Check for updates

Mixed-UNet: Refined class activation mapping for weakly-supervised semantic segmentation with multi-scale inference

Yang Liu[1,2†], Lijin Lian[1,3†], Ersi Zhang[4], Lulu Xu[1,3], Chufan Xiao[1,3], Xiaoyun Zhong[1,3], Fang Li[1,3], Bin Jiang[2], Yuhan Dong[2], Lan Ma[1,3], Qiming Huang[1], Ming Xu[5], Yongbing Zhang[6], Dongmei Yu[2], Chenggang Yan[4] and Peiwu Qin[1,3]*

[1]Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen, China, [2]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China, [3]Institute of Tsinghua-Berkeley Shenzhen, Tsinghua Shenzhen International Graduate School, Shenzhen, China, [4]School of Artificial Intelligence, Hangzhou Dianzi University, Hangzhou, China, [5]Department of Automation, Tsinghua University, Beijing, China, [6]School of Computer Science and Technology, Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen, China

Deep learning techniques have shown great potential in medical image processing, particularly through accurate and reliable image segmentation on magnetic resonance imaging (MRI) scans or computed tomography (CT) scans, which allow the localization and diagnosis of lesions. However, training these segmentation models requires a large number of manually annotated pixel-level labels, which are time-consuming and labor-intensive, in contrast to image-level labels that are easier to obtain. It is imperative to resolve this problem through weakly-supervised semantic segmentation models using image-level labels as supervision since it can significantly reduce human annotation efforts. Most of the advanced solutions exploit class activation mapping (CAM). However, the original CAMs rarely capture the precise boundaries of lesions. In this study, we propose the strategy of multi-scale inference to refine CAMs by reducing the detail loss in single-scale reasoning. For segmentation, we develop a novel model named Mixed-UNet, which has two parallel branches in the decoding phase. The results can be obtained after fusing the extracted features from two branches. We evaluate the designed Mixed-UNet against several prevalent deep learning-based segmentation approaches on our dataset collected from the local hospital and public datasets. The validation results demonstrate that our model surpasses available methods under the same supervision level in the segmentation of various lesions from brain imaging.

KEYWORDS

multi-scale class activation mapping, weakly-supervised semantic segmentation, Mixed-UNet, acute cerebral infarction, conditional random field

Introduction

In recent years, deep learning techniques based on Convolutional Neural Networks (CNNs) have shown great potential in medical image processing (Long et al., 2015; Ronneberger et al., 2015; Yu and Koltun, 2015; Badrinarayanan et al., 2017; Chen et al., 2017a; Zhao et al., 2017; Oktay et al., 2018). Many advanced networks are adopted in the medical imaging field, such as locating lesions in magnetic resonance imaging (MRI) scans or computed tomography (CT) scans through semantic segmentation models, to assist doctors in diagnosing and treating diseases (Lundervold and Lundervold, 2019; Taghanaki et al., 2021; Yang and Yu, 2021). As an application in acute cerebral infarction (ACI), brain tissue necrosis caused by the sudden interruption of cerebral blood supply, the segmentation model can quickly detect the tiny and scattered cerebral infarcts in MRI scans (some infarcts are easily unnoticed by the naked eye) and pinpoint cerebral infarct boundaries, thereby avoiding a substantial expenditure of time and effort of experienced doctors and helping therapeutic decision-making (Warach et al., 1995; Kim and Caplan, 2016). Unfortunately, deep learning applications based on lesion detection are limited because CNNs require many manually annotated images with pixel-level labels, which are time-consuming and labor-intensive. Plenty of works focusing on weakly-supervised semantic segmentation (WSSS) have been implemented to address this limitation and have achieved miraculous performance, even close to supervised learning (Pathak et al., 2014, 2015; Papandreou et al., 2015; Pinheiro and Collobert, 2015; Pourian et al., 2015; Qi et al., 2015; Xu et al., 2015; Bearman et al., 2016; Chen et al., 2017a; Zhao et al., 2017; Wang et al., 2019). These techniques rely on weaker forms of supervision, such as bounding boxes (Dai et al., 2015; Papandreou et al., 2015; Song et al., 2019), points or squiggles (Bearman et al., 2016; Lin et al., 2016), image-level labels (Pathak et al., 2014, 2015; Pinheiro and Collobert, 2015; Kwak et al., 2017; Roy and Todorovic, 2017; Vernaza and Chandraker, 2017; Ahn and Kwak, 2018; Briq et al., 2018; Huang et al., 2018; Wang et al., 2018; Wei et al., 2018; Zhang et al., 2018), etc.

To localize lesions using only image-level labels, Zhou et al. propose class activation maps (CAMs) to excavate the most discriminative regions (Zhou et al., 2016) by locating target objects where pixels with higher evaluation values are more likely to be within the target object. Since image-level labels only indicate whether the target objects exist, they do not contain any object location information (Jiang et al., 2021). Therefore, the localization ability of CAMs can compensate for such image-level labeling problems, which further facilitates ill-posed weakly-supervised semantic segmentation under image-level supervision (Ahn and Kwak, 2018; Li et al., 2018). However, the original CAMs are not enough to directly function as pixel-level supervision. The spatial resolution of the final convolutional layer output is low, causing the fused CAMs only to locate the coarse boundary of the lesion, limiting the accuracy of WSSS for lesion detection.

Inspired by these issues, we propose a novel method to refine the CAMs by fusing the CAMs with multi-scale inference, capturing distinct and complementary features of objects. The model can capture the relative complete object features for small-scale objects, but the model can only perceive parts of the object for large-scale objects. Therefore, it can be inferred that the discriminative features captured by the model should be different for objects at different scales. For the same model and image, if the input image scales are different, the activation regions are also different. Thus, we propose that fusing CAMs with multi-scale achieves better localization performance (Jiang et al., 2021; Wang et al., 2021). Nevertheless, Conditional Random Field (CRF) (Krähenbühl and Koltun, 2011) is widely utilized as an independent post-processing process to refine original CAMs to match the ground truth as much as possible, which assigns the same labels to similar features by using color and position information. The refined CAMs we get after the application of CRF.

Pseudo masks can be obtained from the above refined CAMs, and these masks are utilized as pixel-level supervision to train a new segmentation method. Motivated by U-Net, commonly used in medical image segmentation, we design a Mixed-UNet by combining two U-Nets in a hybrid manner. In order to save more spatial details and reduce the computational complexity, Mixed-UNet has only three downsampling in the encoding stage, which is one layer less than U-Net. The encoding stage shares the same feature extractor, while the decoding stage is divided into two parallel branches with the same structure, effectively helping alleviate the lack of accuracy caused by a single branch in weakly-supervised segmentation on the condition of fewer parameters. Finally, the features extracted by the two branches are fused, and the final segmentation result is obtained through a softmax layer.

Here, we integrate the optimized CAMs as supervision into the Mixed-UNet to locate the lesions on our ACI dataset. The refined CAMs derive the pixel-level pseudo masks from image-level labels and are adopted to Mixed-UNet to segment lesions. Comparison of segmentation accuracy, computational time, and hardware resources with other prevalent intelligent semantic segmentation networks, our Mixed-UNet outperforms well. In addition, we use publicly available datasets associated with brain disease as validation sets to demonstrate that our model is a more effective and universal approach in the segmentation of lesions in brain imaging. The main contributions of this work are summarized as follows:

- We propose a method for avoiding enormous manual annotation in processing medical images by utilizing CAMs to excavate discriminative regions in a weakly

supervised way, helping to locate the lesion only with image-level labels.

