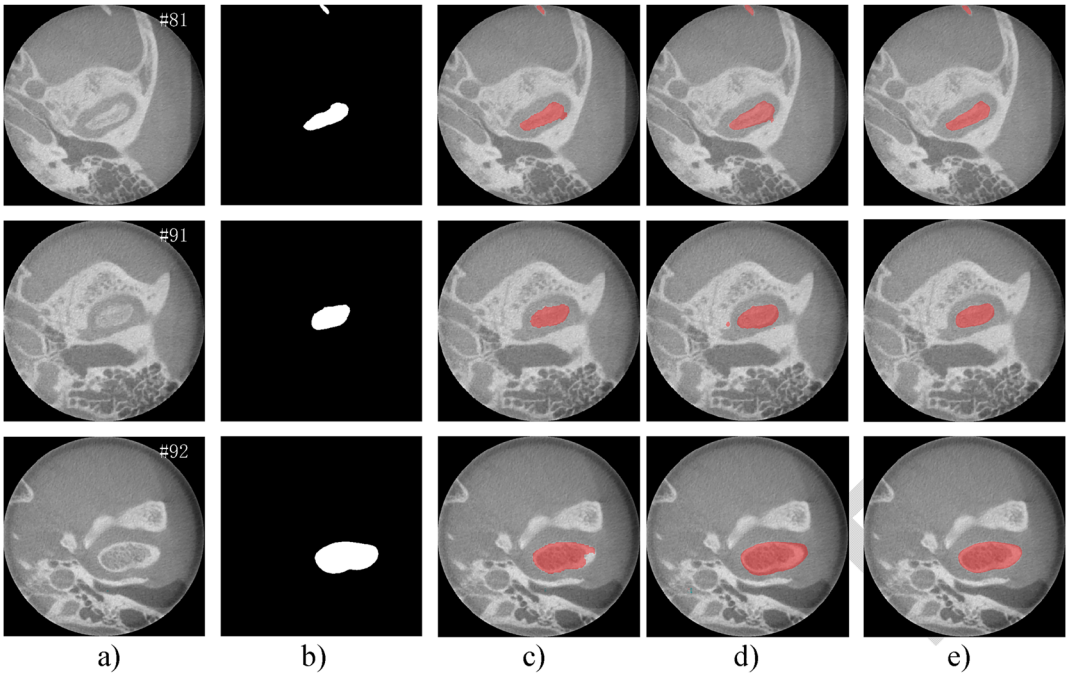


Figure 3: Performance comparison between MIP-Net and multiple different models on TMJ CBCT image segmentation tasks. a) Cross-sectional tomographic images of original CBCT images; b) ground truth of corresponding slices; c) U-Net results; d) V-Net results; e) our results.

Table 2: Performance comparison of different networks on the testing dataset (mean ± std).

Methods	DSC, %	ASD, mm	HD, mm	Time, s	GPU memory, MB
U-Net	97.49 ± 0.49	0.0787 ± 0.0165	0.2283 ± 0.0413	**23.3**	**9,549**
V-Net	96.43 ± 1.26	0.1264 ± 0.0445	0.3499 ± 0.0947	25.5	11,562
MIP-Net (ours)	**98.02 ± 0.65**	**0.0605 ± 0.0229**	**0.1907 ± 0.0627**	48.2	10,365

The bold value indicate the best (i.e., most favorable) performance among the compared methods for each metric.

in DSC, a decrease in ASD to 0.0605 mm, and a decrease in HD to 0.1907 mm.

In addition, we measured the computational cost in terms of inference time and GPU memory usage during training. Inference time was recorded as the average time required to segment a single CBCT image. GPU memory usage was monitored throughout training to assess the resource requirements of the network. A comparison of the computational cost between our proposed MIP-Net and the baseline model is shown in Table 2. The average inference time of the MIP-Net is 48 s per image, and the GPU memory usage during the training is 10,365 MB.

