TC-Net: Triple Context Network for Automated Stroke Lesion Segmentation

Xiuquan Du, Kunpeng Ma

Computer Science and Technology, Anhui University, Hefei, Anhui, China

Abstract

Accurate lesion segmentation plays a key role in the clinical mapping of stroke. Convolutional neural network (CNN) approaches based on U-shaped structures have achieved remarkable performance in this task. However, the single-stage encoder-decoder unresolvable the inter-class similarity due to the inadequate utilization of contextual information, such as lesion-tissue similarity. In addition, most approaches use fine-grained spatial attention to capture spatial context information, yet fail to generate accurate attention maps in encoding stage and lack effective regularization. In this work, we propose a new network, Triple Context Network (TC-Net), with the capture of spatial contextual information as the core. We firstly design a coarse-grained patch attention module to generate patch-level attention maps in the encoding stage to distinguish targets from patches and learn target-specific detail features. Then, to enrich the representation of boundary information of these features, a cross-feature fusion module with global contextual information is explored to guide the selective aggregation of 2D and 3D feature maps, which compensates for the lack of boundary learning capability of 2D convolution. Finally, we use multi-scale deconvolution instead of linear interpolation to enhance the recovery of target space and boundary information in the decoding stage. Our network is evaluated on the open dataset ATLAS, achieving the highest DSC score of 0.594, Hausdorff distance of 27.005 mm, and average symmetry surface distance of 7.137 mm, where our proposed method outperforms other state-of-the-art methods.

Introduction

Stroke is ischemia due to rupture or congestion of blood vessels in the brain (Grysiewicz, Thomas, and Pandey 2008). In clinical practice, physicians need to accurately map out stroke lesions, which is used to calculate the size, shape, and volume of the lesion. This task is not only time-consuming but also relies on the subjective judgment of physicians, is not suitable for larger data. There is a clear need for an automatic method to help physicians automatically segment lesion regions. However, the segmentation of stroke lesions still suffers from some challenges. For instance: (1) Due to the high similarity of normal brain tissue and lesions, they can be misjudged as lesions, such as the modified U-shaped network (MUN) segmented out of misjudged tissue and lesions, as shown in Fig. 1(a). (2) It is easy to confuse the background near the lesion and the border. In the composite image, the tissues on both sides of the green border are similar to the grayscale value of the lesion, so the border of the MUN decomposition result (blue part) only identifies the obvious area, which is far from the green border.

Recently, many U-based networks (Yang et al. 2019; Qi et al. 2019; Xue et al. 2020; Qin et al. 2018) have been proposed for lesion segmentation, and their improvement is mainly to learn effective features in the encoding phase and recover target details and space efficiently in the decoding phase. Unfortunately, because of insufficient extraction of contextual information, these methods unresolvable inter-class similarity challenges. (1) Fine-grained spatial attention (Liu et al. 2020; Wang et al. 2017; Dou et al. 2020; Li, Li, and Fan 2021) methods have been proposed to capture spatial contextual information and locate the location of the target, as show in Fig. 1(c), the yellow color is the attention region. Although it is good in the ideal case, the low-level features learned in the encoding stage and the generation process of the attention map cannot be effectively normalized. (Yu et al. 2020; Yu et al. 2020). Coarse-grained attention in Fig. 1(d) is weighted at the patch level (small red squares represent an attention patch), and the method requires less supporting information. (2) As in Fig. 1(a), MDU only identifies the easily identifiable lesions regions, due to insufficient boundary learning ability of 2D convolution. To obtain better boundary segmentation, previous works (Fang et al. 2019; Wang et al. 2021) used 3D convolution to compensate for the lack of 2D convolution. For example, aggregating features by direct summation or
concatenation is a naïve act (Qin et al. 2018), lacking the interaction of global information between two different tasks and introducing redundant information that can corrupt the information of the original features (Wei, Wang, and Huang 2020). (3) As shown in Fig. 1(e), after four pooling operations, the feature map becomes increasingly abstract and the boundary information, as well as the spatial information of the target, is lost. To recover this information, scholars (Yang et al. 2019; Xue et al. 2020; Zhao et al. 2017) have focused on feature fusion in the encoding and decoding stages with the aim of replenishing the lost information. However, little attention has been paid to the improvement of features during upsampling, and simple linear interpolation of upsampling is detrimental (Gu et al. 2019).

