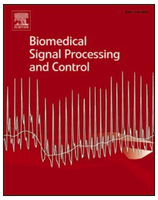
Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



A multi-branch hierarchical attention network for medical target segmentation

Yongtao Yu [a,*], Yifei Tao [b], Haiyan Guan [c], Shaozhang Xiao [a], Fenfen Li [a], Changhui Yu [a], Zuojun Liu [a], Jonathan Li [d]

[a] *Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, Jiangsu 223003, China*
[b] *Women's Nutriology Department, Huaian Maternal and Child Health Care Center of Jiangsu Province, Huaian, Jiangsu 223002, China*
[c] *School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, China*
[d] *Department of Geography and Environmental Management, University of Waterloo, Waterloo, Ontario N2L3G1, Canada*

ARTICLE INFO

Keywords:
Medical target segmentation
Medical image analysis
Convolutional neural network
Feature attention
Clinical diagnosis

ABSTRACT

Medical imaging techniques have been widely used in modern clinical disease diagnosis and treatment programs. The captured medical images can well reflect the conditions of the human body tissues, which are significantly helpful to the doctors to determine the existence or the severity of the disease. In this paper, we develop a hierarchical attentive high-resolution convolutional network (AttHRNet) for segmenting targets of interest from medical images aiming to improve the automated processing standard and the intelligent interpretation quality of the medical images. The AttHRNet is an improved version of the high-resolution network (HRNet) structure with three novel modules. First, built with an improved HRNet structure assisted by a multiscale context augmentation (MSCA) module as the feature extraction backbone, the AttHRNet can produce a set of high-quality, strong-semantic feature maps at different resolutions. The MSCA module functions to reduce the information loss during feature downsampling. Second, designed with an effective feature attention principle, the feature encoding quality in each branch can be significantly promoted by concentrating on the informative and salient feature encodings across both channels and spatial locations. Furthermore, formulated with a hierarchical segmentation scheme, the output feature maps can be further augmented by including the semantic-level category exploitation (SLCE) module with a global perspective. The SLCE module allows the information from lower resolution segmentations to inform higher resolution segmentations. Through quantitative examinations, visual verifications, and comparative evaluations on four medical image datasets, we convince the promising applicability and competitive superiority of the AttHRNet in medical target segmentation issues.

1. Introduction

Medical images act as a crucial role in current clinical disease checking, diagnosis, and treatment programs, as well as in many health examinations for assisting in evaluating the physical conditions. They can visually reflect the situations and changes of the interior tissues of human bodies in a non-invasive way, which contributes significantly to the early detection of diseases and the supervision of therapeutic schedule formulations. The common techniques for medical image collection include computed tomography (CT), X-ray imaging, magnetic resonance imaging (MRI), ultrasonography, etc. The frequent use of these medical imaging equipment in daily inspections results in a large volume of different-type and different-pattern medical images, which cost the doctors considerable time and energy to read and analyze these images for disease diagnoses and severity estimations. As the importance of medical images keeps growing, many researches have focused on the automated processing and interpretation of medical images aiming at providing the pre-analyzed results to improve the efficiency and accuracy of clinical disease diagnoses [1,2]. Thereinto, medical image segmentation, as a widely studied topic, dedicates to locate and segment the medical targets of interest, such as organs and lesions, which provides essential evidence to supervise the detection, determination, observation, and evaluation of the diseases.

In the literature, there are numerous works being conducted with increasing enhanced performances for medical image segmentation tasks. However, despite the success achieved so far in segmentation

[*] Corresponding author.
E-mail address: allennessy@hyit.edu.cn (Y. Yu).

https://doi.org/10.1016/j.bspc.2022.104021
Received 5 February 2022; Received in revised form 23 June 2022; Accepted 20 July 2022
Available online 30 July 2022

accuracy, it still faces great challenges to attain human-level qualities [3] caused by the unique properties of the medical targets in the images captured with different imaging protocols, including heterogeneities in colors and textures, variations in sizes and shapes, low contrasts or similarities between the targets and the surrounding tissues, uncertainties resulted in by obscure borders, etc. Therefore, investigating advanced and effective techniques to promote the segmentation performance is greatly meaningful and significantly favorable to upgrade the detection and precision rates in clinical disease diagnoses.

In this paper, aiming at upgrading the intelligent interpretation quality of the medical images for facilitating clinical disease diagnoses, we develop a novel attentive high-resolution convolutional network architecture. This architecture is an improved version of the high-resolution network (HRNet) structure with three novel modules for improving the medical image segmentation quality. This architecture is made of a multi-branch feature extraction backbone for extracting multiscale strong-semantic feature maps, a feature attention module for boosting the feature representation quality, and a hierarchical segmentation head for gradually producing a high-quality, high-resolution segmentation map. Benefitting from the powerful combination and advantageous integration of these three functional components, the designed model behaves excellently in processing medical targets of different sizes and shapes, varying appearances and self-conditions, and diverse boundary properties and surrounding scenarios. To sum up, the contributions of this paper mainly consist in the following three parts. (1) An improved HRNet structure is built for extracting high-level semantically strong feature representations at different resolutions. To well alleviate the feature detail loss during feature downsampling, an effective multiscale context augmentation (MSCA) module is designed to supervise the cross-branch multi-resolution feature propagation. (2) A novel feature attention module is developed for repeatedly promoting the feature encoding quality at each feature resolution. To well highlight the informative feature semantics and weaken the contributions of the helpless ones, a channel-specific attention unit and a spatial-specific attention unit are cascaded to, respectively, emphasize the informative feature channels and the important spatial locations. (3) A hierarchical segmentation pipeline is designed to progressively refine the output feature maps with semantic-level contextual information for producing a high-quality segmentation map. To well characterize semantic-level properties and allow the information from lower resolution segmentations to inform higher resolution segmentations, a semantic-level context exploitation (SLCE) module is proposed to provide a uniform semantic representation for each individual category with a global perspective. The specific contributions and their progresses in this paper are listed in detail in Table 1.

2. Related work

2.1. Handcrafted feature based strategies

In the early days, prior knowledge, empirical rules, and handcrafted features were widely used to segment the medical targets of interest. Typical strategies included intensity thresholding schemes, edge detection operators, and pixel classification models. Patra et al. [4] developed a multi-level intensity thresholding approach to locate breast lesions by considering the strong intensity contrast of the lesions. Chakraborty and Mali [5] applied the morphological reconstruction technique to suppress the noise interferences, followed by a block-based intensity binarization for lung region determination. The intensity-based approaches usually showed excellent processing efficiencies; however, their performances were easily to be affected by the variations of the target conditions and data sources. In addition, the intensity threshold was task and target sensitive. Wang et al. [6] combined the geometric active contour model and the Hough transform to segment vessel lumens. Specifically, manual intervention was selectively conducted to adjust inaccurate delineations. Likewise, an adaptive active contour model was designed in [7] for segmenting lung images. Lu et al. [8] developed a two-stage level set model constrained with shape priors to segment cardiac ventricles. First, regions of interest (ROI) were estimated through adjacent sequence subtraction and intensity thesholding. Then, Hough transform and level set were applied to detect endocardium and epicardium with circle primitives. The active contour models and Hough transform operations performed effectively to delineate the target contours. Nevertheless, the contrast qualities between the targets and their surrounding background also affected the delineation accuracy.

In order to provide a target-level feature interpretation, the raster images were converted into high-order structures and semantic models. Chen et al. [9] proposed an improved version of the graph cuts model, which was optimized with adaptive shape priors for supervising the accurate extraction of the target boundaries. Filali et al. [10] constructed a graph formulation to rank the distinguishabilities between the skin and the lesion regions, which were eventually fused for lesion segmentation. Jia et al. [11] designed a hierarchical workflow comprising snake model, watershed, and shape fitting to segment cell instances. In this framework, snake model and watershed were cooperated to provide initial segmentations, which were refined through shape fitting. Fan et al. [12] adopted an improved Mumford-Shah model to carry out medical target segmentation. Specifically, dimensionality reduction of the image was initially conducted to improve the processing efficiency and a Chambolle-Pock pairwise algorithm was applied to optimize the Mumford-Shah model. To handle the issue of limited labels, Huang et al. [13] proposed a Chan-Vese model for medical image segmentation with an unsupervised manner. It employed an iterative segmentation scheme with the cooperation of the weight maps generated by the Markov chain. An advantage of the high-order structures or semantic models lies in that the discrete low-semantic pixel primitives constituting the foreground and background were selectively organized to improve the distinguishability and highlight the target components.

Aiming at promoting the feature representation quality and the target-specific feature semantic uniqueness, some machine learning approaches were developed accordingly to serve medical image segmentation tasks. To alleviate noise impacts, Tavakoli-Zaniani [14] presented an improved fuzzy C-means model by weighted integrating the noisy and denoised images. Pereira et al. [15] integrated local binary

Table 1
Main contributions and their progresses in this paper.

Contribution	Function	Progress	Assumption
MSCA module	Reduction of feature detail loss during downsampling	Preserved more feature details	Importance of contextual properties to semantic targets
Feature attention module	Recalibration of channel and spatial feature semantics	Highlighted feature significance	Importance of channel and spatial feature saliencies
SLCE module	Exploitation of category-aware feature semantics	Enhanced target semantic contrast	Importance of target semantics from different categories
Improved HRNet backbone	Extraction of semantically-strong feature representations	Promoted feature representation quality	Importance of feature representation robustness and distinguishabilities
Hierarchical segmentation strategy	Augmentation of segmentation results	Improved segmentation accuracy	Importance of target details at different granularities

patterns into the *k*-means clustering approach for precisely determining the lesion borders in skin images. The extracted local binary pattern properties served to enhance the saliency of the lesion areas. Schneider et al. [16] designed a pair of Hough forest models to classify the image patch features to, respectively, segment vessel regions and extract vessel centerlines. The patch features were depicted using steerable filters at varying scales and orientations. Huang et al. [17] employed a bag-of-visual-words (BoVWs) model to represent the semantic regions, which were generated based on superpixel segmentation. These semantic regions were further recognized to segment the breast tumors through a BoVWs feature based classifier. In addition, dictionary learning [18], wavelet transform [19], feature vector [20], and random forest [21] were also investigated for medical image segmentation.

Generally, the handcrafted feature based methods are easy to be implemented and show promising processing efficiencies. However, they are limited to specific applications and sensitive to the variations and qualities of the image data.

2.2. *Deep feature based strategies*

In recent years, deep learning architectures have encountered unprecedented prosperity in a broad range of vision tasks [22]. An advantageous property of deep learning models is reflected in the automated abstraction of high-level feature representations. As a result, intensive attempts have also been made to introduce deep learning models into medical image segmentation tasks [23]. An and Liu [24] presented a convolutional neural network (CNN) with a feedback philosophy to conduct medical image segmentation. The feedback optimization process was directed via a greed-based pruning and recovering strategy. Liang et al. [25] proposed a region-based CNN architecture to improve the target-level segmentation accuracy. In this architecture, guided anchoring techniques and fusioned box score measures were cooperated for obtaining tight boundaries between the adhered and clustered targets. Gu et al. [26] embedded a comprehensive attention scheme into the CNN to promote the feature semantics. The attention module comprised three parts for recalibrating the spatial, channel, and scale-level feature semantics, respectively. Zhang et al. [27] designed a fully convolutional network (FCN) architecture stacked by compressed dense blocks. Specifically, these blocks employed dilated convolutions to rapidly access large spatial contexts. Wang et al. [28] developed a hybrid network architecture, which involved three task-specific branches sharing the same encoder. Specifically, two of them functioned for pixel-level segmentation and the other one served for patch-level classification. With the gradual exploitation of the deep and high-level feature semantics, the CNN models performed promisingly in the medical image segmentation tasks. However, the output low-resolution feature representations used for prediction sometimes cannot meet the requirements of fine-grained segmentations.

