IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 2, April, 2024, pp. 771-780
Journal homepage: http://iieta.org/journals/ts

MFP-DeepLabv3+: A Multi-scale Feature Fusion and Parallel Attention Network for Enhanced Bone Metastasis Segmentation

Hongwen Gu[1], Pengju Wang[2], Yu Li[1], Nan Bao[2], Hongwei Wang[1], Yanchun Xie[1], Anwu Xuan[1], Yuanhang Zhao[1], Hailong Yu[1*], He Ma[2]

[1] Department of Orthopedics, General Hospital of Northern Theater Command, Shenyang 110016, China
[2] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110179, China

Corresponding Author Email: yuhailong118@aliyun.com

Copyright: ©2024 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.410218

Received: 12 December 2023
Revised: 26 February 2024
Accepted: 10 March 2024
Available online: 30 April 2024

Keywords:
bone metastasis segmentation, DeepLabv3+, AFFP, PSCAN, multi-layer skip connection

ABSTRACT

Bone metastasis segmentation is a crucial task in the field of medical image processing, aimed at automatically and accurately identifying regions of bone metastatic lesions in medical imagery. In recent years, with the rapid development of deep learning technology, various deep learning models have been widely applied to the task of bone metastasis segmentation. This paper proposes a multi-scale feature fusion and parallel attention network based on DeepLabv3+ called MFP-DeepLabv3+, with the following main contributions: (1) Introducing adaptive feature fusion and pooling (AFPP) to enhance the multi-scale feature extraction capability of the network; (2) Introducing parallel spatial-channel attention network (PSCAN) enhances the simultaneous attention of the network to both channel and spatial information; (3) Introducing a multi-layer skip connection strategy to better integrate global semantic information. Experimental results on the BM-Seg dataset demonstrate that MFP-DeepLabv3+ achieves mIoU, mPA, mPrecision, and Dice scores of 83.97%, 93.90%, 87.97%, and 90.50%, respectively, outperforming various mainstream semantic segmentation networks. This study effectively improves the accuracy and efficiency of bone metastasis segmentation, offering valuable auxiliary tools for clinical diagnosis.

1. INTRODUCTION

Bone metastasis is a complex biological process, whereby malignant tumor cells migrate from the primary tumor site through the bloodstream or lymphatic system to the bone tissue, forming secondary tumors. This process involves multiple steps including invasion, shedding, entry into the bloodstream, colonization, and growth of tumor cells [1]. According to statistics, approximately 70% of cancer patients experience bone metastasis, with breast, prostate, lung, and renal cancers being the most common. Bone metastasis not only severely damages skeletal structure and function but also leads to clinical symptoms such as bone pain, fractures, and hypercalcemia, significantly impacting patients' quality of life and survival rates [2].

Currently, the predominant diagnostic methods for bone metastasis in clinical practice primarily rely on medical imaging examinations, such as X-ray radiography, CT scanning, magnetic resonance imaging (MRI), as well as serum biomarker detection [3]. However, traditional medical imaging analysis methods are constrained by physicians' subjective experience and workload, leading to issues of diagnostic accuracy and consistency. Consequently, automated bone metastasis pathology segmentation technology has become a focal point of research, aiming to utilize computer vision and deep learning technologies to achieve automatic identification and segmentation of regions of bone metastatic lesions in medical imagery.

Deep learning models have been widely used in medical segmentation tasks [4], which can be trained on vast amounts of medical imaging data to automatically extract features from images and achieve accurate identification and segmentation of regions of bone metastatic lesions. Compared to traditional manual segmentation methods, automated segmentation techniques based on deep learning offer advantages such as rapid recognition speed, high accuracy, and strong reproducibility [5]. These techniques can alleviate the workload of physicians, improve medical efficiency and diagnostic accuracy, and provide an important auxiliary diagnostic tool for clinical medicine [6].

In recent years, significant progress has been made in the field of bone metastasis through deep learning-based medical image analysis techniques. Song et al. [7] optimized the holistically-nested edge detection (HED) network to enhance the recognition capability of tiny bone metastatic regions in CT images. By removing the terminal pooling layers and introducing additional lateral connection layers, more precise edge detection was achieved, thereby improving the perception and capture of small targets. Ntakolia et al. [8] proposed a lightweight network called LB-FCN light, focusing on the classification of bone metastases in prostate cancer patients. This network, through multi-scale feature extraction

and residual connection techniques, effectively classified bone metastases, emphasizing its lightweight nature in terms of parameters and computational resources, making it suitable for resource-constrained scenarios. Lin et al. [9] proposed a semi-supervised segmentation method based on deep learning, capable of automatically detecting and delineating metastatic lesions in bone scan images. This method utilizes a small amount of manually labeled samples for training, significantly reducing the human resources required for annotation, and providing an effective solution for medical image analysis tasks with high demands for annotated data. Noguchi et al. [10] proposed a bone segmentation network, candidate region segmentation network, and false-positive reduction network using deep convolutional neural networks such as U-Net and ResNet, aiming to achieve automatic segmentation of bone metastatic tumors in CT images. Liu et al. [11] developed an improved UNet3+ network model for the automatic segmentation of bone metastasis lesions on SPECT bone scan images. The model enhanced the feature fusion by modifying the full-scale deep supervision module and introduced an attention mechanism to focus on focal regions.

Although the aforementioned methods have made significant strides in detecting and segmenting bone metastases, they still exhibit several limitations:

1. The network architecture exhibits a notably intricate structure, leading to substantial time consumption during both training and inference phases, alongside heightened computational demands. Such complexities impose limitations on the applicability of these networks, particularly in resource-constrained medical and clinical environments.

2. During feature extraction, there is a potential oversight regarding the interaction between feature channels, resulting in inadequate extraction of channel information.

3. Insufficient extraction and integration of deep features across various hierarchical levels culminate in the loss of crucial semantic information.

