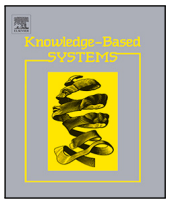
Knowledge-Based Systems 311 (2025) 113005

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Aggregated Mutual Learning between CNN and Transformer for semi-supervised medical image segmentation

Zhenghua Xu [a,b], Hening Wang [a], Runhe Yang [a,b], Yuchen Yang [d], Weipeng Liu [a,c,*], Thomas Lukasiewicz [e,f]

[a] State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin, China
[b] School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China
[c] School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
[d] Department of Applied Mathematics and Statistic, The Johns Hopkins University, Baltimore, United States
[e] Institute of Logic and Computation, Vienna University of Technology, Vienna, Austria
[f] Department of Computer Science, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Keywords:
Medical image segmentation
Semi-supervision learning
Mutual learning
Vision Transformer

ABSTRACT

Recent Advances show that both Convolutional layers and Transformer blocks have their own advantages in the feature learning tasks of medical image analysis. However, the existing models combining both CNN and Transformers cannot effectively integrate the features extracted by both networks. In this work, we propose a new semi-supervised medical image segmentation method which can effectively aggregate mutual learning between CNN and Transformer, denoted AML-CT, which consists of an auxiliary module and a main network. Specifically, the auxiliary module consists of two segmentation subnetworks based on CNN and Transformer, aiming at extracting features from different perspectives, where, to enhance integration of image features from distinct segmentation networks, a Cross-Branch Feature Fusion module is proposed to effectively fuses local and global information via internal cross-fusion of feature maps between networks. Then, to aggregate the extracted image features from the auxiliary module, a three-branch network (TB-net) structure is further proposed to learn the extracted joint features and facilitate aggregation of multi-source information. Experimental results on two public datasets demonstrate that: (i) AML-CT successfully accomplishes medical image segmentation tasks with limited labeled data, outperforming recent mainstream semi-supervised segmentation methods; (ii) Ablation studies confirm the effectiveness of each module in the AML-CT model for performance improvement.

1. Introduction

Medical image semantic segmentation plays a pivotal role in computer vision, serving various applications, especially in medical image analysis. Proficient automatic segmentation models enable rapid identification of anomalies in medical images, aiding accurate diagnoses [1–3]. Recent advancements in deep learning, primarily using Convolutional Neural Network (CNN) and Transformer [4], have yielded impressive results in fully supervised settings, benefiting from rich semantic labels. However, these models share a common limitation—dependency on a substantial amount of pixel/voxel-level semantic annotations [5–7]. Annotating medical images is a specialized task typically carried out by experienced radiologists. Limited availability of specialized professionals, time constraints, and annotation efficiency pose challenges in creating large-scale medical image dataset with accurate labels. This limitation impedes the practical application of supervised learning in clinical settings. The advent of semi-supervised learning (SSL) introduces a new paradigm for training segmentation models. SSL leverages a small labeled dataset alongside a large pool of unlabeled data to train deep learning models [8–11], effectively mitigating the demand for extensive labeled data while maintaining satisfactory performance. The nature of semi-supervised learning expedites clinical data annotation and model development, facilitating informed medical diagnosis.

In the realm of semi-supervised learning, the Cross Pseudo Supervision [12] (CPS) model has garnered attention due to its remarkable performance and lightweight attributes. CPS involves the training of two identical CNNs with different initializations. Each branch network's output is employed to supervise the other network, serving as an additional signal. Expanding upon the CPS model, Luo et al. introduced the Cross Teaching between CNN and Transformer [13]

* Correspondence to: Hebei University of Technology, Tianjin, 300130, China.
E-mail address: liuweipeng@hebut.edu.cn (W. Liu).

https://doi.org/10.1016/j.knosys.2025.113005
Received 17 April 2024; Received in revised form 10 October 2024; Accepted 11 January 2025
Available online 18 January 2025