- Multi-scale CAMs are easy to build based on standard CNNs, making them have strong extensibility. Instead of single-scale inference, multi-scale CAMs can capture more complementary object features, thus reducing the loss of details and helping locate lesion boundaries more clearly.
- Our Mixed-UNet is a semantic segmentation network with two parallel branches in the upsampling phases, characterizing the architecture with robust scalability and flexibility. A series of empirical studies show that the Mixed-UNet achieves higher performance and is more computationally efficient than the popular brain lesion localization models.

Related work

Weakly-supervised semantic segmentation

Semantic segmentation refers to classifying each pixel of an image as an instance, where each instance corresponds to a class. In the field of medical image processing, image segmentation can be used for image-guided intervention, radiotherapy, radiodiagnosis, etc. Different levels of supervision are available when training deep segmentation models, ranging from pixel-level annotations (supervised learning) and image-level and bounding box annotations (semi-supervised learning) to completely unannotated objects (unsupervised learning), where the last two levels of annotations belong to weak supervision (Liang et al., 2020; Taghanaki et al., 2021). Training the architecture relies on a large amount of pixel-level labeled data, which is time-consuming and expensive, especially the pixel-level labels in medical images. However, a large number of images with the image-level label can be obtained in a relatively fast and inexpensive manner. Many weakly supervised semantic segmentation methods have emerged in recent years to alleviate the considerable burden of pixel-level annotation and have achieved miraculous performance, even close to supervised learning (Husain et al., 2012; Pathak et al., 2014, 2015; Papandreou et al., 2015; Pinheiro and Collobert, 2015; Pourian et al., 2015; Qi et al., 2015; Xu et al., 2015; Bearman et al., 2016; Chen et al., 2017a, 2022; Zhao et al., 2017; Chang et al., 2020; Peng et al., 2020; Wang X. et al., 2020; Wang Y. et al., 2020; Zhang et al., 2020; Chan et al., 2021; Ridnik et al., 2021; Jo et al., 2022; Kim et al., 2022; Li et al., 2022; Xie et al., 2022).

The general process of WSSS is conducted as follows: pixel-level pseudo-masks need to be generated by a weakly supervised algorithm at first. The images are then trained through a deep convolutional neural network. Finally, the output results and pseudo-masks are backpropagated to minimize the loss function and improve the model's performance. These techniques rely on weaker forms of supervision, such as

bounding boxes (Dai et al., 2015; Papandreou et al., 2015; Song et al., 2019), points or squiggles (Bearman et al., 2016; Lin et al., 2016; Qu et al., 2020; Xu and Lee, 2020), image-level labels (Pathak et al., 2014, 2015; Pinheiro and Collobert, 2015; Kwak et al., 2017; Roy and Todorovic, 2017; Vernaza and Chandraker, 2017; Ahn and Kwak, 2018; Briq et al., 2018; Huang et al., 2018; Wang et al., 2018; Wei et al., 2018; Zhang et al., 2018; Araslanov and Roth, 2020; Chamanzar and Nie, 2020; Cole et al., 2021; Jo and Yu, 2021; Lanchantin et al., 2021; Yun et al., 2021), etc.

Among them, the image-level label is the simplest form of weak labeling, which is relatively easy to obtain. Training images are labeled only by the classes they belong to, not by their location in the image. However, this also makes it challenging to use image-level labels to train segmentation networks, so many researchers start to consider building the correlations between image- and pixel-level labels.

Vezhnevets et al. proposed a novel multi-image model (MIM) to recover the pixel labels of the training images based on the similarity of appearance (Vezhnevets et al., 2011). In Wei et al. (2016a), have made significant contributions to develop a flexible deep CNN infrastructure, where a shared CNN is connected. The output results are aggregated with max-pooling to produce the ultimate multi-label predictions. Based on such a flexible model, not only does the training network no longer need ground-truth bounding box information, but the network is robust to random noisy and redundant assumptions. Similar literature (Wei et al., 2016b) explored a new framework consisting of two components. Reliable localization maps are first generated by combining hypothesis-aware classification and cross-image contextual refinement. Then, segmentation networks can be trained in a supervised manner from these generated localization maps. Besides, the authors explore two network training strategies to achieve good segmentation performance. The first strategy proposes a novel multi-label cross-entropy loss by directly using multiple localization maps to train the network, where each pixel contributes to each class with different weights. The second strategy infers a coarse segmentation mask from the localization map and uses the resulting mask to optimize the network based on a single-label cross-entropy loss. Kolesnikov et al. conducted CAMs to excavate the segmented objects, expand the regions according to the seeds, and get the segmentation results using CRF and boundary constraints (Kolesnikov and Lampert, 2016). Consequently, image-level labels are widely adopted and followed in this work due to their effortlessness to obtain.

Class activation mapping

Weakly supervised semantic segmentation uses image-level labels to locate objects in images. Among them, some previous studies have proposed techniques of applying class activation mapping (Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhay et al., 2018; Wang X. et al., 2020; Jiang et al.,

2021). Whether it's CAM, Grad-CAM, or Score-Cam, they all follow a similar pipeline to generate CAMs. CNNs are trained corresponding to the classification objective at first, and then the CAMs are generated by global average pooling (GAP) on feature maps. Finally, the obtained seed region is thresholded based on the maximum value of the CAMs, and applied as weak supervision to the segmentation network to obtain the segmented target. These methods differ in the way of each feature map's weights generation. CAM (Zhou et al., 2016) obtains weights from fully connected layers. Grad-CAM (Selvaraju et al., 2017; Chattopadhay et al., 2018) streams class-specific gradients to each feature map, and the gradient of each feature map is averaged as its weight. Whereas Score-CAM (Wang H. et al., 2020) eliminates the dependence on gradients and generates weights for each feature map through its forwarding score. They all generate reliable class activation maps from the final convolutional layer. Other works like CAM-GMP (Oquab et al., 2015) and ACoL (Zhang et al., 2018) add convolutional classification layers (Conv-Cls) on top of the backbone to generate CAMs directly to improve the integration and computational efficiency. In Choe and Shim (2019), the optimized dropout layer integrating erasure and discovery operations by attention mechanism improves the processing efficiency.

These approaches can identify the object regions by extracting the discriminative areas in class activation maps. However, due to the low spatial resolution of the output from the final convolutional layer, the generated CAMs can locate rough object regions, which cannot meet the accuracy requirements of weakly supervised semantic segmentation tasks. Significantly, the application performance in medical image processing is limited because the inherent characteristics of lesions are tiny and scattered (Lee et al., 2018).

To solve the problem of semantic segmentation of lesions, we turn our attention to some more advanced computer vision tasks, which benefit from the semantic knowledge of different feature scales (Lin et al., 2017; Zhao et al., 2017). Shen et al. made full use of multi-scale features and fused them for better tracking results (Shen et al., 2019). LayerCAM is exploiting hierarchical semantic knowledge from different layers (Jiang et al., 2021). CAMs generated from shallow layers tend to capture fine-grained details of target objects, while CAMs generated from deep layers usually locate coarse spatial object regions, which means CAMs with multi-scale inherence all help to locate the lesions in target images.

Encoder–decoder semantic image segmentation networks

A fully convolutional network (FCN) is a CNN-based segmentation network firstly proposed by Pathak et al. (2014).