Aiming at improving both the quality and resolution of the output feature semantics, some modified architectures have also been elaborately developed. A representative was the U-Net architecture. Ronneberger et al. [29] pioneered a U-Net architecture composed of a contracting pathway and a symmetric expanding pathway, resulting in a U-shape formulation. To be specific, the contracting pathway functioned to extract different-level feature semantics at different scales and the expanding pathway functioned to gradually recover a high-resolution feature representation augmented by the feature semantics from the contracting pathway. Yang et al. [30] presented a modified U-Net formulation to integrate multilevel feature encodings for improving the segmentation accuracy. In this network, residual and dilated blocks were leveraged for feature boosting. Similarly, aiming at enhancing the representation quality of the U-Net, Huang et al. [31] designed a hierarchical channel-oriented feature attention scheme. Badshah and Ahmad [32] extended the U-Net architecture by embedding the residual blocks, batch normalization, and bidirectional ConvLSTM to construct a ResBCU-Net architecture. Besides, asymmetric U-Net [33], ensemble U-Net [34], and UNet++ [35] were also designed for segmenting medical targets. To well handle blurred boundaries, Zhou et al. [36] suggested an encoder-decoder architecture to segment low-contrast targets. The encoder-decoder architecture can be viewed as a generalization and relaxation version of the U-Net architecture. In this network, multiscale skip connections and dilated connections were combined to achieve high-resolution feature representations.

An alternative for multi-level feature fusion was the formulation of a feature pyramid network (FPN) architecture [37]. The FPN employed a bottom-up pathway to exploit multilevel and multiscale feature semantics, which were gradually fused through a top-down pathway, resulting in different-resolution promoted feature representations at different stages. Hsiao et al. [38] applied the FPN architecture to segment kidneys from CT images with a specifically-designed hyperparameter optimization process. Gridach [39] proposed a pyramid dilated network (PyDiNet) to capture the small and complex variations in the medical images while preserving the spatial details. This was achieved by the integration of a multi-branch dilated convolution structure. As a novel architecture design paradigm, HRNet [40] adopted a parallel, rather than a cascade, feature exploitation structure. High-level feature semantics were concurrently extracted under different subspaces with the repeated exchanges among them for feature semantic augmentation. Wan et al. [41] developed a coarse-to-fine segmentation framework based on a capsule HRNet architecture, named as HR-CapsSegNet. Specifically, full attention mechanism and dilated convolution operations were embedded for boosting the feature representation quality.

Zhang et al. [42] stacked a generative adversarial network (GAN) to conduct lesion segmentation for COVID-19 analysis. This GAN employed a dense-block formulation and a multi-layer attention strategy for feature semantic augmentation. As improvements, unpaired GAN [43], one-shot GAN [44], and multiscale GAN [45] were also constructed for medical target segmentations. The superiority of the GAN-based architectures lies in that they can also achieve surprising segmentation results even with limited samples. Pang et al. [46] presented a two-stage segmentation pipeline composed of two parallel networks for spine parsing. These two networks operated successively to, respectively, provide initial segmentations and conduct segmentation refinement. In addition, graph convolutional network (GCN) [47], mask R-CNN [48], capsule network (SegCaps) [49], feature fusion attention network (FFANet) [50], pairwise learning [51], multi-task learning [52], weakly supervised learning [53], and transfer learning [54] models were also intensively exploited for medical image segmentation applications.

Comparatively, deep learning models are not only limited to specific segmentation tasks or data sources. Instead, they can be easily retrained and applied to different segmentation tasks and different data sources with little or even no architecture modifications.

2.3. *Feature attention mechanisms*

Aiming at further promoting the feature representation quality to improve the prediction accuracies of the vision tasks, many attempts have been recently made to strengthen the contributions of the useful feature semantics [55]. Roughly speaking, existing techniques generally focus on the recalibrations of the channel feature semantics to highlight the task-specific channels and the recalibrations of the spatial feature semantics to emphasize the task-specific regions. Hu et al. [56] developed a squeeze-and-excitation (SE) block to adaptively recalibrate the channel feature semantics. The SE block determined the channel-wise significances by modelling the interdependencies among the channels. As a modification, Zhang et al. [57] proposed a pyramid squeeze attention (PSA) module for channel feature promotion under different scales. Specifically, the PSA module took the multiscale channel features as the input and accomplished feature recalibration based on the SE block. Differently, Qin et al. [58] presented a frequency channel attention module to exploit channel features under different frequencies. In this module, the feature semantics from different frequencies were concatenated and comprehensively considered to determine the channel feature saliencies. Hou et al. [59] designed a coordinate attention block by embedding positional attributes into channel attentions. The coordinate attention block investigated the feature significances along the horizontal and vertical directions, respectively, which were finally combined to form the position-aware feature encodings. To integrate both local and global contents, Zhong et al. [60] constructed a squeeze-and-attention (SA) module. Different from the SE block, the SA module employed a non-fully-squeezed scheme to parse the local feature details.

To highlight spatial feature saliencies, Jaderberg et al. [61] pioneered a spatial transformer network (STN) architecture to recalibrate feature semantics in the spatial domain. The STN contained three main components, including a localization net, a grid generator, and an image sampler, to determine an affine-transformation-invariant feature representation of the semantic target. Almahairi et al. [62] developed a dynamic capacity network (DCN) formulation to adaptively assign the feature significances to different image portions. The selection was determined based on a gradient-based attention mechanism. To improve localization accuracy, Mayo et al. [63] proposed a spatial embedding principle by using attention mechanisms. Through reinforcement learning, the built attention probability map was applied to infer the spatial information. Ulutan et al. [64] employed a spatial graph network structure to exploit the relative spatial and structural correlations between the semantic objects. The spatial attention was achieved by learning the spatial interaction patterns between the object pairs. Aiming at realizing relative saliency encodings to highlight the foreground regions, Fang et al. [65] suggested a position-preserved attention strategy. The attention module comprised a position embedding stage for enriching the feature semantics with positional properties and a feature interaction stage for making use of the mutual features between object proposals.

As a hybrid type of feature attention mechanisms, multiple feature attention schemes have been combined in some researches. Generally, they demonstrated more advantageous performances compared with those relying on a single feature attention mechanism. Woo et al. [66] designed a convolutional block attention module (CBAM) to simultaneously attend to the semantic-related channel and spatial features. The two subparts were cascaded to sequentially recalibrate the channel and spatial feature semantics. As an alternative, Fu et al. [67] developed a dual-attention (DA) module by paralleling a position attention unit and a channel attention unit. These two units served, respectively, to emphasize the task-aware spatial and channel feature semantics, which were eventually fused to enhance the feature representation quality. Differently, Zhao and Wu [68] applied the channel and spatial attention mechanisms, respectively, to different levels of features to conduct feature recalibrations separately. The attentive multilevel feature semantics were finally combined for directing predictions. Chen et al. [69] combined the feature attention with the confidence attention to optimize the model robustness. Specifically, the confidence attention scheme was applied to formulate the loss function for supervising the model training. To model long-range dependencies, Wiles et al. [70] suggested a co-attention module to match feature semantics with precise spatial location evidences. The attention information was computed by comparing the similarities between feature pairs. In addition, residual attention [71], depth-sensitive attention [72], domain attention [73], and vision transformers [74] were also investigated to perform feature attentions.

3. Method

3.1. Architecture overview

Fig. 1 presents the overview of the proposed attentive high-resolution convolutional network (AttHRNet) designed for medical image segmentation. The AttHRNet employs a fully convolutional network architecture and involves three primary functional elements: a feature extraction backbone, a feature attention module, and a segmentation head. To be specific, the feature extraction backbone follows an improved four-branch HRNet [40] structure to produce a set of multiscale high-level strong feature semantics. The feature attention module cascades two feature recalibration units to emphasize channel and spatial specific informative feature semantics for feature representation quality promotion. The segmentation head adopts a bottom-up hierarchical formulation to progressively refine the multiscale feature maps with semantic-level augmentations for accurate segmentation map generation.

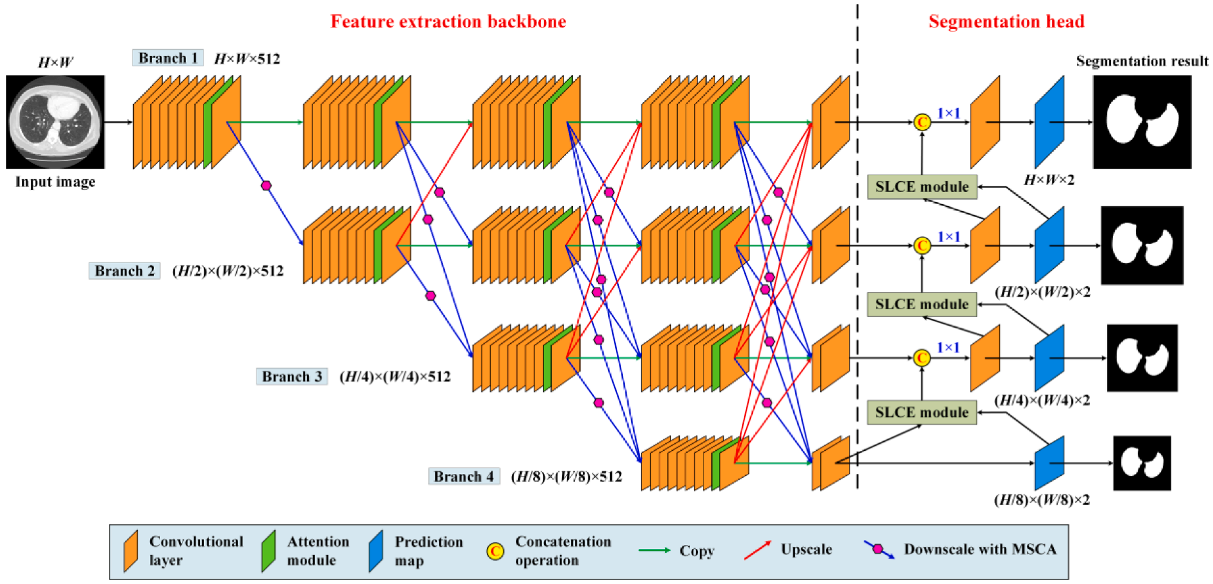


Fig. 1. Overview of the designed attentive high-resolution convolutional network (AttHRNet). The AttHRNet consists of an improved HRNet structure as the feature extraction backbone and a hierarchical segmentation head. The novel additions to the AttHRNet involve the multiscale context augmentation (MSCA) module, the feature attention module, and the semantic-level context exploitation (SLCE) module, which serve to, respectively, reduce the feature detail loss, promote the feature representation quality, and improve the segmentation accuracy.

3.2. *Feature extraction backbone*

Compared with the popularly used deep network architectures, such as U-Net and FPN, which depend on the low-resolution high-level feature maps to recover a high-resolution representation for providing per-pixel task-specific feature semantics, the development of the HRNet [40] innovates a novel network design paradigm. A unique property of the HRNet is reflected in the parallel formulation rather than the cascaded formulation. That is, it parallels multiple convolutional branches for concurrently exploiting high-quality feature representations at varying scales. Specifically, it contains a high-resolution branch across the whole network to maintain the high-resolution feature representation. Besides, the multiscale features are repeatedly fused across the branches to augment the feature quality at each scale. Therefore, due to the advanced multiscale feature encoding characteristic of the HRNet, we formulate the feature extraction backbone as an improved HRNet structure aiming at producing strong feature semantics to improve per-pixel segmentation accuracy.