The results suggest that our MIP-Net, while achieving higher segmentation accuracy, incurs higher computational costs, especially in terms of inference time and GPU memory usage. The trade-off between accuracy and computational efficiency is a common challenge for deep learning models for medical image analysis. In clinical practice, a slight increase in computational time may be acceptable given the significant improvement in segmentation accuracy.

To evaluate the effectiveness of the residual fusion approach, we conducted an ablation study comparing it with several common feature fusion techniques: direct channel merging, convolutional channel merging, and point-to-point addition. (a) Direct channel merging, frequently used in U-Net models, concatenates deep and shallow features along the channel dimension. (b) Convolutional channel merging enhances this by first aligning the channel dimensions using convolutions before concatenation, followed by a $1 \times 1 \times 1$ convolution for more effective fusion. (c) Point-to-point addition, as employed in FCN models, sums corresponding elements after matching channel dimensions through $3 \times 3 \times 3$ convolutions.

We applied each of these methods within our network architecture and compared their performance with the residual fusion approach. The results of these comparisons are summarized in Table 3, highlighting the strengths and weaknesses of each fusion technique in terms of segmentation accuracy, computational efficiency, and overall model performance.

Table 3 shows the comparison results of the segmentation performance of different feature fusion methods. By comparing the dice similarity coefficient (DSC), average surface distance (ASD), and Hausdorff distance (HD), it can be seen that Residual Fusion is the most superior in overall performance. Although its DSC (98.02 %) is slightly lower than that of direct channel splicing (98.04 %), it achieves the best results in both spatial accuracy metrics of ASD (0.0605 mm) and HD (0.1907 mm), which indicates that it is more accurate in boundary reconstruction, and it can effectively retain shape information. The point-to-point addition method also performs better in the distance index, indicating that it has certain advantages in boundary processing. In contrast, the convolutional fusion method performs relatively poorly in all the metrics, especially reaching the highest on HD, suggesting that it may have information loss in processing detailed features. In addition, although the direct splicing method performs slightly better on DSC, its HD is on the high side, which may lead to the existence of bias in the prediction boundary. In summary, the residual fusion method achieves the best balance between boundary localization accuracy and overall performance while maintaining high segmentation accuracy, which verifies its effectiveness and superiority as a feature fusion strategy.

Building on the original network design, we conducted an ablation experiment to assess whether fusing low-resolution features not only to Net-2 but also further to Net-1 can improve segmentation performance. The modification aims to evaluate the impact of incorporating additional global spatial information on the model's performance.

Table 3: Performance of segmentation results corresponding to different feature approaches.

Methods	DSC, %	ASD, mm	HD, mm
Direct channel merging	**98.04 ± 0.56**	0.0664 ± 0.0300	0.2026 ± 0.0718
Convolutional channel merging	97.94 ± 0.65	0.0659 ± 0.0270	0.2497 ± 0.2121
Point-to-point addition	97.96 ± 0.47	0.0624 ± 0.0228	0.1907 ± 0.0627
Residual fusion	98.02 ± 0.43	**0.0605 ± 0.0229**	**0.1907 ± 0.0627**

The bold value indicate the best (i.e., most favorable) performance among the compared methods for each metric.

Table 4: Performance comparison between the original multilevel image pyramid network and the modified architecture.

Network structure	DSC, %	ASD, mm	HD, mm
MIP-Net	98.02 ± 0.65	**0.0605** ± **0.0229**	**0.1907** ± **0.0627**
Net-1+(Net-3)	**98.19** ± **0.58**	0.0638 ± 0.0183	0.2085 ± 0.0508

The bold value indicate the best (i.e., most favorable) performance among the compared methods for each metric.