Facing these challenges, in this work, we design three modules that are integrated into a new network TC-Net to efficiently learn the spatial and boundary information of the target in the encoding phase, and fully retain and recover it in the decoding phase. To capture spatial contextual information, we design a coarse-grained patch attention module and generate patch-level attention maps in a supervised manner, as shown in Fig. 1(d), the regions of the attention method are weighted in the patches, which enhances the distinction between the target region and the background and solves their similar problems. Meanwhile, a cross-feature fusion module is introduced to fuse 2D and 3D features to obtain more accurate boundary segmentation, for capturing global contextual information between all channels in both dimensions, adaptively fusing complementary features of 2D and 3D features, suppressing each other’s noise, and providing a detailed detail representation for the decoding stage. All the above modules are limited to the encoding stage, and the simple upsampling operation in the decoding stage still cannot effectively retain and recover the detailed information of the target. Finally, we further propose a multi-scale deconvolution upsampling module instead of linear interpolation to capture multi-scale information, better recover the lost spatial information and detailed features of the target assembled in the decoding stage, and accurately segment the boundaries of the target. The contributions of our work are summarized as follows:

- The coarse-grained patch attention is presented to drive the network to focus on the target region and better mine the features of the target region.
- We design cross-feature fusion to adaptively filter and fuse features of different dimensions, which can fuse the complementary parts between two dimensional features, enrich detailed features in the encoding stage.
- For each stage of the decoder, we build a multi-scale deconvolution upsampling module, which replaces the parameterless linear interpolation, and allow better recovery of detailed features and spatial information of the target.

**Related Works**

In recent years, contextual information extraction has been widely used in medical image segmentation. To enable CNN to extract target-specific features, the spatial domain attention mechanism (Liu et al. 2020; Wang et al. 2017; Dou et al. 2020; Li, Li, and Fan 2021) is widely used, which suppresses irrelevant tissue regions and enhances the features of the segmented lesion regions. DRANet (Xue et al. 2020) used feature mapping noise from the attention mapping encoding stage to provide high-quality target features for the high-level decoding stage. However, the encoder partly tends to learn detailed features and lacks regularization, which is not sufficient to support the network to produce accurate pixel-based fine-grained target locations. For this reason, some networks incorporate a supervision mechanism, such as CPNet (Yu et al. 2020), which introduced an a priori knowledge inter-class affinity map to supervise the generation of the attention map and ensure its accuracy, but it still supervises only one stage of the attention map. Since the relative positions of the targets in the feature maps do not change, the coarse-grained attentional approach of prior knowledge can supervise the generation of attention maps for each layer. Although the above method helps the network to extract target-specific features, 2D convolution has insufficient boundary learning capability in volumetric data.

In semantic segmentation, the fusion of features containing different information can enrich the feature representation and obtain better segmentation performance. The combination of 2D and 3D convolution can maintain the accuracy of the target boundary, GGPFN (Fang et al. 2019) fused the segmentation results of 2DCNN and 3DCNN directly, but this lack of information interaction. M2D3DNet (Wang et al. 2021) fused 2D and 3D features at the encoding stage to facilitate feature complementary interactions. The features of the same task have different distribution characteristics, this equal treatment of all feature maps does not consider the global information of different dimensions and ignores the dependencies between the features of different tasks. Therefore, it is necessary to design a method to learn global information of different task features to guide the fusion. However, the boundary information is lost with multiple pooling, along with the spatial information, so this information needs to be effectively remediated during the encoding phase.

CNN are able to recover details of target features by capturing multi-scale features containing rich spatial and semantic information with different convolutional kernel sizes (Zhang et al. 2021). The atrous spatial pyramid pooling (ASPP) module was proposed in Deeplab V2 (Chen et al. 2017) to learn multi-scale contextual information. PSPNet (Zhao et al. 2017) used the pyramid pooling module (PPM) to enhance the network’s ability to exploit global contextual information. Multi-scale has also been widely integrated into U-shaped network structures. COPE-Net (Wang et al. 2020) used ASPP modules in bottleneck regions to enrich multi-scale features, and CLCI-Net tied together feature maps from different phases of the network based on ASPP. All these networks utilize multi-scale information in the encoding phase, few of them use multi-scale in the decoding phase. However, successive pooling in the early stages of U-shaped networks leads to increasingly abstract feature maps and loss of spatial and detail information that cannot
be effectively recovered by simple linear interpolation and deconvolution.