As illustrated by Fig. 1, the feature extraction backbone includes four parallel branches serving for mining high-level feature representations at different resolutions. It is an improved version of the HRNet structure, which contains two novel modules for, respectively, reducing the feature detail loss during feature downscaling and promoting the feature representation quality by recalibrating the channel and spatial feature informativeness, and follows the same network architecture and the same dimensions in each branch as those of the original HRNet structure. It begins with a high-resolution branch (Branch 1) and progressively connects up lower-resolution branches with a downscaling step of 0.5. With the connection of the lower-resolution branches, larger contexts can be accessed to exploit feature semantics from a broader perspective. Then, through cross-branch feature fusion, the feature semantics in each branch can be significantly augmented by comprehensively aggregating the different-resolution feature representations from all the branches. Noteworthily, the feature maps in each branch maintain the same size and spatial resolution, which can effectively alleviate the localization bias issue. To sum up, all the above novel design philosophies of the HRNet guarantee the remarkable feature representation capability and build up its position in pixel-wise segmentation tasks.

3.3. *Multiscale context augmentation module*

There is a fly in the ointment in the HRNet with regard to the feature downscaling operation either when connecting up a new lower-resolution branch or when conducting cross-branch feature aggregation, as well as the feature addition operation when fusing the multi-resolution feature semantics. The feature downscaling operation might cause feature detail loss and the feature addition operation might suppress the distinctions of the feature semantics from different resolutions. To solve this issue, we propose a multiscale context augmentation (MSCA) module to perform feature downscaling. As depicted by Fig. 2, the MSCA module comprises five parallel branches serving for exploiting contextual properties at different scales. Each branch (except the top branch) involves a channel reduction operation, a feature aggregation operation, and a context exploitation operation. To be specific, for each branch, a 1 × 1 convolution is first operated on the input feature map to perform channel reduction aiming at reducing the computation overhead. Then, the reduced feature map is concatenated and fused with the output feature map from the previous branch through a 1 × 1 convolution for feature semantic augmentation. Finally, a dilated convolution is applied to the augmented feature map to exploit contextual properties with different-size receptive fields. The aggregation of the output feature map from the previous branch can effectively promote the feature encoding quality by including the contextual information from the smaller-size receptive fields, thereby resulting in a multiscale perspective in each branch rather than a single scale perspective. Specifically, the feature semantics with the original size of receptive field are well conveyed to the following branches by introducing a connection

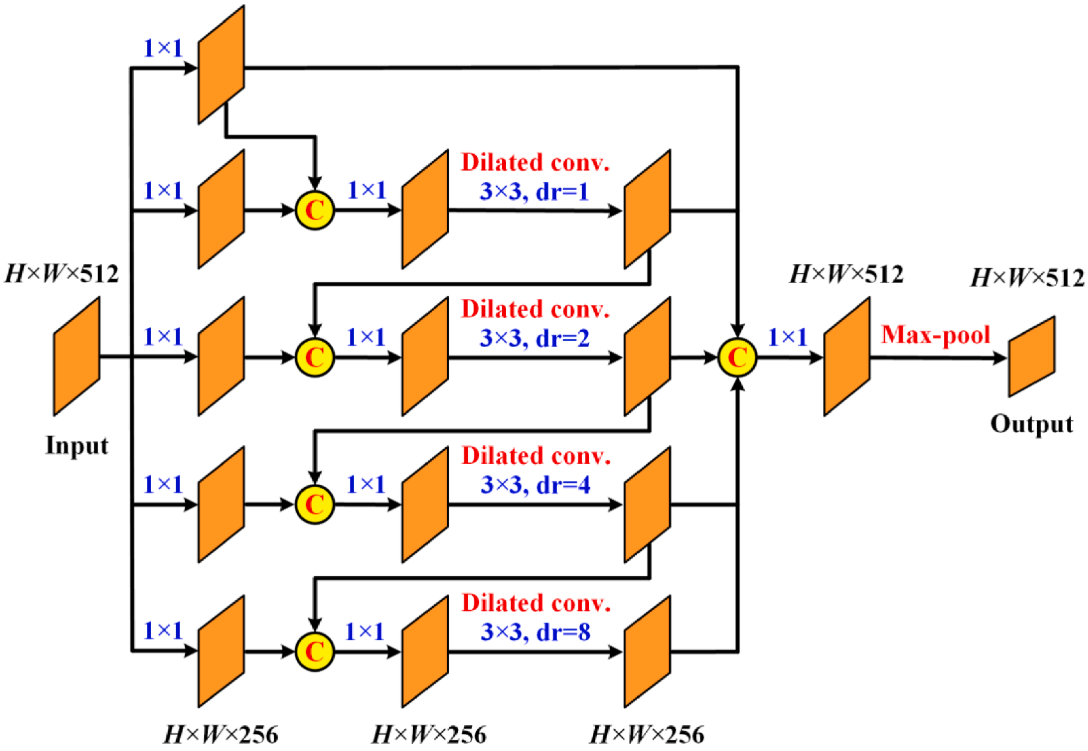


Fig. 2. Architecture of the multiscale context augmentation (MSCA) module. The MSCA module parallels five branches for exploiting contextual properties at different scales. The output feature semantics in each branch are conveyed downward for feature semantic augmentation to achieve a multiscale perspective in each branch.

between the first and the second branches. In our architecture, we apply a 3 × 3 dilated convolution with dilated rates {1, 2, 4, 8} to the last four branches. Next, the feature maps encoding different scales of contextual information from all the branches are concatenated and further integrated via a 1 × 1 convolution. Eventually, a max-pooling operation is applied to convert the feature map to the desired size. Based on the MSCA module for feature downscaling, the feature details from higher-resolution branches can be well maintained.

As illustrated by Fig. 3(a), when paralleling a new lower-resolution branch, we first downscale the feature maps from the existing branches into the identical size expected by the new branch based on the MSCA module. Then, the downscaled feature maps are concatenated, rather than summed up like that in the HRNet, and aggregated via a 1 × 1 convolution to produce the primary feature map in the new branch. By using concatenation operation, the feature semantic distinctions from different resolutions can be well maintained, thereby improving the feature representation quality. Similarly, as illustrated by Fig. 3(b) to (d), when carrying out cross-branch feature fusion, first, the feature maps from the remaining branches are downscaled or upscaled into the identical size desired by the target branch. Next, the scale-modulated feature maps alongside with the feature map in the target branch are concatenated and aggregated via a 1 × 1 convolution, resulting in a semantic-augmented feature representation in the target branch. Here, the downscaling and upscaling operations are, respectively, implemented by the MSCA module and the deconvolution operation.

3.4. Feature attention module

As demonstrated in the literature, the pure convolution operations individually behave less excellently to characterize the channel informativeness and the spatial saliency, thereby going against to obtain high-quality robust feature representations [55]. In this paper, targeting at further upgrading the feature semantics in all branches of the feature extraction backbone, we design an effective feature attention module for emphasizing the important semantics and suppressing the helpless ones. The attentive feature map in each branch is further leveraged for cross-branch feature augmentation. As shown by Fig. 4, the feature attention module cascades a channel-specific attention unit and a spatial-specific attention unit. The channel-specific attention unit functions to emphasize the informative channels by exploiting the intra-channel correlations, while the spatial-specific attention unit functions to salient the important positions by exploiting the inter-channel dependencies.

As shown by the first part in Fig. 4, the channel-specific attention unit first performs two 1 × 1 convolutions on the input feature map, resulting in two feature representations $\mathrm{F}_1 \in \mathrm{R}^{H\times W\times C}$ and $\mathrm{S}_\mathrm{W} \in \mathrm{R}^{H\times W\times 1}$, where H, W, and C represent the height, width, and number of channels. The positions in F_1 encode the feature responses corresponding to the same positions in the input feature map, while S_W can be treated as a spatial weight map reflecting the feature significance of different positions. For facilitating exploiting channel-wise feature correlations, we reshape F_1 into a feature matrix $\mathbf{V}_1 \in \mathrm{R}^{C\times N}$ and reshape S_W into a column vector $\boldsymbol{W} \in \mathrm{R}^{N\times 1}$, where N=$H \times W$. Next, $\mathbf{V}_1$ is multiplied with $\boldsymbol{W}$ to produce a channel attention vector $\boldsymbol{CA} \in \mathrm{R}^{C\times 1}$ by weighted aggregating the feature semantics in each channel with the comprehensive consideration of their correlations. Specifically, $\boldsymbol{W}$ is activated with a softmax function before performing matrix multiplication to normalize the contribution of the feature semantic at each position. Here, being activated by a sigmoid function, the elements in $\boldsymbol{CA}$ encode the feature informativeness related to the channels in the input feature map. Eventually, by channel-wisely multiplying the input feature map with the channel attention vector $\boldsymbol{CA}$, we attain the quality-enhanced feature representation $\mathrm{FC} \in \mathrm{R}^{H\times W\times C}$, whose informative channels are explicitly attended and emphasized.

As shown by the second part in Fig. 4, the spatial-specific attention unit takes the output of the channel-specific attention unit (i.e. FC) as the input to further attend to spatially salient feature semantics. Concretely, first, two 1 × 1 convolutions are performed on FC to produce two feature representations $\mathrm{F}_2 \in \mathrm{R}^{H\times W\times C}$ and $\mathrm{F}_\mathrm{Q} \in \mathrm{R}^{H\times W\times C}$. Then, a global average pooling (GAP) operation is operated on F_Q, resulting in a feature vector $\boldsymbol{CW} \in \mathrm{R}^{1\times C}$ by combining the channel-wise feature semantics. Similarly, the positions in F_2 encode the feature responses corresponding to the same positions in FC, while $\boldsymbol{CW}$ can be treated as a channel weight map reflecting the feature importance of different channels. For facilitating exploiting cross-channel feature dependencies, we reshape F_2 into a feature matrix $\mathbf{V}_2 \in \mathrm{R}^{C\times N}$. Then, after being activated with the softmax function for normalizing the contribution of each channel, $\boldsymbol{CW}$ is multiplied with $\mathbf{V}_2$ to comprehensively take into account the channel dependencies. Next, after conducting reshaping on the product matrix, we obtain the spatial attention map $\mathrm{SA} \in \mathrm{R}^{H\times W\times 1}$. Here, being activated by a sigmoid function, the elements in SA indicate the feature saliencies related to the positions in FC. Eventually, by element-wisely multiplying FC with the spatial attention map SA channel by channel, we attain the quality-enhanced feature representation, whose spatially salient feature semantics are explicitly attended and highlighted.

