Computerized Medical Imaging and Graphics 112 (2024) 102323

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Rethinking automatic segmentation of gross target volume from a decoupling perspective

Jun Shi [a], Zhaohui Wang [a], Shulan Ruan [a], Minfan Zhao [a], Ziqi Zhu [a], Hongyu Kan [a], Hong An [a,b,*], Xudong Xue [c], Bing Yan [d]

[a] School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230026, China
[b] Laoshan Laboratory Qingdao, Qindao, 266221, China
[c] Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430074, China
[d] Department of radiation oncology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, 230001, China

ARTICLE INFO

Dataset link: Code and Data for GTV AutoSeg (Original data)

Keywords:
Radiation therapy
Gross target volume segmentation
Decoupling perspective
Spatial alignment
Deep learning

ABSTRACT

Accurate and reliable segmentation of Gross Target Volume (GTV) is critical in cancer Radiation Therapy (RT) planning, but manual delineation is time-consuming and subject to inter-observer variations. Recently, deep learning methods have achieved remarkable success in medical image segmentation. However, due to the low image contrast and extreme pixel imbalance between GTV and adjacent tissues, most existing methods usually obtained limited performance on automatic GTV segmentation. In this paper, we propose a Heterogeneous Cascade Framework (HCF) from a decoupling perspective, which decomposes the GTV segmentation into independent recognition and segmentation subtasks. The former aims to screen out the abnormal slices containing GTV, while the latter performs pixel-wise segmentation of these slices. With the decoupled two-stage framework, we can efficiently filter normal slices to reduce false positives. To further improve the segmentation performance, we design a multi-level Spatial Alignment Network (SANet) based on the feature pyramid structure, which introduces a spatial alignment module into the decoder to compensate for the information loss caused by downsampling. Moreover, we propose a Combined Regularization (CR) loss and Balance-Sampling Strategy (BSS) to alleviate the pixel imbalance problem and improve network convergence. Extensive experiments on two public datasets of StructSeg2019 challenge demonstrate that our method outperforms state-of-the-art methods, especially with significant advantages in reducing false positives and accurately segmenting small objects. The code is available at https://github.com/shijun18/GTV_AutoSeg.

1. Introduction

Radiation Therapy (RT) is an important and effective treatment for most cancers (Jaffray, 2012), using high doses of radiation to kill cancer cells and shrink tumors. Precise delineation of Gross Target Volume (GTV), the location and extent of the primary tumor, from Computed Tomography (CT) or other functional morphology images play an essential role in RT planning, which determines the accuracy of the radiation delivery (Weiss and Hess, 2003). Although it is time-consuming and labor-intensive, manual delineation still forms the majority of oncologists' work in clinical practice. In particular, due to the ambiguous boundary, the delineation accuracy of GTV depends heavily on the subjective experience of the operator, which can easily lead to inter-observer errors. These clinical challenges underline the need to introduce automated image segmentation technology into rapid and reliable RT planning.

Recently, deep learning methods, like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have been widely used in various medical image segmentation tasks with promising results. In the early research, much effort was devoted to exploring *end-to-end* structures, resulting in a series of segmentation algorithms represented by U-Net (Ronneberger et al., 2015) and its variants (Milletari et al., 2016; Chen et al., 2018b; Zhu et al., 2019; Oktay et al., 2018). Most of these methods employ an encoder–decoder structure, where the encoder is used to extract features, and the decoder fuses the multi-level features to restore resolution. Extensive studies (Zhu et al., 2019;

* Corresponding author.
E-mail addresses: shijun18@mail.ustc.edu.cn (J. Shi), wangzh95@mail.ustc.edu.cn (Z. Wang), slruan@mail.ustc.edu.cn (S. Ruan), zmf@mail.ustc.edu.cn (M. Zhao), ta1ly@mail.ustc.edu.cn (Z. Zhu), honeyk@mail.ustc.edu.cn (H. Kan), han@ustc.edu.cn (H. An), xuexudong511@163.com (X. Xue), bingyan29618@ustc.edu.cn (B. Yan).

https://doi.org/10.1016/j.compmedimag.2023.102323
Received 18 May 2023; Received in revised form 19 October 2023; Accepted 12 December 2023
Available online 29 December 2023

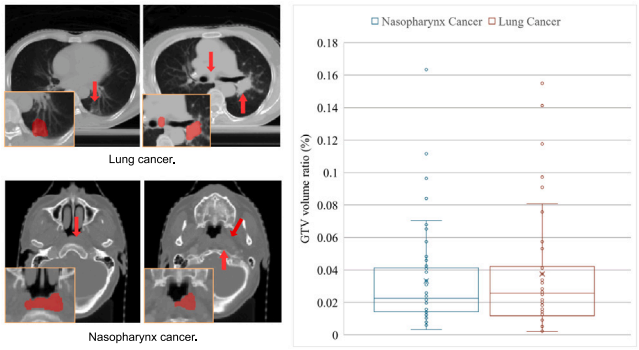


Fig. 1. Visualization and volume ratio statistics of GTV in the public StructSeg2019 CT dataset. The upper and lower rows on the left are imaging representations of GTV (red region) in lung and nasopharynx cancer, respectively, with low image contrast to adjacent tissues. The right shows the statistics results of the volume ratio of GTV in CT images, most of which are less than 0.1%.

Oktay et al., 2018; Isensee et al., 2021; Xiao et al., 2018; Ibtehaz and Rahman, 2020) have confirmed that existing methods can achieve segmentation accuracy comparable to human experts for large targets with clear boundaries in the image, such as the heart and lungs. However, there still remain some unresolved challenges when dealing with GTV segmentation. First, as shown in Fig. 1, the low image contrast between GTV and surrounding soft tissues dramatically increases the difficulty of extracting discriminative features and leads to a mass of false positives. Second, GTV regions are typically small and of varying shapes, with extreme pixel imbalance from the background. Fig. 1 shows the distribution of GTV volume ratio on StructSeg2019 CT dataset, most of which are less than 0.1%, making it difficult to optimize the network. Third, the inherent problem of spatial information loss caused by multi-level downsampling also limits the performance of many existing methods (Jiang et al., 2021).

To address the above problems, many researchers have focused on improving the feature representation capability of the network. Various attention mechanisms including spatial attention (Mei et al., 2021; Fu et al., 2021; Jiang et al., 2021; Sinha and Dolz, 2020) and non-local attention (Wang et al., 2020), are proposed to refine the hidden features. In Seo et al. (2019), Jiang et al. (2018), Zhou et al. (2019b), Chen et al. (2018a) and Zhou et al. (2019a), they tended to make the network capture richer contextual features by modifying or redesigning the skip connections between low- and high-resolution features. Chi et al. (2021), Zhou et al. (2022) and Yu et al. (2019) advocated using multi-branch structures to extract multi-scale features to enhance network representation. Furthermore, transformer structures (Vaswani et al., 2017; Dosovitskiy et al., 2020; Liu et al., 2021) derived from Natural Language Processing (NLP) have recently been introduced into medical image segmentation tasks (Hatamizadeh et al., 2022; Gao et al., 2021; Wang et al., 2021; Chen et al., 2021; Peiris et al., 2022), employing the multi-head attention mechanism to model long-term dependencies and detailed features. However, the performance gain from optimizing the network representation capability is limited, and further exploration is still needed to reduce false positives and improve network convergence.