As shown in Table 4, the improved structure (Net-1+(Net-3)) improves the DSC metric by 0.17 % from 98.02 % in the original network to 98.19 %, but the HD metric increases slightly (from 0.1907 to 0.2085 mm), which suggests that although the segmentation accuracy has been slightly improved, the boundary accuracy has not been improved, or even slightly decreased. This result indicates that the boundary accuracy has not been improved and even decreased slightly despite the slight increase in segmentation accuracy. This result indicates that the fusion of low-resolution features at an earlier stage (e.g., Net-2) can provide sufficient global spatial context information for the model, and further fusion at a shallower level (e.g., Net-1) will bring limited improvement, and may even affect the boundary discrimination ability due to redundant information or over-smoothing of features. Meanwhile, the improvement also brings a small increase in model complexity, but the performance improvement is not obvious, indicating that the original structure has achieved a better balance between accuracy and efficiency. Taken together, further fusion strategies have limited gains in this task, supporting the rationality of our original design.

Discussion

Automatic medical image segmentation is a crucial topic in the medical domain and successively a critical counterpart in the computer-aided diagnosis paradigm. Early traditional approaches to medical image segmentation mainly focused on edge detection, template matching techniques, region growing, graph cuts, active contour lines, machine learning, and other mathematical methods. Since 2012, CNNs have successfully implemented feature representation extraction for medical image processing, including the segmentation task, and their superior performance and accuracy make them the main choice in this field. U-Net is the most widespread image segmentation architecture due to its flexibility, optimized modular design, and success in all medical image modalities. Over the years, the U-Net model has received tremendous attention from academic and industrial researchers. Several extensions of this network have been proposed to address the scale and complexity created by medical tasks [33].

Our work addresses this fundamental limitation through the novel MIP-Net, which uniquely integrates three innovations. (1) Down-sampling the image volume to a lower resolution. Zhang et al. [32]proposed a 3D U-Net model for TMJ-CBCT image segmentation. The down-sampled CBCT images are used as the network input, and the segmentation accuracy is higher than the traditional segmentation methods under the limited training dataset. While this approach improves segmentation accuracy compared to traditional methods, it sacrifices image detail due to the inherent loss of resolution. Furthermore, the shallow depth of such models often hinders the effective fitting of input data, preventing high-precision 3D segmentation results. In addition, the shallow depth of the network makes it difficult to fit the input data effectively, which is unable to obtain high-precision 3D image segmentation results. (2) To overcome the influence of the loss of detail information caused by image down-sampling, Oda et al.24 performed 2D image segmentation on three sectional planes of CT image respectively in the segmentation of the abdominal artery CT image and combined them to generate 3D image segmentation results [34]. The multi-plane U-Net model proposed by Perslev et al. combined cross-sectional segmentation information obtained by multiple 2D U-Net networks, which achieved excellent performance in many medical image segmentation tasks such as brain magnetic resonance [35]. However, although these “pseudo-3D” methods can consume less computing resources to achieve 3D data segmentation of medical images, the performance improvement is limited because of the insufficient use of global image information [36]. (3) Crop into overlapping 3D patches. V-Net and nnU-Net [37] have adopted this pre-processing for network training and prediction, which significantly reduces the demand for GPU memory, and the network can be designed further deeply to obtain higher segmentation accuracy. However, the lack of target context information and unreasonable image block processing led to a decline in segmentation performance. In addition, block iterative calculation aggravates the time burden of the test phase.

After extensive testing, we observed that incorporating global image context significantly enhances segmentation accuracy. Moreover, higher-resolution input images, which contain richer spatial and semantic information, lead to better segmentation performance when processed by the U-Net model, even when the same network parameters are used.

The U-Net segmentation algorithm has some problems in the processing results, as shown in the second row of