3.5. Segmentation head

As illustrated by Fig. 1, the feature extraction backbone outputs a group of feature maps having different resolutions to supervise the generation of the segmentation map. Generally, the low-resolution feature map can well suppress the impacts of the noises, thereby favorable to handle the interior texture heterogeneities of the medical targets. By contrast, the high-resolution feature map has excellent properties to characterize details, thereby beneficial to locate accurate boundaries of the medical targets. Hence, taking advantage of the multi-resolution feature maps, the segmentation head is formulated as a hierarchical structure to progressively refine the higher-branch feature

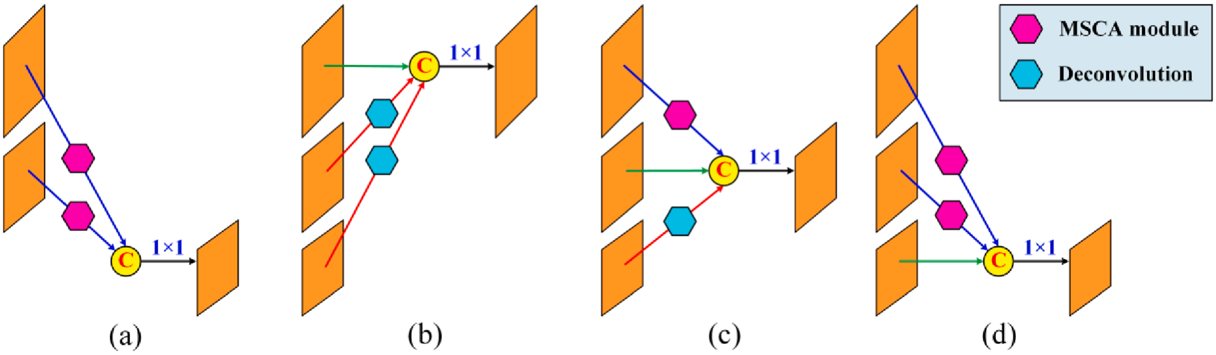


Fig. 3. Illustration of (a) connecting a lower-resolution branch and cross-branch feature fusion to generate (b) high-resolution, (c) medium-resolution, and (d) low-resolution feature maps with the MSCA module and the deconvolution operation.

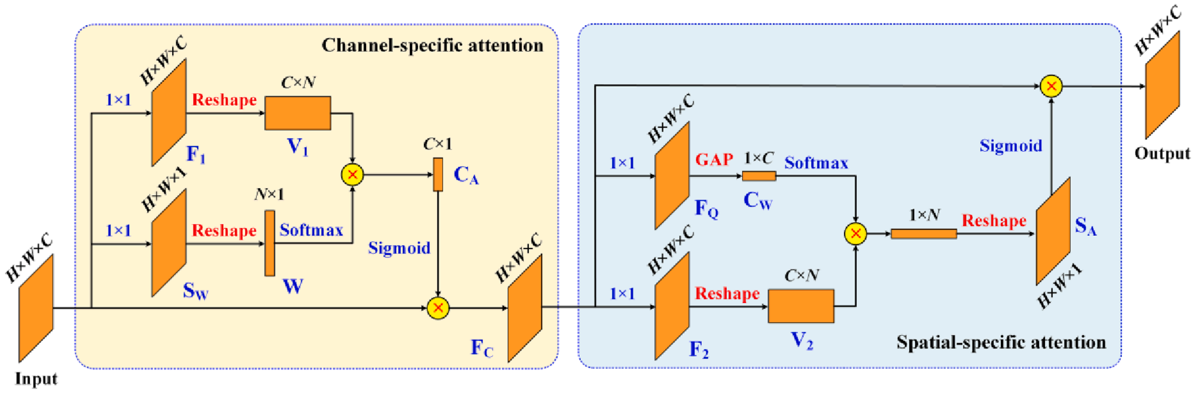


Fig. 4. Structure of the feature attention module. The feature attention module cascades a channel-specific attention unit and a spatial-specific attention unit for highlighting the important channels and salient regions, respectively.

maps in a bottom-up manner based on the lower-branch predictions. To be specific, first, an initial prediction map is generated based on the feature map output by Branch 4. The prediction map involves two channels with a binary classification output: one for the foreground and one for the background. That is, for each position in the prediction map, the channel with the larger output indicates the category label of this position. Then, the prediction map alongside with the feature map from the current branch are fed into a semantic-level context exploitation (SLCE) module to produce a contextual category attentive semantic feature map. This feature map is further concatenated to and integrated with the feature map exported by the upper branch to serve feature augmentation. Afterwards, the augmented feature map is leveraged to produce a higher-resolution prediction map in the upper branch. As illustrated by Fig. 1, the above process is repeated bottom-up to gradually augment the feature maps in the upper branches. Eventually, a high-quality prediction map is produced in Branch 1 to produce the final segmentation output.

The SLCE module functions to aggregate the contextual properties from individual categories with a global perspective to provide a uniform semantic representation for the individual categories across the entire feature map. As illustrated by Fig. 5, the SLCE module takes the current-branch prediction map P and feature map F as the input and outputs a contextual category attentive semantic feature map $\mathrm{F_S}$. First, two index sets ψ_F and ψ_B are defined to encapsulate the positions on the prediction map P that are predicted as the foreground and the background categories, respectively. Then, to provide a uniform semantic representation for each category, we aggregate the semantic-level contextual properties in each category as follows:

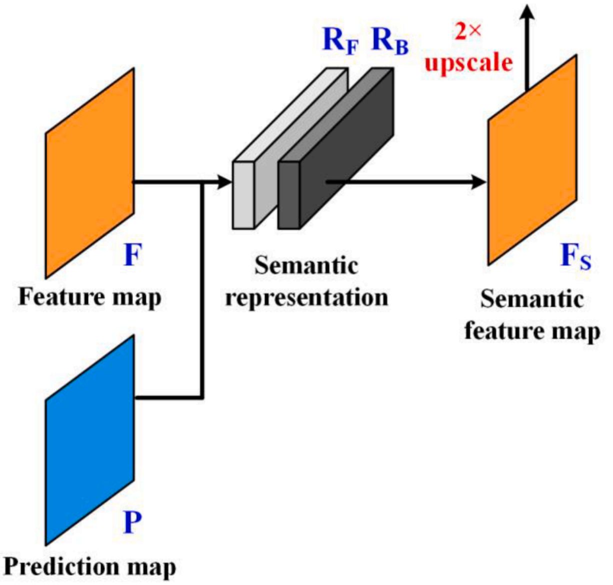


Fig. 5. Structure of the semantic-level context exploitation (SLCE) module. The SLCE module takes the feature map and prediction map as the input to compute the semantic representations for both the foreground and the background categories and outputs a contextual category attentive semantic feature map.

$$\mathrm{R_F} = \sum_k \frac{e^{\mathrm{P}[\psi_\mathrm{F}[\mathrm{k}],1]}}{\sum_n e^{\mathrm{P}[\psi_\mathrm{F}[\mathrm{n}],1]}} \mathrm{F}[\psi_\mathrm{F}[\mathrm{k}]] \tag{1}$$

$$\mathrm{R_B} = \sum_k \frac{e^{\mathrm{P}[\psi_\mathrm{B}[\mathrm{k}],0]}}{\sum_n e^{\mathrm{P}[\psi_\mathrm{B}[\mathrm{n}],0]}} \mathrm{F}[\psi_\mathrm{B}[\mathrm{k}]] \tag{2}$$

where $\psi_\mathrm{F}[\mathrm{k}]$ and$\psi_\mathrm{B}[\mathrm{k}]$, respectively, denote the k-th element in the corresponding index set; $\mathrm{R_F}$ and $\mathrm{R_B}$ represent the obtained semantic representations shown in Fig. 5 for the foreground and the background, respectively. Finally, these two semantic representations are leveraged to construct the semantic feature map $\mathrm{F_S}$ as follows:

$$\begin{cases} \mathrm{F_S}[\psi_\mathrm{F}[\mathrm{k}]] = \mathrm{R_F}, \ \mathrm{k} = 1, 2, 3, \cdots \\ \mathrm{F_S}[\psi_\mathrm{B}[\mathrm{k}]] = \mathrm{R_B}, \ \mathrm{k} = 1, 2, 3, \cdots \end{cases} \tag{3}$$

The semantic feature map FS output by the SLCE module will be upscaled to its twice spatial size to serve feature augmentation in the upper branch.

3.6. Loss function

As depicted by Fig. 1, the AttHRNet outputs a separate prediction map in each branch, which can generate an individual segmentation result at that resolution. Since the prediction map contributes to the feature augmentation in the upper branch to supervise the generation of a finer prediction map, the quality of the prediction map impacts significantly on the final segmentation performance. Thus, each branch of the AttHRNet should be explicitly supervised during the network optimization process. To this end, each branch is assigned with a binary ground-truth segmentation map, where the positions with a value of 1 indicate the foreground areas and the positions with a value of 0 indicate the background areas. The loss function used to direct the optimization of the AttHRNet is formulated as the weighted summation of the losses from all the branches as follows:

$$L = \sum_{i=1}^{4} 2^{i-4} \left(L_{\mathrm{FL}}^{i} + L_{\mathrm{IoU}}^{i} \right) \tag{4}$$

where L_{FL}^{i} denotes the focal loss [75] item defined by the softmax predictions corresponding to the annotated ground truths in the i-th branch; L_{IoU}^{i} denotes the intersection over union (IoU) loss [76] item between the binary segmentation results and the ground-truth segmentation map in the *i*-th branch. Considering the size difference of the output in each branch, the weighting coefficient $2^{i\text{-}4}$ is introduced to balance the contributions of different branches to the summation of the total loss.

3.7. Implementation details

All branches of the AttHRNet were concurrently constructed using the Adam optimizer on a cloud computing platform equipped with a 128-GB memory, ten 16-GB GPUs, and a 16-core CPU. The reason of the selection of the Adam optimizer is that it usually behaves promisingly with excellent convergence performance on the data with complex distributions and at the saddle point. Before training, the network parameters in each layer of the AttHRNet were initialized at random by getting parameters from a zero-mean Gaussian distribution with a standard deviation of 0.01. Then, 1000 epochs were intently optimized to construct the AttHRNet according to the loss function in Eq. (4) with six image patches per batch on each GPU. During training, the learning rate was initially configured to be 0.001 and gradually decreased by the coefficient of 0.1 every 400 epochs. Specifically, data augmentation was also conducted on the training data aiming at promoting the model robustness. First, each training patch was flipped horizontally to produce a horizontal duality image. Then, the training patch alongside with its coupled horizontal duality image were, respectively, rotated anticlockwise with a step interval of 90 degrees. Consequently, eight training samples were obtained for each training patch.

4. Results and discussions

4.1. Datasets

To provide convincing evidence to examine the segmentation performance of the constructed AttHRNet, we conducted intensive confirmatory experiments on four publicly released medical image segmentation datasets, which are all suitable and large enough to construct and test deep learning models.

The first dataset is the lung image database consortium and image database resource initiative (LIDC-IDRI) dataset [77]. The LIDC-IDRI dataset consists of 885 diagnostic and lung cancer screening thoracic CT scans, which were captured by using four different types of scanner models. Each scan involves hundreds of slices with thickness ranging from 0.6 mm to 5 mm and with marked-up annotations of the lung regions, as well as the marked lesions belonging to three nodule categories. It was established collaboratively by eight medical imaging companies and seven academic centers. Each image in the LIDC-IDRI dataset has the identical size of 512 × 512 pixels. This dataset was used to conduct lung instance segmentation.

The second dataset is the 2019 kidney and kidney tumor segmentation challenge (KiTS19) dataset [78]. The KiTS19 dataset includes tens of thousands of CT scan images collected from 300 patients who were treated with partial or radical nephrectomy at the University of Minnesota Medical Center between 2010 and 2018. All the cross sectional CT images, more than half of which were acquired across more than 50 referring institutions, in the KiTS19 dataset have the same size of 512 × 512 pixels and were manually annotated with semantic segmentation masks of the kidney areas and the kidney tumor regions. This dataset was used to conduct kidney instance segmentation.