Orthogonal to optimize the network structure, some studies tried to improve the segmentation performance by adding specific auxiliary tasks to the algorithm pipeline. In Isensee et al. (2021), Isensee et al. proposed a general automated framework for building medical image segmentation pipelines, dubbed nnU-Net, which adopts a coarse-to-fine segmentation paradigm. It constructs a two-stage *cascade* structure based on naive 3D U-Net, where the first stage serves as an auxiliary task and aims to segment objects for localization roughly, and the second stage performs fine-grained segmentation of the extracted region-of-interest (ROI). This method can effectively reduce irrelevant information about the target, thereby alleviating pixel imbalance and reducing false positives. Furthermore, Li et al. (2018), Wang et al. (2017) and Zhang et al. (2018) designed specific auxiliary tasks and more complex cascading ways according to task characteristics and obtained promising results. Despite their effectiveness, most of these methods still suffer from two significant drawbacks. First, 3D networks bring higher computational complexity, which means that the depth and width of the encoder are limited by GPU memory, resulting in an insufficient receptive field and weak representation ability. Second, the information loss in the previous stage will continue to propagate backward, causing the risk of damaging the segmentation performance of small objects. As an alternative, He et al. (2020) introduced a *multi-task learning* (MTL) structure, adding an extra classification branch to the backbone of the segmentation network to suppress false positives. The auxiliary classification task in this method is used to determine whether the slice contains candidate objects to revise the segmentation results. Zhou et al. (2021) and Zhang et al. (2021) advocated end-to-end joint learning of multiple subtasks based on the shared encoder. The premise of multi-task joint learning is that different subtasks rely on consistent features, but segmentation tasks require more detailed semantic features than classification tasks. Therefore, for challenging tasks such as GTV segmentation of lung cancer, it is difficult for the *multi-task learning* structure to converge to the optimal solution.

In our preliminary experiments of existing methods for GTV segmentation, we found that many false positives are distributed in the normal slices adjacent to GTV regions. Inspired by this observation, we added an independent classifier in front of the segmentation network to filter normal slices to reduce false positives, achieving the expected performance gain. To further improve the segmentation performance, this study redesigns the GTV automatic segmentation pipeline from a decoupling perspective, which provides the merit of reduced false positives and accurate small objects segmentation. The contributions of this study are summarized as follows:

- We propose a two-stage Heterogeneous Cascade Framework (HCF) that decomposes GTV segmentation into independent recognition and segmentation subtasks. The former screens out the abnormal slices containing GTV, and the latter performs pixel-wise segmentation of these slices. This method can effectively filter the normal slices to reduce false positives.
- We design a novel multi-level Spatial Alignment Network (SANet) based on the feature pyramid structure, which introduces a Spatial Alignment Module (SAM) in the decoder to compensate for the information loss caused by downsampling operations beneficial to improve the segmentation performance of small objects.
- We propose a Combined Regularization (CR) loss and Balance-Sampling Strategy (BSS) to the alleviate pixel imbalance problem during training.
- Extensive experiments on two public CT datasets of StructSeg2019 challenge demonstrate that our method outperforms existing methods, especially with significant advantages in reducing false positives and accurately segmenting small objects.

2. Related work

2.1. General segmentation structure

2.1.1. End-to-end structure

As a fully convolutional network (FCN), U-Net (Ronneberger et al., 2015) is the most representative end-to-end architecture for medical image segmentation, employing a symmetric encoder–decoder design. The encoder and decoder are used to extract features and restore resolution, respectively, and the corresponding features between the two parts are fused through skip connections. Extensive studies have confirmed that improving the feature quality of the encoder is beneficial to segmentation performance, leading to various attention mechanisms (Mei et al., 2021; Fu et al., 2021; Jiang et al., 2021; Wang

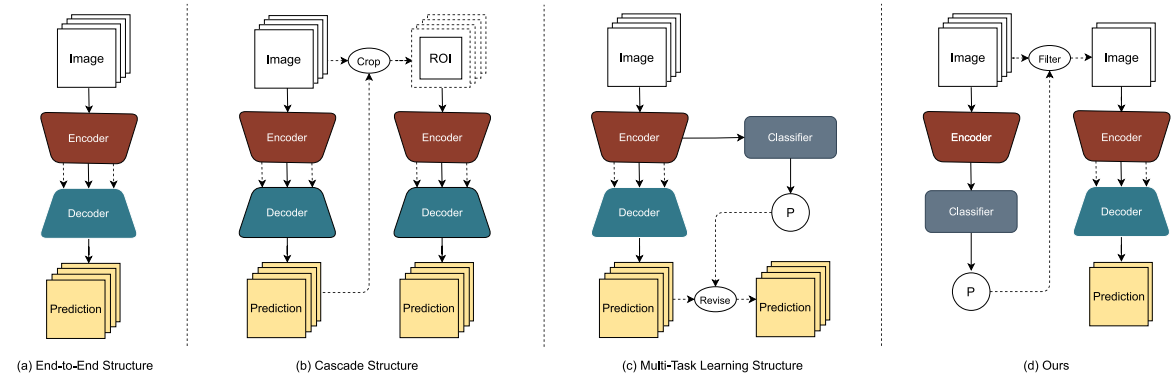


Fig. 2. Illustrations of structures for medical image segmentation. The encoder and decoder are used to extract features and restore resolution, respectively, and the classifier determines whether the input contains candidate targets.

et al., 2020; Sinha and Dolz, 2020) designed to refine hidden features. In addition, some researches (Seo et al., 2019; Jiang et al., 2018; Zhou et al., 2019b; Chen et al., 2018a; Chi et al., 2021; Zhou et al., 2022; Yu et al., 2019; Zhou et al., 2019a) focused on optimizing the way of feature fusion and capturing multi-scale contextual information to enhance the representation ability of networks. Based on the advantages of transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020) in modeling long-range dependencies and detailed features, recent works (Hatamizadeh et al., 2022; Gao et al., 2021; Wang et al., 2021; Chen et al., 2021; Peiris et al., 2022) have proposed the hybrid structures of CNNs and ViTs for improving segmentation performance. These methods have achieved promising results on specific organ and soft tissue segmentation tasks. However, the performance gain achieved by optimizing the end-to-end structure is limited when facing the challenges of GTV segmentation. It is still necessary to redesign the segmentation pipeline according to the data characteristics.

2.1.2. Cascade structure

A naive idea to further improve segmentation performance is to apply a multi-stage framework. The nnU-Net (Isensee et al., 2021) adopted a coarse-to-fine segmentation paradigm and built a two-stage cascade structure based on 3D U-Net, achieving state-of-the-art (SOTA) results on multiple benchmarks and challenges. The structure contains two isomorphic segmentation sub-networks: the former performs coarse segmentation to locate candidate objects, and the latter performs pixel-level segmentation of previously extracted ROIs. This method can reduce the redundant information in the image, thereby alleviating the pixel imbalance. Some similar approaches (Li et al., 2018; Wang et al., 2017; Zhang et al., 2018), focusing on the design of sub-networks and cascading patterns, also achieve slight performance gains. However, these isomorphic cascade structures are not only limited by GPU memory but also suffer from inherent defects of error propagation, which means that insufficient localization accuracy in the previous stage will hurt the final performance.

2.1.3. Multi-task learning structure

He et al. (2020) introduced an auxiliary classification branch into the network to correct the segmentation results achieving satisfactory performance in the multi-organ segmentation task. This design relies on the basic assumption that the classification network is more trustworthy than the segmentation network. In the prediction stage, the corresponding output of the segmentation network will be regarded as false positives and omitted when the classification network determines that there are no candidate targets in the input. Some studies (Zhou et al., 2021; Zhang et al., 2021) employed end-to-end joint learning of multiple subtasks to simplify the structure design. Most multi-task learning structures are based on shared encoders, which means that different subtasks rely on the same feature representations. However, segmentation tasks usually require more high-level and detailed semantic features than classification tasks, especially for complex objects. Therefore, the feature requirement gap between different subtasks will invalidate the multi-task joint learning strategy.

Given the data characteristics of GTV segmentation, this study proposes a heterogeneous 2D cascade structure from a decoupling perspective, consisting of a classifier and a segmentation network, as shown in Fig. 2(d). In the inference stage, the classifier aims to filter normal slices to reduce false positives, while the segmentation network makes predictions on the screened abnormal slices containing GTV. Compared with the isomorphic 3D cascade structures, our method has lower computational complexity and memory requirements. In addition, the independence between the classifier and the segmentation network is beneficial to avoid the feature inconsistency problem of multi-task joint optimization.