It computes pixel-level outputs by upsampling the output activation maps. The context and spatial information in deep network images can be preserved by fusing the outputs with shallower layers' outputs. Encoder–decoder segmentation networks such as SegNet (Badrinarayanan et al., 2017) are introduced based on FCN. The decoder network maps low-resolution encoder features to full-input resolution feature maps for pixel-level classification. The SegNet's decoder upsamples the lower-resolution input feature map. Specifically, the decoder performs non-linear upsampling using the pooling indices computed in the max-pooling step of the corresponding encoder. The architecture consists of a series of non-linear processing layers (encoders) and a corresponding set of decoder layers, followed by pixel classifiers. Typically, each encoder consists of one or more convolutional layers with batch normalization and ReLU non-linearity, followed by non-overlapping max-pooling and subsampling (Ioffe and Szegedy, 2015). U-Net (Ronneberger et al., 2015) consists of a contraction path to capture context and a symmetric expansion path to achieve precise localization. In U-Net, skip connections are added to the Encoder–decoder image segmentation network to improve the model's accuracy and solve the problem of vanishing gradients. DenseNet (Huang et al., 2017) is a densely connected segmentation network architecture by adapting a U-Net-like Encoder–decoder skeleton for more accurate segmentation. The feature fusing operation is modified using a spatial pyramid pooling module based on FCN. The spatial pyramid networks are able to encode multi-scale contextual information by probing the incoming features with filters or pooling operations at multiple rates and multiple effective fields-of-view, while the latter decoding networks can capture sharper object boundaries by gradually recovering the spatial information (Zhao et al., 2017). DeepLabV3+ combines the advantages of dilated convolution and feature pyramid pooling, adding a simple yet effective decoder module on top of DeepLabV3 to refine segmentation results, especially along object boundaries using dilated convolution and pyramid features, making it outperform many advanced segmentation networks (Chen et al., 2017a,b, 2018).

It is essential to compress the model depth to improve processing efficiency for medical image segmentation with larger volumes and resolutions, such as CT, MRI, and histopathology images. A neural architecture search method is applied to U-Net, and a smaller network with better organ/tumor segmentation performance is obtained on CT, MRI, and ultrasound images (Weng et al., 2019). Brügger et al. redesigned the U-Net architecture to make the network more memory efficient for 3D medical image segmentation by exploiting group normalization and the leaky ReLU function (Wu and He, 2018; Brügger et al., 2019). Dilated Residual Network (DRN) has fewer parameters for medical image segmentation, making the architecture less prone to overfitting. Smaller networks with fewer parameters and higher efficiency, such a network structure is the future

development direction of medical image segmentation (Bonta and Kiran, 2019).

Methods

We evaluate the proposed Mixed-UNet segmentation method and provide quantitative and qualitative analysis. The ACI dataset collected from the hospital is used for network training and the public database for validation. We first compare the efficiency of different CNN classification networks and choose the most efficient ResNet50 as the classification network. Based on ResNet50, multi-scale CAMs are applied to obtain pseudo-masks, which are applied as supervision in Mixed-UNet for image segmentation.

Dataset

As shown in Figure 1, our ACI training dataset comprises 637 MRI scans from patients with ACI and 609 normal brain MRI slices from people without ACI, selected by experienced doctors. Among the slices with ACI, 310 slices are used as the training dataset and 133 slices as the testing dataset. Similarly, 427 slices are used for training, and 182 slices are used for testing among the normal slices. The splitting of the training and testing dataset coincides at the patient level, i.e., no brain slices from the same patient exist in training and testing datasets. The dataset is collected from the South University of Science and Technology Hospital in Guangdong Province with ethical approval and patients' notification.

The publicly available dataset about hemorrhage represents an independent dataset for validating the robustness and generalizability of our Mixed-UNet. Intracerebral hemorrhage (ICH) refers to sudden hemorrhage in brain tissue, ventricles, or both caused by rupture of a nontraumatic parenchymal blood vessel. This brain CT-hemorrhage dataset contains 100 normal and 100 hemorrhagic CT images from Kaggle. There is no difference in the type of bleeding. We resize the original CT slice into 256*256 and select 80 slices from the normal set and 80 slices from the hemorrhage as the training dataset. Then, 20 normal and 20 intracerebral hemorrhage slices are used as the test dataset.

Multi-scale reasoning of CAMs

We can get the most discriminative target area for a well-trained classification network through CAMs (Zhou et al., 2016). The classification network's full connection and softmax layers are replaced by global average pooling (GAP). The average value of all pixels in the feature map returns the whole feature map. To be more specific, let A_{ij}^k represents the active unit k at the spatial

position (i, j) of the feature map before the GAP layer and each feature map has a corresponding weight w_k^c. Z represents the size of the input image, where a and b represent length and width, respectively. Then, we can obtain an activation map Cam^c and class score Y^c of the corresponding class c as follows:

$$Cam^c = \sum_k w_k^c \times A^k \tag{1}$$

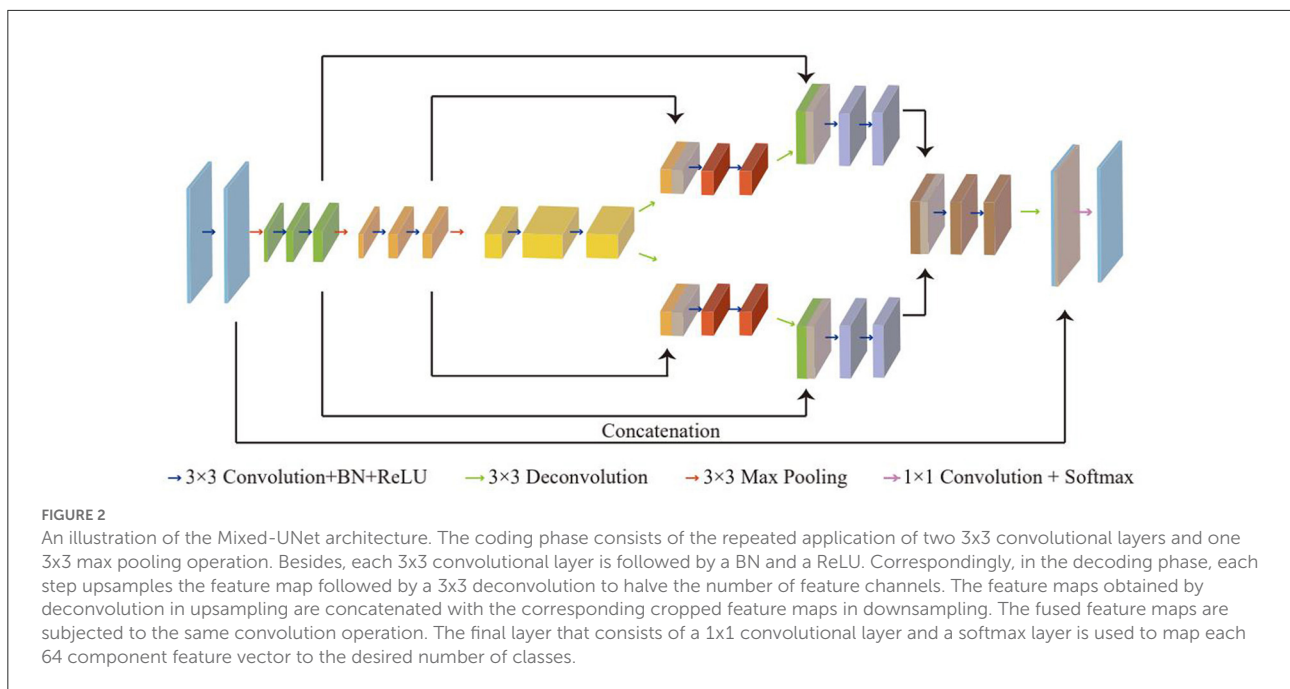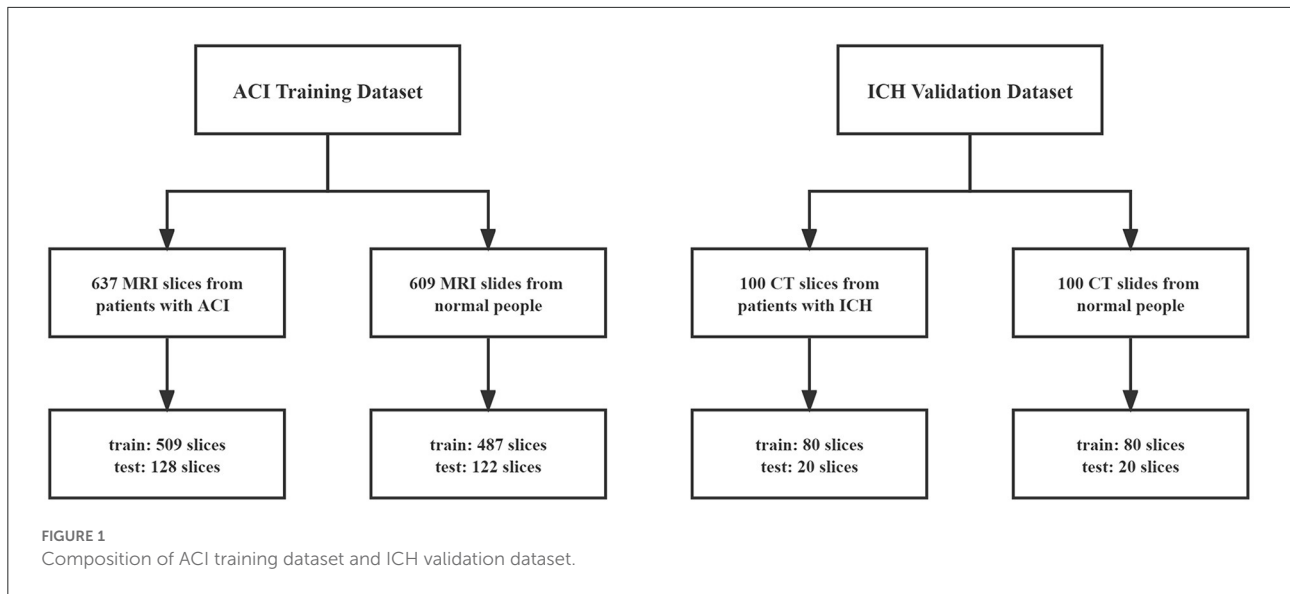$$Y^c = \sum_k w_k^c \frac{1}{Z} \sum_{i,j} A_{ij}^k, Z = a \times b \tag{2}$$