Proposed Method

In this work, we propose a novel network capturing three kinds of contextual information, as shown in Fig. 2. Firstly, with spatial contextual information as the core, we introduce a coarse-grained patch attention module (CPA) to partition the target and the background in the 2D encoding phase to help extract target-specific features. Secondly, to enrich the boundary representation capability of these features, a cross-feature fusion (CFF) module is proposed at the intersection of the 2D and 3D encoding layers to achieve more effective multi-dimensional feature fusion. Finally, the multi-scale deconvolution upsampling module (MDU) is designed to capture the multi-scale contextual information and recover the target location information and boundary information lost after multiple pooling from the abstract feature map. Next, we describe these modules in detail.

Coarse-grained Patch Attention (CPA)

In the 2D encoding layer of the network, we design coarse-grained patch attention with prior knowledge to enhance target cues in the feature map and reduce redundant noise information in the background, to generate high-quality features. Here we present the details of the coarse-grained patch attention (CPA) module.

As shown in Fig. 3, the feature maps generated from the features of each phase of 2D encoding in the network are used as input to the module to generate a coarse-grained attention map of the same resolution size, as shown in the following operation:

\[
F_g = \text{Avgpool}(\text{Avgchannel}(F_i)) \quad i \in 1, 2, 3, 4, 5
\]

\[
F_i \in \mathbb{R}^{C \times H \times W} \text{ represents the feature maps outputted by two layers of convolution at each encoding stage, the Avgchannel represents the average pooling along the channel axis, and Avgpool represents the average pooling over each channel. We first use the feature maps obtained using Avgchannel in the shape of 1*H*W, which highlights the information region (Zhou et al., 2018). Then the global information of each patch in the feature map is obtained using Avgpool to generate } F_g \in \mathbb{R}^{1 \times H \times W}, \text{ covering one patch per pooling step.}
\]

We use a multilayer perceptron to model the interdependence of the global information of each patch, as in Eq. 2. \(f_1, f_2\) denote the two-layer perceptron, and \(\varphi\) denotes the Sigmoid activation function. Therefore, \(F_g\) is expanded into \(F_g' \in \mathbb{R}^{36}\) and fed into the multilayer perceptron with 18 neurons in the middle layer and 36 neurons in the last layer, and \(\omega\) is obtained by significantly mapping the Sigmoid activation function with weights of continuous values between 0 and 1. The output of each neuron represents the probability of having lesioned regions for each patch block.

\[
\omega = \varphi(f_2(f_1(F_g')))
\]

The obtained \(\omega\) reshape into a 1*6*6 probability map \(\omega'\), and finally recovered into a 1*H*W coarse-grained attention map \(A\) by linear interpolation operation, where the upsampling operation is described as Eq. 3 and multiplying the input feature map element-by-element level by the element, as shown in Fig. 4, we can see that in the upsampled attention map \(A\), the weight scores of all pixel points in each patch are the same, and unlike the weight scores in the fine-grained attention maps are based on each pixel point, we are based on the weight scores of patches. The final output of each stage of the encoding is given by \(Y = \omega_i * F_i + F_i\). In this way provide a coarse-grained focus on the location of lesions, so that the network focuses on patches with high weight scores, and improves the target feature representation by making the features more focused on features in the target region. Note that finally, we use the idea of a residual network to add the post-attentional feature maps to the input feature maps to reduce the learning difficulty of the attentional maps while improving the error tolerance of the attentional maps.

\[
A_{x \times h+m, y \times w+n} = \omega'_{x, y} \\
x, y \in 1, 2, 3, 4, 5, 6 \quad h, w = \frac{H}{6} \quad m, n \in 1, 2, 3, \ldots \frac{H}{6}
\]

The module learns in a supervised manner, we generate a 1*6*6 binary coarse-grained lesion location map using the max pooling pair ground truth on the channel, the pool size
is 32*32, each 32*32 patches as long as there is a lesion region is the one that needs attention, we use the binary coarse-grained lesion location map as prior knowledge and use the binary cross-entropy loss as coarse-grained loss function to supervise the generation of coarse-grained attention map in the network to obtain more accurate predictions.