2.2. Semantic flow alignment

Most CNN-based encoders of segmentation methods contain multiple downsampling operations to expand the receptive field, resulting in the resolution of feature maps decreasing as the network deepens. For example, the feature map size in the final stage of ResNet (He et al., 2016) is only 1/32 the size of the input image. In the decoding part, the low-resolution features of deep layers are upsampled and concatenated with the high-resolution features of shallow layers to increase the resolution. However, this naive method cannot reduce the loss of spatial information caused by multiple downsampling operations, which is not conducive to pixel-level prediction, especially for segmenting small objects. Li et al. (2020a) propose the concept of semantic flow based on optical flow (Zhu et al., 2017), which means that the relationship between two semantic features of different resolutions from the same input can be represented with the motion of each pixel from one feature map to the other. They design the Flow Alignment Module (FAM) to model the flow field between features at arbitrary resolutions for feature alignment. Inspired by this research, we introduce the Spatial Alignment Module (SAM) to extract the spatial offset information between features at different resolutions to guide the reconstruction of high-resolution semantic features dynamically. Compared with the existing methods, our method can better compensate for the loss of pixel-level spatial information caused by downsampling and further improve the segmentation performance of small objects.

3. Methods

3.1. Heterogeneous cascade framework

A complete CT image usually consists of nearly a hundred 2D slices with high content continuity. Compared with the background, the volume ratio of GTV is tiny (less than 0.1%), resulting in only a portion of

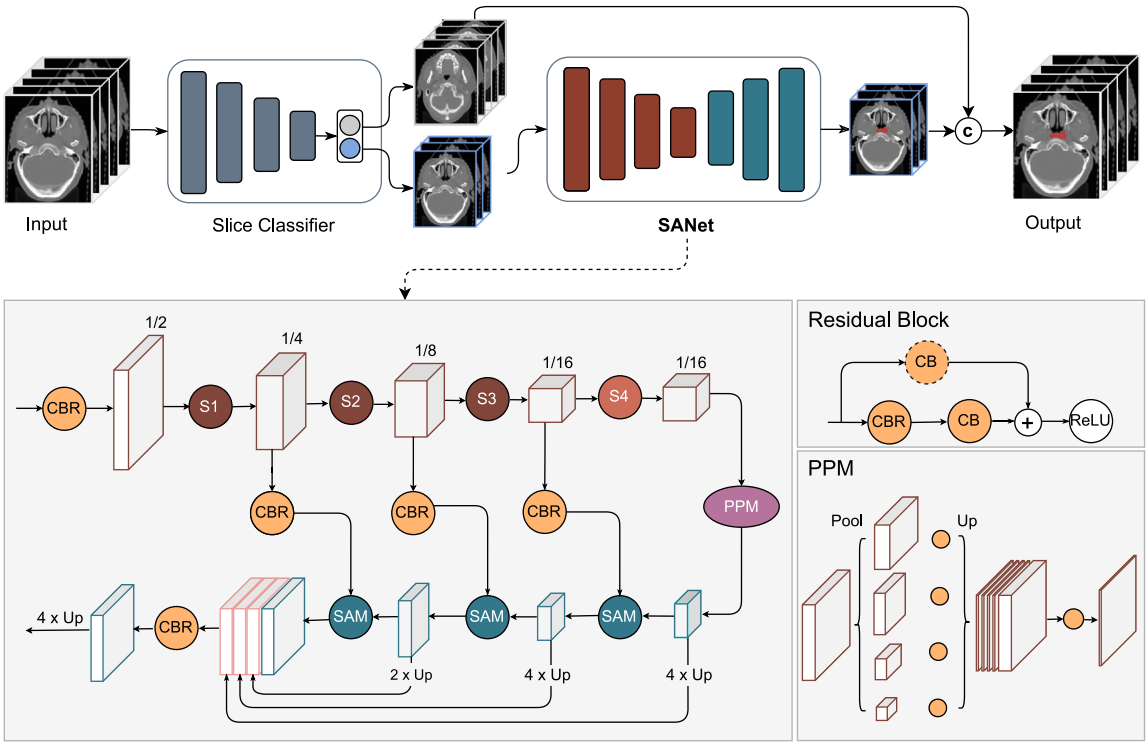


Fig. 3. Overview of the proposed method, the HCF pipeline consists of a slice classifier and SANet in tandem. During the inference phase, SANet only makes predictions on the abnormal slices containing GTV regions filtered out by the classifier. S1-4 represent the four stages of the SANet encoder, each containing two residual blocks. Unlike S1-3, S4 adopts atrous convolution to capture higher-resolution semantic features. CBR refers to the Conv-BN-ReLU operator cluster, and CB represents the form after removing ReLU.

the slices containing visible GTV regions. The large-scale noisy samples (slices without GTV) can seriously interfere with the training of 2D segmentation networks, while extreme pixel imbalance also potentially limits the performance of 3D segmentation networks. Furthermore, most existing methods suffer from the drawback of predicting a mass of false positives due to the low contrast between GTV and surrounding soft tissues. We observe that these false positives are mainly distributed in normal slices adjacent to the GTV regions. Inspired by this, we propose a heterogeneous 2D cascade framework to decompose the GTV segmentation task into independent recognition and segmentation subtasks. Fig. 3 illustrates the overall architecture of HCF, composed of a classifier and a segmentation network in tandem. In the inference stage, the segmentation network only predicts the abnormal slices containing GTV identified by the classifier. In contrast, the segmentation results of the slices without GTV are all-zero mapping by default, and then the two results are concatenated to obtain the final output. This decoupling framework can effectively filter normal slices and reduce false positives. Besides, it facilitates the replacement and optimization of either sub-network since the classifier and segmentation network remain independent. Considering the computational complexity and available data scale, we choose ResNet18 (He et al., 2016) as the slice classifier. To further improve the segmentation performance, we propose a Spatial Alignment Network (SANet) based on the feature pyramid structure. The following sections present the details of each network component and the motivations behind them.

3.2. Spatial alignment network

Most existing segmentation networks adopt an encoder–decoder architecture, where the encoder provides feature representation and the decoder serves for resolution reconstruction. Standard CNN-based encoders employ multiple downsampling operations in the form of max-pooling or large-stride convolution to extract high-level semantic features. Although the downsampling operation can enlarge the receptive field, it also leads to losing local spatial information. In contrast to the encoder, the decoder aims to reconstruct high-resolution semantic features through custom upsampling operations such as bilinear interpolation and deconvolution. However, the simple upsampling methods are deficient in making up for the loss of detailed spatial information caused by multiple downsampling operations, which is not conducive to pixel-level prediction. In other words, obtaining high-resolution features with strong semantic representation is critical in improving semantic segmentation performance. To alleviate this problem, we propose SANet equipped with several SAMs based on the feature pyramid structure, as shown in Fig. 3. In addition, according to the data characteristics of GTV segmentation, we also make adaptive adjustments in other parts of the network structure to further improve the segmentation performance.

3.2.1. Encoder part

Given the computational complexity, we still utilize ResNet18 as the encoder backbone by removing the last fully connected layer. As shown in Fig. 3, the encoder backbone has an embedding head and four stages. The embedding head consists of a 7×7 convolutional (Conv) layer with a stride of 2, a Batch Normalization (BN) layer, and a ReLU activation layer. Each stage contains two basic residual blocks, each composed of two consecutive clusters of Conv-BN-ReLU operators and residual connections. Notably, a convolutional layer with stride 2 precedes the first residual block of each stage in the original ResNet18 to downsample the feature map chasing for a larger receptive field. However, five downsampling operations make the final feature map only $1/32$ the size of the original image leading to a massive loss of spatial information, which increases the difficulty of subsequent resolution reconstruction, especially for small sizes and varying shapes of GTV regions. Therefore, we adjust the last stage by setting the stride of all convolutional layers to 1 while replacing the standard convolution with atrous convolution (Chen et al., 2017) with a dilation rate of 2 to expand the receptive field. The reason is that higher-resolution semantic features can retain more spatial information, which

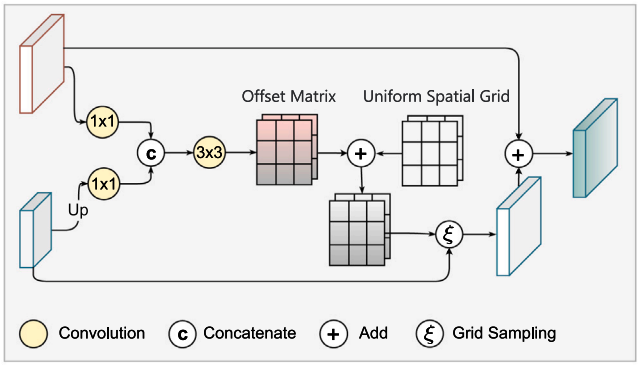


Fig. 4. Overview of the proposed SAM. SAM is used to reconstruct high-resolution features by embedding spatial offset information between features with different resolutions.

is beneficial for pixel-level prediction. Besides, we further optimize obtained semantic features using a Pyramid Pooling Module (PPM) to capture multi-scale contextual information.