The Deeplabv3+ network [12] architecture is notably lightweight, leveraging an encoder-decoder framework, wherein the encoder network extracts deep features from input images. The atrous spatial pyramid pooling (ASPP) module utilizes dilated convolutions with varying rates to capture multi-scale contextual information, thereby enhancing the semantic representation capability of features. Subsequently, the decoder network is employed to restore the resolution of feature maps, enabling precise pixel-level segmentation. To overcome the aforementioned limitations and take advantage of DeepLabv3+, this paper proposes a multi-scale feature fusion and parallel attention network based on DeepLabv3+ (MFP-DeepLabv3+). The main contributions are as follows:

1. To address issues in the atrous spatial pyramid pooling (ASPP) module of DeepLabv3+, such as overlapping information extraction and detail loss, the adaptive feature fusion and pooling (AFFP) module is proposed to achieve multi-scale feature extraction more efficiently, thereby enhancing model performance.

2. To comprehensively extract channel information, the parallel spatial-channel attention network (PSCAN) is proposed to empower the network in intensifying its focus on both channel and spatial information simultaneously during image feature extraction.

3. To meet practical demands, this paper selected the lightweight MobileNetv2 [13] as the backbone network. Considering that different network layers convey distinct depths of information, a multi-layer skip connection strategy is proposed, the incorporation of multi-layer skip connections effectively integrates global semantic information, thereby enhancing the network's capability to tackle diverse image segmentation tasks.

2. METHOD

In this paper, we propose a multi-scale feature fusion and parallel attention network (MFP-DeepLabv3+) for enhanced bone metastasis segmentation. As shown in Figure 1.

MFP-DeepLabv3+ utilizes the lightweight MobileNetv2 as its backbone network, incorporates AFFP for multiscale feature extraction, and introduces PSCAN for weighting the features obtained from AFFP. These weighted features are subsequently fused with the features from deep, intermediate, and shallow layers of the MobileNetv2 backbone network to enhance the performance of image segmentation.

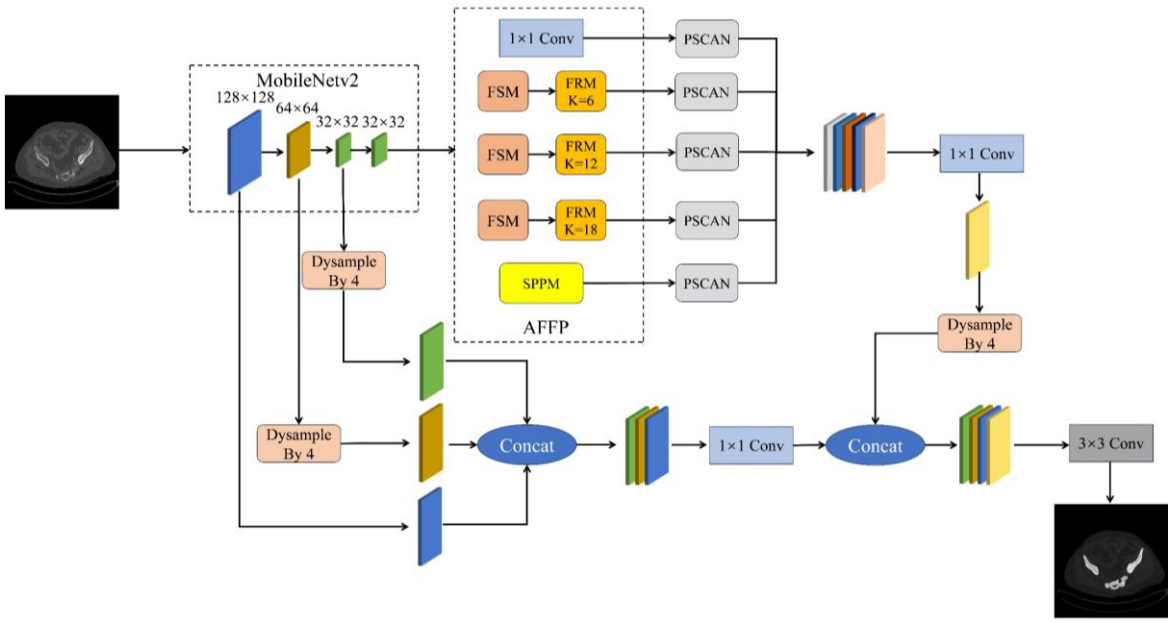

Figure 1. The overall framework of MFP-Deeplabv3+

2.1 AFFP

The ASPP module primarily relies on a series of dilated convolutional layers [14] with varying sampling rates for multiscale feature extraction. However, this approach can result in the extraction of overlapping information, leading to the generation of redundant features. These redundancies not only augment the computational burden of the model but also escalate the time and resource costs associated with both training and inference, thereby compromising the model's generalization capability. Moreover, ASPP employs global average pooling to aggregate information across the entire feature map, which is susceptible to losing detailed information pertaining to edges and local regions, consequently impeding the effectiveness of global feature integration.

To address the aforementioned problems, this paper proposes the AFFP. AFFP comprises three components: the feature selection module (FSM), the feature reconstruction module (FRM), and the spatial pyramid pooling with max-pooling (SPPM). Initially, FSM adopts a cross-reconstruction approach to manage features of varying information densities, obtaining spatially reconstructed features. This technique aims to preserve crucial feature information while mitigating redundancy. Subsequently, the spatially reconstructed features are fed into FRM, facilitating efficient multiscale feature extraction. Additionally, SPPM is utilized to effectively retain intricate image details, thereby enhancing model performance.

2.1.1 FSM

FSM employs a cross-reconstruction approach for feature reconstruction, thereby obtaining spatially reconstructed features, as shown in Figure 2. Initially, it utilizes group normalization to assess the information content of different feature maps using scaling factors, thereby quantifying and evaluating the importance of each feature map. Subsequently, the obtained information content weights undergo normalization to obtain W_C, reflecting the significance of various feature mappings. Following this, the feature maps are reweighted using W_C, and the weights are mapped to the (0,1) range using the sigmoid function, with a threshold gating process (threshold set to 0.5). We assign weights above the threshold to 1, obtaining information weights W_{up}, and weights below the threshold to 0, obtaining non-information weights W_{low}. Then, we multiply the input features X by W_{up} and W_{low} separately, obtaining two weighted features: feature-rich information X^{up} and feature-scarce information X^{low}. We partition X^{up} and X^{low} equally based on the number of channels and enhance the information flow between them by employing cross-reconstruction operations. Finally, the two cross-reconstructed features are concatenated to obtain spatially reconstructed features.