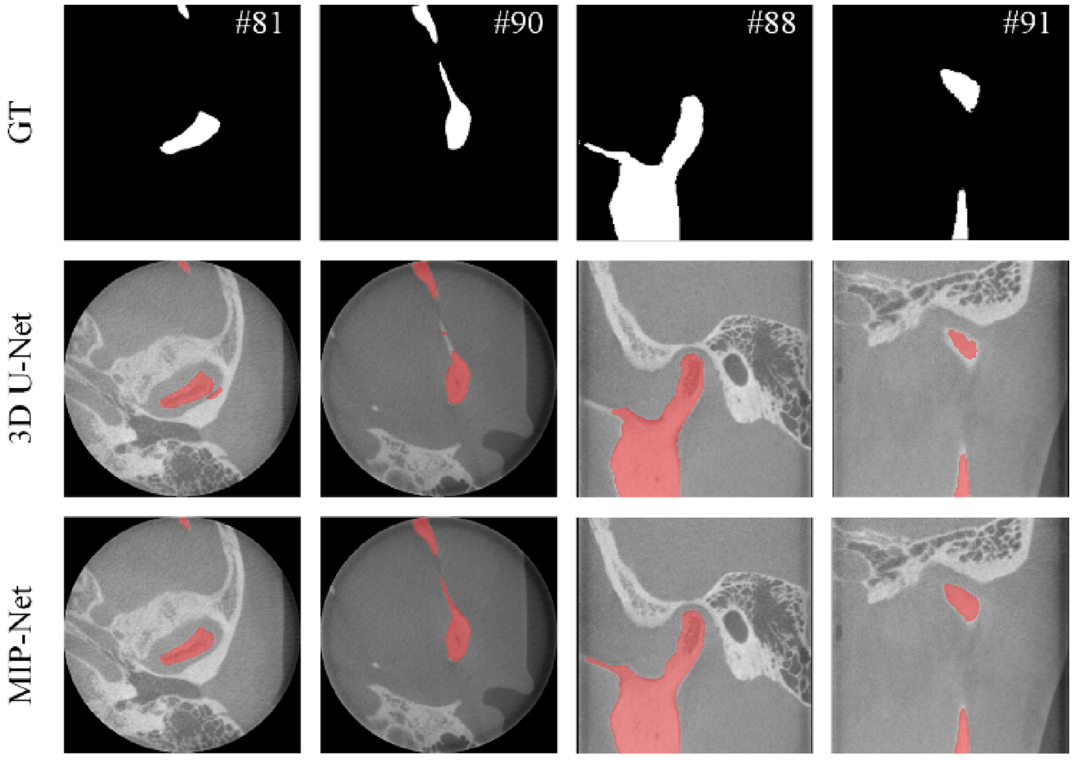


Figure 4: MIP-Net segmentation performance on poorly segmented images of the U-Net method.

Figure 4, where incorrect segmentation occurs at the condyle joint and bone structure tips. To address this issue, this paper proposes the MIP-Net segmentation network based on the image pyramid structure to solve the problems of the U-Net segmentation algorithm. The segmentation results of MIP-Net shown in the third row of Figure 4 can verify that the multi-stage segmentation model based on the image pyramid structure proposed in this paper can achieve high-precision segmentation of TMJ in CBCT images.

While MIP-Net demonstrates significant improvements in segmentation accuracy, its potential benefits for clinical practice require further exploration. The integration of MIP-Net into clinical workflows could substantially enhance diagnostic precision and treatment planning, especially for conditions like TMJ-OA where accurate TMJ segmentation is crucial. For instance, by providing more precise delineation of the mandible and condyle, MIP-Net could aid in more accurate diagnosis and better-targeted treatments. To ensure its successful integration, the model will need to be thoroughly validated in diverse clinical settings. This includes assessing its performance with various imaging conditions and evaluating its impact on clinical workflow efficiency. Key aspects of validation will involve testing the model's ability to handle real-world data variations and its compatibility with existing medical imaging systems and protocols.

Conclusions

To address the issue of image detail loss caused by global down-sampling, we proposed a multi-stage CBCT image segmentation approach for TMJ using the concept of an image pyramid. By leveraging the gradually increasing image resolution and feature sharing across subnetworks, our method significantly improves segmentation performance. Experimental results have demonstrated that, compared to global down-sampling segmentation approaches, our proposed MIP-Net network is effective in reducing the loss of image detail information and improving the segmentation accuracy of the approach.