3.2.2. Pyramid pooling module

PPM was first proposed by Zhao et al. (2017) to provide effective global context priors for scene parsing. To further improve semantic feature quality, we use PPM as the bridge component between the encoder and the decoder. Fig. 3 shows the detailed structure of the PPM module, including four pooling paths and an identity mapping path. The core idea is to use multiple pooling operations in parallel to extract multi-scale contextual information and fuse them. The four pooling paths use different pooling factors $n_i \in \mathbf{N}$, where $\mathbf{N} = \{1, 2, 3, 6\}$ in our experiments. Noted that the number of pyramid levels and the pooling kernel size at each level can be modified, and we followed the default configuration in Zhao et al. (2017) according to the experimental results. Different pooling operations generate feature maps with different resolutions, followed by a 1×1 convolutional layer and a BN layer for dimensionality reduction. Then the feature maps are upsampled to get the same size as the original resolution via bilinear interpolation. Finally, the four feature maps are spliced with the original feature map in the channel dimension and scaled to the specified size by a Conv-BN-ReLU operation cluster, forming the final pyramid pooling global feature. Compared with the original features, the semantic features processed by PPM have richer contextual information, which benefits the segmentation of GTV regions with variable shapes.

3.2.3. Spatial alignment module

To generate high-resolution features with both strong semantic representation and rich spatial information, the feature maps of deep layers are usually upsampled and fused with the feature maps of shallow layers in the form of skip connections. As the most common upsampling operations, bilinear interpolation and deconvolution are widely used in U-like and FCN-like segmentation methods. However, these upsampling methods cannot compensate for the loss of detailed spatial information caused by repeated downsampling. For example, bilinear interpolation reconstructs the high-resolution feature maps by interpolating a set of uniformly sampled locations, which only works for the fixed pattern. In this case, there is a misalignment problem between the two features used for fusion, resulting in insufficient spatial information and poor boundary details of the acquired features. To address this problem, Li et al. (2020a) proposed the concept of semantic flow and designed the FAM to model the flow field between features at different resolutions. Inspired by this work, we extend FAM in this paper to introduce a learnable SAM to explicitly represent the spatial offset information between two features to dynamically guide the reconstruction of high-resolution semantic features and alleviate the feature misalignment. The essence of SAM is to transfer the spatial information from shallow layers to deep layers to improve the quality of fusion features further. Fig. 4 illustrates the structural details of the SAM.

Given two feature maps from different depths of the network $\mathbf{X}_i \in \mathbb{R}^{H_i \times W_i \times C}$ and $\mathbf{X}_j \in \mathbb{R}^{H_j \times W_j \times C}$, they have the same channel depth C. Where $i, j \in [1, 2, 3, 4]$ is the resolution size factor, H and W represent the height and width of the original image, respectively, $H_i = \frac{H}{2^i}$, $W_i = \frac{W}{2^i}$. When $j > i$, we upsample $\mathbf{X}_j$ to the same resolution as $\mathbf{X}_i$ via a bilinear interpolation. Then, we use two 1×1 convolutional layers to compress the channels of $\mathbf{X}_i$ and $\mathbf{X}_j$ to $\frac{C}{2}$, respectively, and concatenate them in the channel dimension. A subsequent 3×3 convolutional layer takes the concatenated feature maps as input and generates a spatial offset matrix $\Delta_i \in \mathbb{R}^{H_i \times W_i \times 2}$, which refers to the pixel-by-pixel spatial position offset of the feature map after downsampling. Finally, add Δ_i and the uniformly sampled spatial grid $\Omega_i \in \mathbb{R}^{H_i \times W_i \times 2}$ bitwise to get the calibrated sampling grid $\mathbf{G}_i$. Mathematically, the above steps can be written as:

$$\Delta_i = \delta_{3\times3}(cat(\delta_{1\times1}(\mathbf{X}_i), \delta_{1\times1}(\zeta(\mathbf{X}_j)))), \tag{1}$$

$$\mathbf{G}_i = \Omega_i + \Delta_i, \tag{2}$$

where $\zeta(\cdot)$, $\delta(\cdot)$ and $cat(\cdot,\cdot)$ represent bilinear interpolation, convolution and concatenation operations, respectively. After obtaining the sampling grid $\mathbf{G}_i$, we use the grid sampling to reconstruct the high-resolution feature map based on $\mathbf{X}_j$. The final output of SAM is:

$$\tilde{\mathbf{X}}_i = \xi(\mathbf{X}_j, \mathbf{G}_i), \tag{3}$$

where $\xi(\cdot,\cdot)$ represents the grid sampling method, which takes the low-resolution feature map and sampling grid as input and generates the output at the specified resolution. Compared with standard bilinear interpolation, the grid sampling method is more flexible, and the sampling pixel position can be specified by customizing the sampling grid instead of uniform sampling. Overall, the proposed SAM is a learnable upsampling way to achieve high-quality feature reconstruction by embedding the spatial offset information between features at different resolutions. Furthermore, SAM is lightweight that brings only a minor additional computational overhead. In the experimental part, we conduct an ablation study to verify the effectiveness of SAM.

3.2.4. Decoder part

As shown in Fig. 3, the decoding path of SANet contains three SAMs for recovering high resolution. In particular, each upsampled feature map is added to the corresponding feature map of the shallow layer through a skip connection to enhance the representation further. The feature maps of shallow layers are processed by a 1×1 Conv-BN-ReLU operator cluster to reduce the computational complexity before being transmitted to the decoder. In our experiments, the channel depth of each dimension-reduced feature map is set to 128, and the size of the feature map after three SAMs is 1/4 the size of the original image. Then, we upsample the other three feature maps on the decoding path to the same resolution as the feature maps of the last layer and concatenate them to get a fusion feature. Finally, the segmentation result is obtained by upsampling the fusion feature by a factor of 4.

3.3. Combined regularization loss

As mentioned above, the extreme pixel imbalance between GTV and the background leads to unstable network convergence and failure to find an optimal solution. Many studies (Hossain et al., 2021; Li et al., 2020b; Nasalwai et al., 2021; Taghanaki et al., 2019; Kervadec et al., 2019) have tried to alleviate this problem by optimizing the loss function, such as introducing Focal loss (Hossain et al., 2021) and Tversky loss (Nasalwai et al., 2021), but the performance gain in our task is limited. In this paper, we design a novel combined loss with

a regularization term, named CR loss, to optimize the segmentation network. The proposed loss function is defined as follows:

$$\zeta_{CR} = \zeta_{dice} + \zeta_{ce-\epsilon}. \tag{4}$$

ζ_{dice} represents the Dice loss, which is region-dependent and tends to focus on foreground pixels. Its mathematical formula is as follows:

$$\zeta_{dice} = 1 - \frac{1}{K}\sum_{k=1}^{K}\frac{2\sum_{i=1}^{N} p_i^k q_i^k}{\sum_{i=1}^{N} p_i^k + q_i^k}, \tag{5}$$

where K refers to the number of segmentation classes, N refers to the total number of pixels in the image, p_i^k and q_i^k are the predicted probability and ground truth of pixel i in category k, respectively. However, when the segmentation target is small, the gradient of Dice loss will oscillate wildly. Thus, we introduce an additional cross-entropy loss with a regularization term, denoted $\zeta_{ce-\epsilon}$, to stabilize the training process. It is defined as follows:

$$\zeta_{ce-\epsilon} = (1-\alpha)\zeta_{ce} + \alpha \cdot \epsilon. \tag{6}$$

ζ_{ce} is the standard cross-entropy loss, which can be expressed as follows:

$$\zeta_{ce} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} -q_i^k log(p_i^k). \tag{7}$$

It explicitly describes the distribution difference between the predicted result and the ground truth, but the extreme pixel imbalance causes its gradient to degenerate rapidly. Therefore we design a regularization term ϵ to avoid this problem. Its mathematical definition is as follows:

$$\epsilon = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N} -log(p_i^k). \tag{8}$$

Unlike ζ_{ce}, ϵ describes the difference between the predicted result and the standard uniform distribution. Specifically, when predictions deviate from the uniform distribution, the overall loss function is penalized by a factor α (set to 0.1). This method can suppress the inductive preferences of the network for background category, thereby improving network convergence. Besides, the loss function tends to be smooth, which helps avoid overfitting.