**Cross-Feature Fusion (CFF)**

We design the cross-feature fusion module at the fusion point of the 2D encoding layer and 3D coding layer features for an adaptive fusion of features in two different dimensions to refine the boundary feature representation of 2D feature maps instead of naive and direct addition or concatenate.

The module takes 2D feature map $2D_{fm} \in \mathbb{R}^{C \times H \times W}$ and 3D feature map $3D_{tfm} \in \mathbb{R}^{C \times H \times W \times D}$ as input. In each CFF module, we first downscale the 3D feature maps, then concatenate the feature maps between different dimensions and model the correlation between the combined feature channels to enhance the sensitivity of the network to features between different dimensions.

To be more specific, we first compress 3D feature map $3D_{tfm} \in \mathbb{R}^{C \times H \times W \times D}$ to $2D_{tfm} \in \mathbb{R}^{C \times H \times W \times 1}$ using a single channel 1*1*1 3D convolution. Then, we use C $3 \times 3$ 2D convolutions to convert the 3D feature map $2D_{tfm}$ after squeeze to the 3D transformed feature map $3D_{tfm} \in \mathbb{R}^{C \times H \times W \times D}$. Then, the 3D feature map and the 3D transformed feature map are then fed into the subsequent module, as shown in Fig. 5.

We use global average pooling to efficiently learn the global information on each channel of $3D_{tfm}$ and $2D_{fm}$. In order to better utilize the information between different dimensions, the network is phased to obtain discriminative features and reweight the features of each dimension, which helps 2D features to reconstruct finer boundary details. We use the global information on different dimensions in series to get $g \in \mathbb{R}^{2C}$, and feed $g$ into the multilayer perceptron, $f_1^2, f_2^3, f_3^3$ denote the two-layer perceptron, $\varphi$ denotes the Sigmoid activation function, $\odot$ denotes concatenate. For this purpose, $g$ is fed into the multilayer perceptron after modeling the combined different dimensional feature channel correlations, the middle layer the number of neurons is C/8 and the number of neurons in the last layer is C. The channel weights $\omega_2, \omega_3 \in \mathbb{R}^C$ in each dimension are obtained after significant mapping of the Sigmoid activation function, as Eq. 5 and Eq. 6.

$$g = \text{Avgglobal}(3D_{tfm}) \odot \text{Avgglobal}(2D_{fm})$$

(4)

$$\omega_2 = \varphi(f_2^3(f_1^2(g)))$$

(5)

$$\omega_3 = \varphi(f_3^3(f_1^2(g)))$$

(6)

The obtained weights $\omega$ are used to remap $3D_{tfm}$ and $2D_{fm}$ to obtain the feature map output $\in \mathbb{R}^{C \times H \times W}$ with fused 2D and 3D information, as Eq. 7. CFF applies contextual information of features in different dimensions, and if the weight score of another corresponding channel is high at the time of fusion, the deep features of that channel should be emphasized, which mutually emphasizes important features in different dimensions at the time of fusion and suppresses irrelevant features, thus enhancing the contextual semantic dependence between channel feature maps and finally enhancing the discriminative power of feature maps at the time of fusion.

$$output = 3D_{tfm} \ast \omega_3 + 2D_{fm} \ast \omega_2$$

(7)

**Multi-scale Deconvolution Upsampling (MDU)**

In U-shaped networks, parameterless linear interpolation and simple deconvolution does not recover well the spatial and detail information lost by successive pooling in the encoding phase, making segmentation challenging. To address this problem, we propose a multi-scale deconvolution upsampling module that uses multi-scale contextual information to recover lesion region features.