Research ethics: This study adhered to the guiding principles of the World Medical Association's Helsinki Declaration, conducting biomedical research involving human subjects. The research received approval from the Biomedical Ethics Committee (BEC) of Peking University School and Hospital of Stomatology (PKUSSIRB-201520035).

Informed consent: Informed consent was obtained from all individuals included in this study.

Author contributions: The authors declare their responsibility for the entire content of the manuscript. Individual contributions of each author, if applicable, are as follows: Piaolin Hu: Conceptualization, methodology, investigation, formal analysis, writing - original draft preparation. Jupeng Li*: Supervision, project administration, writing - review and editing. Ruohan Ma: Methodology, investigation, data curation, writing - original draft preparation. Kai Zhang: Resources, data analysis, writing - review and editing. Yong Guo: Methodology, resources, investigation, data curation. Gang Li: Conceptualization, validation, writing - review and editing.

Use of Large Language Models, AI and Machine Learning Tools: None declared.

Conflict of interest: The authors declare no competing interests.

Research funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62172029, the Beijing Natural Science Foundation under Grant L232029, and the Capital's Fund for Health Improvement and Research (CFH 2024-4-4107) from the Beijing Municipal Health Commission.

Data availability: Research data for this study are not publicly available. Access to the data can be provided upon request and is subject to the conditions specified by the authors. Please contact the authors for further inquiries regarding data usage.

References

1. Li W, Zhao S, Yang H, Zhang C, Kang Q, Deng J, et al. Potential novel prediction of TMJ-OA: miR-140-5p regulates inflammation through Smad/TGF-β signaling. Front Pharmacol 2019;10:15.
2. Lu K, Ma F, Yi D, Yu H, Tong L, Chen D. Molecular signaling in temporomandibular joint osteoarthritis. J Orthop Translat 2022;32:21–7.
3. Zhang J, Pi C, Cui C, Zhou Y, Liu B, Liu J, et al. PTHrP promotes subchondral bone formation in TMJ-OA. Int J Oral Sci 2022;14:37.
4. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 2016;39:640–51.
5. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference. Munich, Germany, October 5–9, 2015, proceedings, part III 18: Springer; 2015.
6. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image computing and computer-assisted intervention–MICCAI 2016: 19th International conference. Athens, Greece, October 17–21, 2016, Proceedings, Part II 19: Springer; 2016.
7. Egger J, Pfarrkirchner B, Gsaxner C, Lindner L, Schmalstieg D, Wallner J. Fully convolutional mandible segmentation on a valid ground-truth dataset. In: 2018 40th Annual International conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu, HI, USA: IEEE; 2018.
8. Torosdagli N, Liberton DK, Verma P, Sincan M, Lee JS, Bagci U. Deep geodesic learning for segmentation and anatomical landmarking. IEEE Trans Med Imag 2018;38:919–31.
9. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW). Honolulu, HI, USA: IEEE; 2017.
10. Qiu B, Guo J, Kraeima J, Glas HH, Borra RJ, Witjes MJ, et al. Automatic segmentation of the mandible from computed tomography scans for 3D virtual surgical planning using the convolutional neural network. Phys Med Biol 2019;64:175020.
11. Qiu B, Guo J, Kraeima J, Glas H, Borra R, Witjes M, et al. Recurrent convolutional neural networks for mandible segmentation from computed tomography. arXiv 2020. arXiv preprint arXiv:200306486.
12. Belikova K, Rogov OY, Rybakov A, Maslov MV, Dylov DV. Deep negative volume segmentation. Sci Rep 2021;11:16292.
13. He W, Zhang C, Dai J, Liu L, Wang T, Liu X, et al. A statistical deformation model-based data augmentation method for volumetric medical image segmentation. Med Image Anal 2024;91:102984.
14. Liang X, Li N, Zhang Z, Xiong J, Zhou S, Xie Y. Incorporating the hybrid deformable model for improving the performance of abdominal CT segmentation via multi-scale feature fusion network. Med Image Anal 2021;73:102156.
15. Liang X, Bibault JE, Leroy T, Escande A, Zhao W, Chen Y, et al. Automated contour propagation of the prostate from pCT to CBCT images via deep unsupervised learning. Med Phys 2021;48:1764–70.
16. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). Stanford, CA, USA: IEEE; 2016.
17. Kim YH, Shin JY, Lee A, Park S, Han S-S, Hwang HJ. Automated cortical thickness measurement of the mandibular condyle head on CBCT images using a deep learning method. Sci Rep 2021;11:14852.
18. Wu C, Zhou X, Chen G, editors. Coarse-to-fine tranformer for articular disc of the temporomandibular joint segmentation. In: 2022 15th International congress on image and signal processing, BioMedical Engineering and Informatics (CISP-BMEI). Beijing, China: IEEE; 2022.
19. Kumar skV, Baskaran S. Tri-stage deep learning model for mandibular condyle segmentation and osteoarthritis identification. In: 2023 International conference on network, multimedia and information technology (NMITCON). Bengaluru, India: IEEE; 2023:1–6 pp.
20. Liu Y, Jiao Z, Yao B, Li Q. A three-branch jointed feature and topology decoder guided by game-theoretic interactions for temporomandibular joint segmentation. Comput Biol Med 2024;173:108373.
21. Liu Y, Lu Y, Fan Y, Mao L. Tracking-based deep learning method for temporomandibular joint segmentation. Ann Transl Med 2021;9:467.
22. Vinayahalingam S, Berends B, Baan F, Moin DA, van Luijn R, Bergé S, et al. Deep learning for automated segmentation of the temporomandibular joint. J Dent 2023;132:104475.
23. Choi YT, Song HJ, Lee JS, Im YG. Temporomandibular joint segmentation using deep learning for automated three-dimensional reconstruction. J Oral Med Pain 2024;49:109–17.
24. Lv B, Liu F, Gou F, Wu J. Multi-scale tumor localization based on priori guidance-based segmentation method for osteosarcoma MRI images. Mathematics 2022;10:2099.

