Original Article

Metastasis lesion segmentation from bone scintigrams using encoder-decoder architecture model with multi-attention and multi-scale learning

Ailing Xie[1,2], Qiang Lin[1,2,3]^, Yang He[2,3], Xianwu Zeng[4], Yongchun Cao[1,2,3], Zhengxing Man[1,2,3], Caihong Liu[1,2,3], Yusheng Hao[1,2,3], Xiaodi Huang[5]

[1]School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, China; [2]Gansu Provincial Engineering Research Center of Multi-modal Artificial Intelligence, Lanzhou, China; [3]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, China; [4]Department of Nuclear Medicine, Gansu Provincial Cancer Hospital, Lanzhou, China; [5]School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, Australia

Contributions: (I) Conception and design: Q Lin, A Xie; (II) Administrative support: Q Lin; (III) Provision of study materials or patients: X Zeng; (IV) Collection and assembly of data: X Zeng, Q Lin; (V) Data analysis and interpretation: Q Lin, Y Cao, Z Man; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Qiang Lin, PhD. School of Mathematics and Computer Science, Northwest Minzu University, 1 Xibei Xincun Rd., Lanzhou 730000, China; Gansu Provincial Engineering Research Center of Multi-modal Artificial Intelligence, Lanzhou, China; Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, China. Email: qiang.lin2010@hotmail.com.

Background: The limitation in spatial resolution of bone scintigraphy, combined with the vast variations in size, location, and intensity of bone metastasis (BM) lesions, poses challenges for accurate diagnosis by human experts. Deep learning-based analysis has emerged as a preferred approach for automating the identification and delineation of BM lesions. This study aims to develop a deep learning-based approach to automatically segment bone scintigrams for improving diagnostic accuracy.

Methods: This study introduces a deep learning-based segmentation model structured around an encoder-decoder architecture. The model employs a multi-attention learning scheme to enhance the contrast of the skeleton outline against the background and a multi-scale learning strategy to highlight the hotspots within skeletal areas. The multi-attention strategies include the Non-local Attention scheme and the vision transformer (ViT), while the multi-scale learning incorporates the multi-scale feature learning strategy and the multi-pooling learning strategy. This combination enables the proposed model to accurately detect and extract lesions of varying sizes with high randomness in location and intensity.

Results: Experimental evaluation conducted on clinical data of single photon emission computed tomography (SPECT) bone scintigrams showed the superior performance of the proposed model, achieving the highest-ever dice similarity coefficient (DSC) score of 0.6720. A comparative analysis on the same dataset demonstrated increased scores of 5.6%, 2.03%, and 7.9% for DSC, Precision, and Recall, respectively, compared to the existing models.

Conclusions: The proposed segmentation model can be used as a promising tool for automatically extracting metastasis lesions from SPECT bone scintigrams, offering significant support to the development of deep learning-based automated analysis for characterizing BM.

Keywords: Tumor bone metastasis; bone scintigram; lesion segmentation; multi-attention scheme; multi-scale feature learning

^ ORCID: 0000-0002-3842-2634.

Submitted Jun 20, 2024. Accepted for publication Nov 25, 2024. Published online Dec 30, 2024.
doi: 10.21037/qims-24-1246
View this article at: https://dx.doi.org/10.21037/qims-24-1246

Introduction

Patients with malignant tumors may develop bone metastasis (BM) when a solid tumor invades the bone (1). The bone ranks as the third most affected site after the lung and liver (2). Clinical events indicate that breast cancer accounts for over 60% of BM cases (3,4), while the remaining cases metastasized from thyroid, lung, and kidney cancers (5). Patients suffering from BMs often experience a series of skeletal-related events (6,7), significantly affecting both survival time and quality of life. Therefore, early detection of BMs is crucial for making treatment decisions and improving survivability (6).

Medical imaging offers a noninvasive, routine method for surveying BMs (8). Currently, a variety of medical imaging modalities are available, including bone scintigraphy (also known as a bone scan), whole-body magnetic resonance imaging (MRI), and positron emission tomography (PET) (9). Specifically, single photon emission computed tomography (SPECT) is the routine clinical technique for bone scintigraphy due to its cost-effectiveness and high detection sensitivity (10,11). By using SPECT bone scintigraphy (12), metastasized bone sites can be detected and displayed as hotspots in the scintigrams with a high uptake of the radionuclide of technetium-99 methylene diphosphonate (^{99m}Tc-MDP) (13).

SPECT bone scintigraphy is characterized by its limited spatial resolution (14) with a distance of 2.26 mm between two adjacent elements in a typical 1,024 (row) × 256 (column) whole-body bone scintigram. Unlike natural images, in which the pixel values have a relatively narrow range, the element values of the bone scintigram, which represent the captured counts of ^{99m}Tc-MDP, can span from zero to hundreds even thousands, depending on patients' metabolic activity. Moreover, some normal bones, such as the vertebrae, may exhibit high radionuclide uptake (15). These factors, combined with the unpredictable and irregular nature of metastasis lesions in terms of location, size, and shape, pose significant challenges for manual diagnosis in nuclear medicine practices (16).

Medical professionals and researchers have recently turned to automated image analysis, where machine learning, especially deep learning, plays a crucial role in the diagnostic process (17). Deep learning-based medical image analysis has immense potential for accurately characterizing BM and improving diagnosis (18). Convolutional neural networks (CNNs) are among the most widely used deep learning architectures (19,20), possessing the capacity to automatically learn significant image features without requiring human supervision.

Most of the CNN-based analysis of SPECT bone scintigrams concentrates on determining whether BM is present in scintigrams by classifying them (21-24). Several pioneering studies (25) are relevant to our work, such as BM Segmentation. Based on an Improved U-Net Algorithm, which enhances U-Net with an attention mechanism in the skip connections to improve feature selection and segmentation accuracy, the proposed model achieves superior performance in segmenting bone scintigraphy images (26). Another example is a study on Semantic Segmentation in Radiation Therapy for Prostate Cancer, which uses a 2D U-Net with additional bladder and rectum channels to increase segmentation accuracy on non-contrast computed tomography (CT) images (27). Shimizu *et al.* (28) introduced a dual-stage approach to first extract metastasis lesion hotspots from whole-body bone scintigrams and then classify these hotspots to identify true metastasis lesions by using the existing model BtrflyNets (29). Saito *et al.* (30) focused on hotspot extraction of metastasis lesions from whole-body bone scintigrams using the BtrflyNets model, along with symmetry-based hotspot assessment. In addition to these, custom segmentation models (31,32) based on the encoder-decoder structure of the U-Net model (33) were developed specifically for segmenting BM lesions from regional SPECT bone scintigrams in the thorax. Some efforts have also been made towards building semi-supervised models to mitigate the scarcity of annotated images for CNN-based supervised segmentation. MaligNet (34), for instance, is a two-stage semi-supervised model that first segments hotspots and then classifies them to diagnose BM. An end-to-end semi-supervised model was proposed by Lin *et al.* (35) for segmenting BM lesions in bone scintigrams. Additionally, Huang *et al.* (36) proposed a semi-supervised model that balances unsupervised label construction and supervised learning, offering valuable insights for weakly labeled image analysis. Yu *et al.* (37) explored the effect of bone segmentation on an internal dataset by comparing three CNN-based models. Chen *et al.* (38) introduced a pairwise attention-enhanced adversarial model that performs bone segmentation through pairwise attention graphs and semantic fusion. Morita *et al.* (39) employed U-Net to automatically segment facial bones into eight regions. Li *et al.* (40) proposed a new semi-supervised learning method to leverage the unique bone structure in CT scans, along with a patch-shuffle-based data augmentation method for bone segmentation. Zhan *et al.* (41) introduced the Smart Assisted Framing Network (SEAGNET), which captures contextually relevant information in the feature graph through hybrid attention and uses edge attention to guide the network's focus on boundaries. Afnouch *et al.* (16) developed Hybrid-AttUnet++ with dual decoders for simultaneous segmentation of BMs and bone regions. Zhou *et al.* (42) proposed a two-stage whole-body bone SPECT scan residual artifact image restoration algorithm based on contextual attention. The first stage performs bone segmentation, and the second stage restores artifacts in the segmented image. Liu *et al.* (43) presented a deep learning-based automatic analysis method for bone scan images to detect BMs, including a bone scan classification model, a region segmentation model, a tumor burden assessment model, and a diagnostic report generation model. Li *et al.* (44) proposed an efficient segmentation model for MR spine images, called Inception-CBAM Unet++, which enhances feature extraction with Inception and attention modules. The model significantly improves segmentation metrics, achieving an IoU of 83.16% and a DSC of 90.32% on the SpineSagT2W dataset. Wang *et al.* (45) proposed an efficient method for medical image segmentation, introducing residual interconnections, attention modules, and an adaptive denoising learning strategy (ADL) to mitigate noisy labels. Their model demonstrated competitive performance on spinal CT datasets. Latif *et al.* (46) introduced a multi-inception-UNET model to improve the scalability of U-Net for brain tumor segmentation. Their model demonstrated superior performance on the BraTS datasets, achieving the best results for complete, core, and enhancing tumor regions. Qiu *et al.* (47) introduced AgileFormer, a spatially agile ViT-UNet model designed for medical image segmentation, incorporating deformable patch embedding, spatially dynamic multi-head attention, and deformable positional encoding. Their experiments across three segmentation tasks demonstrated the model's effectiveness.

However, it is worth noting that the CNN-based segmentation of SPECT bone scintigrams is still in its infancy, and the performance of the aforementioned studies requires further improvement. Based on an in-depth analysis of the characteristics of SPECT bone scintigraphy, this work presents a CNN-based segmentation model for automatically separating metastasis lesions from a bone scintigram.

We adopted an encoder-decoder structure to address the significant size disparity between the background and foreground regions. To address the variability in lesion size, location, shape, and radiance values, we employed a multi-scale feature learning strategy at each layer and a multi-pooling learning strategy during the initial down-sampling. Additionally, to tackle the issues of poorly defined lesion contours and inconspicuous hotspots in BMs, we integrated multiple attention mechanisms, including non-local attention and vision transformer (ViT), in the model's bottom layer.