![Figure 5: The architecture of the proposed cross-feature fusion. 2D feature maps with coarse-grained patch attention and transposed 3D feature maps are used as input, and each channel feature map in different dimensions is filtered and mapped before being summed and fused.](image)

![Figure 6: (a, b, c): The 3*3 deconvolution moving process. (d): The coverage of the missed information by the 5*5 convolution on the 3*3 convolution.](image)

As shown in Fig. 6, the 9*9 white and blue feature maps are the expanded feature maps during the deconvolution process, the white area is the 0 value expanded by the deconvolution operation, the blue region being the valid region of
the original input feature map, and the gray rectangle representing the 3*3 convolution operation. In the 3*3 convolution, less than half of the valid information is undoubtedly detrimental to the recovery of detailed features in the lesion region, and this problem is fatal for small targets. By comparing the three figures a,b and c we can find that during the convolution transformation of 3*3 size, the relationship between some pixel points is ignored due to the limitation of sensory field, for example, the relationship between red pixel points 1, 2, 3 and 4 in the d figure is ignored. However, these pixel points can be covered by the 5*5 convolution. Similarly, 5*5 convolution has this drawback, but 7*7 convolution can make up for it. So we use multiple convolution kernels with different sizes of deconvolution to compensate for these drawbacks, which can improve the overall structure of the region by more multi-scale contextual and better recover the lesion region features. Considering the computational effort of the downsampling stage and the size of the feature map, we only use 3*3, 5*5, 7*7, and 9*9 deconvolution.

Figure 7: The architecture of the proposed multi-scale deconvolution upsampling.

As in Fig. 7, the module \( \text{input} \in \mathbb{R}^{C\times h \times w} \), we use different convolution kernel size of the inverse convolution to enlarge the feature map, specifically the convolution kernel size is 3*3, 5*5, 7*7, 9*9 convolution, and the padding size is 1, 2, 3, 4, respectively, at all yield an eigenmap of \( \mathbb{R}^{C\times 2H \times 2W} \). Then use to concatenate them together as \( \text{output} \in \mathbb{R}^{C\times 2H \times 2W} \), each convolution is followed by BN, Relu and Dropout operations. This not only avoids overfitting and GPU memory usage but also increases the richness of the learned representations, as demonstrated by similar designs.

Experiments and Results

Datasets and Preprocessing

We evaluate our method on the open-source dataset ATLAS (anatomical tracing of post-stroke lesions), this dataset is used for segmentation of post-stroke lesions (Liew et al. 2018), which contains T1-weighted MRI scans of 229 stroke patients. The 3D image of each case contains 189 slices of 233*197. To accommodate the needs of the network, we took the 3D images of each case and sliced them along the Z-axis, and cropped the images to 180*180. Next, the image is resized to 194*194, and the first two adjacent images of each slice are selected images and the next adjacent slice of each slice, for a total of 4 channels, as the input to each network. When training was performed, 40 patients in the training set were randomly selected, and the single slices of which were lesion-free regions were removed to speed up the network training.

Implementation Details

Our proposed TC-Net takes as input a multi-channel image of 192*194*4. The network is initialized using the Kaiming Initialization method. The optimizer used to train the network in the training phase is Adam, with an initial learning rate of 10-3. Also, for the stability of the network training, our learning rate decay strategy uses exponential decay, with a decay floor of 0.96. The deep learning framework we use is PyTorch, version Linux 1.2.0. Each layer uses a different Dropout to prevent overfitting of the network, and is trained for 150 epochs before the network is tested. The tests for the training of the experiments were run on NVIDIA Tesla V100 GPUs.

Comparison with State-of-the-Art Methods

In this section, we compare our approach with other well-known methods: Segnet mentioned in (Ito, Kim, and Liew 2019); CLCI-Net, X-Net, Psnet, U-Net, Deeplab v2, and Unet++ (Zhou et al. 2018). For fairness and comparability, we compare them on the same computational environment as well as on the same data processing. Table 1 compares the DSC, ASD, and HD values of the eight methods in detail. ASD and HD reflect the goodness of the segmentation result boundary, and the best value of each metric is highlighted in black in the table. DSC(global) provides a more visual representation of the overall DSC score based on voxels, and the remaining metrics are based on the mean of each patient in the test set mean ± standard deviation. For the ATLAS stroke dataset segmentation, TC-Net obtained the highest mean DSC score of 0.594, an improvement of 1.3% compared to the current best CLCI-Net. Also comparing to all networks, DSC( Global) is better than all the networks, which reflects the metric between recall and precisions. Compared with the conventional 2D CNN, our work is in the best condition in ASD and HD, indicating that our segmented boundaries are better than their methods, showing that our method has good extraction ability for boundary features. Although Psnet, Deeplab v2, and CLCI-Net all use multi-scale contexts, and U-net++ better combines the encoding and decoding phase features, it is still weaker than us in terms of segmentation effect, this reflects the effectiveness of using multi-scale information in the decoding stage.