3.4. Balance-sampling strategy

Despite the effectiveness of the proposed method, the training process of the two sub-networks still needs to overcome two challenges: on the one hand, a large proportion of normal slices negatively affects the inductive preferences and convergence of both the classifier and the segmentation network. On the other hand, as a two-stage method, the proposed HCF pipeline also suffers from the risk of error propagation, which means that the classification accuracy of the first stage is correlated with the overall performance. In previous experiments, as shown in Table 3, we found that adjusting the data ratio of normal and abnormal slices used for training directly affects the performance of the sub-network since we achieved the best performance with the partial rather than full dataset. Therefore, we design the BSS to control the sample distribution during training to improve overall performance. Given the high content similarity between adjacent normal slices, we achieve dynamic adjustment of the data ratio by randomly undersampling the normal slice set. For the sake of description, we use $\Theta = \{\theta_0, \theta_1, \theta_2, \theta_4, \theta_n\}$ to denote the set of data ratio of normal and abnormal slices, defined as follows: $\theta_1, \theta_2, \theta_4$ respectively indicate that the data ratio of abnormal slice and normal slice is 1, $\frac{1}{2}$ and $\frac{1}{4}$. In particular, θ_0 indicates that the training data contains only abnormal slices as a special case for segmentation network training, and θ_n represents the original ratio. In the experimental part, we analyze the impact of the BSS on the overall performance and obtain some anti-empirical observations and conclusions.

Table 1
The statistics and division of the experimental dataset.

Classes	Training set	Test set	In total
Lung cancer			
# scans (slices)	40 (3816)	10 (959)	50 (4775)
Normal slices	3134	796	3930
Abnormal slices	682	163	845
Nasopharynx cancer			
# scans (slices)	40 (4933)	10 (1218)	50 (6151)
Normal slices	4368	1091	5459
Abnormal slices	565	127	692

4. Experiments

4.1. Dataset and metrics

The experimental data we used comes from the open challenge StrcutSeg2019,[1] which provides two annotated GTV CT datasets of lung and nasopharynx cancer from Zhejiang Cancer Hospital. Each dataset contains 50 CT scans, each of which was annotated by one experienced oncologist and validated by another. Specifically, each CT scan of nasopharynx cancer consists of 100 to 150 slices with a slice thickness of 3 mm, and that of lung cancer consists of 80 to 130 slices with a slice thickness of 5 mm. All CT slices have a resolution of 512 × 512 pixels with a pixel spacing of around 1.0×1.0 mm^2. The specific data distribution is shown in Table 1, where abnormal slices refer to slices containing visible GTV regions, otherwise normal slices. We randomly split each dataset into a training set of 40 scans and an independent test set of 10 scans. In particular, the training set was further randomly divided into five folds at the patient-level for cross-validation. To more objectively evaluate the accuracy and robustness of different models, we conduct the final evaluation on the fixed independent test set. For quantitative analysis, we use Recall to evaluate classification accuracy, while the Dice Similarity Coefficient (DSC), Jaccard Index (JI), and 95% Hausdorff Distance (HD95) are used to evaluate segmentation performance. A better segmentation will have smaller HD95 and larger values for DSC and JI.

4.2. Implementation details

We use Pytorch to implement the baselines and our proposed method. During the training phase, the AdamW optimizer (Loshchilov and Hutter, 2018) with an initial learning rate of 10^{-3} and a weight decay of 10^{-4} is adopted to minimize the loss functions, i.e., the cross-entropy loss of the slice classifier and our proposed CR loss of all segmentation networks. We use the cosine annealing strategy (Loshchilov and Hutter, 2016) to control the change in the learning rate with an initial restart cycle of 20 epochs. All the models are trained from scratch using two NVIDIA A100 GPUs. Specifically, the batch size in the classifier and segmentation network is set to 128 and 64, respectively, and the input size is fixed to 512 × 512 pixels by default. Each model is evaluated on the validation set at the end of each training epoch. We use an early-stopping strategy with a tolerance of 30 epochs to search for the best model within 120 epochs to alleviate the overfitting problem. Besides, we preprocess the data as follows: first, we extract the foreground binary mask by operations such as erosion and threshold segmentation, then obtain the bounding box of the foreground by preserving the largest connectivity region, then crop the original image based on the bounding box, and finally resize the cropped image to 512 × 512. During training, we perform online data augmentation, including random erasure, zooming, distortion, rotation, vertical flip, and noise.

[1] https://structseg2019.grand-challenge.org/.

Table 2
Comparison with state-of-the-art approaches on the independent test set. We show the mean±std (standard deviation) scores averaged over models trained with different folds. ↓ means the lower, the better; while ↑ means the higher, the better.

Methods	Structures	Params (M)	Lung cancer			Nasopharynx cancer		
			DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑	DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑
U-Net (Ronneberger et al., 2015)	End-to-End	14.32	52.6 ± 2.5	34.2 ± 3.4	39.1 ± 9.2	63.1 ± 2.3	4.9 ± 3.7	45.3 ± 2.2
U-Net++ (Zhou et al., 2019b)	End-to-End	15.97	52.7 ± 3.9	28.5 ± 2.7	40.2 ± 2.2	63.4 ± 2.3	5.6 ± 3.4	45.8 ± 2.5
Deeplabv3+ (Chen et al., 2018b)	End-to-End	12.32	51.9 ± 2.7	15.6 ± 1.5	38.8 ± 2.1	61.3 ± 2.9	6.7 ± 2.8	44.2 ± 1.8
AttU-Net (Oktay et al., 2018)	End-to-End	13.81	54.2 ± 2.6	26.3 ± 2.5	40.1 ± 2.4	63.6 ± 2.4	5.8 ± 2.4	46.7 ± 1.9
ResU-Net (Xiao et al., 2018)	End-to-End	14.58	54.6 ± 2.3	18.4 ± 1.9	41.4 ± 3.7	63.8 ± 2.4	5.7 ± 3.1	48.6 ± 2.1
SFNet (Li et al., 2020a)	End-to-End	12.71	54.3 ± 3.1	26.8 ± 2.5	41.2 ± 2.8	63.9 ± 2.6	5.9 ± 2.5	48.5 ± 2.3
3D V-Net (Milletari et al., 2016)	End-to-End	12.95	45.6 ± 2.6	17.2 ± 3.8	34.5 ± 2.1	63.9 ± 2.0	5.2 ± 2.2	47.3 ± 2.9
TransU-Net (Chen et al., 2021)	End-to-End	105.91	48.5 ± 3.2	32.1 ± 2.3	34.9 ± 2.7	56.9 ± 2.8	6.2 ± 2.7	41.6 ± 1.8
UTNet (Gao et al., 2021)	End-to-End	10.02	46.8 ± 2.3	30.1 ± 2.7	34.6 ± 2.7	62.4 ± 2.3	6.1 ± 2.2	45.2 ± 2.4
3D nnU-Net (Isensee et al., 2021)	Cascade	22.24	55.1 ± 1.9	**13.4 ± 1.2**	41.2 ± 2.1	64.8 ± 2.3	4.8 ± 2.7	48.9 ± 1.9
3D VB-Net (Shi et al., 2022)	Cascade	25.88	54.9 ± 2.2	15.8 ± 1.9	40.8 ± 2.8	64.9 ± 2.4	5.3 ± 2.6	48.2 ± 2.7
SANet	End-to-End	12.82	56.4 ± 2.1	21.3 ± 2.3	41.6 ± 2.4	65.1 ± 1.6	5.9 ± 2.3	50.3 ± 2.9
SANet + MTL	MTL	12.82	43.5 ± 2.9	24.2 ± 3.8	35.2 ± 2.6	63.1 ± 2.7	5.2 ± 3.1	47.6 ± 2.4
Ours	**HCF**	23.99	**58.6 ± 1.6**	18.8 ± 1.8	**45.6 ± 1.9**	**68.5 ± 1.7**	**4.3 ± 2.2**	**53.1 ± 1.5**