The third dataset is the Kvasir polyp segmentation (Kvasir-SEG) dataset [79]. The Kvasir-SEG dataset was collected from different colorectal cancer screening patients through colonoscopy by experienced gastroenterologists from Vestre Viken Health Trust in Norway. It comprises 1000 polyp image samples and the associated instance-level polyp annotations including the pixel-wise annotations of the polyp regions and the bounding box annotations of the polyp instances. The image size varies within the range of 332 × 487 pixels to 1920 × 1072 pixels. This dataset was used to conduct polyp instance segmentation.

The last dataset is the international skin imaging collaboration (ISIC) dataset [80]. The ISIC dataset comprises 2750 dermoscopy images collected with different devices in several international clinical institutions for melanoma diagnosis. The image size varies within the range of 540 × 722 pixels to 4499 × 6748 pixels. All the images in the ISIC dataset were annotated with region-wise lesion masks and grouped into different categories of melanoma conditions. This dataset was used to conduct skin lesion segmentation.

4.2. Quantitative evaluation metrics

To provide quantitative verifications on the segmentation performance of the proposed AttHRNet when dealing with different medical image segmentation tasks, we employed the following four widely used quantitative assessment indexes: precision, recall, Jaccard index (JI), and Dice coefficient (DC). Thereinto, precision and recall indexes, respectively, examine the capabilities of the segmentation model in correctly reducing the background interferences and in completely retrieving the foreground contents. JI and DC indexes measure the segmentation model from an overall perspective by comprehensively evaluating its capability in the suppression of both the false negatives and false positives. These four quantitative assessment indexes are computed as follows:

$$precision = \frac{TP}{FP + TP} \times 100\backslash\% \tag{5}$$

$$recall = \frac{TP}{FN + TP} \times 100\backslash\% \tag{6}$$

$$JI = \frac{TP}{FP + FN + TP} \times 100\backslash\% \tag{7}$$

$$DC = \frac{2 \times precision \times recall}{precision + recall} \times 100\backslash\% \tag{8}$$

where the numbers of true positives, false positives, and false negatives are, respectively, denoted by *TP*, *FP*, and *FN*.

4.3. Medical image segmentation

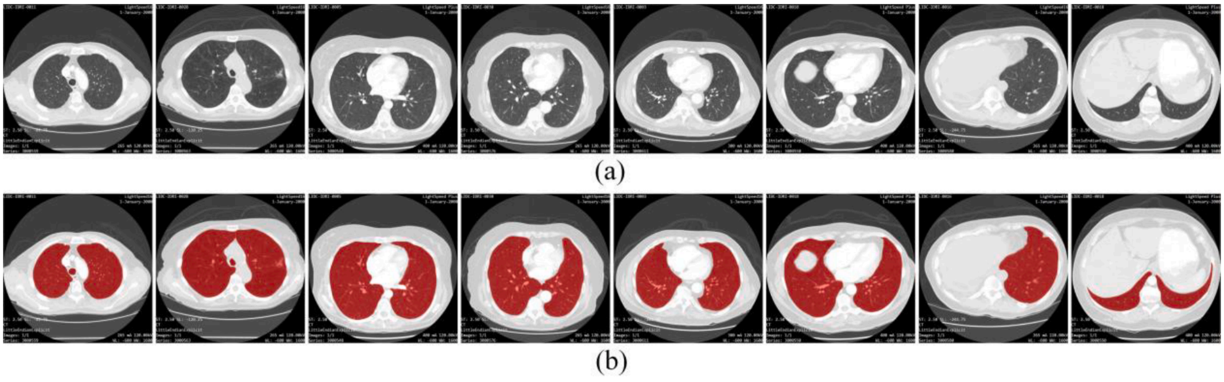
The medical image segmentation results on the four test datasets quantitatively evaluated by the four indexes are recorded in Table 2. As reflected in Table 2, the AttHRNet attained excellent and competitive performance in segmenting different types of medical targets on the four test datasets. Specifically, a segmentation accuracy with a precision of 98.91 %, a recall of 98.52 %, a JI of 97.46 %, and a DC of 98.71 %, respectively, was obtained on the LIDC-IDRI dataset in segmenting lungs. A segmentation accuracy with a precision of 98.52 %, a recall of 97.94 %, a JI of 96.52 %, and a DC of 98.23 %, respectively, was achieved on the KiTS19 dataset in segmenting kidneys. For the Kvasir-SEG dataset, a segmentation performance with a precision, a recall, a JI, and a DC of 93.13 %, 94.87 %, 88.66 %, and 93.99 %, respectively, was obtained in segmenting polyps. For the ISIC dataset, a segmentation performance with a precision, a recall, a JI, and a DC of 90.35 %, 95.98 %, 87.06 %, and 93.08 %, respectively, was achieved in segmenting skin lesions. Comparatively, the best segmentation performance appeared on the LIDC-IDRI dataset with a DC of 98.71 %, while the worst segmentation performance appeared on the ISIC dataset with a DC of 93.08 %. Nevertheless, the proposed AttHRNet still behaved promisingly with an acceptable segmentation accuracy on the ISIC dataset due to the remarkably challenging and complex conditions of the skin lesions in comparison with the medical targets in the other three datasets. Note that, for the LIDC-IDRI and KiTS19 datasets, the value of the precision index was slightly higher than that of the recall index, whereas the value of the recall index was slightly higher than that of the precision index on the Kvasir-SEG and ISIC datasets. In fact, the precision index indicates

Table 2
Segmentation performances of different models.

Model	Dataset	Precision (%)	Recall (%)	JI(%)	DC (%)
AttHRNet	LIDC-IDRI	**98.91**	**98.52**	**97.46**	**98.71**
	KiTS19	**98.52**	**97.94**	**96.52**	**98.23**
	Kvasir-SEG	**93.13**	**94.87**	**88.66**	**93.99**
	ISIC	**90.35**	**95.98**	**87.06**	**93.08**
CA-Net [26]	LIDC-IDRI	97.54	97.03	94.71	97.28
	KiTS19	97.46	96.98	94.59	97.22
	Kvasir-SEG	90.86	92.68	84.78	91.76
	ISIC	87.90	93.97	83.21	90.83
MultiResUNet [30]	LIDC-IDRI	96.88	96.31	93.41	96.59
	KiTS19	96.62	96.15	93.02	96.38
	Kvasir-SEG	89.03	91.36	82.12	90.18
	ISIC	85.23	93.34	80.34	89.10
HMEDN [36]	LIDC-IDRI	93.06	91.33	85.51	92.19
	KiTS19	92.98	91.27	85.39	92.12
	Kvasir-SEG	83.15	89.41	75.70	86.17
	ISIC	79.24	91.97	74.11	85.13
PyDiNet [39]	LIDC-IDRI	96.56	95.88	92.71	96.22
	KiTS19	96.12	95.67	92.11	95.89
	Kvasir-SEG	88.74	91.05	81.62	89.88
	ISIC	84.92	93.11	79.90	88.83
HR-CapsSegNet [41]	LIDC-IDRI	98.72	98.34	97.10	98.53
	KiTS19	97.94	97.21	95.26	97.57
	Kvasir-SEG	92.50	94.66	87.91	93.57
	ISIC	89.76	95.31	85.96	92.45
SegCaps [49]	LIDC-IDRI	97.22	96.74	94.14	96.98
	KiTS19	97.13	96.66	93.98	96.89
	Kvasir-SEG	89.37	91.54	82.55	90.44
	ISIC	86.59	93.62	81.77	89.97
FFANet [50]	LIDC-IDRI	93.79	92.05	86.76	92.91
	KiTS19	93.58	91.96	86.50	92.76
	Kvasir-SEG	85.71	90.33	78.51	87.96
	ISIC	80.41	92.83	75.71	86.17

the rate of the falsely identified background elements. That is, the higher the value of the precision index, the less the false positives. In contrast, the recall index indicates the rate of the correctly identified foreground elements. That is, the higher the value of the recall index, the less the false negatives. Thus, comparatively speaking, the AttHRNet generated more false negatives on the LIDC-IDRI and KiTS19 datasets, while introducing more false positives on the Kvasir-SEG and ISIC datasets. The false negatives generated on the LIDC-IDRI and KiTS19 datasets were mainly caused by the presence of large-size nodule-like structures inside the lung and kidney regions or due to the blurred boundaries exhibiting quite low contrasts at some sections. As a result, the integrities of these targets were not well maintained, thereby leading to the decline of the recall index. On the contrary, the false positives generated on the Kvasir-SEG and ISIC datasets were mainly caused by the indistinguishable borderlines or the extremely similar textural properties of the targets to their surroundings. As a result, some background elements were falsely recognized as the foreground, thereby leading to the decline of the precision index. Even so, the proposed AttHRNet still performed effectively on the four test datasets as reflected by the overall quantified assessment results with respect to the JI and DC indexes.

To sum up, the challenging conditions of the four test datasets were reflected in the following aspects. (1) The medical targets vary greatly in size and shape within the same dataset, such as the polyps in the Kvasir-SEG dataset (Fig. 8) and the skin lesions in the ISIC dataset (Fig. 9). Specifically, some polyps and skin lesions have very small sizes. (2) The same type of medical targets exhibits severe appearance inconsistencies with different colors and textural properties, such as the greenish, reddish, yellowish, or whitish polyps (Fig. 8) and the blackish, reddish, brownish, or purplish skin lesions (Fig. 9). (3) Some medical targets show extremely low contrasts with their surrounding tissues, such as the polyps in the Kvasir-SEG dataset (Fig. 8) and the skin lesions in the ISIC dataset (Fig. 9). Specifically, some polyps and skin lesions show quite similar appearances in color and texture to their surrounding tissues. (4) Some medical targets present considerably obscure boundaries. For instance, some skin lesions exhibit a fade-away pattern at the border areas (Fig. 9) and some polyps are directly connected to the intestine with a protuberance with no borderlines (Fig. 8). (5) Some medical targets suffer from interior anomaly contaminations, such as the nodule-like structures in the lung (Fig. 6) and kidney images (Fig. 7) and the hairs in the skin lesion images (Fig. 9). (6) Some background regions have similar properties to the medical targets, such as the folds of the intestines in the polyp images (Fig. 8). All of the above issues bring about remarkable ordeals to the correct localization and accurate segmentation of the medical targets, thereby impeding the upgradation of the segmentation performance. Fortunately, the proposed AttHRNet still behaved excellently with competitive segmentation accuracies when processing the different types of medical images towards instance segmentation. The segmentation superiority of the proposed AttHRNet benefitted from the following factors. First, by stacking an improved HRNet structure augmented with the MSCA module as the feature extraction backbone, a set of multiscale high-level feature maps with strong and spatially accurate semantics are obtained for supervising the segmentation map generation. Second, by designing an effective feature attention module and integrating it into all branches of the feature extraction backbone, the feature encoding quality can be repeatedly promoted by simultaneously attending to the informative feature channels and the salient spatial locations. Last but not least, by employing a hierarchical strategy as the segmentation head, the multi-


(a)

(b)

Fig. 6. Segmentation results on the LIDC-IDRI dataset. (a) Test images and (b) overlaid segmentation results.