We also visually compare the stroke segmentation results of our method with those of other methods. As shown in Fig. 8, our proposed method accurately identifies the location of lesion regions in the confusing lesion and background environments. Although, multi-scale information in Psnet and Deeplab V2, X-Net got feature similarity module captured the location information but confusing lesions and normal tissues. Even with low contrast between tissue and lesions near the border, our network still obtains the best border segmentation results, the red part of the figure is the boundary of the ground truth, and we can see that the result of our
Table 1: Comparison of the accuracy of our proposed method with other well-known methods on the Atlas dataset. "-" indicates that the corresponding result is not given or cannot be measured.

| Method   | DSC   | DSC(global) | Recall  | Precision  | ASD       | HD        |
|----------|-------|-------------|---------|------------|-----------|-----------|
| U-Net    | 0.485±0.298 | 0.712 | 0.469±0.296 | 0.574±0.328 | 15.145±20.374 | 52.972±34.855 |
| Segnet   | 0.451±0.293 | 0.6629 | 0.440±0.304 | 0.545±0.301 | 12.226±17.091 | 45.671±31.776 |
| Pspnet   | 0.468±0.283 | 0.6723 | 0.443±0.284 | 0.539±0.307 | 15.406±26.172 | 44.910±34.688 |
| Deeplab v2 | 0.415±0.263 | 0.6411 | 0.399±0.266 | 0.492±0.299 | 14.195±22.463 | 41.596±32.93 |
| Unet++   | 0.511±0.291 | 0.728 | 0.494±0.484 | 0.638±0.352 | 13.964±23.576 | 37.544±32.748 |
| X-Net    | 0.499±0.300 | 0.6724 | 0.466±0.303 | 0.633±0.312 | 11.611±16.543 | 46.499±34.197 |
| CLCI-Net | 0.581 | - | 0.581 | 0.649 | - | - |
| Ours     | 0.594±0.273 | 0.754 | 0.579±0.287 | 0.713±0.275 | 7.137±14.988 | 27.005±30.353 |

Figure 8: Comparisons of our method, X-net, Pspnet, Segnet, Deeplab v2, Unet++, and U-Net on five different patients.

segmentation is closest to the real boundary. Although our approach is patch-level attention, the segmentation of small targets is still superior to other networks.

Ablation Studies of Our Method

In order to analyze the impact of each module, we performed ablation experiments using these modules and their combinations, and the results of the ablation experiments are shown in Table[2] to verify that the design modules are effective for improving segmentation accuracy. Specifically, we use CPA, CFF, and MDU to represent the three modules respectively.

It can be seen that adding any of the modules results in a DSC score higher than 0.57. We propose the CPA module, which adaptively highlights the target location and extracts features closely related to the lesion area. And after fusing the feature maps of different dimensions by CFF module, the boundary expression of the network is optimized, and the accurate boundary also provides the location information of the lesion, and the DSC reaches 0.581, although the precision is reduced to 0.667, the recall is improved to 0.569, and the recall is more meaningful in medical images, which can improve the accuracy of predicting the presence of disease. The above two modules are limited to the encoding stage only. We recover the image size while better recovering the detailed information of high-level semantic features in the decoding stage by the multi-scale deconvolution upsampling module. With the combination of all modules we achieve the best results, which shows the effectiveness of all components.