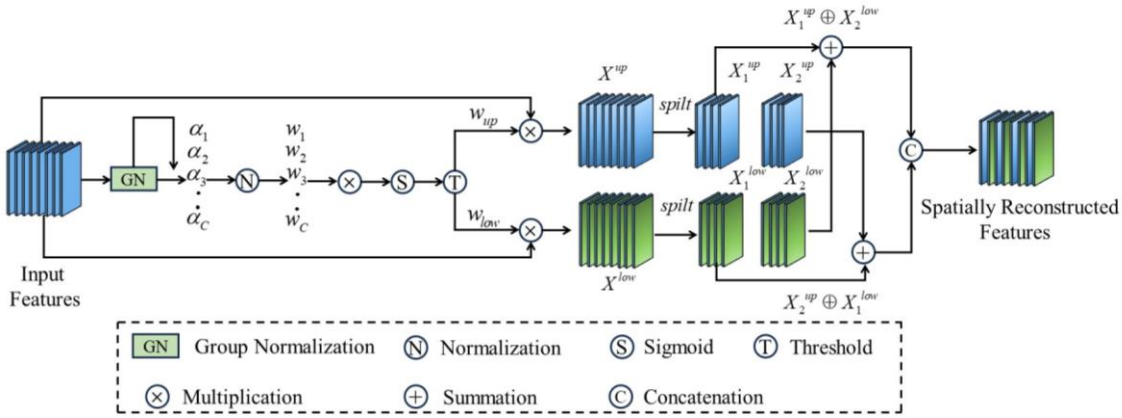

Figure 2. The overall framework of FSM

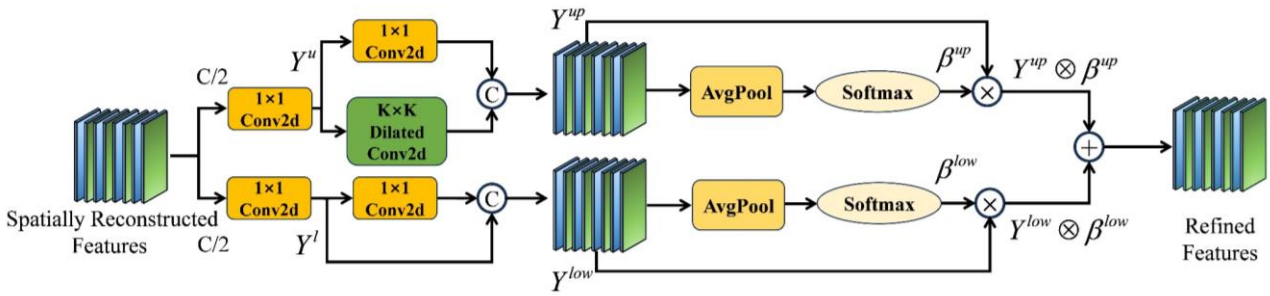

Figure 3. The overall framework of FRM

2.1.2 FRM

FRM employs multi-scale rich feature extraction on the spatially reconstructed features. As shown in Figure 3, initially, the spatially reconstructed features are evenly divided into two parts according to the number of channels, A 1×1 convolution kernel is then applied for channel compression to obtain features Y^u and Y^l respectively. For feature Y^u, a dual-branch feature extraction process is utilized. One branch employs dilated convolution to expand the receptive field for capturing broader spatial information, while the other branch employs a 1×1 pointwise convolutional layer. The outputs of these branches are concatenated to obtain feature Y^{up}. As for feature Y^l, a single branch with a 1×1 convolutional layer is employed. This branch is then concatenated with the original

feature residual branch to obtain feature Y^{low}. Subsequently, both Y^{up} and Y^{low} undergo global average pooling to aggregate global spatial information and channel-wise statistics. Softmax is then applied to the pooled results to derive feature weight vectors β^{up} and β^{low}. Finally, the output is obtained by weighting the original features with the feature weight vectors, resulting in $Y^{up} \otimes Y^{up}$ and $Y^{low} \otimes Y^{low}$. These refined features are then combined, effectively reducing common spatial and channel redundancies found in standard convolutions, thereby enhancing model efficiency and performance.

2.1.3 SPPM

ASPP employs global average pooling layers to conduct averaging operations across the entire feature map, aiming to extract comprehensive contextual information from the image and integrate it into the feature representation. However, this conventional pooling approach, which directly averages feature values, leads to blurring or overlooking of fine-grained details within edges and local regions. Consequently, there is a loss of emphasis on the intricate details of image features. To address this challenge, this paper proposes SPPM, which aims to enhance the capture of detailed global feature information.

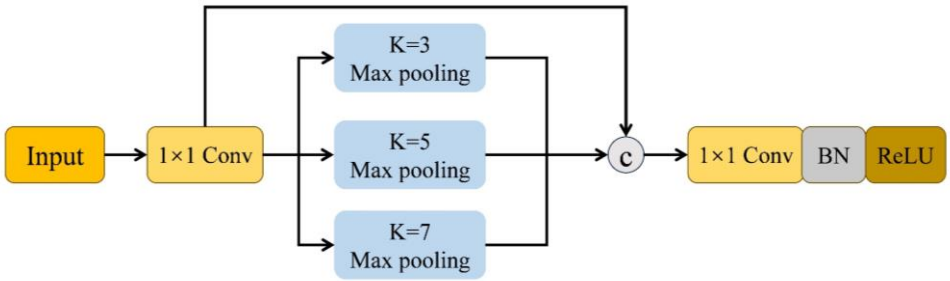

Figure 4. The overall framework of SPPM

As shown in Figure 4, the input features undergo dimensionality reduction through a 1 × 1 convolutional layer initially. Subsequently, these features are fed into three max-pooling modules with different kernel sizes for further processing. To maintain the output size matching the input size and avoid cropping of the input feature boundaries, padding is applied during the pooling operation. Each max-pooling module is dedicated to extracting the most prominent features from individual regions. The processed features are organized into a list and merged with the residual branch, which preserves the original features through a 1 × 1 convolutional layer. Compared to traditional global average pooling, SPPM demonstrates enhanced preservation of image details, resulting in improved model performance.