The rest of this paper is organized as follows. In Methods, we present the proposed BM lesion segmentation model. Results reports the experimental evaluation conducted on clinical SPECT bone scintigrams. In Discussion, we provide a brief discussion of the ablation study regarding the pros and cons of the proposed model. In Conclusions, we conclude this work and point out future research directions.

Methods

In this study, we propose a BM lesion segmentation model based on an encoder-decoder architecture, integrating with a non-local attention scheme, ViT, multi-pooling, and multi-scale feature learning strategies. The network structure of the proposed model is illustrated in *Figure 1*.

As shown in *Figure 1*, during the encoding stage, a composite convolution (i.e., Conv 3×3 + Conv 3×3) is used to double the number of channels apart from the input (i.e., 256^2×1) and a pooling operation (i.e., MaxPooling) is used to halve the size of a feature map. On the contrary, during the decoding stage, a transpose convolution (i.e., Transpose Conv) is used to recover the feature maps layer by layer.

The encoding process is used to learn image features, gradually focusing on target areas from large to small. The decoding process then reconstructs the image, generating the final output through step-by-step restoration. This structure not only meets the demands of image segmentation but also effectively extracts and utilizes multi-scale features within the images, thereby enhancing segmentation performance.

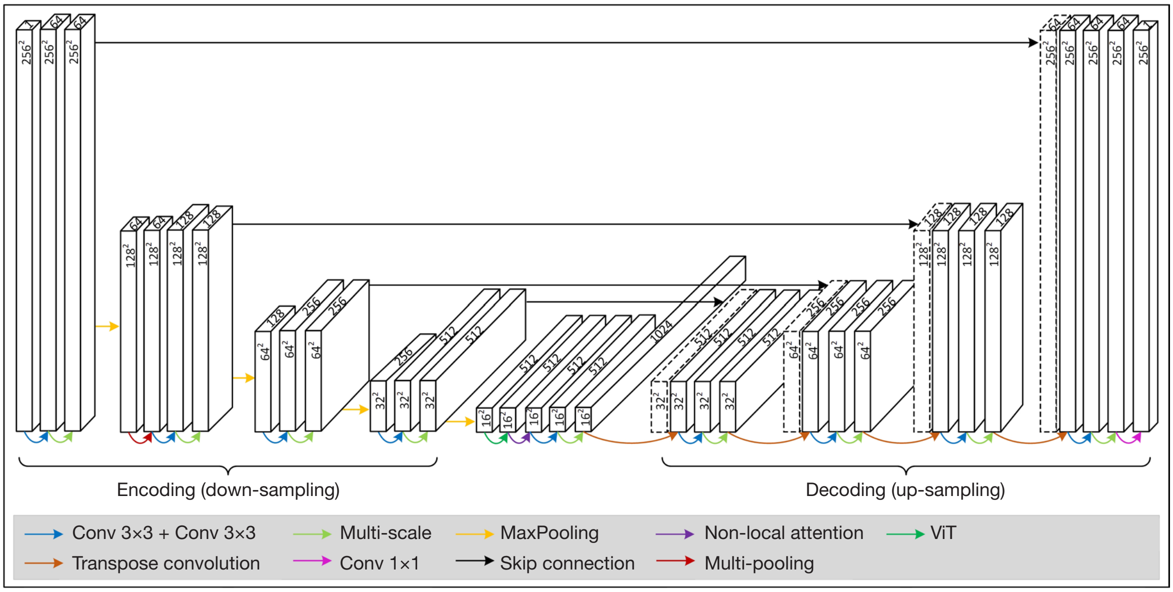


Figure 1 The network structure of the proposed encoder-decoder architecture segmentation model with multi-attention and multi-scale feature learning strategies. ViT, vision transformer.

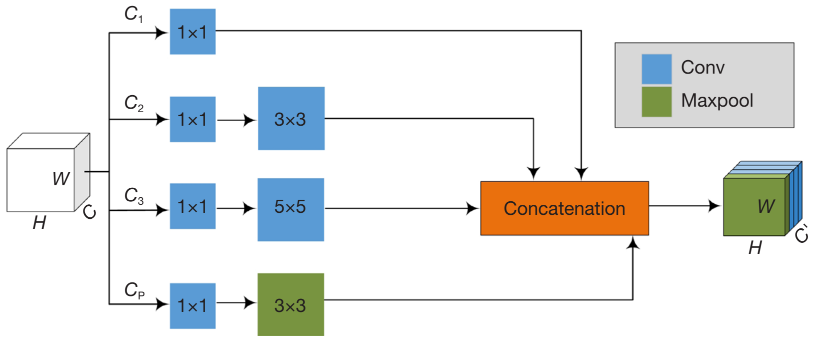


Figure 2 An illustration of the multi-scale feature learning block used in the proposed model. *C*, channels; *H*, height; *W*, width; C_1, number of channels in branch 1; C_2, number of channels in branch 2; C_3, number of channels in branch 3; C_p, number of channels in branch *p*; *C*', output channels.

The core components of the model that enhance segmentation performance include multi-scale feature learning applied across all layers, multi-pooling feature learning implemented in the second layer, and the Non-local attention mechanism along with the ViT utilized in the final layer. These components will be elaborated upon in the following sections.

Multi-scale feature learning strategy

As illustrated in *Figure 1*, we use multi-scale learning block in every layer of the proposed model to help the model capture information about size-varied lesions. *Figure 2* depicts the structure of the multi-scale learning block used.

Following the structure of Inception model (48), the multi-scale learning block contains five branches with different kernel sizes to focus on size-varied lesions. The outputs of these branches are then concatenated according to Eq. [1].

$$C' = C_1 + C_2 + C_3 + C_P$$ [1]

where *C*', C_1, C_2, C_3, and C_P indicate the number of the channel in the output feature map, the number of the channel in the feature map with a kernel size of 1×1, the

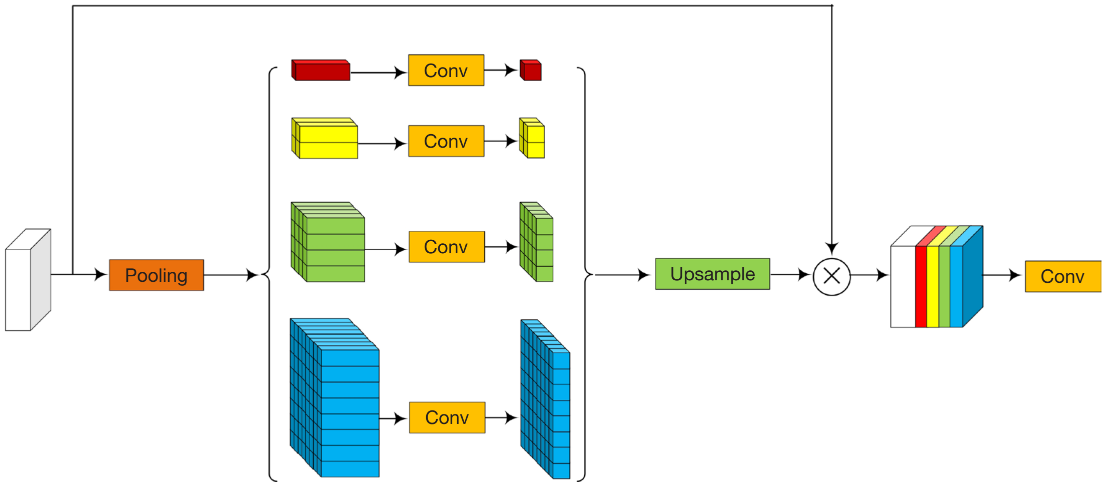


Figure 3 An illustration of the multi-pooling learning block used in the second layer.

number of channels in the feature map with a kernel size of 3×3, the number of channels in the feature map with a kernel size of 5×5, and the number of channels in the feature map with max pooling, respectively.

Let $U \in \mathbb{R}^{H \times W \times C}$ be the input feature cube, the output feature cube U' can be formally represented as follows.

$$U' = H \times W \times \left(C_1 + C_2 + C_3 + C_P\right) \quad [2]$$

where H and W is the height and width of a feature, respectively.

Multi-pooling learning strategy

In this work, we employ a multi-pooling learning mechanism to help the model capture objects (i.e., metastasis lesions) of different sizes and details in a SPECT bone scintigram. *Figure 3* illustrates the detail of the multi-pooling learning block used.

As seen in *Figure 3*, the multi-pooling module (49) contains four pooling layers with window sizes of 1×1, 2×2, 4×4, and 8×8. In order to maintain the weight of the global features, we use a 1×1 convolutional layer after each pooling layer to reduce the dimensions of the contextual representations to 1/N of the original dimensions:

$$F_i = Conv_{1\times1}\left(AdaptiveAvgPool_{s_i}\left(F_{input}\right)\right), s_i \in \{1,2,4,8\} \quad [3]$$

where S_i is the window size of the ith pooling layer, F_{input} is the input feature map, F_i is the feature map after being processed by a 1×1 convolution. The feature maps output from each pooling layer are then upsampled to the original input feature map size by bilinear interpolation:

$$F_i^{upsampled} = BilinearUpsample\left(F_i, size\left(F_{input}\right)\right), s_i \in \{1,2,4,8\} \quad [4]$$

where $F_i^{upsampled}$ is the feature map after sampling on the ith pooling layer. The original input feature map is then spliced with the feature maps after sampling on each pooling layer:

$$F_{concat} = Concat\left(F_{input}, F_1^{upsampled}, F_2^{upsampled}, F_3^{upsampled}, F_4^{upsampled}\right) \quad [5]$$

where F_{concat} is the spliced feature map.

Multi-pooling extracts features at different scales through adaptive average pooling operations with multiple pooling windows (1×1, 2×2, 4×4, 8×8). This multi-scale feature extraction allows the model to capture objects and details of various sizes, aiding in handling lesions of different scales and shapes, thereby reducing false negatives and false positives. Additionally, pooling at different scales captures richer contextual information, which helps the model better understand the relationship between local and global features, thus improving the accuracy of segmentation boundaries. We place this module after the first down-sampling step in the model. This design enables the integration of multi-scale information at an early stage, thereby enhancing the model's ability to detect lesion areas.

ViT learning strategy

We utilize the ViT architecture to aid the model in capturing global contextual information and enhancing feature representation capabilities. *Figure 4* details the ViT learning block used.

The ViT (50) is an innovative approach that adapts the Transformer architecture for computer vision tasks.