People use the original CAMs to conduct the follow-up operation. However, the problem that can not be ignored is that the original CAMs can only identify the image regions most relevant to the particular category, which is highly dissimilar to ground truth. Applying different rescaling transformations on the input images will not be able to obtain the same transformations on the generated CAMs, which is mainly due to the supervision gap between fully and weakly supervised semantic segmentation. Inspired by this inequivalence, we propose a method of multi-scale reasoning to refine CAMs. In detail, for an input image i, we sample it m times by setting different sampling rates. Then, $S_{i,c}^j$ represents the CAM of category c corresponding to the scale j of the input image i. Finally, we can obtain the fused CAM $S_c^{'}$ as following:

$$S_c^{'} = \sum_j^m \frac{S_{i,c}^j}{m} \tag{3}$$

Medical image segmentation with Mixed-UNet

Most of the current semantic segmentation methods adopt an Encoder–decoder architecture based on deep learning, such as FCN (Long et al., 2015), UNet (Ronneberger et al., 2015), Attention UNet (Oktay et al., 2018), SegNet (Badrinarayanan et al., 2017), DeepLab (Chen et al., 2017a), and PSPNet (Zhao et al., 2017). The deep semantic information is extracted in the encoding phase and then is mapped to the given category region in the decoding phase. However, these methods use one branch in the decoding stage, affecting the performance of weakly supervised semantic segmentation because of inaccurate labels obtained from CAMs. We, therefore, put forward a new segmentation model for weakly supervised semantic segmentation as illustrated in Figure 2, namely Mixed-UNet.

There are three downsampling in the coding stage for the segmentation because the lesion area is always tiny and scattered, which is one layer less than U-Net. It can save more spatial details and reduce computational complexity, contributing to high efficiency, which is essential in medical

FIGURE 1
Composition of ACI training dataset and ICH validation dataset.

FIGURE 2
An illustration of the Mixed-UNet architecture. The coding phase consists of the repeated application of two 3x3 convolutional layers and one 3x3 max pooling operation. Besides, each 3x3 convolutional layer is followed by a BN and a ReLU. Correspondingly, in the decoding phase, each step upsamples the feature map followed by a 3x3 deconvolution to halve the number of feature channels. The feature maps obtained by deconvolution in upsampling are concatenated with the corresponding cropped feature maps in downsampling. The fused feature maps are subjected to the same convolution operation. The final layer that consists of a 1x1 convolutional layer and a softmax layer is used to map each 64 component feature vector to the desired number of classes.

images processing. Unlike double U-Net (Pham et al., 2019), cascaded V-Net (Casamitjana et al., 2017), where two networks are connected in a cascade way, our method combines two U-Nets in a mixed way. The coding phase shares a feature extractor and is divided into two parallel branches in the decoding phase, which have the same structure. In detail, the downsampling follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutional layers and one 3x3 max pooling operation, and each 3x3 convolutional layer contains a 3x3 convolution, a batch normalization(BN) and a rectified linear unit(ReLU). Every blue

arrow represents a convolutional layer and every red arrow represents a max pooling layer, as shown in the figure. In particular, each downsampling step is distinguished by blocks of different colors. In the decoding phase, each step upsamples the feature map and followed by a 3x3 deconvolutions. Then, the feature maps obtained by deconvolution in upsampling are concatenated with the corresponding cropped feature maps in downsampling. The cropping is necessary due to the loss of border pixels in every convolution. The fused feature maps are subjected to the same convolution operation. And after the fuse of the features extracted from the two branches, the fused

feature map then is processed by two convolutional layer and one deconvolutional layer. The final layer that consists of a 1x1 convolutional layer and a softmax layer is used to map each feature vector to the desired number of classes, obtaining the final segmentation result. Non-linear activation as the softmax layer is added to improve the expression ability of the network.

For an input image x_i, ef represents the semantic feature information extracted in encoding stage and df_i represents the feature extracted by the ith branch in the decoding phase. y_i represents the final predicted segmentation result.

We can use the following formulas to describe the overall operations:

$$ef = encoding\,(x_i) \tag{4}$$

$$df_1 = decoding\,(ef) \tag{5}$$

$$df_2 = decoding\,(ef) \tag{6}$$

$$y_i = softmax\,(conv(concat\,(df_1, df_2))) \tag{7}$$

To train the proposed Mixed-UNet model, we combine seeding loss (Krizhevsky et al., 2012) using Refined_mask and cross-entropy loss function using CRF_mask.

Supposing X is the input image, T is the category of classification, Y is the output of the model, and u represents any position in the space, S_c is a set of locations that are labeled with class c by the weak localization procedure. We define the following loss function:

$$L_{seed}\,(Y, T, S_c) = -\frac{1}{\sum_{c \in T} |S_c|} \sum_{c \in T} \sum_{u \in S_c} \log Y_{u,c} \tag{8}$$

$$L_{CE}\,(Y, T) = -\sum_{i=1}^{T} T_i \log Y_i \tag{9}$$

$$L_{loss} = L_{seed}\,(f\,(X)\,, T, S_c) + L_{CE}\,(Y, T) \tag{10}$$

Experiment and result

Classification of medical image

We use the CNN models to generate CAMs and test the classification performance of ACI with popular CNN models of AlexNet, GoogLeNet, VGG, ResNet, and DenseNet, respectively.