4.3. Comparison with the SOTA methods

To evaluate the effectiveness of the proposed method, we compare the performance of our method with ten existing state-of-the-art methods on two challenging datasets. We train all networks from scratch under the same experimental configurations for a fair comparison. The experimental results are shown in Table 2. We observe that the proposed method achieves the best performance on both DSC and JI in the two tasks with significant improvements. Among all end-to-end models, our proposed SANet also achieves the highest DSC and JI, indicating that the network structure optimization is effective and outperforms existing methods. The hybrid structures equipped with transformer blocks, such as TransU-Net (Chen et al., 2021) and UTNet (Gao et al., 2021), fail to get the expected results due to the small size of the available data. Given the more severe pixel imbalance, 3D V-Net (Milletari et al., 2016) performs worse compared to most 2D networks, especially on the lung cancer dataset. In contrast, 3D VB-Net (Shi et al., 2022) and 3D nnU-Net (Isensee et al., 2021) with a cascaded structure achieve significant performance gains on both tasks but also bring much extra computation. Furthermore, the combination of our SANet and multi-task learning structure yields competitive results on the nasopharynx cancer dataset, while the performance on the lung cancer dataset plummets. The reason is that identifying lung cancer GTV is more complicated, further widening the feature requirements gap between different subtasks resulting in the failure of joint training to converge to the optimal solution. Similarly, we also analyze the difference in results between the proposed HCF and MTL structure and find that HCF outperforms MTL in both performance and robustness. Overall, the experimental results in Table 2 demonstrate the superiority of our method in GTV segmentation, and subsequent sections will further clarify the effectiveness of each key module.

4.4. Impact of balance-sampling strategy

Table 3 quantitatively analyzes the impact of BSS with different sampling ratios on the overall performance. We can see that the accuracy of the slice classifier is positively correlated with the overall segmentation performance, which means that the higher the Recall, the better the overall segmentation performance. It is worth noting that the sampling ratio has different effects on the recall of the slice classifier in different tasks. Specifically, when the ratios of normal and abnormal slices are θ_1 and θ_2, the classifier achieves the highest recall in both tasks. Likewise, in the case of a fixed slice classifier, the segmentation performance also varies with the sampling ratio. Another counter-intuitive finding is that our method achieves the best results on both tasks under the ratio combination of $\theta_1 + \theta_1$ and $\theta_4 + \theta_4$ instead of $\theta_n + \theta_n$. The reason is that the slice classifier trained with complete data has an inductive preference for the normal category, which leads to many abnormal slices being incorrectly filtered out, thus resulting in a hard loss of performance. At the same time, a large proportion of normal slices can also interfere with the convergence of the segmentation network. Experimental results show that BSS can alleviate these problems to a certain extent and further improve the overall segmentation performance.

Table 3
Comparison of mean DSC under different sampling ratios.

Lung cancer						
Slice classifier		SANet				
Ratio	Recall (%) ↑	θ_0	θ_1	θ_2	θ_4	θ_n
θ_1	**87.7**	58.1	**58.6**	56.4	57.9	56.7
θ_2	86.1	51.8	50.1	49.2	48.7	49.4
θ_4	80.0	44.2	43.8	40.7	43.5	41.4
θ_n	79.4	43.8	44.1	42.6	41.2	42.0
Nasopharynx cancer						
Slice classifier		SANet				
Ratio	Recall (%) ↑	θ_0	θ_1	θ_2	θ_4	θ_n
θ_1	91.2	55.7	61.9	64.2	65.7	64.8
θ_2	**94.7**	59.8	63.2	66.2	68.3	66.2
θ_4	94.2	61.7	63.9	66.3	**68.5**	66.1
θ_n	87.9	63.3	64.4	64.9	66.4	64.9

Table 4
Ablation study on the proposed HCF.

Methods (+HCF)	Lung cancer		Nasopharynx cancer	
	DSC (%) ↑	HD95 (mm) ↓	DSC (%) ↑	HD95 (mm) ↓
U-Net	55.8 ± 2.7	27.1 ± 2.3	65.6 ± 2.4	4.7 ± 3.1
U-Net++	54.6 ± 2.3	24.2 ± 3.1	64.8 ± 2.6	5.4 ± 3.2
Deeplabv3+	52.3 ± 2.3	**15.4** ± 1.7	62.8 ± 2.4	5.3 ± 2.1
AttU-Net	55.9 ± 2.8	18.2 ± 2.4	64.2 ± 2.2	5.3 ± 2.3
ResU-Net	54.9 ± 2.0	17.3 ± 1.8	64.6 ± 2.3	5.5 ± 2.8
SFNet	56.2 ± 2.7	24.4 ± 2.2	65.6 ± 2.4	5.5 ± 2.3
TransU-Net	51.9 ± 2.0	23.1 ± 2.3	63.2 ± 2.6	4.9 ± 2.5
UTNet	49.2 ± 1.9	23.3 ± 2.9	63.1 ± 2.4	5.4 ± 2.2
SANet	**58.6 ± 1.6**	18.8 ± 1.8	**68.5 ± 1.7**	**4.3 ± 2.2**

4.5. Ablation study

4.5.1. Impact of heterogeneous cascade framework

Using the same slice classifier, we combine different end-to-end models with the proposed HCF to verify its effectiveness. The results are shown in Table 4. We observe varying degrees of improvement in DSC and HD95 for each model after adding HCF. In particular, the model with poorer performance, such as TransU-Net (Chen et al., 2021),

Table 5
Ablation study on the proposed components (w/o denotes without) on the independent test.

Lung cancer			
Methods	DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑
w/o SAM and HCF	54.3 ± 2.5	24.5 ± 2.8	40.8 ± 2.3
w/o SAM	56.3 ± 2.9	22.7 ± 1.9	41.4 ± 2.4
w/o HCF	56.4 ± 2.1	21.3 ± 2.3	41.6 ± 2.4
w/o PPM	57.9 ± 2.0	19.4 ± 2.4	44.8 ± 1.8
Ours	**58.6 ± 1.6**	**18.8 ± 1.8**	**45.6 ± 1.9**
Nasopharynx cancer			
Methods	DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑
w/o SAM and HCF	64.7 ± 2.5	6.7 ± 3.6	48.7 ± 2.1
w/o SAM	64.9 ± 1.8	6.5 ± 3.4	50.2 ± 2.2
w/o HCF	65.1 ± 1.6	5.9 ± 2.3	50.3 ± 2.9
w/o PPM	67.5 ± 1.8	4.4 ± 2.0	52.3 ± 1.7
Ours	**68.5 ± 1.7**	**4.3 ± 2.2**	**53.1 ± 1.5**

Table 6
Ablation study on different loss functions.