2.2 PSCAN

At the forefront of computer vision research, attention mechanisms, crafted to emulate the selective focus of the human visual system on particular elements of a visual scene, have markedly augmented the efficacy of various visual tasks. Channel attention mechanisms and spatial attention mechanisms, as two prevalent attention mechanisms, delve into the channel dimension and spatial dimension of image features, respectively, thereby effectively enriching the complexity and discriminative capability of feature representations.

Channel attention mechanisms evaluate the importance of various channels and employ a weighted approach to augment channels harboring pivotal information while dampening irrelevant ones. This empowers the model to adeptly apprehend abstract and nuanced features. Spatial attention mechanisms concentrate on directing focus towards diverse regions within an image, enabling them to discern and intensify attention upon pivotal objects or areas of interest, thereby optimizing computational resource allocation and augmenting the model's acuity to local features.

To empower the model to concurrently extract spatial and channel feature information from images, thereby facilitating comprehensive analysis and showcasing heightened resilience in tackling complex tasks, this paper proposes PSCAN. PSCAN consists of two parallel branches: spatial self-attention and channel self-attention. The spatial self-attention branch is tasked with capturing interactions among features within the spatial dimensions H and W, while the channel self-attention branch is dedicated to capturing interactions among feature channels.

As shown in Figure 5, in the spatial self-attention branch, the input feature X is initially divided into Q (C/2 × H × W) and V (C/2 ×H × W) utilizing separate 1 × 1 convolutions to average the channels. For Q, both global average pooling and global standard deviation pooling operations are employed to aggregate features across the spatial dimensions H × W, obtaining a C/2 × 1 × 1 matrix. Subsequently, this matrix is then reshaped into a 1 × C/2 format for further processing. Due to the potential information loss from this compression operation, it is essential to enhance the compressed Q by applying the softmax function to retain crucial features. For V, its channel count remains C/2, and it is reshaped into a C/2 × HW format. Matrix multiplication is then performed between the enhanced Q (1 ×C/2) and the reshaped V (C/2× HW), resulting in a reshaped matrix of 1 × 1 × HW. The sigmoid activation function is employed to ensure that the weight parameter values are constrained within the range of 0 to 1. This obtains a vector containing weighted information from various spatial channels, which is subsequently utilized to weight the original feature X, resulting in spatially enhanced information.

The calculation equations for spatial self-attention are as follows:

$$A^{CH}(X) = F_{sigmoid}\delta_3 \left[F_{softmax}\delta_1\left(F_{avg}F_{std}\left(W_{qw}(X)\right) \otimes \delta_2\left(W_{qv}(X)\right)\right)\right] \quad (1)$$

$$F_{avg} = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F(i,j) \quad (2)$$

$$F_{std} = \sqrt{\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(F(i,j) - F_{avg}\right)^2} \quad (3)$$

$$F^{HW} = A^{HW}(X) \odot^{HW} X \quad (4)$$

where, F_{avg}, F_{std} denote global average pooling and global standard deviation pooling, δ_1, δ_2 denote dimensionality reduction operation, δ_3 denotes dimensionality expansion operation, $F_{sigmoid}$, $F_{softmax}$ denote sigmoid and softmax functions, W_{qw}, W_{qv} denote 1×1 convolution operation, $\odot^{HW}$ denotes spatial multiplication operator.

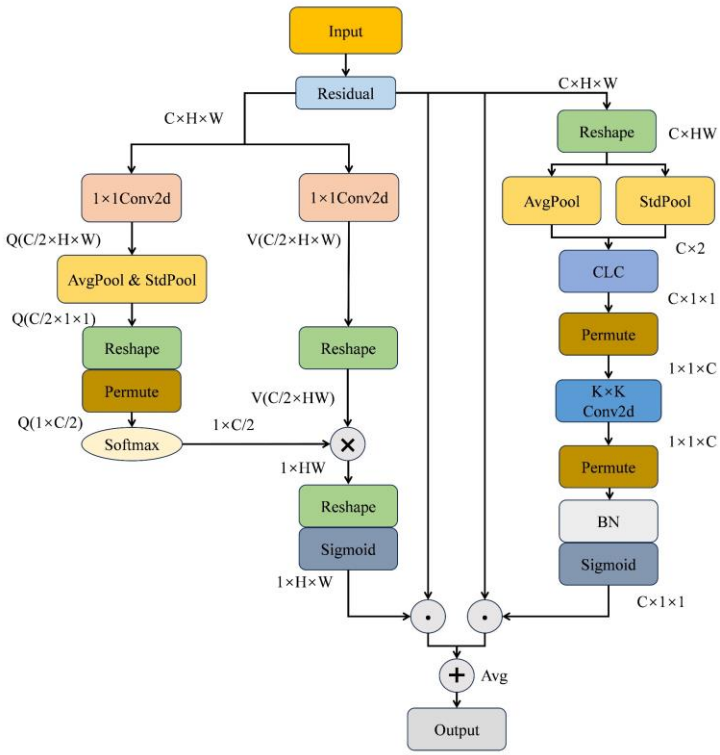

Figure 5. The overall framework of PSCAN

In the channel self-attention branch, the initial step involves reshaping the input feature X into the shape of $C \times HW$. Subsequently, global average pooling and global standard deviation pooling operations are separately executed on each channel, resulting in two tensors with a length equivalent to the number of channels C, which are then merged. This merged tensor is fed into a fully connected layer at the channel level, obtaining an output of size $C \times 1 \times 1$. To facilitate subsequent convolutional operations, this output is reshaped into the shape of $1 \times 1 \times C$. Following this, convolutional operations are conducted along the channel dimension, generating a new feature map containing inter-channel interaction information. The output of the convolutional operation is normalized and processed using the sigmoid activation function, resulting in a weight vector containing weights for each channel. These weights are applied to the original feature X, resulting in channel-enhanced information.