Table 1 reports the diagnosis classification of the diagnosis of acute cerebral infarction from brain MRI. While all the methods achieve promising performance, ResNet50 has fewer parameters and faster convergence speed. The pre-trained ResNet50 has higher accuracy and faster convergence speed. Therefore, we choose pre-trained ResNet.

We compare the performance of ResNet50 pre-trained with that without pre-trained, as illustrated in Figure 3. The pre-trained ResNet50 has higher accuracy and faster convergence

TABLE 1 Classification results for ACI and normal MRI slices.

Model	Accuracy
AlexNet	0.99
GoogLeNet	0.99
VGG	1.00
ResNet50(without pre-trained)	0.987
ResNet50(pre-trained)	1.00
DenseNet	1.00

speed, which is selected to be used in our architecture for the following CAMs generation.
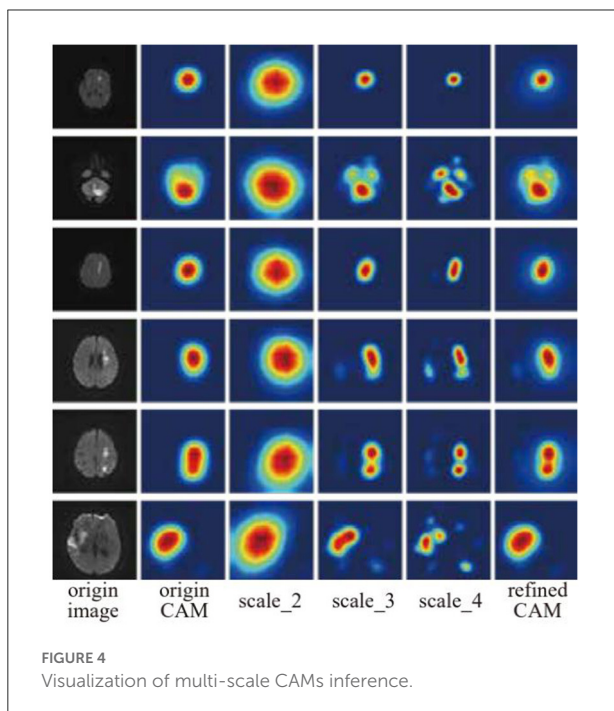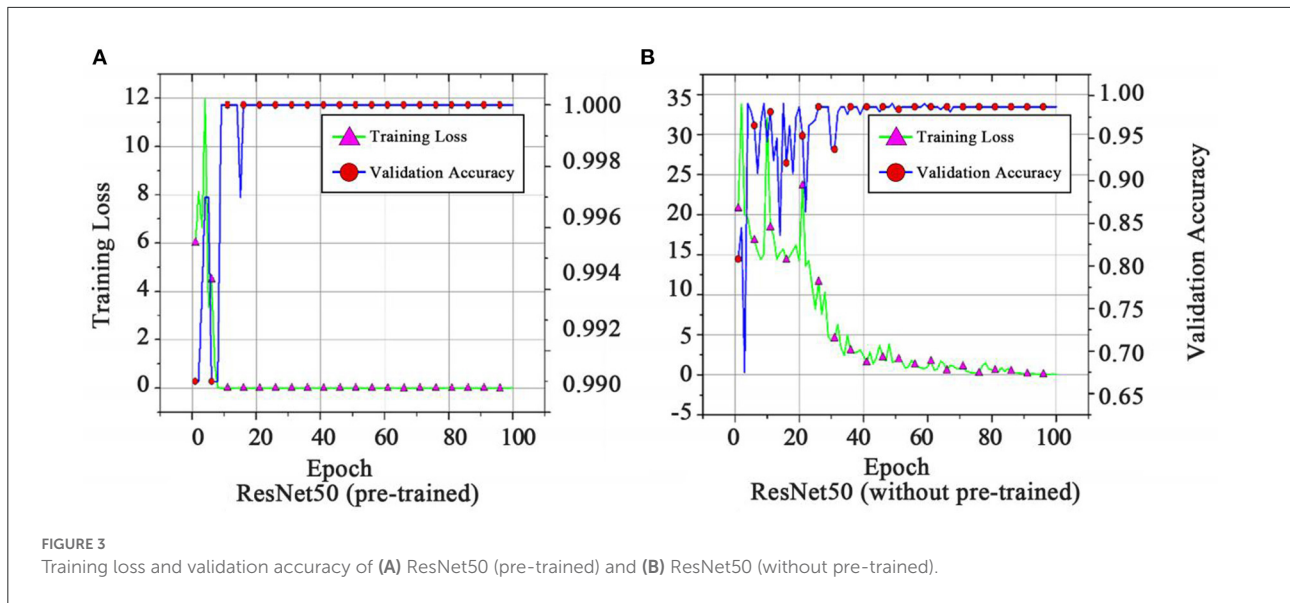
Results of multi-scale inference

After training, we can get the CAMs through the classification model to locate the most significant area of the target. We sample the original image four times with different sizes, and the sampling rates are 0.5 (scale_2), 1.0 (origin_CAM), 1.5 (scale_3) and 2.0 (scale_4), respectively. Figure 4 shows the result of multi-scale inference, and refined CAMs refer to CAMs after fusing four scales of CAMs. The original CAMs can only locate the general area of the lesion, but it differs significantly from the ground truth in size and boundary. In addition, for multiple scattered lesions, origin CAMs can only identify the most apparent lesions while ignoring other difficultly identified lesions. The shape and size of the refined CAMs are more consistent with the lesion area, which can find multiple lesion areas. Then, we normalize the refined CAMs. Let x_c be any position of CAM of class c. By using the following normalization formula, x_c can be normalized into a number between [0,1]:

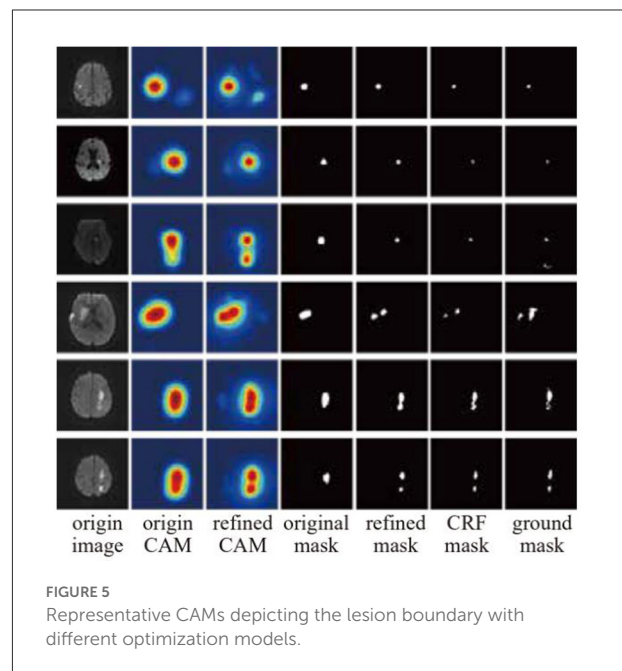$$x_c^{'} = \frac{x_c - x_{c,\min}}{x_{c,\max} - x_{c,\min}} \tag{11}$$

$x_{c,max}$ and $x_{c,min}$ refer to the maximum and minimum value of CAM of different class c. We take the following threshold segmentation method to obtain the pseudo masks used for semantic segmentation:

$$f^{'}\,(x, y) = \begin{cases} 1, & f\,(x, y) \geq T \\ 0, & f\,(x, y) < T \end{cases} \tag{12}$$

Where T is the threshold depending on the lesions, we set T to be 0.35. Here, the threshold needs a special explanation. When the lesion range is large, such as the lesion of ACI, we need to adjust the threshold value to a lower level to obtain a better segmentation effect. A higher threshold is required for small focal areas, such as ICH's small and scattered bleeding sites. $f\,(x, y)$ refers to the pixels whose spatial position is (x, y)

FIGURE 3
Training loss and validation accuracy of **(A)** ResNet50 (pre-trained) and **(B)** ResNet50 (without pre-trained).