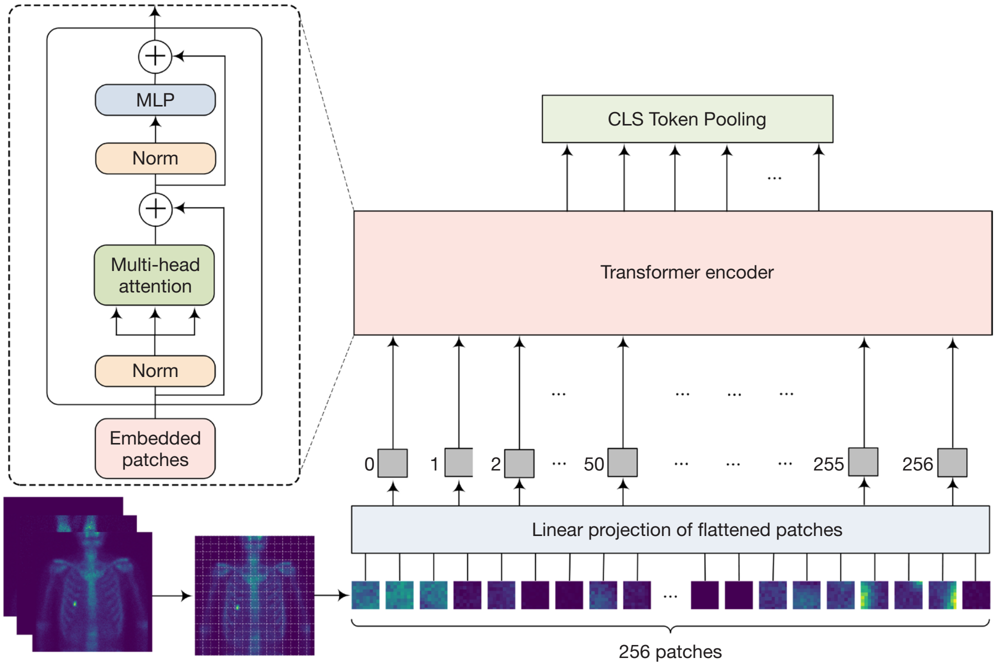


Figure 4 An illustration of the ViT learning block used in the last layer. ViT, vision transformer; CLS Token Pooling, Classification Token Pooling; MLP, multilayer perceptron.

The ViT module comprises several key components: image segmentation and embedding, positional encoding, Transformer encoder layers, and Classification Token Pooling (CLS Token Pooling). Below is a detailed description of the ViT module.

As illustrated in *Figure 4*, the input image $X \in \mathbb{R}^{H \times W \times C}$ (where H is the height, W is the width, and C is the number of channels) is partitioned into non-overlapping patches, each of size P×P. Given an image of size H×W, this division results in N patches. The number of image blocks is calculated as Eq. [6].

$$N = \left[\frac{H}{P}\right] \times \left[\frac{W}{P}\right] \qquad [6]$$

Each patch is flattened into a vector, denoted $X_p \in \mathbb{R}^{P \times P \times C}$, where C is the number of channels. Subsequently, each patch is mapped to a vector space of fixed dimension D using a linear transformation:

$$z_o^i = Ex_p^i + p_i \qquad [7]$$

where $E \in \mathbb{R}^{P \times P \times C \times D}$ is a trainable linear matrix, and $p_i \in \mathbb{R}^D$ represents the position encoding vector of the i-th image block, z_0^i is the embedding vector of the i-th image patch. Additionally, a classification marker (CLS) is appended at the beginning of the input sequence. The initial state of this marker is a trainable vector, designed to aggregate information from all patches in the image. Consequently, the initial input sequence z_0 is represented as follows:

$$z_0 = \left[z_0^0; z_0^1; z_0^2; \cdots; z_0^N\right] \qquad [8]$$

The input sequence is then passed through the Transformer Encoder module. Initially, layer normalization is performed on the input z_{l-1}:

$$\hat{z}_{l-1} = LayerNorm\left(z_{l-1}\right) \qquad [9]$$

For each layer l and each head h within the multi-head self-attention mechanism, the Query, Key, and Value matrices are computed:

$$Q_h = \hat{z}_{l-1} W_Q^h, K_h = \hat{z}_{l-1} W_K^h, V_h = \hat{z}_{l-1} W_V^h \qquad [10]$$

where $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{D \times D_k}$ represents the trainable weight matrix of the h-th head, D denotes the dimension of the input vector, and D_k signifies the dimension of each head. Q_h is the query, K_h is the key, V_h is the value, which are used to compute the self-attention mechanism. The computational self-attention is subsequently performed as follows:

Table 1 ViT parameter settings

Parameter	Value
Image_size	256/[$2^{\text{len(features)}}$]
Patch_size	16/[$2^{\text{len(features)}}$]
Num_classes	512
Dim	512
Depth	6
Heads	8
Dropout	0.1

ViT, vision transformer; Num_classes, number of output classes; Dim, dimension; len(features), length of the feature sequence in the input.

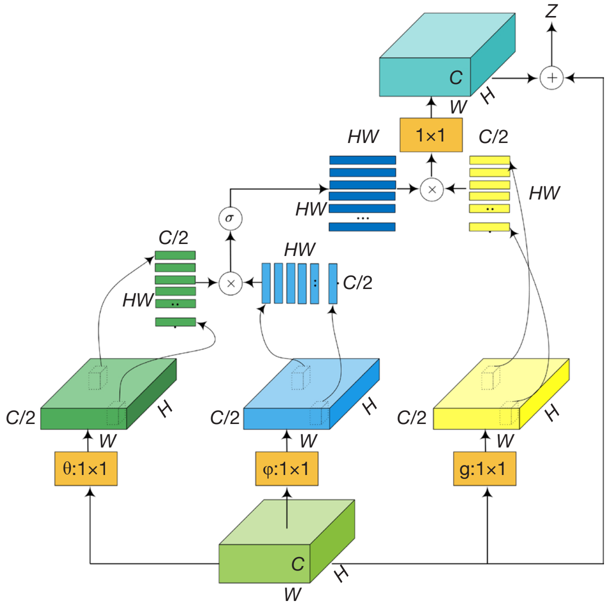


Figure 5 A non-local attention learning mechanism used in the last layer. *C*, channels; *H*, height; *W*, width; *Z*, output feature map; *θ*, produces the query feature map; *φ*, produces the key feature map; *g*, produces the value feature map.

$$Attention\left(Q_h, K_h, V_h\right) = softmax\left(\frac{Q_h K_h^T}{\sqrt{D_k}}\right) V_h \qquad [11]$$

The outcome of a multi-head self-attention mechanism, combined across *H* heads, is as Eq. [12].

$$MultiHead\left(Q, K, V\right) = Concat\left(head_1, head_2, \cdots, head_H\right) W_o \qquad [12]$$

where $head_h = Attention\left(Q_h, K_h, V_h\right)$, $W_o \in \mathbb{R}^{H \cdot D_v \times D}$ is the trainable output weight matrix. By applying residual connection and layer normalization to the output of each layer, we obtain Eq. [13].

$$\hat{z}_l = LayerNorm\left(z_{l-1} + MultiHead\left(Q, K, V\right)\right) \qquad [13]$$

It is then applied to each position via a feed-forward neural network.

$$FFN\left(z\right) = MLP\left(GELU\left(zW_1 + b_1\right)\right) W_2 + b_2 \qquad [14]$$

where W_1, W_2 and b_1, b_2 are trainable weights and biases and GELU is the activation function. Then applied residual connection and layer normalization:

$$z_l = LayerNorm\left(\hat{z}_l + FFN\left(\hat{z}_l\right)\right) \qquad [15]$$

Finally, the feature map is extracted by CLS Token Pooling:

$$h = z_{[CLS]}^{L} \qquad [16]$$

where *h* is the feature map after processing through the ViT module. *Table 1* below shows the description of the ViT parameters.

ViT excels at capturing global contextual information within an image. Through the self-attention mechanism, it can focus on the entire image from each position of the feature map. Incorporating the ViT module at the lowest level of the codec effectively combines the local feature extraction capability of CNNs with the global feature extraction capability of the Transformer. This integration enhances segmentation accuracy, particularly in images with complex backgrounds or similar regions.

Non-local attention

In this work, a non-local attention learning mechanism (51) is employed to enhance the model’s focus on SPECT bone metastatic lesions. *Figure 5* illustrates the detailed architecture of the non-local attention module.

The input feature map *x* has dimensions (*B*, *C*, *H*, *W*), where *B* is the batch size, *C* is the number of channels, and *H* and *W* are the height and width of the feature map, respectively. The model channels the input feature map through three 1×1 convolution layers (θ, φ, *g*) to reduce the input channels from *C* to *C/*2.

$$\theta\left(x\right) = Conv2d\left(x\right), \varphi\left(x\right) = Conv2d\left(x\right), g\left(x\right) = Conv2d \qquad [17]$$

Spread the transformed feature map:

$$Y = Y.view\left(B, C/2, H \times W\right), Y \in \left(\theta_x, \varphi_x, g_x\right) \qquad [18]$$

Table 2 An overview of patient characteristics

Patient characteristics	Male (N=103)	Female (N=65)	Total (N=168)
Age (years)	59.11±9.84 [29.0–81.0]	56.68±9.98 [36.0–75.0]	58.17±9.93 [29.0–81.0]
Number of anterior lesions	2.65±2.44 [0–13]	3.20±2.43 [0–12]	5.69±36.86 [0–481]
Number of posterior lesions	3.09±3.56 [0–17]	3.34±3.33 [0–14]	6.33±41.05 [0–535]
Total number of lesions	591	425	1,016

Data are presented as mean ± standard deviation and, where applicable, range [minimum – maximum].

where the dimension of Y is (B, $C/2$, N) at this point $N = H{\times}W$. θ_x represents the 1×1 convolution operation in the first branch, used to generate the query features (Query). φ_x represents the 1×1 convolution operation in the second branch, used to generate the key features (Key). g_x represents the 1×1 convolution operation in the third branch, used to generate the value features (Value). The similarity matrix f between θ_x and φ_x is then computed:

$$f = torch.matmul\left(\left(\theta_{x\cdot} permute\left(0,2,1\right)\right),\varphi_x\right) \quad [19]$$

The similarity matrix is then softmax normalized:

$$f_{div} = F.softmax\left(f, dim = -1\right) \quad [20]$$

The normalized similarity matrix f_{div} is subsequently used to multiply with g_x:

$$y = torch.matmul\left(f_{div}, g_x, permute\left(0,2,1\right)\right) \quad [21]$$

The resultant matrix y is transposed back to its original dimensions:

$$y = y.permute\left(0,2,1\right).contiguous\left(\cdot\right) \quad [22]$$

Residual connection is performed after reshaping back to the original feature map shape and restoring the original number of channels. The final output feature map is z, which is the sum of the residual output and the original input feature map.

$$z = Conv2d\left(y.view\left(B, C/2, H, W\right)\right) + x \quad [23]$$

Placing the Non-Local Block module after the ViT module effectively combines the global feature extraction capability of ViT, which performs global feature extraction on the feature map through the self-attention mechanism, focusing on the structure and features of the image as a whole. The Non-Local Block further enhances this global contextual information, allowing each position on the feature map to more accurately focus on other positions. This combination can improve the accuracy of lesion segmentation.