Lung cancer			
Loss functions	DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑
Dice loss	50.7 ± 2.7	30.3 ± 2.6	34.6 ± 2.9
CE loss	53.2 ± 2.6	28.3 ± 2.3	41.4 ± 2.4
Focal loss (Hossain et al., 2021)	54.7 ± 2.1	26.2 ± 1.9	42.9 ± 2.0
Dice + CE loss	56.8 ± 2.2	22.5 ± 2.2	43.6 ± 2.1
Ours (CR loss)	**58.6 ± 1.6**	**18.8 ± 1.8**	**45.6 ± 1.9**
Nasopharynx cancer			
Loss functions	DSC (%) ↑	HD95 (mm) ↓	JI (%) ↑
Dice loss	64.2 ± 2.3	5.8 ± 2.6	47.7 ± 2.1
CE loss	67.8 ± 2.1	4.6 ± 2.3	51.9 ± 1.8
Focal loss (Hossain et al., 2021)	67.4 ± 1.9	4.8 ± 2.6	51.6 ± 2.1
Dice + CE loss	67.6 ± 2.1	4.6 ± 2.4	51.8 ± 1.9
Ours (CR loss)	**68.5 ± 1.7**	**4.3 ± 2.2**	**53.1 ± 1.5**

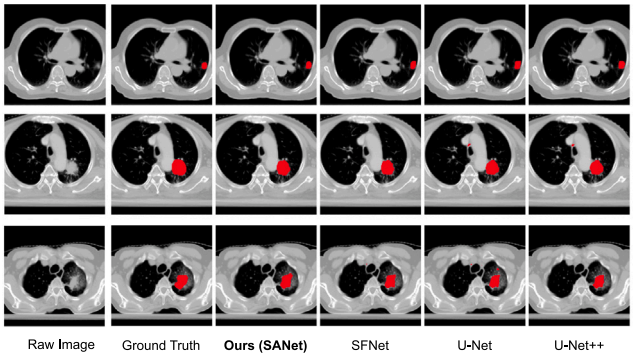
achieves a more significant performance gain. These results reveal our design idea that the overall segmentation performance can be effectively improved when the performance gain achieved by filtering normal slices is greater than the hard loss caused by the classifier.

4.5.2. Impact of different modules

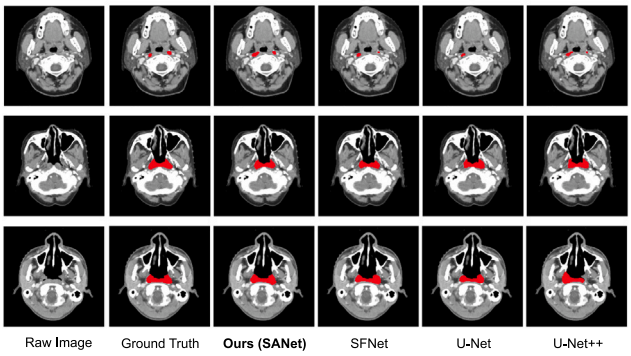
The above results demonstrate the superiority of our method, but it is unclear which module plays a more critical role in performance improvement. Therefore, we perform ablation experiments on existing PPM, our proposed SAM and HCF to analyze the impact of each module on the overall performance. Table 5 compares the performance of the methods under different configurations. We can observe different degrees of performance degradation of the proposed method after removing PPM, SAM and HCF. Compared with PPM and HCF, the proposed SAM has a more significant impact on the overall performance, further proving its effectiveness.

4.5.3. Impact of combined regularization loss

To further improve the segmentation performance, we propose a novel CR loss to alleviate the pixel imbalance problem in training. Table 6 presents the performance comparison of different loss functions, and we can observe that the proposed method achieves the best performance on both tasks by applying CR loss. Compared with the Dice loss, CE loss and Focal loss (Hossain et al., 2021), the CR loss brings significant improvements in all three evaluation metrics. In addition, the CR loss can suppress the network's preference for background categories by introducing a regularization term, outperforming the naive combination of Dice and CE loss. Overall, the experimental results confirm the effectiveness of the CR loss.



(a) Lung cancer.



(b) Nasopharynx cancer.

Fig. 5. Results visualization of different models.

4.6. Visualization analysis

To visually assess the segmentation quality of different models, Fig. 5 shows the segmentation results for three examples from the two tasks. Unlike the quantitative results reported above, we can obtain some interesting findings from the qualitative results. First, we observe that compared with the three existing models, the proposed SANet is more sensitive to boundary information and more accurate in segmenting smaller objects. The reason is that SAM can achieve the high-quality reconstruction of high-resolution features, which preserves detailed and rich spatial information, effectively compensating for the inherent defects of downsampling. Second, for the more challenging lung cancer dataset, existing methods tend to predict false positives due to the low contrast of GTV with surrounding tissues. Our method can alleviate this problem to a certain extent, thereby reducing false positives. These visualization results demonstrate the superiority of the proposed method.

4.7. Discussion

Despite the effectiveness of our approach, several issues remain to be further explored in the future. First, the inherent error propagation problem of the proposed two-stage framework still needs to be further resolved. A potential improvement solution is to use the spatial continuity between slices to improve the identification accuracy of abnormal slices. Second, we use two of the most challenging small-scale public datasets during the experiments, which may result in a large standard deviation, so the proposed method needs to be extended to larger-scale datasets to complete the evaluation of the method. Third, considering the computational complexity, we adopt a lightweight slice classifier and backbone network to construct our method. The following research direction is to use more complex and deeper structures such

as ResNet50 (He et al., 2016) and ViT (Dosovitskiy et al., 2020) to improve the segmentation accuracy further. Finally, extensive experimental results confirm the superiority of the proposed method in GTV segmentation, suggesting that it may also have the potential to segment other similar objects with small sizes and low contrast, such as tumors, which requires more experiments to verify.

5. Conclusion

In this study, we proposed a novel two-stage cascade framework for the automatic segmentation of GTV from a decoupling perspective driven by data characteristics. Our core idea is to achieve an overall performance gain by filtering the normal slices to reduce false positives. To further improve segmentation accuracy, we designed the SANet, equipped with learnable SAMs to provide the high-quality reconstruction of high-resolution features. Furthermore, we proposed a Combined Regularization (CR) loss and Balance-Sampling Strategy (BSS) to alleviate the pixel imbalance problem and improve network convergence. We conducted extensive ablation experiments to evaluate the impact of each key component. The experimental results on two open-source datasets demonstrate that our method outperforms existing methods in automatic GTV segmentation, especially in reducing false positives and accurately segmenting small objects.

CRediT authorship contribution statement

Jun Shi: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Zhaohui Wang:** Supervision, Writing – review & editing. **Shulan Ruan:** Writing – review & editing. **Minfan Zhao:** Writing – review & editing. **Ziqi Zhu:** Writing – review & editing. **Hongyu Kan:** Writing – review & editing. **Hong An:** Writing – review & editing. **Xudong Xue:** Discussion, Writing – review & editing. **Bing Yan:** Writing – review & editing.

Declaration of competing interest

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

Data availability

I have shared the link to my data/code at the attach file step.

Code and Data for GTV AutoSeg (Original data) (RunMyCode)
.

Acknowledgments

The work is supported by the Fundamental Research Funds for the Central Universities, China (GrantsNo. YD2150002001), National Key Research and Development Program of China, China (GrantsNo. 2016YFB1000403) and Laoshan Laboratory, China (GrantsNo. LSKJ202300305).

References

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2018a. DRINet for medical image segmentation. IEEE Trans. Med. Imaging 37 (11), 2453–2462, URL: https://doi.org/10.1109/tmi.2018.2835303.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587. URL: https://doi.org/10.48550/arXiv.1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. Eur. Conf. Comput. Vis.. pp. 801–818, URL: https://doi.org/10.1007/978-3-030-01234-2_49.

Chen, J., et al., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. URL: https://doi.org/10.48550/arXiv.2102.04306.

Chi, J., Han, X., Wu, C., Wang, H., Ji, P., 2021. X-Net: Multi-branch UNet-like network for liver and tumor segmentation from 3D abdominal CT scans. Neurocomputing 459, 81–96, URL: https://doi.org/10.1016/j.neucom.2021.06.021.

Dosovitskiy, A., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent.. URL: https://doi.org/10.48550/arXiv.2010.11929.

Fu, X., Bi, L., Kumar, A., Fulham, M., Kim, J., 2021. Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. IEEE J. Biomed. Health Inform. 25 (9), 3507–3516, URL: https://doi.org/10.1109/jbhi.2021.3059453.

Gao, Y., Zhou, M., Metaxas, D., 2021. UTNet: a hybrid transformer architecture for medical image segmentation. In: Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. pp. 61–71, URL: https://doi.org/10.1007/978-3-030-87199-4_6.

Hatamizadeh, A., et al., 2022. UNETR: Transformers for 3D medical image segmentation. In: Proc. IEEE Wint. Conf. Applicat. Comput. Vis.. pp. 574–584, URL: https://doi.org/10.1109/wacv51458.2022.00181.