(CTCT) model, wherein one of the convolutional segmentation networks in CPS is replaced by a Transformer-based segmentation network (Swin-Unet [14]). Wang et al. presented the S4CVnet [15] model, which harnesses the feature learning capabilities of CNNs and Vision Transformers [16] (ViT) to enhance performance through dual-view collaborative training and consistency-aware supervision. Given CNN's proficiency in local feature extraction [17–21] and the specialized global image feature capturing capabilities of Transformer-based segmentation networks [22–24], the outputs of these networks can serve as pseudo-labels for mutual learning. This approach facilitates a deeper understanding of medical images. Experimental results demonstrate that the introduction of Transformers in both models surpasses CPS models that solely rely on CNN. However, we posit that relying solely on their outputs as mutual supervisory signals for mutual learning between CNN and Transformers in these models is insufficient for effectively fusing the local and global features extracted from the images. This limitation impedes further enhancements in segmentation performance.

In this paper, we propose a method to enhance the segmentation performance of medical images by leveraging a large amount of unlabeled data. We call this method Aggregated Mutual Learning between CNN and Transformer (AML-CT), a neural network model meticulously designed for seamless integration of the local and global information extracted by CNN and Transformer. The model comprises an auxiliary module with CNN and Transformer segmentation pathways to extract shared image features, alongside a main module to aggregate these features. Specifically, a Cross-Branch Cross-Fusion (CBC-Fusion) mechanism is introduced within the auxiliary module. Input images undergo concurrent processing by the CNN and Transformer encoders to generate two outputs. These serve as the basis for two independent decoder paths, each employing skip connections carrying encoder feature maps from both networks. This robustly fuses the local and global information extracted by the two network types. Currently, commonly used Transformer blocks include Vision Transformer and Swin Transformer [25] modules. However, Swin Transformer introduce shifted window modules to process block-wise information, which increases computational overhead and irreversibly alters the image blocks, hindering integration with CNN. Therefore, our model uses a segmentation network based on Vision Transformer to retain the original image blocks. By utilizing a pyramid structure similar to CNN, this approach facilitates hierarchical interaction.

Our proposed CBC Fusion module conceptually resembles the previous TransFuse [26] model, as both employ interconnected CNN and Transformer branches to merge locally extracted and globally attended features, thereby enhancing image representation. However, our approach achieves a unique cross-branch fusion structure by linking the feature maps of one branch's encoder to the decoder of the other branch. This design better preserves the integrity of features encoded by each path, allowing for more comprehensive utilization of the extracted information. In contrast, the TransFuse model connects two different structured segmentation models at each encoder layer, achieving fusion by combining feature maps of varying sizes. We hypothesize that the proposed cross-branch encoder–decoder connection and fusion strategy offer greater advantages for feature representation.

Nonetheless, selecting the optimal network as the primary segmentation model in this dual-branch design is not straightforward. To address this challenge, we leverage the superior performance of CNN to establish a three-branch network structure. This structure comprises two auxiliary branches that integrate CNN and Transformer features, as well as a U-Net [27] model serving as the core segmentation path. The U-Net uses the output from cross-branch feature fusion as supervisory signals to learn joint representations, thereby systematically integrating the local and global information extracted by the auxiliary branches. Subsequent experimental results will demonstrate the enhanced segmentation performance of this three-branch structure. Additionally, we have designed related loss functions to maintain output consistency among the three branches, effectively achieving consistency regularization during the training process.

The contributions of this paper can be summarized as follows:

- Existing models combining CNN and Transformers face challenges in effectively integrating the features extracted by both networks. To address this limitation, we propose AML-CT, a high-performance semi-supervised segmentation model enabling deep integration between CNN and Transformer networks.
- To enhance integration of image features from distinct segmentation networks, we introduce a Cross-Branch Feature Fusion module, which effectively fuses local and global information via internal cross-fusion of feature maps between networks. In parallel, a three-branch structure is devised that utilizes a U-Net with strong representational capabilities to learn the extracted joint features, thereby facilitating aggregation of multisource information.
- We conducted comprehensive experiments on two publicly available medical image segmentation datasets. The results reveal that: (i) With partial labeling, AML-CT achieves segmentation performance on par with fully supervised methods and superior to mainstream semi-supervised approaches; (ii) Ablation studies validate the efficacy and rationality of each incorporated module in improving segmentation with lower computational cost.