This work was approved by the Ethics Committee of Gansu Provincial Cancer Hospital (No. A202106100014). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The used bone SPECT images were de-identified before the authors received the data. The fully anonymised image data was received by the authors on June 01, 2021. A requirement for informed consent was waived for this study because of the anonymous nature of the data.

Results

In this section, we present the experimental results of the segmentation performance obtained by the proposed model on clinical data of SPECT BM scintigrams, beginning with a description of the experimental data used.

Experimental data

The data of the planar bone scintigrams used in this retrospective study were collected from the Department of Nuclear Medicine, Gansu Provincial Cancer Hospital, China, spanning the period from January 2016 to December 2019. The bone scintigrams were acquired using single-head imaging equipment (GE SPECT Millennium MPR) that captured both the anterior- and posterior-view scintigrams from each patient who was intravenously injected with ^{99m}Tc-MDP (20–25 mCi).

Patient information

The statistical information of patients in this dataset is presented in *Table 2* below, including the total number of male and female patients, age range, number of anterior view lesions, and number of posterior view lesions.

The study comprises 168 patients diagnosed with BMs secondary to lung cancer, with an average age of approximately 58.17 years (age at the time of imaging). As shown in *Table 2*, the total number of lesions is 1,016, with

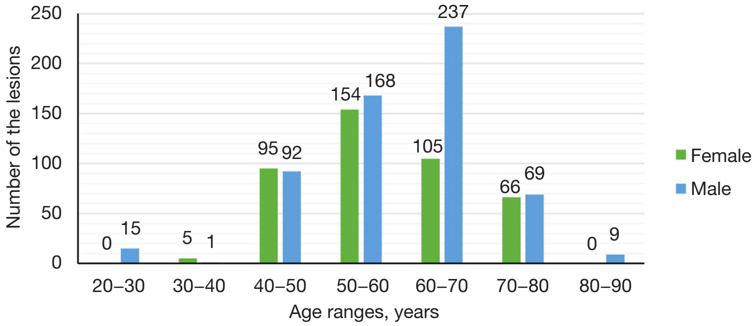


Figure 6 Statistical distribution of lesions across age groups.

males having 591 lesions and females having 425 lesions.

The statistics regarding the number of individuals suffering from lung cancer BMs in each age group, based on their age at the time of imaging, are depicted in the chart (see *Figure 6*).

As illustrated in *Figure 6*, males exhibit a significant risk of developing the disease between the ages of 60 and 70 years. For both males and females aged 50 to 60 years, the number of foci is notably higher. The number of foci generally decreases in the older age group. Overall, the number of foci increases with age for both males and females until it begins to decline after the age of 60 to 70 years.

Data preprocessing

Step 1: normalization

SPECT images differ significantly from standard grayscale or color images, in which pixel values typically range from 0 to 255. To address the impact of variations in radiation intensity on image feature extraction, each DICOM file undergoes adaptive normalization. This normalization adjusts pixel intensities within a fixed range based on the observed maximum and minimum values in the dataset.

In this study, whole-body SPECT images of size m=256 and n=1,024 are used. The dataset is obtained after normalization.

Step 2: bladder region removal

Non-lesion hot spots, characterized by high radiation values, can significantly affect the accurate diagnosis of true lesions by medical professionals. When bones are affected by disease, changes in density and structure lead to high uptake areas of radiopharmaceuticals. However, the concentration of radiopharmaceuticals in lesion areas is significantly lower compared to regions like the injection site and bladder. Extremely high radiation values in non-lesion hot spots can lead to the "big numbers eat small numbers" phenomenon, making it difficult to identify lesions and resulting in missed diagnoses.

Step 3: labelling

In order to obtain the manual labels (i.e., ground truth) for each image, we asked three experienced nuclear medicine physicians (one chief physician, one associate senior technologist, and one technologist in charge) within our research group to manually label the metastasized lesions using an ITK-SNAP-based system (http://www.itksnap.org/).

Step 4: image segmentation

To facilitate the model's processing, the input images are cropped into manageable segments. The original images, sized at 256×256, undergo cropping using a sliding window approach with a stride of 128. This process results in the creation of several sub-images, each sized at 256×256 pixels, from a single 256×1,024-pixel image. Subsequently, the corresponding label files are cropped in the same manner.

Images containing BM are then identified and isolated from the cropped images. These identified images, along with their manually crafted annotations, constitute the training dataset for our proposed model. *Table 3* presents a statistical overview of the data utilized in this study.

We divided the 838 images used into two subsets, i.e., the training set consisting of 590 images and the test set consisting of 248 images (see *Table 3*). Each subset was divided based on patients rather than individual images to prevent data leakage between the training and test sets.

Experimental setup

The experimental evaluation metrics used in the experiment include dice similarity coefficient (DSC), Recall, and Precision, defined in Eqs. [24-26].

$$DSC = \frac{2TP}{2TP + FP + FN} \qquad [24]$$

Table 3 The statistics of the data of SPECT BM scintigrams used

Statistics	Value
Number of images	838
Image size	256×256
Training set (~70%)	590
Test set (~30%)	248

SPECT, single photon emission computed tomography; BM, bone metastasis.

Table 4 Parameter settings of the proposed segmentation model

Model parameter	Value
Learning rate	1e−4
Optimizer	Adaptive moment estimation
Batch size	16
Epoch	200

Table 5 The experimental scores of evaluation metrics obtained by the proposed model on test samples

Evaluation metric	Score
DSC	0.6720
Recall	0.6671
Precision	0.6771
Time	1 h + 3 min

DSC, dice similarity coefficient.

$$Recall = \frac{TP}{TP + FN} \qquad [25]$$

$$Precision = \frac{TP}{TP + FP} \qquad [26]$$

Here, TP represents true positives, FP denotes false positives, and FN indicates false negatives.

The parameter settings of the segmentation models used in the experiments are detailed in *Table 4*. All experiments were conducted on the PyTorch 1.11.0 platform, using an Intel Core i7-9700 PC with 32GB RAM running Windows 10. All models were trained on NVIDIA RTX 3090 GPU.

Experimental results

Table 5 reports the experimental results in terms of the defined evaluation metrics on samples in the test set.

Table 6 Overview of the existing image segmentation models for comparison

Model	Alias	Purpose
U-Net (33)	M #1	General
U-Net++ (52)	M #2	General
Attention U-Net (53)	M #3	General
Res_U-Net (54)	M #4	General
AMSUnet (55)	M #5	General
CMUNeXt (56)	M #6	General
Multi-inception-UNet (46)	M #7	General
RAR-U-NET (45)	M #8	General
ICUnet++ (44)	M #9	General
Model in (32)	M #10	Specialized1
Model in (43)	M #11	Specialized2

The results in *Table 5* demonstrate that our model performs well on segmenting BM lesions in low-resolution SPECT bone scintigrams, achieving a DSC metric score of 0.6720.

To benchmark our model, we tested a group of classical image segmentation models using the same dataset outlined in *Table 3*. *Table 6* lists the existing models used for comparison, and the experimental results are illustrated in *Figure* 7.

The experimental results presented in *Figure* 7 demonstrate that our model outperforms all existing models in terms of the DSC and Recall metrics, and achieves competitive Recall. Notably, our proposed model surpasses the specialized model (32), achieving increased scores of 5.6%, 2.03%, and 7.9% for DSC, Precision, and Recall, respectively. Additionally, when compared to the existing full-body segmentation model (43), our model shows improvements of 5.7%, 5.3%, and 2.9% in DSC, Precision, and Recall, respectively. While our model does not achieve the highest Precision across all comparative experiments, it demonstrates a balanced performance that excels in both DSC and Recall. This balance makes it more robust in reducing false positives and enhancing segmentation precision, which is essential for practical clinical applications. Therefore, we can conclude that our model is an effective automated tool for accurately segmenting BM lesions from low-resolution SPECT bone scintigrams.

The segmentation results are shown in *Figure 8*.

As illustrated in *Figure 8*, our model significantly outperforms the other models in terms of segmentation

Table 7 The experimental results of the ablation studies on our segmentation model

Evaluation metric	Multi-pooling/multi-attention/multi-scale/ViT				
	x/x/x/x	x/x/x/√	x/x/√/√	x/√/√/√	√/√/√/√
DSC	0.6369	0.6446	0.6408	0.6536	0.6720
Recall	0.5911	0.6043	0.5977	0.6261	0.6671
Precision	0.6918	0.6906	0.6912	0.6840	0.6771
Time	18 min	17 min	53 min	33 min	1 h + 3 min

ViT, vision transformer; DSC, dice similarity coefficient.