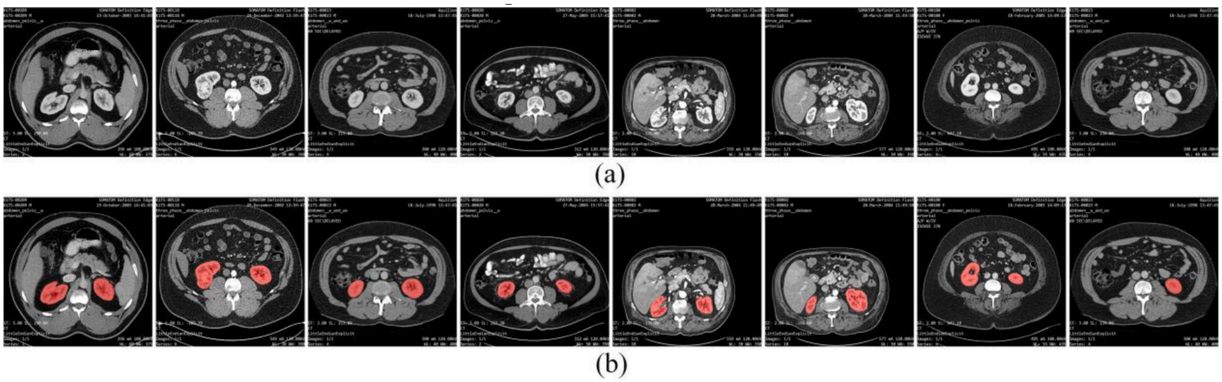


Fig. 7. Segmentation results on the KiTS19 dataset. (a) Test images and (b) overlaid segmentation results.

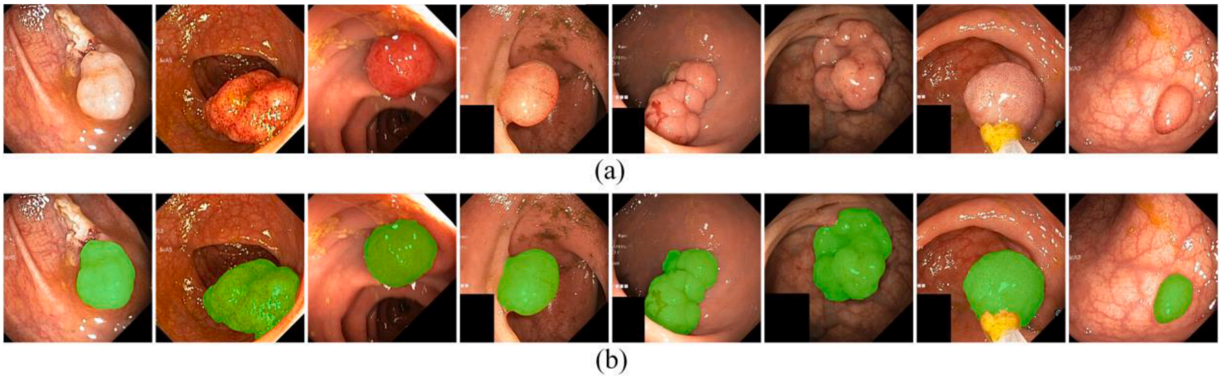


Fig. 8. Segmentation results on the Kvasir-SEG dataset. (a) Test images and (b) overlaid segmentation results.

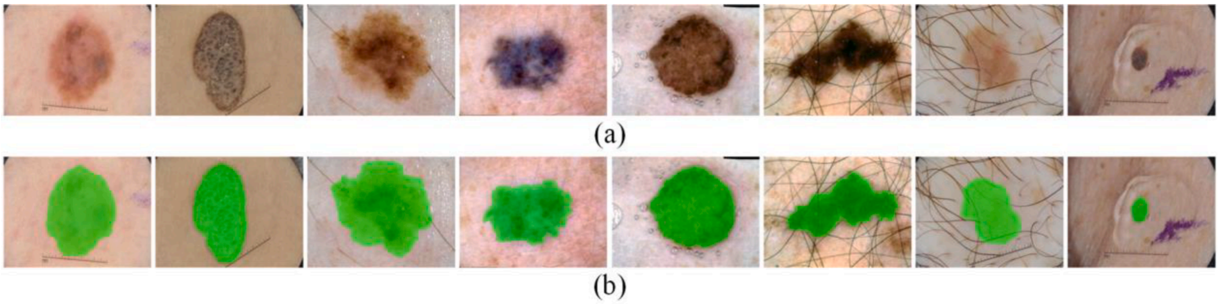


Fig. 9. Segmentation results on the ISIC dataset. (a) Test images and (b) overlaid segmentation results.

resolution feature maps can be progressively refined to focus more on the task-specific semantic regions, thereby producing a high-quality segmentation output.

For qualitative inspection purposes, Figs. 6 to 9 present a subset of medical target segmentation illustrations from the four test datasets. As shown by these figures, the medical targets of different types and varying self and background conditions were nicely located and segmented with a very small quantity of missing detections and incorrect identifications. The outlines were well delineated and the boundaries were nicely adhered. Specifically, as shown in Figs. 6 and 7, the disturbances of the small-size nodule-like structures inside the lung and kidney regions were well suppressed, guaranteeing the solidness of these instances. As shown in Figs. 8 and 9, the polyps and skin lesions were excellently distinguished from their surrounding tissues without introducing too many false positives in spite of the obscure boundary properties. Moreover, the small-size polyps and skin lesions were also correctly recognized and promisingly segmented. However, as shown in Figs. 6 and 7, some lungs and kidneys contained quite large-size nodule-like structures in some slices, which formed quite strong color contrasts and showed extremely similar texture properties to the background. Consequently, the completeness of these instances was not successfully guaranteed, resulting in hole-like phenomena in the segmentation results. Overall speaking, through quantitative and qualitative verifications, it confirmed the promising performance of the proposed AttHRNet in medical image segmentation tasks.

4.4. Comparative studies

Aiming at providing more convincing evidences to testify the feasibility and effectiveness of the proposed AttHRNet in medical target segmentation tasks, we also performed a group of intensive segmentation tests using the recently developed state-of-the-art deep learning models. The models involved in the comparative analyses included the comprehensive attention network (CA-Net) [26], the dilated multiresidual blocks network (MultiResUNet) [30], the high-resolution multiscale encoder-decoder network (HMEDN) [36], the PyDiNet [39], the

HR-CapsSegNet [41], the SegCaps [49], and the FFANet [50]. Amongst these models, the MultiResUNet and SegCaps employed the U-Net architecture with a contraction pathway for feature abstraction and an expansion pathway for segmentation map recovery. However, the MultiResUNet was constructed with scalar primitives, whereas the SegCaps was constructed with capsule primitives. They represented two different styles of the U-Net architecture. The CA-Net, HMEDN, and FFANet followed a general encoder-decoder architecture with different network architecture styles and functional modules. Specifically, a high-resolution pathway stacked by residual dilated convolution blocks was designed in the HMEDN for feature skip connection. Attention mechanism was positively integrated into the CA-Net for feature semantic promotion, and multi-level feature fusion techniques were leveraged in the FFANet for feature semantic augmentation. The PyDiNet and HR-CapsSegNet were formulated with the FPN and HRNet architectures, respectively, for integrating multilevel feature semantics. To achieve fair comparisons on the same baseline, all these models were optimized and tested on the four datasets used in this paper. Likewise, the same quantitative evaluation indexes were leveraged to examine and analyze their segmentation performances. The detailed segmentation results of these models are quantitatively reported in Table 2.

As reflected in Table 2, superior segmentation performances were achieved on the four test datasets by the HR-CapsSegNet, CA-Net, SegCaps, MultiResUNet, and PyDiNet with respect to the overall indexes JI and DC. By contrast, relatively lower segmentation accuracies were obtained by the FFANet, and HMEDN. To be specific, the HR-CapsSegNet outperformed the other models, while the HMEDN behaved less effectively than the other models. Besides, for all the models, the best segmentation performance appeared on the LIDC-IDRI dataset, whereas the worst segmentation performance appeared on the ISIC dataset. The advantageous performance of the HR-CapsSegNet was achieved by employing the HRNet architecture for semantically-strong feature extraction under multiple resolutions. The performance gains of the CA-Net benefitted from the design of the three types of attention modules used for recalibrating the feature semantics at channel, spatial, and scale levels, thereby significantly boosting the feature representation quality. The superior performance of the SegCaps consisted in the use of the tensor-form capsule primitives for abstracting high-order entity-aware feature representations. In addition, the advantageous performance of the MultiResUNet owed to the dense residual blocks and the dilated convolutions for contextual feature exploitation and feature semantic promotion. However, compared with the segmentation performances of these seven models, our proposed AttHRNet showed significantly competitive and distinctly advantageous segmentation performances on all the four test datasets with regard to the overall indexes JI and DC. This performance superiority convinced the powerful architecture of the AttHRNet, which was built with the improved HRNet backbone for multi-resolution feature extraction, the effective feature attention module for feature semantic promotion, and the advanced hierarchical formulation for segmentation map generation. In conclusion, according to contrastive analyses, we confirmed that the developed AttHRNet provided a feasible and effective solution to medical image segmentation tasks.

4.5. Ablation studies

To further evaluate the effectiveness of the key modules in the AttHRNet, we also conducted a group of ablation studies with a series of modified models based on the AttHRNet. As the first set of ablation studies, we examined the contributions of the MSCA module, the feature attention module, and the SLCE module to the performance gains of the AttHRNet. To this end, first, we removed the MSCA module in the feature extraction backbone and used the original strided convolutions in the HRNet to perform feature downscaling. We named the modified architecture as the AttHRNet-1. Second, we removed the feature attention module in the feature extraction backbone to cancel the channel and spatial feature recalibrations. We named the modified architecture as the AttHRNet-2. Third, we removed the SLCE module along with the hierarchical segmentation structure, and directly applied the feature map generated in Branch 1 to predict the segmentation result. We named the modified architecture as the AttHRNet-3. Finally, as a simplified version of the hierarchical segmentation structure, we removed the SLCE module and directly upscaled the prediction map in the lower resolution branch and concatenated it with the feature map in the upper branch to serve feature augmentation. We named the modified architecture as the AttHRNet-4. The quantitative evaluation results of these modified models on the four test datasets are reported in detail in Table 3. As reflected in Table 3, obviously, with the removals or the simplification of these key modules, all the modified models performed less promisingly with significant lower segmentation accuracies compared with the AttHRNet. For the AttHRNet-1, without the MSCA module, the feature detail loss caused by the strided convolutions affected the quality and informativeness of the resultant feature semantics in the lower resolution branches. For the AttHRNet-2, without the feature attention module for channel and spatial feature recalibrations, the important channels and salient regions were not well highlighted, thereby affecting the extraction of high-quality and distinctive feature representations. Note that, the AttHRNet-4 behaved better than the AttHRNet-3. It indicated that the inclusion of the information from lower resolution segmentations to inform higher resolution segmentations meant significantly to provide valuable evidences to upgrade the segmentation accuracy. For clear visual comparisons, Fig. 10 also show the feature saliency maps generated with and without these key modules. Obviously, the MSCA module was beneficial to the identification of the small-size targets, the feature attention module functioned excellently to adhere tighter boundaries of the targets with low contrasts, and the SLCE module performed promisingly to suppress the impacts of the textural inconsistencies. In conclusion, we confirmed that the MSCA module, the feature attention module, and the SLCE module contributed positively and significantly to the performance gains of the AttHRNet by,

Table 3
Segmentation performances of different modified models.