2. Related work

Medical image segmentation, the process of identifying and delineating targeted objects (e.g. organs or lesions) in clinical images, remains challenging due to the difficulty in obtaining precise medical image annotations. As such, semi-supervised learning methods have become increasingly popular. However, the performance of existing semi-supervised learning models remains unsatisfactory, especially when labeled data is limited. To address this, we propose AML-CT, a novel approach for medical image segmentation. Our method leverages unlabeled data through a consistency regularization module and affinity learning module to improve segmentation performance under limited supervision. In this work, we demonstrate AML-CT's effectiveness for medical image segmentation with scarce annotations.

2.1. Medical image segmentation

Semantic segmentation is a dense prediction visual task that aims to classify each pixel into a category [28–30]. Fully Convolutional Networks (FCNs) [31] introduced a groundbreaking approach, presenting an encoder–decoder architecture with all-convolutional networks for per-pixel semantic segmentation. This concept spurred a wealth of dense prediction endeavors employing similar architectures, including traditional CNN-based methods. [18] proposed DeepLab, which used atrous convolution for improved multi-scale segmentation accuracy. [19] introduced SegNeXt, a flexible architecture that enhanced segmentation performance with a hierarchical design. [20] presented HRNet, maintaining high-resolution representations for better localization in segmentation tasks. [21] proposed PSPNet, which employed a pyramid pooling module to improve segmentation results by capturing global context. Recently, with the substantial success of Transformers [4], various attempts have emerged to harness the potent attention mechanism to capture distant contextual information for semantic segmentation. [22] introduced SegFormer, a lightweight framework that captured both local and global context using a transformer-based architecture. [32] proposed HRFormer, which enhanced segmentation performance by leveraging high-resolution representations and transformer blocks. [24] presented SegViT, integrating vision transformers to focus on self-attention and pixel-wise classification for improved results. SETR, a framework that employed transformers to directly map image features to segmentation maps, was then proposed in [33]. However, the outstanding performance of these methods heavily relies on

comprehensive annotation supervision and often demands significant time investment to acquire annotations.

Currently, CNN-based medical image segmentation methods have been extensively studied for many years, with most being based on U-Net [27] and its variants. These models have shown promising results across various tasks. However, due to the inherent locality of convolutional operations, they lack the ability to model global and long-range semantic interactions. Recently, architectures based on self-attention (such as Vision Transformer [16]) have been introduced into visual recognition tasks to model long-range dependencies. Subsequently, numerous variants of Vision Transformers have achieved remarkable success in natural image recognition tasks. Benefiting from the immense representation power of Transformers, some works attempt to replace or combine CNNs with Transformers to achieve improved medical image segmentation. Swin-Unet was introduced in [14], a hierarchical segmentation model that combined the strengths of the Swin Transformer with U-Net architecture, enabling effective multi-scale feature extraction and enhancing segmentation performance in medical imaging tasks. [23] proposed TransUNet, which integrated transformers into the U-Net framework, allowing the model to leverage both global context and local detail for improved semantic segmentation accuracy. [26] presented TransFuse, a model that fused features from convolutional neural networks and transformers, facilitating enhanced representation learning and achieving superior performance in various segmentation tasks. These efforts highlight that Transformer can further enhance performance beyond CNNs and suggest that the Transformer architecture deserves more attention in the future. Despite the impressive representational capacity of Transformer, they remain data-hungry solutions for recognition tasks, often requiring even more data than CNNs. Training Transformers in a semi-supervised manner is an intriguing yet challenging problem, especially for data-limited tasks like medical image analysis.

2.2. Semi-supervised image segmentation

Creating pixel-wise labels for semantic image segmentation is substantially more labor-intensive compared to image-level labeling for classification, with costs estimated to be up to 25 times higher [5,11]. This prohibitive annotation burden has motivated extensive research into semi-supervised semantic segmentation methods. Most existing techniques are based on extensions of popular Mean Teacher [34] (MT) frameworks, which use an ensemble of weighted teacher models to generate pseudo-labels for unlabeled data. While there is growing interest in leveraging multiple input modalities for semi-supervised learning across domains like medical imaging, videos, and speech, this remains relatively unexplored for semantic segmentation. Our proposed approach applies semi-supervised learning to multi-modal semantic segmentation and extends the mean teacher framework to enhance performance and improve model robustness to missing modalities. By harnessing different input modalities during training, the model can better exploit unlabeled data.