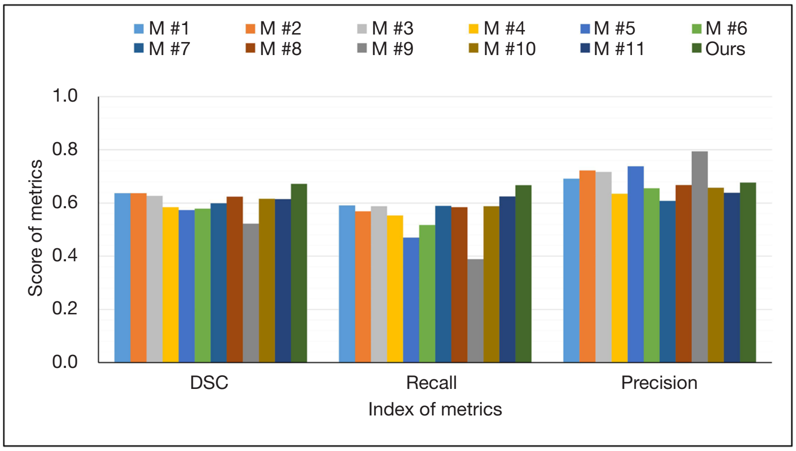


Figure 7 Comparative analysis of segmentation performance between our model and the existing ones. M, model; DSC, dice similarity coefficient.

accuracy. Not only does it achieve superior results compared to general segmentation models, but it also surpasses models specifically designed for bone scintigraphy segmentation. This demonstrates the robustness and effectiveness of our approach in accurately identifying and delineating BM lesions.

Discussion

This section provides a brief discussion of an ablation study, analyzing the effects of the network structure of the model on lesion segmentation performance.

SPECT lesion imaging presents several challenges due to its low resolution, leading to blurry contours, and a significant disparity between large background areas and small foreground regions. The lesions themselves vary greatly in size, location, and shape, and the radiotracer uptake varies across different skeletal sites, causing variability in radiometric values. These factors result in issues such as extensive background areas, small foreground regions, varied lesion shapes and sizes, unclear contours, and inconspicuous hotspots in BMs.

To address these challenges, our proposed model follows an encoder-decoder architecture to learn hierarchical features from the scintigrams. This structure is particularly effective for handling the significant size disparity between the large background and small foreground regions. The encoder progressively reduces the spatial dimensions of feature maps while increasing their depth, capturing essential information and high-level semantics. The decoder then reconstructs these down-sampled images through up-sampling, gradually restoring the feature maps to the original input image's spatial dimensions and generating precise segmentation boundaries.

To enhance the model's ability to recognize lesions of varying sizes, locations, shapes, and radiometric values, we incorporate a multi-scale feature learning strategy and a multi-pooling learning strategy. The multi-scale learning blocks in each layer use convolutional kernels of different sizes to capture both local details and global information. The multi-pooling learning strategy processes features

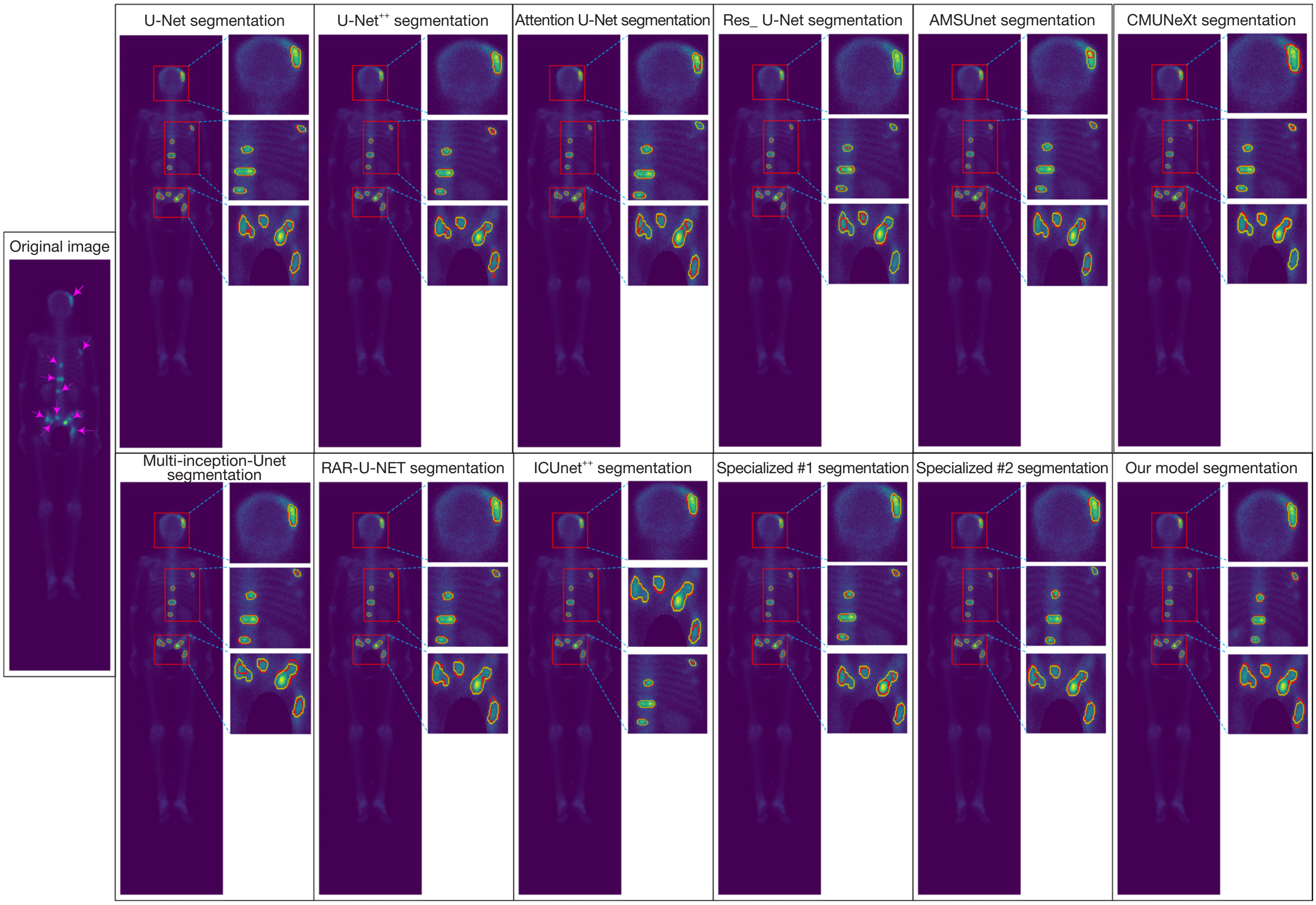


Figure 8 An illustration of segmentation results by existing models and our model with yellow and red curves denoting ground truth and model prediction respectively.

through multiple pooling layers of different scales during the first down-sampling, which is crucial for retaining critical information and capturing both global and local features of lesions.

At the model's bottom layer, we employ a multi-attention mechanism, including non-local attention and ViT. Non-local attention establishes relationships between any two positions in the image to capture global features, enhancing boundary clarity and making inconspicuous lesion areas more prominent. ViT treats the image as a series of patches, using self-attention to process relationships between these patches and capture global contextual information.

These strategies collectively improve the model's performance by ensuring it captures comprehensive and detailed features necessary for accurate lesion segmentation. The combination of multi-scale learning, multi-pooling, and multi-attention mechanisms allows the model to effectively handle the complex and variable nature of BM lesions in SPECT images.

The main contribution of this paper can be summarized as follows:

- ❖ Comprehensive analysis: we conduct an in-depth analysis of SPECT bone scintigraphy to gain deeper insights into the characteristics of BM. This analysis enables us to establish a mapping from size-varied, location-random lesions to the network configuration of our proposed model.
- ❖ Proposed segmentation model: we propose a CNN-based segmentation model that integrates multi-attention and multi-scale feature learning schemes. This integration significantly enhances the performance of automated segmentation.

- ❖ Experimental validation: we conduct extensive experiments, including ablation studies on clinical data of SPECT bone scintigraphy. These experiments demonstrate the usefulness and effectiveness of the proposed model.

As illustrated in *Figure 1*, we use multi-attention and multi-scale learning to improve the ability of the model to learn features of BM lesions. To examine how these aspects affect the segmentation performance, we provide experimental results of ablation studies in *Table* 7.

We can observe from the results in *Table* 7 that the best performance is achieved when both multi-attention and multi-scale learning schemes are used simultaneously.

In *Figure 9*, we present two cases of SPECT BM scintigrams to illustrate the impact of our strategy on extracting image features through multi-attention and multi-scale learning. Feature maps from different layers are displayed for the following cases:

- ❖ Case #1: patient with metastasis in the temporoparietal bones, thoracic, lumbar, and sacral vertebrae, ribs, bilateral iliac bones, and acetabulum.
- ❖ Case #2: patient with metastasis in the cervical spine, lumbar spine, proximal radial ulnar joint, ribs, and iliac acetabulum.

Two observations can be made from the visual presentations of the extracted feature maps in *Figure 9*. First, the skeleton outline that indicates the foreground of an image has been greatly enhanced against the background. This enables the segmentation model to exclusively focus on learning more useful information from the images. Second, the regions of interest that represent the hotspots have been remarkably highlighted against the skeleton. This facilitates the segmentation model by paying more attention to the lesion areas in the images. These two points demonstrate the usefulness and validity of the multi-attention and multi-scale learning used in the proposed segmentation model.

To demonstrate the clinical value of our model, we present its application in lesion detection. The model can accurately identify and delineate BM lesions, providing clinicians with a powerful tool to enhance diagnostic accuracy and efficiency. This is particularly significant for the early detection and treatment of BM lesions.

As depicted in *Figure 2*, the multi-scale learning module uses several kernels with various sizes to help the model focus on size-varied lesions. The experimental results reported in *Table 8* show the reasonableness of using 1×1, 3×3, 5×5 kernels.

The reason we chose three types of kernel sizes is derived from a statistical analysis of the sizes of actual BM lesions, as depicted in *Figure 10*.

The statistics in *Figure 10* reveal that most BM lesions range in size from 6 to 16 cm. This wide range of lesion sizes necessitates the use of kernel sizes. Therefore, employing 1×1, 3×3 and 5×5, kernels achieves better performance than the other configurations as shown in *Table 8*.

We conducted comparative experiments to assess the impact of multi-scale learning on our model's performance. *Figure 11* presents the comparative graphs, clearly illustrating the benefits of incorporating a multiscale learning strategy.

The results demonstrate that lesions of varying shapes and sizes are significantly better detected when multi-scale learning is applied. This improvement underscores the importance of multi-scale learning in achieving more accurate and robust segmentation outcomes. Specifically, before applying multi-scale training, the model performed poorly in detecting lesions on the spine and ribs. However, after incorporating the multi-scale strategy, the detection of lesions in these areas improved significantly. Additionally, before incorporating the multi-scale strategy, the model could effectively identify lesions in high-intensity regions but struggled to detect lesions that had minimal color contrast with adjacent areas or the overall region. Our multi-scale strategy effectively addressed this issue, enabling the model to more accurately identify lesions of varying intensity and location.