Model	Dataset	Precision(%)	Recall(%)	JI(%)	DC(%)
AttHRNet-1	LIDC-IDRI	98.58	98.17	96.80	98.37
	KiTS19	98.21	97.55	95.85	97.88
	Kvasir-SEG	92.78	94.48	88.01	93.62
	ISIC	89.94	95.53	86.31	92.65
AttHRNet-2	LIDC-IDRI	96.92	96.35	93.49	96.63
	KiTS19	96.67	96.21	93.12	96.44
	Kvasir-SEG	89.11	91.42	82.23	90.25
	ISIC	85.33	93.46	80.52	89.21
AttHRNet-3	LIDC-IDRI	97.68	97.28	95.08	97.48
	KiTS19	97.65	97.14	94.92	97.39
	Kvasir-SEG	91.67	93.51	86.19	92.58
	ISIC	88.62	94.57	84.33	91.50
AttHRNet-4	LIDC-IDRI	98.04	97.66	95.79	97.85
	KiTS19	97.84	97.22	95.18	97.53
	Kvasir-SEG	92.23	93.94	87.05	93.08
	ISIC	89.26	95.06	85.30	92.07
AttHRNet-SE	LIDC-IDRI	97.61	97.22	94.96	97.41
	KiTS19	97.58	97.14	94.85	97.36
	Kvasir-SEG	91.35	93.23	85.67	92.28
	ISIC	88.31	94.46	83.96	91.28
AttHRNet-CA	LIDC-IDRI	97.75	97.34	95.21	97.54
	KiTS19	97.70	97.18	95.01	97.44
	Kvasir-SEG	91.74	93.62	86.34	92.67
	ISIC	88.76	94.73	84.58	91.65
AttHRNet-CBAM	LIDC-IDRI	97.55	97.10	94.79	97.32
	KiTS19	97.52	97.03	94.69	97.27
	Kvasir-SEG	90.88	92.71	84.82	91.79
	ISIC	87.94	94.08	83.33	90.91
AttHRNet-DA	LIDC-IDRI	98.36	97.91	96.34	98.13
	KiTS19	97.95	97.26	95.32	97.60
	Kvasir-SEG	92.57	94.23	87.60	93.39
	ISIC	89.68	95.27	85.86	92.39

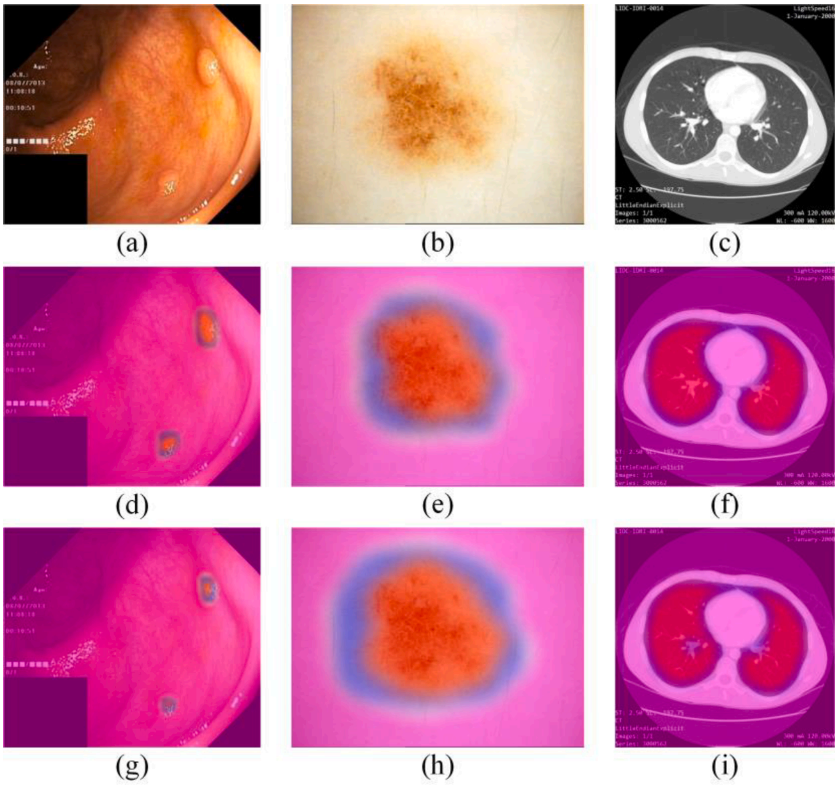


Fig. 10. Illustrations of feature saliency maps generated with different models. (a), (b), and (c) Sample images and feature saliency maps generated (d) with the MSCA model, (e) with the feature attention module, (f) with the SLCE module, (g) without the MSCA module, (h) without the feature attention module, and (i) without the SLCE module.

respectively, reducing the information loss, promoting the feature representation quality, and improving the segmentation accuracy.

As the second set of ablation studies, we further analyzed the superiority of the proposed feature attention module by comparing it with some existing popularly used feature attention mechanisms. The following four attention mechanisms were considered: SE block [56], coordinate attention [59], CBAM [66], and DA module [67]. To be specific, we substituted the proposed feature attention module in the AttHRNet with the SE block, coordinate attention, CBAM, and DA module, respectively, to construct four modified architectures. These modified architectures were named as the AttHRNet-SE, AttHRNet-CA, AttHRNet-CBAM, and AttHRNet-DA, respectively. The quantitative evaluation results of these modified models on the four test datasets are reported in detail in Table 3. As reflected in Table 3, the AttHRNet-DA achieved the best segmentation performance among the four modified models, whereas the AttHRNet-CBAM behaved less promisingly. Besides, the AttHRNet-CA performed slightly better than the AttHRNet-SE due to the embedding of the spatial position information when exploiting the channel feature saliencies. The performance advantage of the AttHRNet-DA benefitted from the simultaneous consideration and integration of both the channel and spatial feature significances like that in our proposed feature attention module, thereby effectively promoting the feature representation quality. Nevertheless, the AttHRNet with the proposed feature attention module demonstrated significant segmentation accuracy improvement compared with these four modified models, which convinced the effectiveness and superiority of the proposed feature attention module. Moreover, designed with a lightweight architecture, the proposed feature attention module also showed a higher efficiency than the DA module, which required a set of complex matrix multiplication operations.

5. Conclusion

This paper has built an advanced network architecture, termed as AttHRNet, for segmenting targets of interest from medical images. The AttHRNet employed a one-stage semantic segmentation pipeline and consisted of three primary components, including a feature extraction backbone, a feature attention module, and a segmentation head. The progresses of the AttHRNet lied in the MSCA module for preserving more feature details, the feature attention module for highlighting significant feature semantics, the SLCE module for exploiting category-aware feature semantics, the improved HRNet backbone for generating high-quality feature representations, and the hierarchical segmentation scheme for improving segmentation accuracies. To be specific, the feature extraction backbone followed an improved HRNet architecture functioned with the MSCA module for reducing the feature detail loss during the cross-branch feature propagation process, thereby favorable to provide multiscale high-level feature maps with strong and spatially accurate semantics. The feature attention module cascaded a channel-specific attention unit and a spatial-specific attention unit for explicitly attending to the informative feature channels and the task-specific spatial locations, thereby beneficial to promote the feature encoding semantics in each branch of the feature extraction backbone. The segmentation head was designed as a hierarchical structure for progressively augmenting the output feature maps with semantic-level contextual properties, thereby finalizing a high-quality prediction map to improve the per-pixel segmentation accuracy. The AttHRNet has been intensively examined on four medical image datasets towards medical target segmentation. Quantitative assessments and visual verifications showed the promising and competitive performance of the AttHRNet in segmenting different-type medical targets with varying self-conditions in diverse background scenarios. Furthermore, comparative analyses with the state-of-the-art models and ablation experiments also demonstrated the significant advantages of the AttHRNet in medical image segmentation tasks.

Funding

This work was supported by the Natural Science Foundation of Jiangsu Province [grant numbers BK20211365, BK20191214]; the National Natural Science Foundation of China [grant numbers 62076107, 41971414, 51975239]; and the Six Talent Peaks Project in Jiangsu Province [grant number XYDXX-098].

CRediT authorship contribution statement

Yongtao Yu: Funding acquisition, Methodology, Software, Writing – original draft. **Yifei Tao:** Conceptualization, Investigation, Methodology, Validation. **Haiyan Guan:** Formal analysis, Funding acquisition, Supervision, Writing – review & editing. **Shaozhang Xiaov: . Fenfen Li:** Investigation, Methodology, Validation, Writing – original draft. **Changhui Yu:** Conceptualization, Investigation, Visualization, Writing – original draft. **Zuojun Liu:** Formal analysis, Methodology, Software, Writing – original draft. **Jonathan Li:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this article are publicly accessed data published by existing works.

References

[1] I. Alnazer, P. Bourdon, T. Urruty, O. Falou, M. Khalil, A. Shahin, C. Fernandez-Maloigne, Recent advances in medical image processing for the evaluation of chronic kidney disease, Med. Image Anal. 69 (2021), 101960.

[2] G. Litjens, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[3] I.R.I. Haque, J. Neubert, Deep learning approaches to biomedical image segmentation, Inf. Med. Unlocked 18 (2020), 100297.

[4] D.K. Patra, T. Si, S. Mondal, P. Mukherjee, Breast DCE-MRI segmentation for lesion detection by multi-level thresholding using student psychological based optimization, Biomed. Signal Process. Control 69 (2021), 102925.

[5] S. Chakraborty, K. Mali, A morphology-based radiological image segmentation approach for efficient screening of COVID-19, Biomed. Signal Process. Control 69 (2021), 102800.

[6] Y. Wang, et al., Segmentation of lumen and outer wall of abdominal aortic aneurysms from 3D black-blood MRI with a registration based geodesic active contour model, Med. Image Anal. 40 (2017) 1–10.

[7] P.P.R. Filho, P.C. Cortez, A.C.D.S. Barros, V.H.C. Albuquerque, J.M.R.S. Tavares, Novel and powerful 3D adaptive crisp active contour method applied in the segmentation of CT lung images, Med. Image Anal. 35 (2017) 503–516.

[8] J. Lu, C. Feng, J. Yang, W. Li, D. Zhao, C. Wan, Segmentation of the cardiac ventricle using two layer level sets with prior shape constraint, Biomed. Signal Process. Control 68 (2021), 102671.

[9] H. Chen, X. Pan, X. Lu, Q. Xie, A modified graph cuts image segmentation algorithm with adaptive shape constraints and its application to computed tomography images, Biomed. Signal Process. Control 62 (2020), 102092.

[10] I. Filali, M. Belkadi, R. Aoudjit, M. Lalam, Graph weighting scheme for skin lesion segmentation in macroscopic images, Biomed. Signal Process. Control 68 (2021), 102710.

[11] D. Jia, C. Zhang, N. Wu, Z. Guo, H. Ge, Multi-layer segmentation framework for cell nuclei using improved GVF snake model, watershed, and ellipse fitting, Biomed. Signal Process. Control 67 (2021), 102516.

[12] L. Fan, L. Shen, X. Zuo, Feature extraction and recognition of medical CT images based on Mumford-Shah model, Adv. Math. Phys. 2021 (2021) 1–13.

[13] Q. Huang, Y. Zhou, L. Tao, W. Yu, Y. Zhang, L. Huo, Z. He, A Chan-Vese model based on Markov chain for unsupervised medical image segmentation, Tsinghua Sci. Technol. 26 (6) (2021) 833–844.