The calculation equations for channel self-attention are as follows:

$$A^{CH}(X) = F_{sigmoid}F_{BN}\beta_3\left[F_{1\times K}\beta_2\left(F_{CLC}\left[\left(F_{avg} + F_{std}\right)\left(\delta_1(X)\right)\right]\right)\right] \quad (5)$$

$$F^{CH} = A^{CH}(X) \odot^{CH} X \quad (6)$$

where, $\odot^{CH}$ denotes channel-wise multiplication operator.

In the final integration stage, the outputs from the channel self-attention and spatial self-attention branches are averaged to obtain the refined feature map. The specific equation is as follows:

$$F = \frac{1}{2} \otimes \left(F^{CH} \oplus F^{HW}\right) \quad (7)$$

2.3 Multi-layer skip connection

MobileNetv2, recognized as an efficient backbone network, typically consists of multiple hierarchical layers, each tasked with extracting features from input images at different levels of abstraction. The shallower layers of the network focus on capturing low-level features such as textures and edges, while the deeper layers are dedicated to extracting more abstract and semantically meaningful high-level features. In the original DeepLabv3+ network, only features from the shallow and deepest layers are typically utilized, under the premise that these layers inherently contain richer spatial information and local details.

To optimally leverage the feature information across all layers of the backbone network, this paper proposes a novel multi-layer skip connection strategy. This strategy promotes effective fusion among features from deep, intermediate, and shallow layers. Not only does this approach retains fine-grained details of the image, but it also integrates global contextual information, thereby significantly improving the

model's ability to comprehend global information within images.

During multi-layer feature fusion, the original DeepLabv3+ network utilizes the nearest-neighbor interpolation method to upsample low-level feature maps to the dimensions congruent with high-level feature maps, facilitating element-wise addition or concatenation operations. However, the nearest-neighbor interpolation method, due to its simplistic replication of nearest-neighbor pixel values, often leads to the loss of detailed information.

To address this limitation, this paper proposes the utilization of the lightweight dynamic upsampling method, DySample [15]. DySample exhibits minimal parameters and reduces computational complexities, thereby ensuring more efficacious retention of fine-grained details within feature maps.

3. EXPERIMENTS

3.1 Dataset

The dataset utilized in this paper is BM-Seg [16], comprising 1517 CT images from 23 patients with bone metastasis, including 9 females and 14 males, ranging in age from 18 to 83 years. These scanning data were collected from November 2020 to June 2022 at the Hedi Chaker University Hospital Center in Tunisia. The dataset categorizes images into infected and non-infected classes, with each CT instance accompanied by corresponding bone and bone marrow (BM) masks.

We performed data preprocessing operations such as CLAHE algorithm, adding salt and pepper noise, and horizontal mirror inversion on the dataset, which can help reduce overfitting and improve the performance of the network in asymmetric scenarios. This ultimately enhances the network's generalization ability by maintaining a consistent data distribution.

The experiments were carried out using 5-fold cross-validation. The dataset was randomly divided into training and validation sets at a ratio of 9:1.

3.2 Experimental configurations

The operating system is Ubuntu 20.04, using the PyTorch 1.10.1 deep learning framework with CUDA version 11.1. The programming language employed is Python 3.8.18. The central processing unit (CPU) is an 8-core PC with an Intel(R) Core (TM) i7-9700 CPU @ 3.00GHz, while the graphics processing unit (GPU) is an Nvidia GeForce RTX 2080Ti with 11.36 GB of memory.

For our experiments, we selected the SGD optimizer with an initial learning rate of 0.01, a weight decay of 0.0005, and a momentum of 0.937 during model training. We resized the input image size to 512×512 for training purposes. Additionally, we utilized 3 worker threads to load data on the GPU GeForce RTX 2080Ti 11.36 GB during model training. The batch size was set to 8, and all models were trained for 200 epochs.

3.3 Evaluation metrics

To validate the performance of the model and compare it with other mainstream segmentation models, we employ multiple evaluation metrics, including mean intersection over union (mIoU), dice coefficient (Dice), mean pixel accuracy (MPA), mean precision (mPrecision). The calculation equations of these evaluation metrics are as follows:

$$mPA = \frac{1}{N+1}\sum_{i=0}^{N}\frac{1}{TP+TN+FP+FN} \tag{8}$$

$$mIoU = \frac{1}{N+1}\sum_{i=0}^{N}\frac{TP}{TP+FP+FN} \tag{9}$$

$$mPrecision = \frac{1}{N+1}\sum_{i=0}^{N}\frac{TP}{TP+FP} \tag{10}$$

$$Dice = \frac{2\times TP}{2\times TP+FN+FP} \tag{11}$$

where, TP denotes the number of true positive samples predicted as positive class, TN denotes the number of true negatives predicted as negative class, FP denotes the number of false positives predicted as positive class, FN denotes the number of false negatives predicted as negative class. N is the total number of samples.

4. RESULTS

4.1 Ablation study of AFFP

To validate the effectiveness of AFFP, we conducted ablation experiments, and the experimental results are shown in Table 1. Through the incremental integration of the AFFP module, all performance metrics exhibited a consistent improvement trend. Compared to the original ASPP module, our proposed AFFP module achieved improvements of 1.34% in mIoU, 1.90% in mPA, 0.28% in mPrecision, and 0.97% in Dice. Moreover, the parameter count of the AFFP module is 4,882,258, which is reduced compared to the original ASPP module with 5,813,266 parameters. These experimental results demonstrate the effectiveness of the AFFP module.