In recent years, semi-supervised learning for medical image segmentation has garnered considerable research attention within the field of image computing. A semi-supervised segmentation approach was proposed in [12] to utilize a consistency-based training framework to effectively leverage both labeled and unlabeled data. [8] introduced a deep convolutional model for semantic segmentation, enhancing feature representation through a dual-channel architecture. [9] employed adversarial training to improve the robustness of segmentation models using limited labeled data. [10] focused on contrastive learning to optimize segmentation performance even with sparse annotations. These techniques combine labeled and unlabeled data during training to develop powerful and robust CNN models for segmentation. Among them, the MT framework has been highly influential. The core concept involves a "Teacher" network generating pseudo-labels to supervise and guide a "Student" network's learning process. Building on the MT model, Chen et al. proposed the CPS [12] model, which trains two networks with identical architectures but distinct initializations. By minimizing discrepancies between the pseudo-labels produced by the two networks, CPS enhances pseudo-label quality and extracts richer information from unlabeled data. Subsequently, Dosovitskiy et al. introduced the Vision Transformer, recognizing the superiority of Transformer models in extracting global image features. Building on the CPS model, Luo et al. proposed the CTCT [13] model, which replaces one network with a Transformer-based segmentation model to integrate semantic features from both CNN and Transformer. Our proposed model is built upon the CTCT framework to achieve more effective integration of the CNN and Transformer features for medical images. This enables more comprehensive analysis of medical imagery by deeply amalgamating the local and global information extracted by the two network types.

2.3. Mutual learning

In recent years, with the rapid advancement of deep learning, the concept of "Mutual Learning" has gained widespread attention as a novel approach for collaborative training within neural networks. This paradigm emphasizes performance improvement through sharing activations, gradients, losses, and other informative signals between distinct sub-networks in a single model. Each sub-network employs different initializations or data augmentation strategies to observe data from various perspectives. By exchanging feedback on their learning processes, the sub-networks can reduce overfitting and explore diverse feature spaces. The key distinction of Mutual Learning is the implementation of collaborative training within one model, establishing internal pathways for information transmission and exchange among sub-networks. This allows for supervision and regularization between networks, enhancing overall performance.

Qiao et al. introduced the Deep Co-Training [35] model, which simultaneously trains two convolutional neural networks. One network uses original images, while the other uses images that are cropped and randomly shuffled. The models exchange high-confidence labeled examples for retraining. Han et al. proposed the Co-teaching [36] model, maintaining two identical neural networks, each utilizing a distinct loss function. They exchange samples with smaller losses for training, which helps reduce the accumulation of errors due to mislabeled data. Building upon the Co-teaching model, Reed et al. introduced the Co-teaching+ [37] model, incorporating consistency regularization and a dynamic adaptive cross-training mechanism. This enhances the model's robustness against high-noise data.

In summary, the 'Mutual Learning' approach offers a novel perspective for semi-supervised learning by enabling information sharing and collaborative training between sub-networks within a single model. Compared to traditional methods like Co-training, Mutual Learning implements the idea of collaborative learning within one model, establishing internal information propagation pathways to enhance overall performance. This approach presents a promising solution to address the inherent challenges of limited labeled data and annotation scarcity in real-world applications. The distinct advantages of Mutual Learning lie in its internal model architecture and mechanism for information exchange across sub-networks. By receiving supervision signals from different viewpoints, the model can mitigate overfitting and explore diverse feature representations.