FIGURE 4
Visualization of multi-scale CAMs inference.

FIGURE 5
Representative CAMs depicting the lesion boundary with different optimization models.

competitive results closer to the ground truth. Table 2 shows the numerical comparison of different masks. CRF_mask has the highest predicted segmentation (PA) and mean intersection over union (MIoU).

Mixed-UNet construction

For semantic segmentation, 60% of data is used for training, 20% of data is used for model selection and hyper-parameters adjustment, and the remaining 20% is used for testing. We

in refined CAMs. In order to better estimate the boundary of lesions, we use the dense CRF model (Chen et al., 2017a) to optimize the pseudo masks, which will assign the same label to the pixels with similar features in the image when inputting the pseudo masks and the original image as illustrated in Figure 5. The original mask is very rough, and if there are multiple lesions, only one lesion can be detected. Refined_mask after multi-scale inference are consistent with the original lesion area in size and shape but can detect additional lesion areas. Based on Refined_mask, CRF_mask optimized by dense CRF can produce

TABLE 2 Numerical comparison results of different masks.

Mask	PA	MIoU
Origin_mask	0.989	0.621
Refined_mask	0.991	0.652
CRF_mask	0.993	0.726

TABLE 3 The quantitative comparison of different models.

Model	PA	MIoU	Dice
FCN	0.991	0.602	0.409
UNet	0.991	0.615	0.487
DeepLab V3	0.991	0.656	0.529
Mixed-UNet	0.992	0.645	0.560
Mixed-UNet (+CRF)	0.992	0.674	0.543
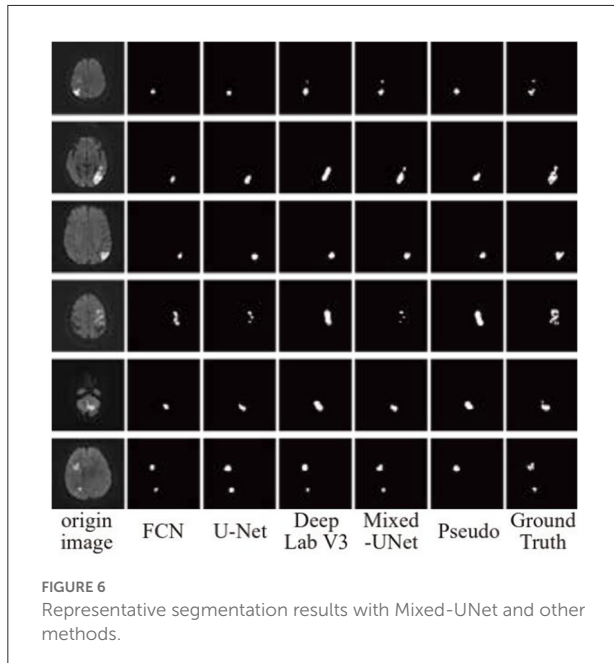
FIGURE 6
Representative segmentation results with Mixed-UNet and other methods.

TABLE 4 A comparison of computational time and hardware resources required for various deep architectures.

Model	# of trainable parameters	Memory	Time(s)
FCN (VGG16)	18,643,713	73M	0.421
U-Net	31,043,521	121M	0.861
DeepLab V3	39,633,729	155M	0.936
Mixed-UNet	10,303,105	50M	1.153

compare our designed Mixed-UNet algorithm with state-of-the-art methods: FCN (Long et al., 2015), U-Net (Ronneberger et al., 2015), and DeepLab V3 (Chen et al., 2017a). The model is trained on 4 NVIDIA TITAN X GPUs with batch size 8 for 100 epochs. The Mixed-UNet is first trained using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate (lr_{init}) of 1e−3. The practical learning rate follows a polynomial decay, to be zero until the max iteration.

The poly policy follows (Pourian et al., 2015) with $\gamma = 0.9$ for decay:

$$lr_{itr} = lr_{init} \left(1 - \frac{itr}{max - itr} \right)^{\gamma} \qquad (13)$$

Due to the limited training images, we employ random flipping (up-down or left-right), random rotation (one of $\pm$ 25, 90, 180, or 270 degrees), and random Gaussian noise addition (σ from 0.3 to 0.7) to augment data. We use $\mathcal{L}_2$ regularization on the network parameters with weight $\lambda = 10^{-4}$. The original MRI slices have a different number of pixels spatially, and we resize them to 256 * 256. The qualitative segmentation results of Mixed-UNet on samples including acute cerebral infarction are shown in Figure 6 Mixed-UNet obtains reasonable predictions

due to the design of parallel branch structure, which can extract more information than a single branch.

The quantitative results in Table 3 show that all the deep architectures have high PA and low Dice. One reason for the overall poor performance is that the focus areas occupy a small part of the image while the background occupies the majority space of the image.

A small error in the background will significantly influence the segmentation results. Our model achieves better optimal results, except MIoU is slightly lower than DeepLab V3. When we join CRF_mask, using Refined_mask and CRF_mask to train the network simultaneously, we find that MIoU increases by 3.1%, but Dice decreases by 1.7%. More optimal settings should exist theoretically, but the grid search is required, which is computational expensive given the considerable inference time for dense-CRFs. Table 4 compares several different deep learning models regarding parameter quantity, memory consumption, and computation time spent on segmenting an image. Compared with other methods, Mixed-UNet has the least number of parameters and minor memory consumption, which is a quarter of Deeplab V3. However, in the case of fewer parameters, Mixed-UNet takes the longest time to infer an image, mainly due to the structure of two parallel branches.

The parallel structure integrates the feature extracted from two branches and pinpoints more tiny details that a single branch can not find, whose cost is the computation time needed to mine the subtle details.

FIGURE 7
The validation of the Mixed-UNet and other methods.
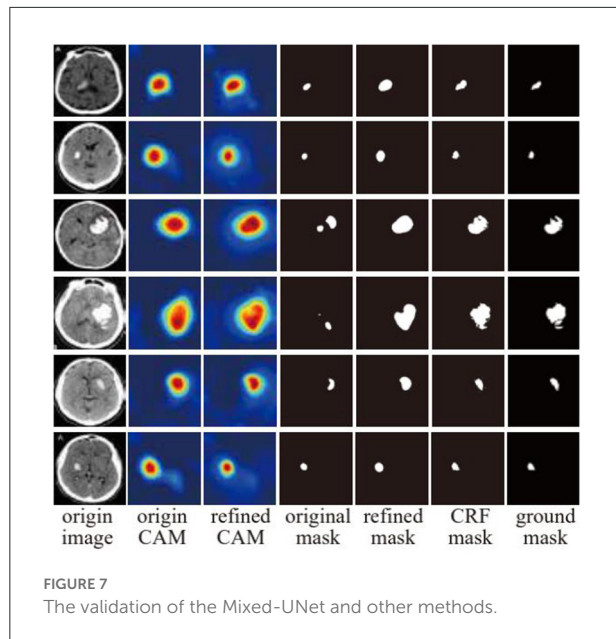
Validation of the mixed-UNet

We further evaluate our model on the ICH dataset and set the segmentation threshold to 0.7 due to the tiny lesions of the ICH dataset. The results show that our method can significantly improve the weakly supervised classification performance by providing refined pixel-level labels. A dense CRF model with refined CAMs shows a remarkable enhancement of depicting the boundary of lesions as illustrated in Figure 7. The original mask is limited to detecting partial lesions and cannot cover the lesions in massive bleeding images. Refined_mask can capture the rough shape of lesion areas. On the premise of accurate location of Refined_mask and utilization of dense CRF, CRF_mask demonstrates an ability to segment details of lesions. The public dataset does not include the expert ground label, which is labeled by the experienced expert for metric evaluation calculation. CRF_mask has the highest PA (0.9831), MioU (0.7706), and Dice (0.6907) among the output masks.