Table 1. Ablation experiments of AFFP

Module	mIoU (%)	mPA (%)	mPrecision (%)	Dice (%)	Parameters
ASSP	81.86	91.54	86.73	88.98	5,813,266
FSM	82.01	91.65	86.84	89.10	4,553,554
FSC+FSM	82.70	92.55	87.04	89.60	4,553,554
AFFP	**82.89**	**92.91**	**87.01**	**89.73**	**4,882,258**

4.2 Comparison of different attention mechanisms

To validate the effectiveness of PSCAN, we conducted comparative experiments. The PSCAN was compared with seven other mainstream attention mechanisms, including CBAM [17], ECA [18], GAM [19], LSK [20], SGE [21], SimAM [22], and ParNet [23]. The experimental results, as shown in Table 2, indicate that the PSCAN demonstrates superior or comparable performance across all evaluation

metrics. Specifically, it achieved mIoU of 83.69%, outperforming the runner-up SGE by 0.33%. This series of improvements demonstrates the effectiveness of the PSCAN in enhancing the extraction and fusion of channel and spatial information. The specific segmentation results are depicted in Figure 6.

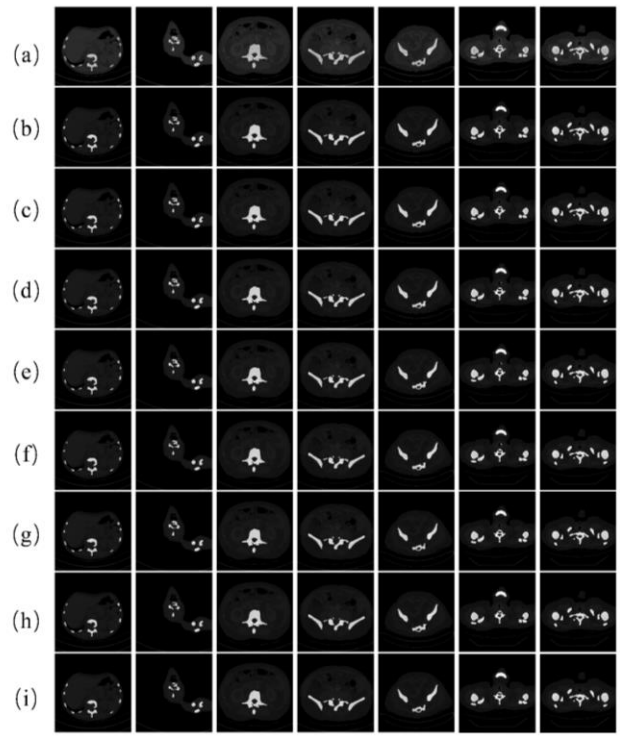

Figure 6. Segmentation results analysis with various attention mechanisms. (a) The ground truth, (b) CBAM, (c) ECA, (d) GAM, (e) LSK, (f) SGE, (g) SimAM, (h) ParNet, (i) PSCAN

Table 2. Comparison results of different attention mechanisms

Attention Mechanism	mIoU (%)	mPA (%)	mPrecision (%)	Dice (%)
CBAM	83.32	93.50	86.99	89.93
ECA	82.98	92.46	87.47	89.80
GAM	83.06	93.40	86.87	89.86
LSK	82.95	93.08	87.12	89.97
SGE	83.36	93.16	87.41	90.07
SimAM	83.00	93.16	86.96	89.82
ParNet	83.38	93.07	87.52	90.09
PSCAN	**83.69**	**93.19**	**87.80**	**90.31**

4.3 Comparison of upsampling methods

To validate the effectiveness of Dysample, we conducted comparative experiments, comparing Dysample with three other upsampling methods: Nearest-neighbor interpolation, DeConvolution (DeConv), and Caraffe [24]. The experimental results, as shown in Table 3, demonstrate that Dysample outperforms the other methods across all metrics while concurrently possessing the fewest parameters.

Table 3. Comparison results of upsampling improvement experiments

Upsampling	mIoU (%)	mPA (%)	mPrecision (%)	Dice (%)	Parameters
Nearest-neighbor	83.69	93.19	87.80	90.31	5,817,224
DeConv	83.88	93.56	87.74	90.44	6,886,984
Caraffe	83.75	93.29	87.04	90.35	6,012,580
Dysample	**83.97**	**93.90**	**87.97**	**90.50**	**5,633,832**

4.4 Comparison of different semantic segmentation networks

We compared MFP-DeepLabv3+ network with various mainstream semantic segmentation networks, including HRNet [25], Non-local [26], EncNet [27], SegFormer [28], Mask-RCNN [29], U-Net [30], and TransU-Net [31]. The experimental results shown in Table 4 are the mean number after 5-fold cross-validation of each network. It is evident from the results that our proposed network outperforms other networks across all metrics. Specifically, compared to the runner-up TransU-Net network, our network achieved improvements of 0.29% in mIoU, 2.5% in mPA, 0.21% in Dice.

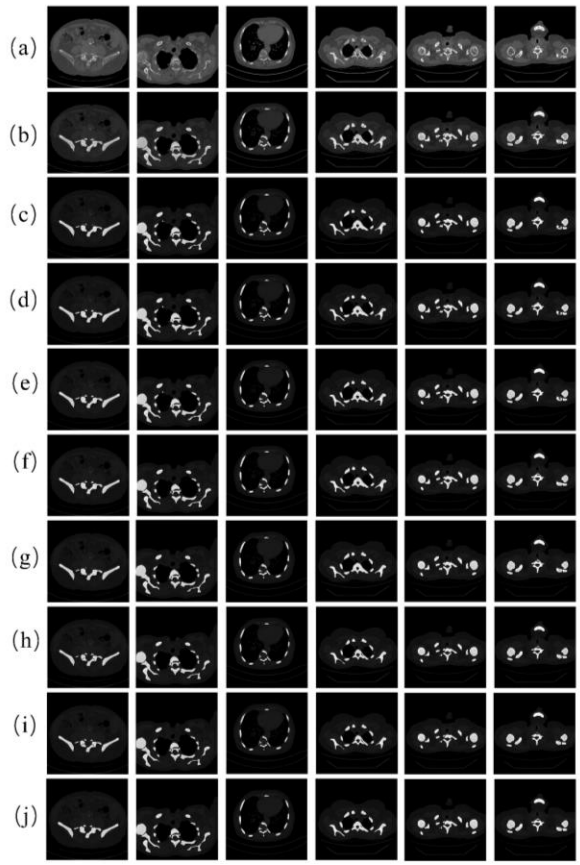

Figure 7. Segmentation results analysis with various networks. (a) The original image, (b) The ground truth, (c) HRNet, (d) Nonlocal, (e) EncNet, (f) Segformer, (g) Mask-RCNN, (h) U-Net, (i) Trans U-Net, (j) MFP-DeepLabv3+

This comprehensive comparison demonstrates the significant superiority of our proposed network in semantic segmentation tasks, providing valuable insights for further

optimization of semantic segmentation networks. The specific segmentation results are depicted in Figure 7.