3. Methodology

For the general semi-supervised segmentation setup, the training set always includes a small portion of labeled data $D_l = \{x_i, y_i\}_{i=1}^{N}$ and a large amount of unlabeled data $D_u = \{x_i\}_{i=N+1}^{N+M}$. In this work, we introduce a model called Aggregated Mutual Learning between CNN and Transformer (AML-CT). This model effectively integrates local

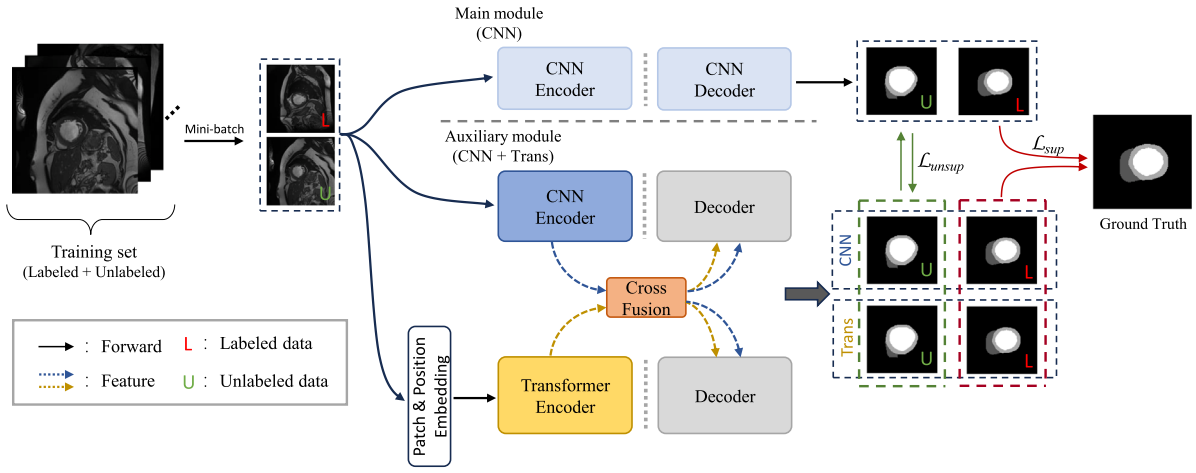


Fig. 1. The overall structure of the **AML-CT**, which consists of the main module and auxiliary modules, forming a three-branch network. The main module is a high-capacity U-Net model for aggregating image features, while the auxiliary module is a cross-branch cross-fusion module designed to extract comprehensive image features, incorporating both a CNN and a Transformer network to extract image features from different perspectives. A cross-branch fusion mechanism is introduced within the auxiliary module to better combine local and global features. Subsequently, the auxiliary module transfers the extracted features to the main module using consistency regularization. In the diagram, red arrows indicate losses supervised by labeled data, while green arrows represent unsupervised losses between unlabeled data.

features extracted from the CNN branch and global attention features from the Transformer branch. The overview of our model is shown in Fig. 1.

As illustrated in Fig. 1, our proposed model consists of two parts: The first component is the Cross-Branch Cross-Fusion (CBC-Fusion) module, which serves as an auxiliary training module. The structure of this auxiliary module is visually depicted in Fig. 2. It consists of two branches, based on CNN and Transformer encoder–decoder structures. The module facilitates deep mutual learning between CNN and Transformer networks by integrating features of various sizes from both pathways. The second component is our devised three-branch network structure. Specifically, capitalizing on the U-net model's robust representation capabilities, we have incorporated a U-net segmentation pathway as the primary segmentation model within the network. This pathway leverages consistency regularization to learn the joint features extracted from the auxiliary training module, thereby streamlining the aggregation of feature information. Within the CBC-Fusion module, the internal components of the two segmentation networks are interconnected. We have substantiated that this structure effectively combines locally extracted features from CNN and globally attended features from Transformer. To facilitate internal connectivity with CNN, we have developed a segmentation network based on the Vision Transformer with a pyramid structure.

3.1. Segmentation network based on Vision Transformer

To extract global attention and enable cross-branch fusion with CNNs, we devised a Vision Transformer (ViT) based segmentation network with a pyramid structure, denoted **ViT-ED** (Fig. 2). Due to the introduction of the shifted window module in Swin-Transformer to handle patch information, there is an increased computational cost and irreversible changes to image patches, hindering integration with CNN. Therefore, our model instead utilizes a Vision Transformer based segmentation network preserving original image blocks.