He, T., Hu, J., Song, Y., Guo, J., Yi, Z., 2020. Multi-task learning for the segmentation of organs at risk with label dependence. Med. Image Anal. 61, 101666, URL: https://doi.org/10.1016/j.media.2020.101666.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. pp. 770–778, URL: https://doi.org/10.1109/cvpr.2016.90.

Hossain, M., Betts, J., Paplinski, A., 2021. Dual Focal Loss to address class imbalance in semantic segmentation. Neurocomputing 462, 69–87, URL: https://doi.org/10.1016/j.neucom.2021.07.055.

Ibtehaz, N., Rahman, M., 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. 121, 74–87, URL: https://doi.org/10.1016/j.neunet.2019.08.025.

Isensee, F., et al., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211, URL: https://doi.org/10.1038/s41592-020-01008-z.

Jaffray, D., 2012. Image-guided radiotherapy: from current concept to future perspectives. Nat. Rev. Clin. Oncol. 9 (12), 688–699, URL: https://doi.org/10.1038/nrclinonc.2012.194.

Jiang, J., et al., 2018. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. IEEE Trans. Med. Imaging 38 (1), 134–144, URL: https://doi.org/10.1109/tmi.2018.2857800.

Jiang, Y., et al., 2021. ALA-Net: Adaptive lesion-aware attention network for 3D colorectal tumor segmentation. IEEE Trans. Med. Imaging 40 (12), 3627–3640, URL: https://doi.org/10.1109/tmi.2021.3093982.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I., 2019. Boundary loss for highly unbalanced segmentation. In: Int. Conf. Med. Imag. Deep Learning. pp. 285–296, URL: https://doi.org/10.1016/j.media.2020.101851.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A., 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging 37 (12), 2663–2674, URL: https://doi.org/10.1109/tmi.2018.2845918.

Li, Z., Kamnitsas, K., Glocker, B., 2020b. Analyzing overfitting under class imbalance in neural networks for image segmentation. IEEE Trans. Med. Imaging 40 (3), 1065–1077, URL: https://doi.org/10.1109/tmi.2020.3046692.

Li, X., et al., 2020a. Semantic flow for fast and accurate scene parsing. In: Proc. Eur. Conf. Comput. Vis.. pp. 775–793, URL: https://doi.org/10.1007/978-3-030-58452-8_45.

Liu, Z., et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE Int. Conf. Comput. Vis.. pp. 10012–10022, URL: https://doi.org/10.1109/iccv48922.2021.00986.

Loshchilov, I., Hutter, F., 2016. SGDR: Stochastic gradient descent with warm restarts. Learning 10, 3. URL: https://doi.org/10.48550/arXiv.1608.03983.

Loshchilov, I., Hutter, F., 2018. Decoupled weight decay regularization. In: Int. Conf. Learn. Represent.. URL: https://doi.org/10.48550/arXiv.1711.05101.

Mei, H., et al., 2021. Automatic segmentation of gross target volume of nasopharynx cancer using ensemble of multiscale deep neural networks with spatial attention. Neurocomputing 438, 211–222, URL: https://doi.org/10.1016/j.neucom.2020.06.146.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Int. Conf. 3D Vis.. IEEE, pp. 565–571, URL: https://doi.org/10.1109/3dv.2016.79.

Nasalwai, N., Punn, N., Sonbhadra, S., Agarwal, S., 2021. Addressing the class imbalance problem in medical image segmentation via accelerated tversky loss function. In: Advances in Knowledge Discovery and Data Mining. pp. 390–402, URL: https://doi.org/10.1007/978-3-030-75768-7_31.

Oktay, O., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. URL: https://arxiv.org/abs/1804.03999.

Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M., 2022. A robust volumetric transformer for accurate 3D tumor segmentation. In: Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. pp. 162–172, URL: https://doi.org/10.1007/978-3-031-16443-9_16.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. pp. 234–241, URL: https://doi.org/10.1007/978-3-319-24574-4_28.

Seo, H., Huang, C., Bassenne, M., Xiao, R., Xing, L., 2019. Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. IEEE Trans. Med. Imaging 39 (5), 1316–1325, URL: https://doi.org/10.1109/tmi.2019.2948320.

Shi, F., et al., 2022. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. Nature Commun. 13 (1), 6566, URL: https://doi.org/10.1038/s41467-022-34257-x.

Sinha, A., Dolz, J., 2020. Multi-scale self-guided attention for medical image segmentation. IEEE J. Biomed. Health Inform. 25 (1), 121–130, URL: https://doi.org/10.1109/jbhi.2020.2986926.

Taghanaki, S., et al., 2019. Combo loss: Handling input and output imbalance in multi-organ segmentation. Comput. Med. Imag. Graph. 75, 24–33, URL: https://doi.org/10.1016/j.compmedimag.2019.04.005.

Vaswani, A., et al., 2017. Attention is all you need. In: Adva. Neural Inf. Process. Syst.. pp. 5998–6008, URL: https://doi.org/10.48550/arXiv.1706.03762.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J., 2021. Transbts: Multimodal brain tumor segmentation using transformer. In: Int. Conf. Med. Image Comput. Comput.-Assisted Intervention. pp. 109–119, URL: https://doi.org/10.1007/978-3-030-87193-2_11.

Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2017. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: Int. MICCAI Brainlesion Workshop. pp. 178–190, URL: https://doi.org/10.1007/978-3-319-75238-9_16.

Wang, Z., Zou, N., Shen, D., Ji, S., 2020. Non-local u-nets for biomedical image segmentation. In: Proc. AAAI Conference, Vol. 34. pp. 6315–6322, URL: https://doi.org/10.1609/aaai.v34i04.6100.

Weiss, E., Hess, C., 2003. The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy. Strahlenther. Onkol. 179 (1), 21–30, URL: https://doi.org/10.1007/s00066-003-0976-5.

Xiao, X., Lian, S., Luo, Z., Li, S., 2018. Weighted res-unet for high-quality retina vessel segmentation. In: Int. Conf. Inf. Tech. Med. Edu.. IEEE, pp. 327–331, URL: https://doi.org/10.1109/itme.2018.00080.

Yu, Q., et al., 2019. Crossbar-net: a novel convolutional neural network for kidney tumor segmentation in CT images. IEEE Trans. Imag. Process. 28 (8), 4060–4074, URL: https://doi.org/10.1109/tip.2019.2905537.

Zhang, J., Saha, A., Zhu, Z., Mazurowski, M., 2018. Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics. IEEE Trans. Med. Imaging 38 (2), 435–447, URL: https://doi.org/10.1109/tmi.2018.2865671.

Zhang, Y., et al., 2021. 3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. IEEE Trans. Med. Imaging 40 (6), 1618–1631, URL: https://doi.org/10.1109/tmi.2021.3062902.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. pp. 2881–2890, URL: https://doi.org/10.1109/cvpr.2017.660.

Zhou, S., Nie, D., Adeli, E., Yin, J., Lian, J., Shen, D., 2019a. High-resolution encoder–decoder networks for low-contrast medical image segmentation. IEEE Trans. Imag. Process. 29, 461–475, URL: https://doi.org/10.1109/tip.2019.2919937.

Zhou, Z., Siddiquee, M., Tajbakhsh, N., Liang, J., 2019b. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39 (6), 1856–1867, URL: https://doi.org/10.1109/tmi.2019.2959609.

Zhou, L., Wang, S., Sun, K., Zhou, T., Yan, F., Shen, D., 2022. Three-dimensional affinity learning based multi-branch ensemble network for breast tumor segmentation in MRI. Pattern Recognit. 129, 108723, URL: https://doi.org/10.1016/j.patcog.2022.108723.

Zhou, Y., et al., 2021. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. Med. Image Anal. 70, 101918, URL: https://doi.org/10.1016/j.media.2020.101918.

Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y., 2017. Deep feature flow for video recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. pp. 2349–2358, URL: https://doi.org/10.1109/cvpr.2017.441.

Zhu, W., et al., 2019. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Med. Phys. 46 (2), 576–589, URL: https://doi.org/10.1002/mp.13300.