Here we visualize the effect of the CPA module, as shown in Fig.[9] The second row shows the attention map generated by our coarse-grained patch attention, where the brightly colored regions represent the patches that the network needs to pay attention to, the target regions have a high probability of being present in these patches. Comparing the contrast between the background and lesion regions of the feature maps before and after attention, the task-relevant regions are significantly enhanced, indicating that the module has good localization ability and the ability to help the network extract features better, and try to exclude the interference of irrelevant features and reduce noise, providing a low-noise feature map during high and low level feature fusion. At the same time, our attention map allows for effective supervision, and as the number of network layers deepens, our attention map becomes more accurate and more consistent with the a priori
Table 2: Verify the impact of each module on the baseline.

| Module                  | DSC     | DSC(global) | Recall   | Precision | ASD     | HD       |
|------------------------|---------|-------------|----------|-----------|---------|----------|
| Baseline               | 0.726   | 0.548±0.264 | 0.530±0.268 | 0.665±0.303 | 8.932±16.565 | 36.552±30.262 |
| Baseline+CPA           | 0.741   | 0.574±0.272 | 0.561±0.290 | 0.682±0.297 | 7.833±13.900 | 33.544±30.346 |
| Baseline+CFF           | 0.734   | 0.572±0.244 | 0.538±0.244 | 0.701±0.271 | 8.670±17.513 | 30.197±29.383 |
| Baseline+MDU           | 0.735   | 0.570±0.276 | 0.541±0.274 | 0.701±0.302 | 7.972±15.818 | 27.220±29.333 |
| Baseline+CPA+CFF       | 0.744   | 0.581±0.269 | 0.569±0.277 | 0.667±0.302 | 7.521±16.156 | 29.174±28.246 |
| Baseline+CPA+CFF+MDU   | 0.754   | 0.594±0.273 | 0.579±0.287 | 0.713±0.275 | 7.137±14.988 | 27.005±30.353 |

Figure 9: The first to fifth columns show the feature map illustrations in the five coarse-grained location attention modules in the network coding stage. The first to the third rows are the input feature maps, coarse-grained attention maps and output features of the module, respectively. The rightmost part of the figure shows the ground truth as well as the a priori binary lesion location map.

Figure 10: The vector of attention values for the 2D branch weights of the two phases with the addition of CFF and with the addition of CFF and CPA are shown. The color bar on the right side is the color corresponding to the weight values.

map generated based on ground truth, although the attention map of the first and second columns is not as accurate as the last three columns, the contrast between the patches of the target region and other patches in the attention map is still obvious, and we borrow the idea of Resnet to make the feature map of important information will not be suppressed because of the shortage of the pre-attentive map.

As shown in Fig. 9, we can see that the feature map is very abstract after four times of ensemble downsampling. Join this time still using simple interpolation upsampling or single deconvolution method is too coarse, so it is necessary to complete the upsampling operation while better recovering the lesion details with the help of multi-scale contextual information. It can also be seen from Table 2 that after adding MDU, the segmentation effect has been improved, the improvement of Recall also reflects the accuracy of the segmentation position of our method, and the ASD and HD reflect that the segmentation boundary effect has also been improved.

To more effectively illustrate the extraction of higher-level features that contribute to the segmentation task, we quantitatively support this by the weights of each channel in the cross-feature fusion module. Because high-quality features allow better identification of segmentation lesions, the variability between each feature increases and each channel is a response to the input feature map to the feature, so the variability between channels increases and the difference between each channel weight value increases (Yang et al. 2020). As shown in Fig. 10 (b) and (d) are the weights of all channels of the 2D feature map after adding the cross-feature fusion module, respectively. Observing the vector map, the color difference of the vector map increases after adding CPA at each stage, both the channel weight value variance increases, especially in (d), the difference of the weight value variance is more obvious, which proves that the network extracts higher quality features after the CPA module. This is because as the network deepens, our CPA module provides more accurate information about the location of the lesion, as shown in Fig. 10 and the extracted features are more differentiated and related to the segmentation target, which will also cause the difference between channels to increase.

Conclusion

In this work, we propose a TC-Net that makes full use of contextual information for ATLAS lesion segmentation of stroke datasets. Our TC-Net integrates CPA and CFF modules in the encoding stage. The former uses a coarse-grained attention map to focus the network on the target region to extract features highly relevant to the target, and the latter fuses feature from 3D for the former for the purpose of refining the boundaries of the 2D feature map. Finally, the MDU are used in the decoding stage to recover the lesioned features from the feature maps that have been downsampled several times. The main advantage of our TC-Net is the use of spatial contextual information between each feature and different patches, as well as global contextual information between different dimensional channels and multi-scale contextual information, showing that the application of contextual information is crucial for performance improvement and demonstrates a great advantage when compared with existing methods.
References

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4): 834–848.