[14] M. Tavakoli-Zaniani, Z. Sedighi-Maman, M.H.F. Zarandi, Segmentation of white matter, grey matter and cerebrospinal fluid from brain MR images using a modified FCM based on double estimation, Biomed. Signal Process. Control 68 (2021), 102615.

[15] P.M.M. Pereira, R. Fonseca-Pinto, R.P. Paiva, P.A.A. Assuncao, L.M.N. Tavora, L. A. Thomaz, S.M.M. Faria, Dermoscopic skin lesion image segmentation based on local binary pattern clustering: Comparative study, Biomed. Signal Process. Control 59 (2020), 101924.

[16] M. Schneider, S. Hirsch, B. Weber, G. Székely, B.H. Menze, Joint 3-D vessel segmentation and centerline extraction using oblique Hough forests with steerable filters, Med. Image Anal. 19 (2015) 220–249.

[17] Q. Huang, Y. Huang, Y. Luo, F. Yuan, X. Li, Segmentation of breast ultrasound image with semantic classification of superpixels, Med. Image Anal. 61 (2020), 101657.

[18] N. Moradi, N. Mahdavi-Amiri, Multi-class segmentation of skin lesions via joint dictionary learning, Biomed. Signal Process. Control 68 (2021), 102787.

[19] J. Ramya, H.C. Vijaylakshmi, H.M. Saifuddin, Segmentation of skin lesion images using discrete wavelet transform, Biomed. Signal Process. Control 69 (2021), 102839.

[20] B. Toptaş, D. Hanbay, Retinal blood vessel segmentation using pixel-based feature vector, Biomed. Signal Process. Control 70 (2021), 103053.

[21] S. Pereira, R. Meier, R. McKinley, R. Wiest, V. Alves, C.A. Silva, M. Reyes, Enhancing interpretability of automatically extracted machine learning features: Application to a RBM-random forest system on brain lesion segmentation, Med. Image Anal. 44 (2018) 228–244.

[22] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, D. Andina, Deep learning for computer vision: A brief review, Computational Intelligence and Neuroscience 2018 (2018) 1–13.

[23] X. Liu, L. Song, S. Liu, Y. Zhang, A review of deep-learning-based medical image segmentation methods, Sustainability 13 (3) (2021) 1224.

[24] F.P. An, Z.W. Liu, Medical image segmentation algorithm based on feedback mechanism convolutional neural network, Biomed. Signal Process. Control 53 (2019), 101589.

[25] H. Liang, Z. Cheng, H. Zhong, A. Qu, L. Chen, A region-based convolutional network for nuclei detection and segmentation in microscopy images, Biomed. Signal Process. Control 71 (2022), 103276.

[26] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, S. Zhang, CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imaging 40 (2) (2021) 699–711.

[27] H. Zhang, W. Zhang, W. Shen, N. Li, Y. Chen, S. Li, B. Chen, S. Guo, Y. Wang, Automatic segmentation of the cardiac MR images based on nested fully convolutional dense network with dilated convolution, Biomed. Signal Process. Control 68 (2021), 102684.

[28] X. Wang, Y. Fang, S. Yang, D. Zhu, M. Wang, J. Zhang, K.Y. Tong, X. Han, A hybrid network for automatic hepatocellular carcinoma segmentation in H&E-stained whole slide images, Med. Image Anal. 68 (2021), 101914.

[29] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234-241.

[30] J. Yang, J. Zhu, H. Wang, X. Yang, Dilated MultiResUNet: Dilated multiresidual blocks network based on U-Net for biomedical image segmentation, Biomed. Signal Process. Control 68 (2021), 102643.

[31] Z. Huang, Y. Zhao, Y. Liu, G. Song, GCAUNet: A group cross-channel attention residual UNet for slice based brain tumor segmentation, Biomed. Signal Process. Control 70 (2021), 102958.

[32] N. Badshah, A. Ahmad, ResBCU-Net: Deep learning approach for segmentation of skin images, Biomed. Signal Process. Control 71 (2022), 103137.

[33] M. Aghalari, A. Aghagolzadeh, M. Ezoji, Brain tumor image segmentation via asymmetric/symmetric UNet based on two-pathway-residual blocks, Biomed. Signal Process. Control 69 (2021), 102841.

[34] V. Sundaresan, G. Zamboni, P.M. Rothwell, M. Jenkinson, L. Griffanti, Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images, Med. Image Anal. 73 (2021), 102184.

[35] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2020) 1856–1867.

[36] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, D. Shen, High-resolution encoder-decoder networks for low-contrast medical image segmentation, IEEE Trans. Image Process. 29 (2020) 461–475.

[37] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[38] C.H. Hsiao, P.C. Lin, L.A. Chung, F.Y.S. Lin, F.J. Yang, S.Y. Yang, C.H. Wu, Y. Huang, T.L. Sun, A deep learning-based precision and automatic kidney segmentation system using efficient feature pyramid networks in computed tomography images, Comput. Methods Programs Biomed. 221 (2022), 106854.

[39] M. Gridach, PyDiNet: Pyramid dilated network for medical image segmentation, Neural Networks 140 (2021) 274–281.

[40] J. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3349–3364.

[41] J. Wan, S. Yue, J. Ma, X. Ma, A coarse-to-fine full attention guided capsule network for medical image segmentation, Biomed. Signal Process. Control 76 (2022), 103682.

[42] J. Zhang, L. Yu, D. Chen, W. Pan, C. Shi, Y. Niu, X. Yao, X. Xu, Y. Cheng, Dense GAN and multi-layer attention based lesion segmentation method for COVID-19 CT images, Biomed. Signal Process. Control 69 (2021), 102901.

[43] S. Nema, A. Dudhane, S. Murala, S. Naidu, RescueNet: An unpaired GAN for brain tumor segmentation, Biomed. Signal Process. Control 55 (2020), 101641.

[44] X. Chen, et al., One-shot generative adversarial learning for MRI segmentation of craniomaxillofacial bony structures, IEEE Trans. Med. Imaging 39 (3) (2020) 787–796.

[45] S.P. Pawar, S.N. Talbar, LungSeg-Net: Lung field segmentation using generative adversarial network, Biomed. Signal Process. Control 64 (2021), 102296.

[46] S. Pang, C. Pang, L. Zhao, Y. Chen, Z. Su, Y. Zhou, M. Huang, W. Yang, H. Lu, Q. Feng, SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation, IEEE Trans. Med. Imaging 40 (1) (2021) 262–273.

[47] J. Zhang, Z. Hua, K. Yan, K. Tian, J. Yao, E. Liu, M. Liu, X. Han, Joint fully convolutional and graph convolutional networks for weakly-supervised segmentation of pathology images, Med. Image Anal. 73 (2021), 102183.

[48] F. Bagheri, M.J. Tarokh, M. Ziaratban, Skin lesion segmentation from dermoscopic images by using mask R-CNN, Retina-Deeplab, and graph-based methods, Biomed. Signal Process. Control 67 (2021), 102533.

[49] R. LaLonde, Z. Xu, I. Irmakci, S. Jain, U. Bagci, Capsules for biomedical image segmentation, Med. Image Anal. 68 (2021), 101889.

[50] J. Yu, D. Yang, H. Zhao, FFANet: Feature fusion attention network to medical image segmentation, Biomed. Signal Process. Control 69 (2021), 102912.

[51] R. Wang, S. Cao, K. Ma, Y. Zheng, D. Meng, Pairwise learning for medical image segmentation, Med. Image Anal. 67 (2021), 101876.

[52] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.T. Yap, D. Shen, Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images, Med. Image Anal. 70 (2021), 101918.

[53] D. Zhang, B. Chen, J. Chong, S. Li, Weakly-supervised teacher-student network for liver tumor segmentation from non-enhanced images, Med. Image Anal. 70 (2021), 102005.

[54] J. Liu, B. Dong, S. Wang, H. Cui, D.P. Fan, J. Ma, G. Chen, COVID-19 lung infection segmentation with a novel two-stage cross-domain transfer learning framework, Med. Image Anal. 74 (2021), 102205.

[55] M. Guo, T. Xu, J. Liu, Z. Liu, P. Jiang, T. Mu, S. Zhang, R.R. Martin, M. Cheng, S. Hu, Attention mechanisms in computer vision: A survey, Computational Visual, Media 8 (2022) 331–368.

[56] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023.

[57] H. Zhang, K. Zu, J. Lu, Y. Zou, D. Meng. ESPANet: An efficient pyramid squeeze attention block on convolutional neural network, arXiv preprint, arXiv: 2105.14447v2, 2021. [Online]. Available: https://arxiv.org/abs/2105.14447v2.

[58] Z. Qin, P. Zhang, F. Wu, X. Li, FcaNet: Frequency channel attention networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 783–792.

[59] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.

[60] Z. Zhong, Z.Q. Lin, R. Bidart, X. Hu, I.B. Daya, Z. Li, W. Zheng, J. Li, A. Wong, Squeeze-and-attention networks for semantic segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13065–13074.

[61] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, in: International Conference on Neural Information Processing Systems, 2015, pp. 2017–2025.

[62] A. Almahairi, N. Ballas, T. Cooijmans, Y. Zheng, H. Larochelle, A. Courville, Dynamic capacity networks, arXiv preprint, arXiv:1511.07838v5, 2016. [Online]. Available: https://arxiv.org/abs/1511.07838v5.

[63] B. Mayo, T. Hazan, A. Tal, Visual navigation with spatial attention, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16898–16907.

[64] O. Ulutan, A.S.M. Iftekhar, B.S. Manjunath, VSGNet: Spatial attention network for detecting human object interactions using graph convolutions, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13617-13626.

[65] H. Fang, D. Zhang, Y. Zhang, M. Chen, J. Li, Y. Hu, D. Cai, X. He, Salient object ranking with position-preserved attention, in: International Conference on Computer Vision, 2021, pp. 16331–16341.

[66] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: European Conference on Computer Vision, 2018, pp. 3-19.

[67] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual-attention network for scene segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[68] T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3085–3094.

[69] X. Chen, X. Yan, F. Zheng, Y. Jiang, S. Xia, Y. Zhao, R. Ji, One-shot adversarial attacks on visual tracking with dual attention, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10176–10185.

[70] O. Wiles, S. Ehrhardt, A. Zisserman, Co-attention for conditioned image matching, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15920–15929.

[71] K. Zhu, J. Wu, Residual attention: A simple but effective method for multi-label recognition, International Conference on Computer Vision (2021) 184–193.

[72] P. Sun, W. Zhang, H. Wang, S. Li, X. Li, Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1407–1417.

[73] X. Wang, Z. Cai, D. Gao, N. Vasconcelos, Towards universal object detection by domain attention, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7289–7298.

[74] C. Chen, Q. Fan, R. Panda, CrossViT: Cross-attention multi-scale vision transformer for image classification, in: International Conference on Computer Vision, 2021, pp. 357–366.

[75] T.Y. Lin, P. Goyal, R. Girshic, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE/CVF International Conference on Computer Vision, 2017, pp. 2999–3007.

[76] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, BASNet: Boundary-aware salient object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7479-7489.

[77] S.G. Armato III, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans, Med. Phys. 38 (2) (2011) 915–931.

[78] N. Heller, et al., The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge, Med. Image Anal. 67 (2021), 101821.

[79] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T.D. Lange, D. Johansen, H.D. Johansen, Kvasir-SEG: A segmented polyp dataset, in: International Conference on Multimedia Modeling, 2020, pp. 451-462.

[80] N.F. Codella, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC), in: IEEE International Symposium on Biomedical Imaging, 2018, pp. 168-172.