Table 4. Comparison results with other semantic segmentation networks

Models	mIoU (%)	mPA (%)	mPrecision (%)	Dice (%)
HRNet	81.85	88.80	87.15	88.97
Non-local	82.37	88.75	87.99	89.36
EncNet	82.02	88.62	87.27	89.36
SegFormer	82.81	90.25	87.12	89.67
Mask-RCNN	82.04	87.92	87.39	89.11
U-Net	82.78	90.66	87.19	89.80
TransU-Net	83.67	91.45	87.97	90.29
MFP-DeepLabv3+	**83.97**	**93.90**	**87.97**	**90.50**

4.5 Ablation experiment

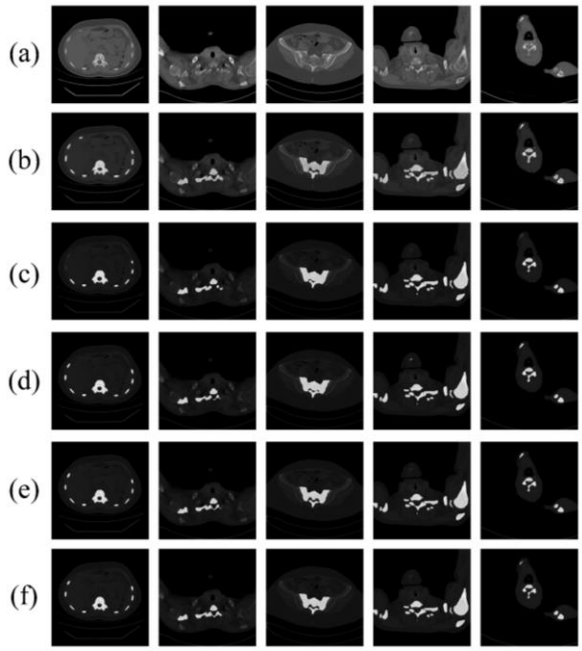

Figure 8. Segmentation results analysis of the improved network. (a) The original image, (b) The ground truth, (c) DeepLabv3+, (d)AFFP, (e)PSCAN, (f) MFP-DeepLabv3+

Table 5. Ablation experiments for improved MFP-DeepLabv3+

Improvement	mIoU (%)	mPA (%)	mPrecision (%)	Dice (%)	Parameters
DeepLabv3+	81.86	91.54	86.73	88.98	5,813,266
AFFP	82.89	92.91	87.01	89.73	4,882,258
PSCAN	83.69	93.19	87.80	90.31	5,325,000
MFP-DeepLabv3+	**83.97**	**93.90**	**87.97**	**90.50**	**5,633,832**

All improvement results in the final analysis are summarized, which is as shown in Table 5. Compared to the original DeepLabv3+, our proposed MFP-DeepLabv3+ achieved a reduction of 5% in parameters, and improvements of 2.11% in mIoU, 2.36% in mPA, 1.24% in mPrecision, and 1.52% in Dice. These experimental results demonstrate the effectiveness of our improvement approach and provide strong support for further research and application. Segmentation results are depicted in Figure 8.

5. CONCLUSION

This paper proposes an MFP-DeepLabv3+ network tailored for bone metastasis segmentation tasks. Initially, we enhance the multi-scale feature extraction capability of the network by introducing the AFFP module. By effectively capturing and integrating features across multiple scales, we enabled more accurate identification and segmentation of both subtle and significant pathological features. Furthermore, we introduce the PSCAN to refine the focus of the network on both channel and spatial information, enabling the network to more sensitively attend to prominent features in the image data. This refinement allows for a more precise differentiation between healthy bone tissue and metastatic lesions. Additionally, we employed a multi-layer skip connection strategy to integrate global semantic information. Experimental results demonstrate that the MFP-DeepLabv3+ network achieves significant improvements of 2.11% in mIoU, 2.36% in mPA, 1.24% in mPrecision, and 1.52% in Dice. Additionally, the GPU memory usage during training is 5.91G, with an inference speed of 0.0521 seconds per image. These results conclusively demonstrate the proposed MFP-DeepLabv3+ network possesses detailed and accurate segmentation capabilities for bone metastatic regions, providing substantial assistance to clinicians in treatment strategy determination. For cancer patients, early and precise detection of bone metastasis is crucial, aiding in timely treatment selection.

Despite the significant improvements achieved by our proposed network, there are still limitations to be addressed. Although the improved network may perform well on specific datasets or tasks, its generalization ability might be limited, making it challenging to maintain efficiency across diverse clinical scenarios or different types of bone metastasis images. This limitation restricts its applicability in clinical practice. Future work will focus on exploring methods to transfer the trained model to other medical imaging tasks or diverse datasets, ensuring the model's performance and generalization across various scenarios.

FUNDING

The research was supported by the Applied Basic Research Program of Liaoning Province (Grant No.: 2022JH2/101300024).