Two cases depicted in *Figure 12* further show that the proposed segmentation model using a multi-scale learning strategy can accurately identify and delineate the size-varied lesions that are often located randomly.

As shown in *Figure 13*, we present a case illustrating how our model assists nuclear medicine doctors in diagnosis. When a patient undergoes imaging, our model can aid in diagnosis by providing accurate segmentation results, allowing doctors to make quick assessments based on the model's output. This is particularly beneficial for detecting lesions of varying locations, sizes, and intensities.

In scenarios where the original model's performance is suboptimal, an advanced or experienced doctor is typically required to reassess or recommend additional imaging to achieve a conclusive diagnosis, which increases costs. However, our model addresses these issues effectively.

Firstly, the enhanced accuracy of our model aids physicians in the diagnostic process by reducing the need for specialist consultations and repeated procedures, thereby lowering overall diagnostic costs and increasing efficiency. Additionally, it provides robust support in cases of unclear

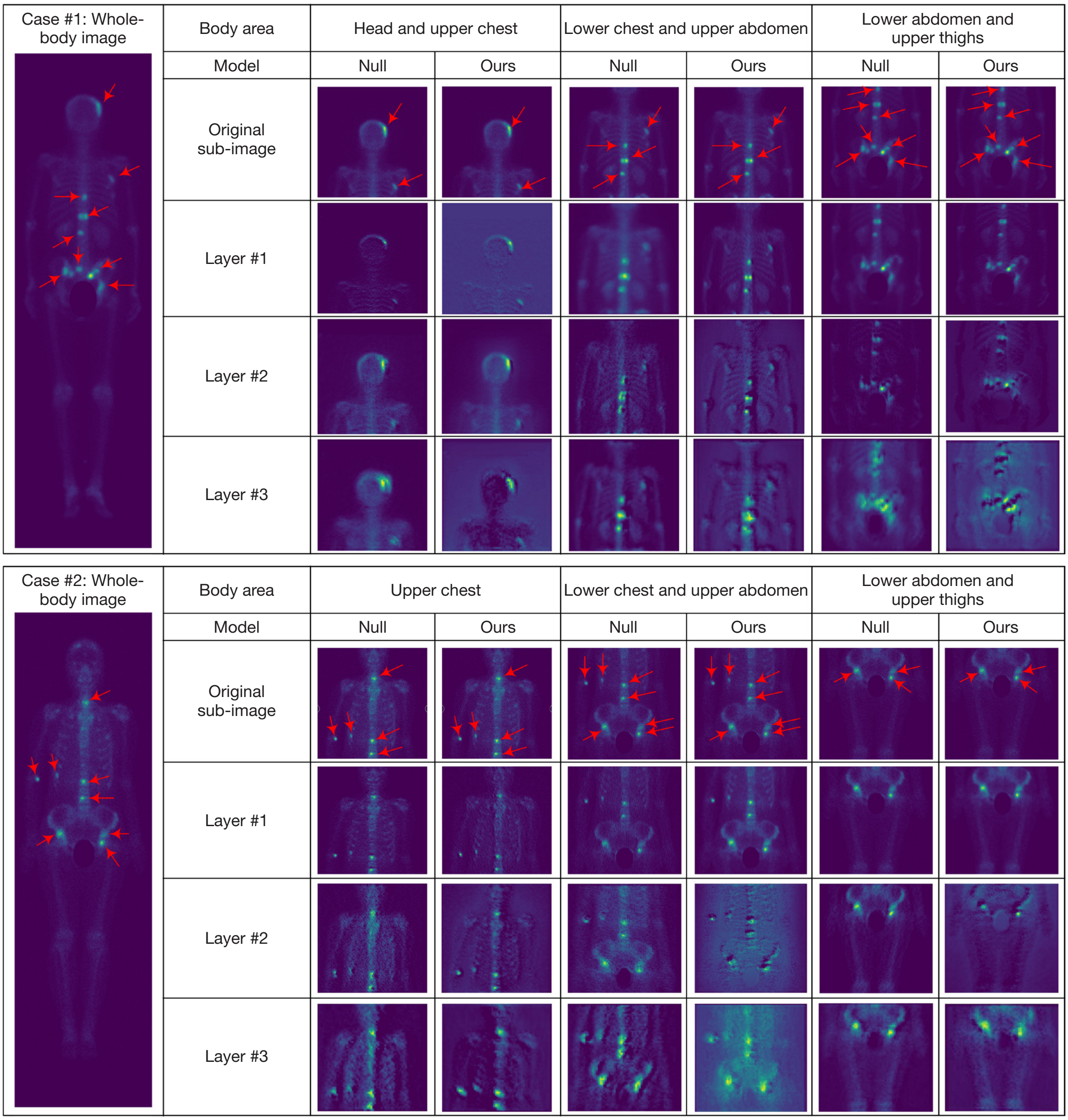


Figure 9 Effect of multi-attention and multi-scale learning on extracting features from SPECT bone scintigrams, displaying feature maps from different layers using samples from two lung cancer patients with BM. Case #1: patient with metastasis in the temporoparietal bones, thoracic, lumbar, and sacral vertebrae, ribs, bilateral iliac bones, and acetabulum. Case #2: patient with metastasis in the cervical spine, lumbar spine, proximal radial ulnar joint, ribs, and iliac acetabulum. SPECT, single photon emission computed tomography; BM, bone metastasis.

lesion positions, which improves diagnostic accuracy and efficiency, and enhances patient management strategies.

Moreover, the ability to effectively segment subtle lesions with low visibility during initial examinations is another significant benefit. This improved segmentation accuracy facilitates earlier intervention, timely treatment adjustments, and personalized care plans, all of which contribute to better patient outcomes and quality of life.

Furthermore, the integration of our model into clinical workflows streamlines patient care management. It enables more effective prioritization of treatment plans, efficient resource allocation, and reduces the burden on healthcare systems. This results in better coordination among medical teams, more informed decision-making, and a higher standard of patient care.

Table 8 The experimental results of the multi-scale learning module with various sizes of kernels

Kernel size of multi-scale learning module	DSC	Recall	Precision	Time
1×1, 3×3	0.6461	0.6466	0.6472	46 min
1×1, 3×3, 5×5	0.6720	0.6671	0.6771	1 h + 3 min
1×1, 3×3, 5×5, 7×7	0.6380	0.6216	0.6555	52 min

DSC, dice similarity coefficient.

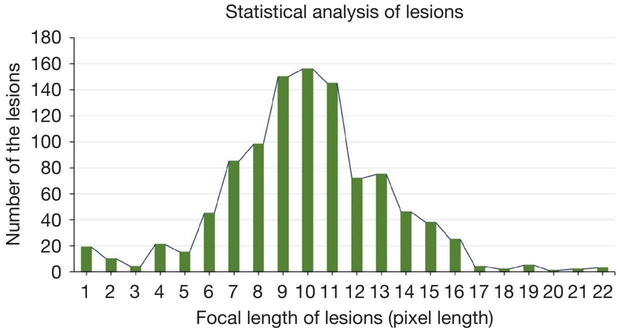


Figure 10 Statistical analysis of the sizes of the true BM lesions in the SPECT bone scintigrams with a pixel distance of 2.26 mm. BM, bone metastasis; SPECT, single photon emission computed tomography.

Overall, these improvements underscore the clinical relevance of our work, offering substantial benefits in terms of cost efficiency, diagnostic accuracy, and patient care management.

In "Multi-pooling learning strategy", we mentioned that the proposed segmentation model employs four types of poolings with the sizes of 1×1, 2×2, 4×4 and 8×8. The rationale behind this choice is elucidated, where we showcase the experimental performance that led to this decision, as demonstrated in *Figure 14*.

The use of pooling sizes 1×1, 2×2, 4×4 and 8×8 resulted in the best performance in terms of the DSC metric, achieving increased scores of 3.10% and 1.55% compared to the other two configurations.

The original dataset was supplemented with images of patients without BM for the experiment, and the final dataset used is shown in *Table 9*.

The experimental results by comparing the above data with the original data on the U-Net model are shown as follows.

As detailed in *Table 10*, our experiments involve manipulating non-bone transfer parts. This entailed isolating the BM regions and introducing non-BM areas. These experiments revealed that including non-BM images had a negative impact on overall segmentation performance and doubled the computation time. Our primary goal is to precisely segment lesions from diseased images to aid medical

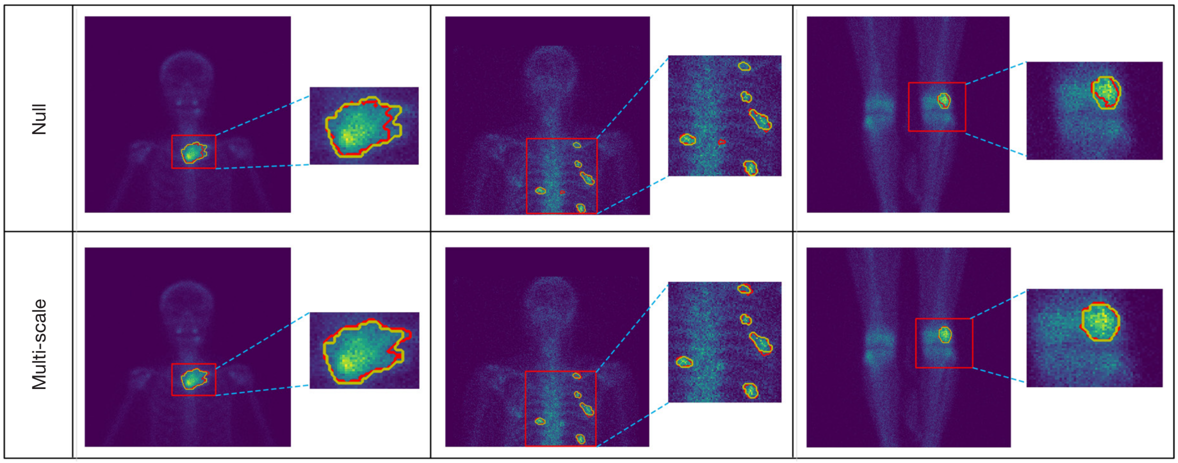


Figure 11 An illustration of segmented metastasis lesions with different sizes and at different locations when the model uses multi-scale learning (multi-scale) or does not use multi-scale learning (Null), where the yellow and red curves denote ground truth and model predictions respectively.