Discussion and conclusion

In this study, we present a weekly supervised semantic segmentation framework that integrates the image classification and segmentation of cerebral infarction. We propose multi-scale inference to optimize origin CAMs. Meanwhile, we develop a new semantic segmentation network named Mixed-UNet, adopting two paralleled branches in the upsampling phase and testing the performance using Refined_Mask and CRF_Mask as the supervision. The experimental results show that our models perform better due to the multi-scale inference and two paralleled branches. The limitation of our method is

the supervised way of training, and it does not dig out the information except the label. In this case, when the label is inaccurate, label dependence is worth further investigation. The model is currently used to pinpoint the brain imaging lesion, which can definitely be extended to other imaging modalities and various lesion detection.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by South University of Science and Technology Hospital in Guangdong Province. The patients/participants provided their written informed consent to participate in this study.

Author contributions

YL and LL: conceptualization, methodology, software, investigation, and formal analysis. EZ: data curation and writing—original draft. LX: visualization and investigation. CX: software and validation. XZ: resources, supervision, and writing—original draft. FL: investigation and funding acquisition. BJ: writing—review and editing. YD and LM: supervision. YZ, DY, and CY: project administration. QH and MX: resources. PQ: conceptualization, funding acquisition, resources, supervision, and writing—review and editing. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahn, J., and Kwak, S. (2018). "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Araslanov, N., and Roth, S. (2020). *Single-Stage Semantic Segmentation from Image Labels*, pp. 4252–4261.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional Encoder–decoder architecture for image segmentation. *IEEE. Trans. Pattern. Anal. Mach. Intell.* 39, 2481–95. doi: 10.1109/TPAMI.2016.2644615

Bearman, A., Russakovsky, O., Ferrari, V., and Fei-Fei, L. (2016). *What's the Point: Semantic Segmentation with Point Supervision*. arxiv:150602106.

Bonta, L. R., and Kiran, N. U. (2019). *Efficient Segmentation of Medical Images Using Dilated Residual Networks Computer Aided Intervention and Diagnostics in Clinical and Medical Images*. Berlin, Germany: Springer, 39–47.

Briq, R., Moeller, M., and Gall, J. (2018). *Convolutional Simplex Projection Network (Cspn) for Weakly Supervised Semantic Segmentation*. arXiv:180709169.

Brügger, R., Baumgartner, C. F., and Konukoglu, E. A. (2019). "Partially reversible u-net for memory-efficient volumetric image segmentation," in *Med Image Comput Comput Assist Interv (MICCAI)*. Berlin, Germany: Springer.

Casamitjana, A., Catà, M., Sánchez, I., Combalia, M., and Vilaplana, V. (2017). "Cascaded V-Net Using Roi Masks for Brain Tumor Segmentation," in *Proceedings of the MICCAI BrainLes Workshops*. Berlin, Germany: Springer.

Chamanzar, A., and Nie, Y. (2020). *Weakly Supervised Multi-Task Learning for Cell Detection and Segmentation*, pp. 513–516.

Chan, L., Hosseini, M., and Plataniotis, K. A (2021). Comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *Int. J. Comput. Vis.* 129, 361–384. doi: 10.1007/s11263-020-01373-4

Chang, Y-. T., Wang, Q., Hung, W-. C., Piramuthu, R., Tsai, Y-. H., Yang, M-. H., et al. (2020). "Weakly-Supervised Semantic Segmentation Via Sub-Category Exploration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8991–9000

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). "Grad-Cam++: generalized gradient-based visual explanations for deep convolutional networks," in *Proceedings of the IEEE Winter Conference Application of Computer Visual (WACV)*. New York City, NY: IEEE.

Chen, L-. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017a). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected Crfs. *IEEE. Trans. Pattern Anal. Mach. Intell.* 40, 834–48. doi: 10.1109/TPAMI.2017.2699184

Chen, L-. C., Papandreou, G., Schroff, F., and Adam, H. (2017b). *Rethinking Atrous Convolution for Semantic Image Segmentation*. arXiv:170605587.

Chen, L-. C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). *Encoder–decoder with Atrous Separable Convolution for Semantic Image Segmentation*. Proc ECCV.

Chen, Z., Wang, C., Wang, Y., Jiang, G., Shen, Y., Tai, Y., et al. (2022). Lctr: on awakening the local continuity of transformer for weakly supervised object localization. *Proc. AAAI Conf. Artif. Intell.* 36, 410–418. doi: 10.1609/aaai.v36i1.19918

Choe, J., and Shim, H. (2019). "Attention-based dropout layer for weakly supervised object localization," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Cole, E., Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N., et al. (2021). *Multi-Label Learning from Single Positive Labels*, pp. 933–942.

Dai, J., He, K., and Sun, J. (2015). "Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference of Computer Visual (ICCV)*

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE*

Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Huang, Z., Wang, X., Wang, J., Liu, W., and Wang, J. (2018). "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Husain, S., Ong, E-. J., Minskiy, D., Bober-Irizar, M., Irizar, A., Bober, M., et al. (2012). *Subcellular Protein Localisation in the Human Protein Atlas Using Ensembles of Diverse Deep Architectures*. arXiv:2205.09841

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. Proc ICML: PMLR.

Jiang, P. T., Zhang, C. B., Hou, Q., Cheng, M. M., and Wei, Y. (2021). Layercam: exploring hierarchical class activation maps for localization. *IEEE. Trans. Image. Process* 30, 5875–88. doi: 10.1109/TIP.2021.3089943

Jo, S., Yu, I., and Kim, K-. S. (2022). *Recurseed and Certainmix for Weakly Supervised Semantic Segmentation*. New York City, NY: IEEE.

Jo, S., and Yu, I-. J. (2021). *Puzzle-Cam: Improved Localization Via Matching Partial and Full Features*, pp. 639–643.

Kim, J. S., and Caplan, L. R. (2016). Clinical stroke syndromes. *Front. Neurol. Neurosci* 40, 72–92. doi: 10.1159/000448303

Kim, Y., Kim, J., Akata, Z., and Lee, J. (2022). *Large Loss Matters in Weakly Supervised Multi-Label. Classification*. 14136–45. p. doi: 10.1109/CVPR52688.2022.01376

Kingma, D. P., and Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. arXiv:14126980.

Kolesnikov, A., and Lampert, C. H. (2016). *Seed Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation*. Proc ECCV. Cham: Springer International Publishing.

Krähenbühl, P., and Koltun, V. (2011). "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*. 24.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst* 25, 1097–105.

Kwak, S., Hong, S., and Han, B. (2017). "Weakly supervised semantic segmentation using superpixel pooling network," in *Proceedings of the Conference AAAI Artificial Intelligence*. New York City, NY: IEEE.

Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y. (2021). *General Multi-Label Image Classification with Transformers*, pp. 16473–16483.