25. Feng S, Zhao H, Shi F, Cheng X, Wang M, Ma Y, et al. CPFNet: context pyramid fusion network for medical image segmentation. IEEE Trans Med Imag 2020;39:3008–18.
26. Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. IEEE Computer Society/IEEE/CVF; 2021.
27. Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:191109516. 2019.
28. Hou Q, Zhou D, Feng J. Coordinate attention for efficient Mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE Computer Society/IEEE/CVF; 2021.
29. Kushnure DT, Talbar SN. HFRU-Net: high-level feature fusion and recalibration unet for automatic liver and tumor segmentation in CT images. Comput Methods Progr Biomed 2022;213:106501.
30. Dong Z, An S, Zhang J, Yu J, Li J, Xu D. L-UNet: a landslide extraction model using multi-scale feature fusion and attention mechanism. Remote Sens 2022;14:2552.
31. Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The importance of skip connections in biomedical image segmentation. In: International workshop on deep learning in medical image analysis. Athens, Greece: Springer; 2016.
32. Zhang K, Li J, Ma R, Li G. An end-to-end segmentation network for the temporomandibular joints CBCT image based on 3D U-Net. In: 2020 13th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI). Chengdu, China: IEEE; 2020.
33. Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, et al. Medical image segmentation review: the success of u-net. IEEE Trans Pattern Anal Mach Intell 2024. https://doi.org/10.1109/tpami.2024.3435571.
34. Oda M, Roth HR, Kitasaka T, Misawa K, Fujiwara M, Mori K. Abdominal artery segmentation method from CT volumes using fully convolutional neural network. Int J Comput Assist Radiol Surg 2019;14: 2069–81.
35. Perslev M, Dam EB, Pai A, Igel C. One network to segment them all: a general, lightweight system for accurate 3d medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference. Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22: Springer; 2019. .
36. Vu MH, Grimbergen G, Nyholm T, Löfstedt T. Evaluation of multislice inputs to convolutional neural networks for medical image segmentation. Med Phys 2020;47:6216–31.
37. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11.