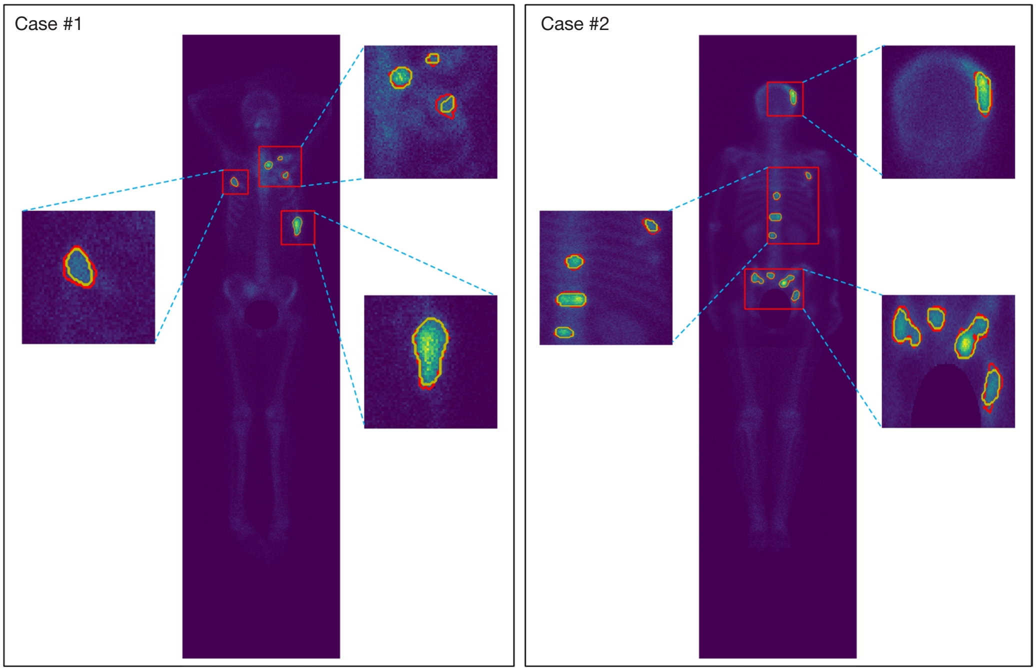


Figure 12 Visual comparison of the segmented lesions (red curves) by the models against the ground truth (yellow curves) manually labelled by nuclear medicine physicians.

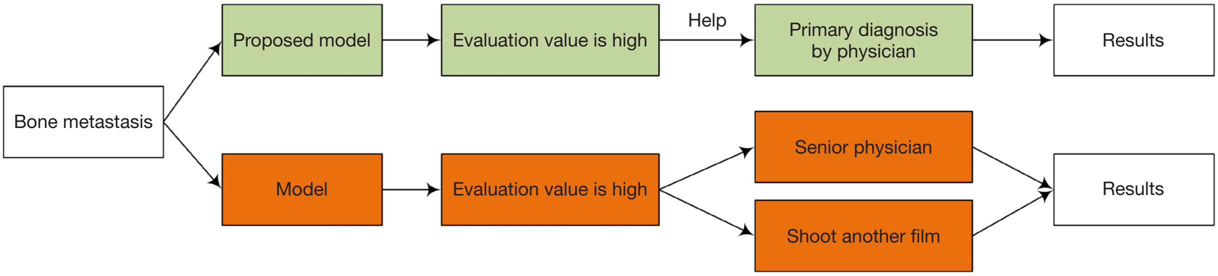
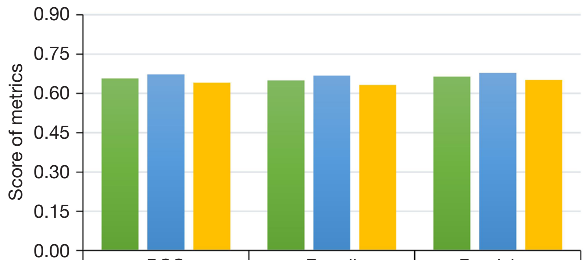


Figure 13 Workflow diagram of diagnosis in medical practice by using the proposed model.



	DSC	Recall	Precision
🟩 [1×1] + [2×2] + [4×4] + [7×7]	0.6565	0.6495	0.6640
🟦 [1×1] + [2×2] + [4×4] + [8×8]	0.6720	0.6671	0.6771
🟨 [1×1] + [3×3] + [4×4] + [8×8]	0.6410	0.6319	0.6510

Figure 14 Comparative analysis of segmentation performance when the model uses mutli-poolings with different sizes. DSC, dice similarity coefficient.

Table 9 The statistics of the data of BM images and non-BM images

	BM images	Non-BM images	All images
Number of images	838	1,115	1,953
Image size	256×256	256×256	256×256
Training set (~70%)	590	803	1,393
Test set (~30%)	248	312	560

BM, bone metastasis; non-BM, non-bone metastasis.

Table 10 Comparison of trial statistics with inclusion of non-BM images

	BM images only	BM + non-BM images
Model	Ours	Ours
DSC	0.6720	0.6236
Recall	0.6671	0.6220
Precision	0.6771	0.6256
Time	1 h + 3 min	2 h + 23 min

non-BM, non-bone metastasis; DSC, dice similarity coefficient; BM, bone metastasis.

professionals. Due to the potentially adverse effects of non-disease images on results and the significant increase in training time, we opted to focus solely on BM lesion images for segmentation.

In our experiments, we included non-disease images and observed that these images indeed interfere with the overall segmentation performance and double the computation time. However, our primary objective is to accurately segment lesions from diseased images to aid clinicians. Given the potential adverse impact of non-disease images on the results and the doubled training time, we conducted the segmentation using only BM lesion images.

Conclusions

In this section, we summarize the outcomes of our experiments, and address both the limitations and potential future directions for this study. The following discussion is structured around two key aspects: (I) interpretation of experiments, and (II) limitations and future directions.

Interpretation of experiments

This study has explored the automated segmentation of BM lesions in SPECT bone scintigrams through the development of a CNN-based segmentation model. We introduced our proposed model, detailing its network structure and key components. Through experimental evaluations using clinical data, we compared our model with classical CNN-based models and achieved an unprecedented DSC score of 0.6720, demonstrating its superior performance. Furthermore, we conducted a thorough ablation study to dissect the strengths and weaknesses of our proposed model.

Limitations and future directions

While our model has demonstrated competitive performance, some inherent challenges remain. One primary limitation is the processing time required to handle large original images (1,024×256). Although dividing these images into smaller segments can reduce the processing time per segment, the increased number of segments still results in a prolonged overall processing time. Additionally, our model's relatively high complexity contributes to a longer runtime compared to other models. However, this complexity enables our model to achieve superior segmentation performance: compared to bone scan-specific models (32), our model outperforms them in DSC, Precision, and Recall by 5.7%, 5.3%, and 2.9%, respectively.

For future work, we plan to optimize the segmentation strategy and explore more efficient model architectures to further reduce the overall training and inference time. We will also investigate lower-parameter and more computationally efficient versions of the model to maintain segmentation performance while reducing complexity. Additionally, we aim to incorporate human domain knowledge to refine feature extraction from imaging data further. These efforts will advance automated segmentation technology in medical imaging, ultimately providing more efficient and accurate support for clinical diagnosis and treatment planning.

Acknowledgments

Funding: This work was supported in part by the National Natural Science Foundation of China (Nos. 62362058 and 61562075), the Key R&D Plan of Gansu Province (Nos. 24YFGA048 and 21YF5GA063), the Gansu Province's Key Provincial Talents Program, the Young and Middle-aged Talents Training Program of State Ethnic Affairs Commission, the Natural Science Foundation of Gansu Province (Nos. 20JR5RA511 and 22JR11RA236), the Fundamental Research Fund for the Central Universities (Nos. 31920230172, 31920240092 and 31920240099), and the Youth Ph.D. Foundation of Education Department of Gansu Province (No. 2021QB-063).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1246/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was approved by the Ethics Committee of Gansu Provincial Cancer Hospital (No. A202106100014). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). A requirement for informed consent was waived for this study because of the anonymous nature of the data.