Lee, S., Lee, J., Lee, J., Park, C-. K., and Yoon, S. (2018). *Robust Tumor Localization with Pyramid Grad-Cam*. arXiv:180511393.

Li, K., Wu, Z., Peng, K-. C., Ernst, J., and Fu, Y. (2018). "Tell me where to look: guided attention inference network," in *Proceedings of the IEEE Conference of Computer Visual Pattern Recognition (CVPR)*.

Li, R., Mai, Z., Trabelsi, C., Zhang, Z., Jang, J., Sanner, S., et al. (2022). *Transcam: Transformer Attention-Based Cam Refinement for Weakly Supervised Semantic Segmentation*. arXiv:2203.07239v1

Liang, B., Liu, Y., He, L., and Li, J. (2020). "Weakly supervised semantic segmentation based on deep learning," in *Proceedings of the IASTED International Conference of Model Identification Control (ICMIC)*; Singapore: Springer.

Lin, D., Dai, J., Jia, J., He, K., and Sun, J. (2016). "Scribblesup: scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Lin, G., Milan, A., Shen, C., and Reid, I. (2017). "Refinenet: multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings*

of the IEEE Conference Computer Visual Pattern Recognition (CVPR). New York City, NY: IEEE.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference of Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Lundervold, A. S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on Mri. *Z. Med. Phys.* 29, 102–27 doi: 10.1016/j.zemedi.2018.11.002

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). *Attention U-Net: Learning Where to Look for the Pancreas*. arXiv:180403999.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2015). "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Papandreou, G., Chen, L-. C., Murphy, K. P., and Yuille, A. L. (2015). "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the IEEE International Conference of Computer Visual (ICCV)*.

Pathak, D., Krahenbuhl, P., and Darrell, T. (2015). "Constrained convolutional neural networks for weakly supervised segmentation," in *Proceedings of the IEEE International Conference of Computer Visual (ICCV)*

Pathak, D., Shelhamer, E., Long, J., and Darrell, T. (2014). *Fully Convolutional Multi-Class Multiple Instance Learning*. arXiv:14127144.

Peng, J., Kervadec, H., Dolz, J., Ben Ayed, I., Pedersoli, M., Desrosiers, C., et al. (2020). Discretely-constrained deep network for weakly supervised segmentation. *Neural. Netw.* 130, 297–308. doi: 10.1016/j.neunet.2020.07.011

Pham, D. D., Dovletov, G., Warwas, S., Landgraeber, S., Jäger, M., Pauli, J., et al. (2019). "Deep segmentation refinement with result-dependent learning," in *Bildverarbeitung Für Die Medizin*. Berlin, Germany: Springer, pp. 49–54.

Pinheiro, P. O., and Collobert, R. (2015). "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE

Pourian, N., Karthikeyan, S., and Manjunath, B. S. (2015). "Weakly supervised graph based semantic segmentation by learning communities of image-parts," in *Proceedings of the IEEE International Conference of Computer Visual (ICCV)*.

Qi, X., Shi, J., Liu, S., Liao, R., and Jia, J. (2015). "*Semantic Segmentation with Object Clique Potential*," in *Proceedings of the IEEE International Conference of Computer Visual (ICCV)*.

Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., et al. (2020). "Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images." *IEEE transactions on medical imaging*, pp.3655–3666.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., et al. (2021). *Asymmetric Loss for Multi-Label Classification,* 82–91.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in https://link.springer.com/book/10.1007/978-3-319-24574-4 *Medical* Image Computing and Computer-Assisted Intervention. Berlin, Germany: Springer.

Roy, A., and Todorovic, S. (2017). "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., et al. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference of Computer Visual (ICCV)*. New York City, NY: IEEE.

Shen, J., Tang, X., Dong, X., and Shao, L. (2019). Visual object tracking by hierarchical attention siamese network. *IEEE. Trans. Cybern* 50, 3068–80. doi: 10.1109/TCYB.2019.2936503

Song, C., Huang, Y., Ouyang, W., and Wang, L. (2019). "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Taghanaki, S. A., Abhishek, K., Cohen, J. P., Cohen-Adad, J., and Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. *Artif. Intell. Rev.* 54, 137–78. doi: 10.1007/s10462-020-09854-1

Vernaza, P., and Chandraker, M. (2017). "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Vezhnevets, A., Ferrari, V., and Buhmann, J. M. (2011). Weakly supervised semantic segmentation with a multi-image model," in *Proceedings of the IEEE International Conference of Computer Visual (ICCV)*. New York City, NY: IEEE.

Wang, B., Yuan, C., Li, B., Ding, X., Li, Z., Wu, Y., et al. (2021). Multi-scale low-discriminative feature reactivation for weakly supervised object localization. *IEEE. Trans. Image. Process* 30, 6050–65. doi: 10.1109/TIP.2021.3091833

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., et al. (2020). "*Score-cam: score-weighted visual explanations for convolutional neural networks*," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Wang, X., Liu, S., Ma, H., and Yang, M-. H. (2020). Weakly-supervised semantic segmentation by iterative affinity learning. *Int. J. Comput. Vis.* 128, 1736–49. doi: 10.1007/s11263-020-01293-3

Wang, X., Ma, H., and You, S. (2019). Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes. *Neurocomputing* 381, 20–28. doi: 10.1016/j.neucom.2019.11.019

Wang, X., You, S., Li, X., and Ma, H. (2018). "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Wang, Y., Zhang, J., Kan, M., Shan, S., and Chen, X (2020). *Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation*, 12272–12281. New York City, NY: IEEE.

Warach, S., Gaa, J., Siewert, B., Wielopolski, P., and Edelman, R. R. (1995). Acute human stroke studied by whole brain echo planar diffusion-weighted magnetic resonance imaging. *Ann. Neurol.* 37, 231–41. doi: 10.1002/ana.410370214

Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y., et al. (2016b). Learning to segment with image-level annotations. *Pattern. Recognit.* 59, 234–44. doi: 10.1016/j.patcog.2016.01.015

Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., et al. (2016a). Hcp: a flexible cnn framework for multi-label image classification. *IEEE. Trans. Pattern. Anal. Mach. Intell.* 38, 1901–7. doi: 10.1109/TPAMI.2015.2491929

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T. S., et al. (2018). "Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Weng, Y., Zhou, T., Li, Y., and Qiu, X. (2019). Nas-Unet: neural architecture search for medical image segmentation. *IEEE. Access* 7, 44247–57. doi: 10.1109/ACCESS.2019.2908991

Wu, Y., and He, K. (2018). *Group Normalization*. Proc ECCV.

Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L., et al. (2022). *Contrastive Learning of Class-Agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation.* arXiv:2203.13505v1

Xu, J., Schwing, A. G., and Urtasun, R. (2015). "Learning to segment under various forms of weak supervision," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE

Xu, X., and Lee, G. (2020). "Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 13706–13715.

Yang, R., and Yu, Y. (2021). Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front. Oncol* 11, 638182. doi: 10.3389/fonc.2021.638182

Yu, F., and Koltun, V. (2015). *Multi-Scale Context Aggregation by Dilated Convolutions*. arXiv:151107122.

Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., Chun, S., et al. (2021). "Re-labeling imagenet: from single to multi-labels, from global to localized labels." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2340–2350.

Zhang, B., Xiao, J., Wei, Y., Sun, M., and Huang, K. (2020). "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach." in *Proceedings of the AAAI Conference on Artificial Intelligence* pp. 12765–12772.

Zhang, X., Wei, Y., Feng, J., Yang, Y., and Huang, T. S. (2018). "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). "Pyramid scene parsing network," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference Computer Visual Pattern Recognition (CVPR)*. New York City, NY: IEEE.