REFERENCES

[1] Clézardin, P., Coleman, R., Puppo, M., Ottewell, P., Bonnelye, E., Paycha, F., Confavreux, C.B., Holen, I. (2021). Bone metastasis: Mechanisms, therapies, and biomarkers. Physiological Reviews, 101(3): 797-855. https://doi.org/10.1152/physrev.00012.2019

[2] Huang, J.F., Shen, J., Li, X., Rengan, R., Silvestris, N., Wang, M., Derosa, L., Zheng, X., Belli, A., Zhang, X.L.,

Li, Y.M. (2020). Incidence of patients with bone metastases at diagnosis of solid tumors in adults: A large population-based study. Annals of Translational Medicine, 8(7): 482. https://doi.org/10.21037/atm.2020.03.55

[3] Ban, J., Fock, V., Aryee, D.N., Kovar, H. (2021). Mechanisms, diagnosis and treatment of bone metastases. Cells, 10(11): 2944. https://doi.org/10.3390/cells10112944

[4] Rahman, S.Z., Singasani, T.R., Shaik, K.S. (2023). Segmentation and classification of skin cancer in dermoscopy images using SAM-based deep belief networks. Healthcraft Front, 1(1): 15-32. https://doi.org/10.56578/hf010102

[5] Yucel, N., Mutlu, H.B., Durmaz, F., Cengil, E., Yildirim, M. (2023). A CNN approach for enhanced epileptic seizure detection through EEG analysis. Healthcraft Frontiers, 1(1): 33-43. https://doi.org/10.56578/hf010103

[6] Liu, S., Feng, M., Qiao, T., Cai, H., Xu, K., Yu, X., Jiang, W., Lv, Z., Wang, Y., Li, D. (2023). Deep learning for the automatic diagnosis and analysis of bone metastasis on bone scintigrams. Cancer Management and Research, 14: 51-65. https://doi.org/10.2147/CMAR.S340114

[7] Song, Y., Lu, H., Kim, H., Murakami, S., Ueno, M., Terasawa, T., Aoki, T. (2019). Segmentation of bone metastasis in CT images based on modified HED. In 2019 19th International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea (South), pp. 812-815. https://doi.org/10.23919/ICCAS47443.2019.8971539

[8] Ntakolia, C., Diamantis, D.E., Papandrianos, N., Moustakidis, S., Papageorgiou, E.I. (2020). A lightweight convolutional neural network architecture applied for bone metastasis classification in nuclear medicine: A case study on prostate cancer patients. Healthcare, 8(4): 493. https://doi.org/10.3390/healthcare8040493

[9] Lin, Q., Gao, R., Luo, M., Wang, H., Cao, Y., Man, Z., Wang, R. (2022). Semi-supervised segmentation of metastasis lesions in bone scan images. Frontiers in Molecular Biosciences, 9: 956720. https://doi.org/10.3389/fmolb.2022.956720

[10] Noguchi, S., Nishio, M., Sakamoto, R., Yakami, M., Fujimoto, K., Emoto, Y., Kubo, T., Iizuka, Y., Nakagomi, K., Miyasa, K., Satoh, K. (2022). Deep learning–based algorithm improved radiologists' performance in bone metastases detection on CT. European Radiology, 32(11): 7976-7987. https://doi.org/10.1007/s00330-022-08741-3

[11] Liu, C., Cao, Y., Lin, Q., Man, Z., He, Y., Peng, L. (2023). Segmentation of metastatic lesions on bone scan images based on improved UNet3+ network. In 2023 4th International Conference on Computer Engineering and Application (ICCEA), Hangzhou, China, pp. 916-920. https://doi.org/10.1109/ICCEA58433.2023.10135231.

[12] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 801-818. https://doi.org/10.48550/arXiv.1802.02611

[13] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 4510-4520. https://doi.org/10.48550/arXiv.1801.04381

[14] Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122. https://doi.org/10.48550/arXiv.1511.07122

[15] Liu, W., Lu, H., Fu, H., Cao, Z. (2023). Learning to upsample by learning to sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 6027-6037. https://doi.org/10.48550/arXiv.2308.15085

[16] Afnouch, M., Gaddour, O., Hentati, Y., Bougourzi, F., Abid, M., Alouani, I., Ahmed, A.T. (2023). BM-Seg: A new bone metastases segmentation dataset and ensemble of CNN-based segmentation approach. Expert Systems with Applications, 228: 120376. https://doi.org/10.1016/j.eswa.2023.120376

[17] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19.

[18] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534-11542.

[19] Liu, Y., Shao, Z., Hoffmann, N. (2021). Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561. https://doi.org/10.48550/arXiv.2112.05561

[20] Li, Y., Hou, Q., Zheng, Z., Cheng, M.M., Yang, J., Li, X. (2023). Large selective kernel network for remote sensing object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16794-16805. https://doi.org/10.48550/arXiv.2303.09030

[21] Li, X., Hu, X., Yang, J. (2019). Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. arXiv preprint arXiv:1905.09646. https://doi.org/10.48550/arXiv.1905.09646

[22] Yang, L., Zhang, R.Y., Li, L., Xie, X. (2021). Simam: A simple, parameter-free attention module for convolutional neural networks. In International Conference on Machine Learning, pp. 11863-11874.

[23] Goyal, A., Bochkovskiy, A., Dengl, J., Koltun, V. (2022). Non-deep networks. Advances in Neural Information Processing Systems, 35: 6789-801.

[24] Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D. (2019). Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3007-3016.

[25] Sun, K., Xiao, B., Liu, D., Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Korea, pp. 5693-5703. https://doi.org/10.48550/arXiv.1902.09212

[26] Wang, X., Girshick, R., Gupta, A., He, K. (2018). Non-local neural networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

[27] Zhang, H., Xue, J., Dana, K. (2017). Deep ten: Texture encoding network. In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, Honolulu, USA, pp. 708-717.
[28] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems. 34: 12077-12090.
[29] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322
[30] Abo-El-Rejal, A., Ayman, S.E., Aymen, F. (2024). Advances in breast cancer segmentation: A comprehensive review. Acadlore Transactions on AI and Machine Learning, 3(2): 70-83. https://doi.org/10.56578/ataiml030201
[31] Chen, J., Lu, Y., Yu, Q., et al. (2021) Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021. https://doi.org/10.48550/arXiv.2102.04306