As illustrated in the orange flowchart of Fig. 2, in the ViT-ED branch's encoder, the input image undergoes image embedding and positional encoding before entering the encoder. Each layer of the encoder includes a multi-head self-attention mechanism, allowing the model to compute self-attention across different heads to capture relationships between various parts of the image. Each layer also contains a feed-forward neural network composed of two linear transformations and a nonlinear activation function, typically GELU. The output of each layer is added back to the input through residual connections, which help in gradient propagation and mitigate the vanishing gradient problem in deep networks. The residual connection output is then layer-normalized to standardize the input of each layer, enhancing training stability and speed. To adapt it for medical image segmentation tasks, we implemented patch merging on the output to reduce image size, thereby expanding the model's receptive field, ultimately forming a pyramid structure. To ensure compatibility with the next ViT+ module layer and enable connectivity with the CNN branch, we deemed it necessary to restore the output of each layer to the image format.

The ViT-ED decoder mirrors U-Net, using upsampling and convolutions to reconstruct full resolution images from the encoded low-level features. By limiting Transformers to the encoder, overall model complexity is reduced substantially. The rationale behind this design is empirically validated in subsequent experiments.

3.2. Cross-Branch feature Cross-Fusion module

The structure of this module is shown in Fig. 2. Specifically, this module consists of two branches, namely the segmentation networks based on CNN and Transformer. The internals of these two networks are interconnected, facilitating the internal feature fusion of the two pathways. We refer to this module as CBC-Fusion. The input image X undergoes encoding through two branches. During the encoding process, the size of the feature maps gradually decreases, and the number of channels increases. Specifically, we describe how the features extracted by the convolutional encoder, X_c, and the Transformer encoder, X_t, are combined:

$$X^0 \in \mathbb{R}^{H\times W\times C} \tag{1}$$

$$X_c^n = \text{CNNEncoder}(X_c^{n-1}) \in \mathbb{R}^{H'\times W'\times D_{\text{conv}}} \tag{2}$$

$$X_t^n = \text{TransEncoder}(X_t^{n-1}) \in \mathbb{R}^{H'\times W'\times D_{\text{trans}}} \tag{3}$$

where, c and t represent two different branches in the auxiliary module, n denotes the number of encoder layers, X^n represents the features extracted by each layer of the encoder in the branch, and X^0 is the input image. The sizes and dimensions of these features are all different. Subsequently, starting from these two underlying features, enter into the Decoder part.

In the decoders of both branches, multi-scale feature fusion and image reconstruction are performed. The specific process is shown in the following formulas:

$$X_{\text{fusion}} = X_c + X_t \tag{4}$$

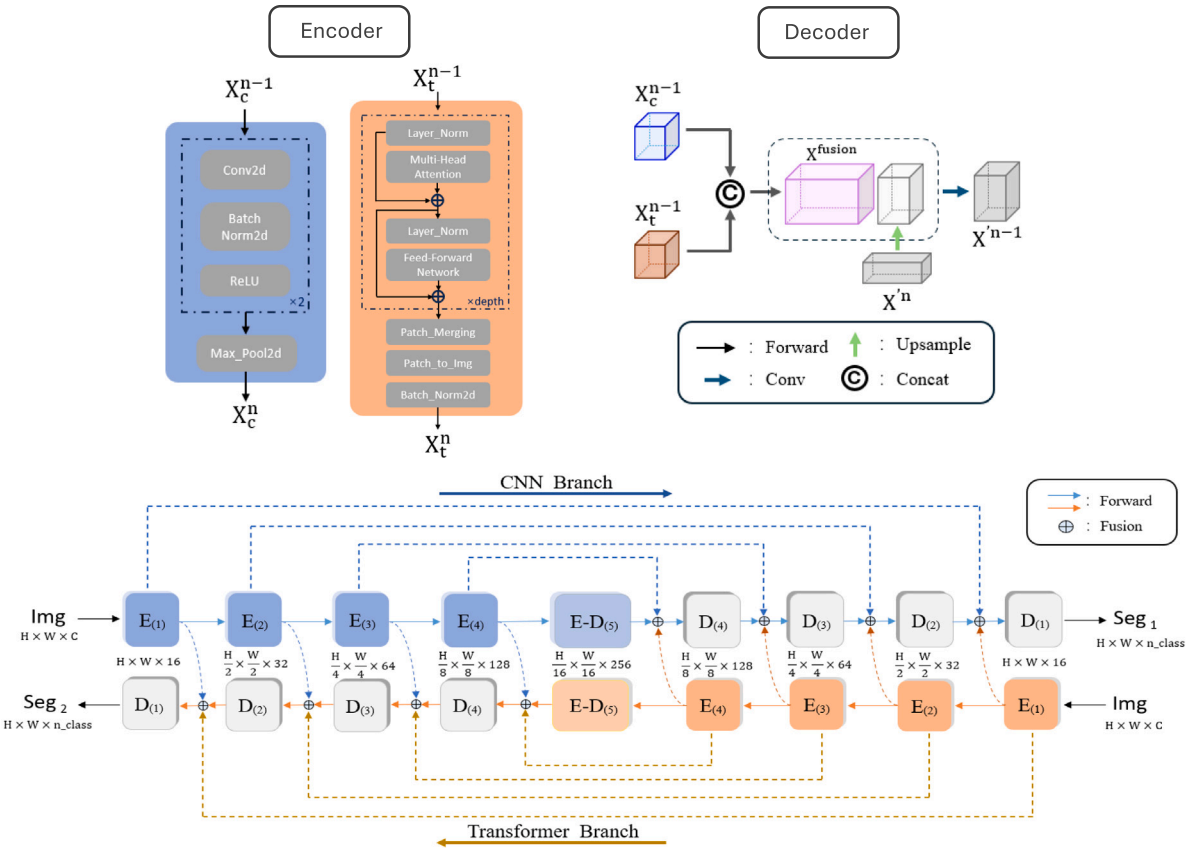


Fig. 2. The overall structure of the **Cross-Branch Cross-Fusion** (CBC-Fusion) module is depicted in the figure. The top section includes diagrams illustrating the workflow of encoder layers based on CNNs and Transformers, as well as the feature fusion process within the decoder. The bottom section presents the overall framework of CBC Fusion, which includes a segmentation model based on convolutional neural networks (blue channel) and a segmentation model based on Transformers (orange channel).

$$X_c'^{n-1} = \text{Decoder}(X_{fusion}^{n-1} + \text{Upsample}(X_c'^{n})) \quad (5)$$

$$X_t'^{n-1} = \text{Decoder}(X_{fusion}^{n-1} + \text{Upsample}(X_t'^{n})) \quad (6)$$

here, X_{fusion} represents the intermediate fusion features obtained by concatenating the features from the convolutional branch (X_c) and the Transformer branch (X_t) along the channel dimension. This fusion step combines the local and global features extracted by both branches. The n denotes the number of decoder layers, and X' represents the output of each decoder layer. The decoder output for the convolutional branch ($X_c'^{n-1}$) is computed by concatenating the upsampled output of the previous decoder layer (Upsample($X_c'^{n}$)) with the intermediate fusion features (X_{fusion}^{n-1}), and then passing it through a decoder block. Similarly, the decoder output for the Transformer branch ($X_t'^{n-1}$) is obtained by adding the upsampled output of the previous decoder layer (Upsample($X_t'^{n}$)) to the intermediate fusion features (X_{fusion}^{n-1}), and then passing it through a decoder block.

Notably, in the auxiliary module of our proposed method, the two networks undergo deep feature fusion before producing their outputs, and the decoders of both networks share the same structure. Therefore, to avoid the issue of the model becoming overly confident in the generated pseudo-labels due to the identical outputs of the two networks' unlabeled data, we have severed the cross-supervision between the outputs of the two networks in the CBC-Fusion module.

3.3. The three-branch network structure

Using only the CBC-Fusion module poses significant challenges in selecting a simple yet effective primary segmentation network, as the cross-branch fusion can introduce chaotic feature combinations. To address this issue, we designed a robust three-branch structure, as illustrated in Fig. 1. This architecture consists of two auxiliary branches that integrate features extracted from CNN and Transformer networks, with a U-Net serving as the core segmentation pathway.