References

1. Wang M, Xia F, Wei Y, Wei X. Molecular mechanisms and clinical management of cancer bone metastasis. Bone Res 2020;8:30.
2. Ashford RU, Benjamin L, Pendlebury S, Stalley PD. The modern surgical and non-surgical management of appendicular skeletal metastases. Orthop Trauma 2012;26:184-99.
3. Manders K, van de Poll-Franse LV, Creemers GJ, Vreugdenhil G, van der Sangen MJ, Nieuwenhuijzen GA, Roumen RM, Voogd AC. Clinical management of women with metastatic breast cancer: a descriptive study according to age group. BMC Cancer 2006;6:179.
4. Yazdani A, Dorri S, Atashi A, Shirafkan H, Zabolinezhad H. Bone Metastasis Prognostic Factors in Breast Cancer. Breast Cancer (Auckl) 2019;13:1178223419830978.
5. Łukaszewski B, Nazar J, Goch M, Łukaszewska M, Stępiński A, Jurczyk MU. Diagnostic methods for detection of bone metastases. Contemp Oncol (Pozn) 2017;21:98-103.
6. Santini D, Galluzzo S, Zoccoli A, Pantano F, Fratto ME, Vincenzi B, Lombardi L, Gucciardino C, Silvestris N, Riva E, Rizzo S, Russo A, Maiello E, Colucci G, Tonini G. New molecular targets in bone metastases. Cancer Treat Rev 2010;36 Suppl 3:S6-S10.
7. Miler SF, Thomas CY, Shiozawa Y. Molecular involvement of the bone marrow microenvironment in bone metastasis. In: Ahmad A, editor. Introduction to Cancer Metastasis. Academic Press; 2017. p. 263-276.
8. Abhisheka B, Biswas SK, Purkayastha B, Das D, Escargueil A. Recent trend in medical imaging modalities and their applications in disease diagnosis: a review. Multimed Tools Appl 2024;83:43035-70.
9. Söderlund V. Radiological diagnosis of skeletal metastases. Eur Radiol 1996;6:587-95.
10. Costelloe CM, Rohren EM, Madewell JE, Hamaoka T, Theriault RL, Yu TK, Lewis VO, Ma J, Stafford RJ, Tari AM, Hortobagyi GN, Ueno NT. Imaging bone metastases in breast cancer: techniques and recommendations for diagnosis. Lancet Oncol 2009;10:606-14.
11. Capanna R, Campanacci DA. The treatment of metastases in the appendicular skeleton. J Bone Joint Surg Br 2001;83:471-81.
12. Hata H, Kitao T, Sato J, Asaka T, Ohga N, Imamachi K, Hirata K, Shiga T, Yamazaki Y, Kitagawa Y. Monitoring indices of bone inflammatory activity of the jaw using SPECT bone scintigraphy: a study of ARONJ patients. Sci Rep 2020;10:11385.
13. Zhang Y, Lin Z, Li T, Wei Y, Yu M, Ye L, Cai Y, Yang S, Zhang Y, Shi Y, Chen W. Head-to-head comparison of (99m)Tc-PSMA and (99m)Tc-MDP SPECT/CT in diagnosing prostate cancer bone metastasis: a prospective, comparative imaging trial. Sci Rep 2022;12:15993.
14. Kuwert T. Skeletal SPECT/CT: a review. Clin Transl Imaging 2014;2:505-517.
15. Britton KE. Nuclear medicine imaging in bone metastases. Cancer Imaging 2002;2:84-6.
16. Afnouch M, Gaddour O, Hentati Y, Bougourzi F, Abid M, Alouani I, Taleb Ahmed A. BM-Seg: A new bone metastases segmentation dataset and ensemble of CNN-based segmentation approach. Expert Syst Appl 2023;228:120376.
17. Wang S, Li C, Wang R, Liu Z, Wang M, Tan H, Wu Y, Liu X, Sun H, Yang R, Liu X, Chen J, Zhou H, Ben Ayed I, Zheng H. Annotation-efficient deep learning for automatic

medical image segmentation. Nat Commun 2021;12:5915.

18. Rayed ME, Islam SMS, Niha SI, Jim JR, Kabir MM, Mridha MF. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics Med Unlocked 2024;47:101504.
19. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
20. Dhillon A, Verma GK. Convolutional neural network: a review of models, methodologies and applications to object detection. Prog Artif Intell 2020;9:85-112.
21. Kao YS, Huang CP, Tsai WW, Yang J. A systematic review for using deep learning in bone scan classification. Clin Transl Imaging 2023;11:271-283.
22. Wang Y, Lin Q, Zhao S, Zeng X, Zheng B, Cao Y, Man Z. Automated Diagnosis of Bone Metastasis by Classifying Bone Scintigrams Using a Self-defined Deep Learning Model. Curr Med Imaging 2024. [Epub ahead of print]. doi: 10.2174/0115734056281578231212104108.
23. Guo Y, Lin Q, Wang Y, Cao X, Cao Y, Man Z, Zeng X, Huang X. Integrating Transfer Learning and Feature Aggregation into Self-defined Convolutional Neural Network for Automated Detection of Lung Cancer Bone Metastasis. J Med Biol Eng 2023;43:53-62.
24. Lin Q, Man Z, Cao Y, Wang H. Automated Classification of Whole-Body SPECT Bone Scan Images with VGG-Based Deep Networks. Int Arab J Inf Technol 2023;20:1-8.
25. Paranavithana IR, Stirling D, Ros M, Field M. Systematic Review of Tumor Segmentation Strategies for Bone Metastases. Cancers (Basel) 2023;15:1750.
26. Zhang J, Huang M, Deng T, Cao Y, Lin Q. Bone metastasis segmentation based on Improved U-NET algorithm. J Phys Conf Ser 2021;1848:012027.
27. Nemoto T, Futakami N, Yagi M, Kunieda E, Akiba T, Takeda A, Shigematsu N. Simple low-cost approaches to semantic segmentation in radiation therapy planning for prostate cancer using deep learning with non-contrast planning CT images. Phys Med 2020;78:93-100.
28. Shimizu A, Wakabayashi H, Kanamori T, Saito A, Nishikawa K, Daisaki H, Higashiyama S, Kawabe J. Automated measurement of bone scan index from a whole-body bone scintigram. Int J Comput Assist Radiol Surg 2020;15:389-400.
29. Sekuboyina A, Rempfler M, Valentinitsch A, Menze BH, Kirschke JS. Labeling Vertebrae with Two-dimensional Reformations of Multidetector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy. Radiol Artif Intell 2020;2:e190074.
30. Saito A, Wakabayashi H, Daisaki H, Yoshida A, Higashiyama S, Kawabe J, Shimizu A. Extraction of metastasis hotspots in a whole-body bone scintigram based on bilateral asymmetry. Int J Comput Assist Radiol Surg 2021;16:2251-60.
31. Lin Q, Luo M, Gao R, Li T, Man Z, Cao Y, Wang H. Deep learning based automatic segmentation of metastasis hotspots in thorax bone SPECT images. PLoS One 2020;15:e0243253.
32. Cao Y, Liu L, Chen X, Man Z, Lin Q, Zeng X, Huang X. Segmentation of lung cancer-caused metastatic lesions in bone scan images using self-defined model with deep supervision. Biomed Signal Process Control 2023;79:104068.
33. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; 2015 Oct 5-9; Munich, Germany. Cham: Springer; 2015. p. 234-241.
34. Apiparakoon T, Rakratchatakul N, Chantadisai M, Vutrapongwatana U, Kingpetch K, Sirisalipoch S, Rakvongthai Y, Chaiwatanarat T, Chuangsuwanich E. MaligNet: semisupervised learning for bone lesion instance segmentation using bone scintigraphy. IEEE Access 2020;8:27047-66.
35. Lin Q, Gao R, Luo M, Wang H, Cao Y, Man Z, Wang R. Semi-supervised segmentation of metastasis lesions in bone scan images. Front Mol Biosci 2022;9:956720.
36. Huang K, Huang S, Chen G, Li X, Li S, Liang Y, Gao Y. An end-to-end multi-task system of automatic lesion detection and anatomical localization in whole-body bone scintigraphy by deep learning. Bioinformatics 2023;39:btac753.
37. Yu PN, Lai YC, Chen YY, Cheng DC. Skeleton Segmentation on Bone Scintigraphy for BSI Computation. Diagnostics (Basel) 2023.
38. Chen C, Qi S, Zhou K, Lu T, Ning H, Xiao R. Pairwise attention-enhanced adversarial model for automatic bone segmentation in CT images. Phys Med Biol 2023.
39. Morita D, Mazen S, Tsujiko S, Otake Y, Sato Y, Numajiri T. Deep-learning-based automatic facial bone segmentation using a two-dimensional U-Net. Int J Oral Maxillofac Surg 2023;52:787-92.
40. Li X, Peng Y, Xu M. Patch-shuffle-based semi-supervised segmentation of bone computed tomography via consistent learning. Biomed Signal Process Control 2023;80:104239.

41. Zhan X, Liu J, Long H, Zhu J, Tang H, Gou F, Wu J. An Intelligent Auxiliary Framework for Bone Malignant Tumor Lesion Segmentation in Medical Image Analysis. Diagnostics (Basel) 2023.
42. Zhou P, He G, Chen Z, Zhao L. A two-stage whole-body bone SPECT scan image inpainting algorithm for residual urine artifacts based on contextual attention. In: Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV 2023; 2023 Oct 13–15; Xiamen, China. Proceedings, Part XIII. Berlin, Heidelberg: Springer-Verlag; 2023. p. 497-508.
43. Liu S, Feng M, Qiao T, Cai H, Xu K, Yu X, Jiang W, Lv Z, Wang Y, Li D. Deep Learning for the Automatic Diagnosis and Analysis of Bone Metastasis on Bone Scintigrams. Cancer Manag Res 2022;14:51-65.
44. Li L, Qin J, Lv L, Cheng M, Wang B, Xia D, Wang S. ICUnet++: an Inception-CBAM network based on Unet++ for MR spine image segmentation. Int J Mach Learn Cybern 2023. [Epub ahead of print]. doi: 10.1007/s13042-023-01857-y.
45. Wang Z, Zhang Z, Voiculescu I. RAR-U-NET: a residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels. In: Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP); 2021 Sep 19-22; Anchorage, AK, USA. IEEE; 2021. p. 21-25.
46. Latif U, Shahid AR, Raza B, Ziauddin S, Khan MA. An end-to-end brain tumor segmentation system using multi-inception-UNET. Int J Imaging Syst Technol 2021;31:1803-16.
47. Qiu P, Yang J, Kumar S, Ghosh SS, Sotiras A. AgileFormer: spatially agile transformer UNet for medical image segmentation. arXiv 2024;2404.00122.
48. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7-12; Boston, MA, USA. IEEE; 2015. p. 1-9.
49. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI, USA. IEEE; 2017. p. 6230-6239.
50. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. An image is worth 16x16 words: transformers for image recognition at scale. arXiv 2020;2010.11929.
51. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. IEEE; 2018. p. 7794-7803.
52. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018) 2018;11045:3-11.
53. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B. Attention U-Net: learning where to look for the pancreas. arXiv 2018;1804.03999.
54. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J Photogramm Remote Sens 2020;162:94-114.
55. Yin Y, Han Z, Jian M, Wang GG, Chen L, Wang R. AMSUnet: A neural network using atrous multi-scale convolution for medical image segmentation. Comput Biol Med 2023;162:107120.
56. Tang F, Ding J, Quan Q, Wang L, Ning C, Zhou SK. Cmunext: an efficient medical image segmentation network based on large kernel and skip fusion. In: Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI); 2024 May 13-16; New York, NY, USA. IEEE; 2024. p. 1-5.

Cite this article as: Xie A, Lin Q, He Y, Zeng X, Cao Y, Man Z, Liu C, Hao Y, Huang X. Metastasis lesion segmentation from bone scintigrams using encoder-decoder architecture model with multi-attention and multi-scale learning. Quant Imaging Med Surg 2025;15(1):689-708. doi: 10.21037/qims-24-1246