Dou, H.; Karimi, D.; Rollins, C. K.; Ortinau, C. M.; Vasung, L.; Velasco-Annis, C.; Oualam, A.; Yang, X.; Ni, D.; and Gholipour, A. 2020. A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal MRI. IEEE transactions on medical imaging, 40(4): 1123–1133.

Fang, C.; Li, G.; Pan, C.; Li, Y.; and Yu, Y. 2019. Globally guided progressive fusion network for 3D pancreas segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 210–218. Springer.

Grzywacz, R. A.; Thomas, K.; and Pandey, D. K. 2008. Epidemiology of ischemic and hemorrhagic stroke: incidence, prevalence, mortality, and risk factors. Neurologic Clinics, 26(4): 871–895.

Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging, 38(10): 2281–2292.

Ito, K. L.; Kim, H.; and Liew, S.-L. 2019. A comparison of automated lesion segmentation approaches for chronic stroke T1-weighted MRI data. Human brain mapping, 40(16): 4669–4685.

Li, Y.; Li, H.; and Fan, Y. 2021. ACEnet: Anatomical context-encoding network for neuroanatomy segmentation. Medical Image Analysis, 70: 101991.

Liew, S.-L.; Anglin, J. M.; Banks, N. W.; Sondag, M.; Ito, K. L.; Kim, H.; Chan, J.; Ito, J.; Jung, C.; Khoshab, N.; et al. 2018. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Scientific data, 5(1): 1–11.

Liu, Y.-C.; Shahid, M.; Sarapugdi, W.; Lin, Y.-X.; Chen, J.-C.; and Hua, K.-L. 2020. Cascaded atrous dual attention U-Net for tumor segmentation. Multimedia Tools and Applications, 1–25.

Qi, K.; Yang, H.; Li, C.; Liu, Z.; Wang, M.; Liu, Q.; and Wang, S. 2019. X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In International conference on medical image computing and computer-assisted intervention, 247–255. Springer.

Qin, Y.; Kamnitsas, K.; Ancha, S.; Nanavati, J.; Cottrell, G.; Criminisi, A.; and Nori, A. 2018. Autofocus layer for semantic segmentation. In International conference on medical image computing and computer-assisted intervention, 603–611. Springer.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3156–3164.

Wang, G.; Liu, X.; Li, C.; Xu, Z.; and Zhang, S. 2020. A Noise-robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions from CT Images. IEEE Transactions on Medical Imaging, PP(99): 1–1.

Wang, H.; Cao, J.; Feng, J.; Xie, Y.; Yang, D.; and Chen, B. 2021. Mixed 2D and 3D convolutional network with multi-scale context for lesion segmentation in breast DCE-MRI. Biomedical Signal Processing and Control, 68: 102607.

Wei, J.; Wang, S.; and Huang, Q. 2020. F²Net: Fusion, Feedback and Focus for Salient Object Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 12321–12328.

Xue, Y.; Farhat, F. G.; Boukrina, O.; Barrett, A.; Binder, J. R.; Roshan, U. W.; and Graves, W. W. 2020. A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images. NeuroImage: Clinical, 25: 102118.

Yang, H.; Huang, W.; Qi, K.; Li, C.; Liu, X.; Wang, M.; Zheng, H.; and Wang, S. 2019. CLCI-Net: Cross-level fusion and context inference networks for lesion segmentation of chronic stroke. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 266–274. Springer.

Yang, Z.; Zhu, L.; Wu, Y.; and Yang, Y. 2020. Gated channel transformation for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11794–11803.

Yu, C.; Wang, J.; Gao, C.; Yu, G.; Shen, C.; and Sang, N. 2020. Context prior for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12416–12425.

Zhang, Y.; Li, H.; Du, J.; Qin, J.; Wang, T.; Chen, Y.; Liu, B.; Gao, W.; Ma, G.; and Lei, B. 2021. 3D Multi-Attention Guided Multi-Task Learning Network for Automatic Gastric Tumor Segmentation and Lymph Node Classification. IEEE Transactions on Medical Imaging, 40(6): 1618–1631.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2881–2890.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In Deep learning in medical image analysis and multimodal learning for clinical decision support, 3–11. Springer.