The two auxiliary branches leverage the complementary strengths of CNN and Transformer networks: CNNs excel at capturing local features and fine details, while Transformers are adept at modeling long-range dependencies and global context. By fusing these features through the CBC-Fusion module, we aim to harness the advantages of both approaches. The U-Net is then employed as the primary segmentation network, utilizing the dual outputs from the CBC-Fusion module as supervisory signals. This enables the U-Net to learn joint feature representations, effectively integrating local and global information extracted by the auxiliary branches. Consequently, the U-Net can be optimally trained, resulting in improved segmentation performance.

This systematic approach not only mitigates the potential issues arising from chaotic feature combinations but also ensures that the model leverages the full spectrum of extracted features, enhancing overall segmentation accuracy.

3.4. The overall of our method

The overall architecture of the proposed model is illustrated in Fig. 1. Specifically, the input image is fed into the Cross-Branch Cross-Fusion (CBC-Fusion) module. The CNN branch extracts local features, while the Transformer branch extracts global attention features. The CBC-Fusion module then generates two distinct outputs from the CNN and Transformer branches. To aggregate the feature information extracted by the CBC-Fusion module, we leverage the robust representational capabilities of the U-net model as the main segmentation pathway. Through consistency regularization, the U-net model learns

Table 1
The information of datasets, where the training set, test set, and validation set are divided in a ratio of 7: 2: 1.

Datasets	Patients	Total slices	Training set		Validation set		Testing set	
			Samples	Slices	Samples	Slices	Samples	Slices
ACDC	100	1902	140	1312	20	210	40	380
BraTs2019	259	17 399	181	11 977	25	1741	53	3681

to systematically integrate the joint features from the CBC-Fusion module, thereby achieving an organized fusion of both local and global information.

The loss function for our proposed method primarily comprises two components: the supervised loss for labeled data and the unsupervised loss for unlabeled data. This model integrates supervised and unsupervised components:

$$L_{\text{total}} = L_{\text{sup}} + \lambda L_{\text{unsup}} \tag{7}$$

here, L_{total} represents the overall loss function used in our model, which is a combination of the supervised loss (L_{sup}) and the unsupervised loss (L_{unsup}). The parameter λ is a weighting factor that balances the contributions of the supervised and unsupervised losses, The calculation formula is defined as:

$$\lambda(t) = 0.1 \cdot \exp\left(-5\left(1 - \frac{\mathrm{t}}{\mathrm{T}}\right)^2\right) \tag{8}$$

where t denotes the current training iteration and T is the total number of training iterations.

$$L_{\text{sup}} = \frac{1}{2|D_l|} \sum_{(x_i, y_i) \in D_l} \sum_{k=1}^{3} \left(L_{ce}(\mathrm{P}_i^k, \mathrm{y}_i) + L_{dice}(\hat{\mathrm{y}}_i^k, \mathrm{y}_i) \right) \tag{9}$$

where the supervised loss (L_{sup}) is calculated on labeled data. It is the average of the cross-entropy loss and Dice loss for each label, where k denotes the three branches in the network. The loss calculated for each branch is summed. P_i^k represents the predicted probability map output by the kth branch, apply argmax to P_i^k to obtain the pseudo segmentation map $\hat{\mathrm{y}}_i^k$, and y_i represents the ground truth corresponding to the input image x_i.

$$L_{\text{unsup}} = \frac{1}{2|D_u|} \sum_{x_i \in D_u} \sum_{k \in aux} \left(L_{dice}(\mathrm{P}_i^{main}, \hat{\mathrm{y}}_i^k) + L_{dice}(\mathrm{P}_i^k, \hat{\mathrm{y}}_i^{main}) \right) \tag{10}$$

where the unsupervised loss (L_{unsup}) is calculated on unlabeled data. It computes the Dice loss between the main module and the auxiliary modules to ensure consistency between the main and auxiliary predictions. Here, k represents the two branches in the auxiliary module. Losses are calculated using the pseudo-labels from each other between the main and auxiliary modules, and the final unsupervised loss is obtained by averaging these values.

4. Experiments

We conducted an extensive set of experiments to evaluate the performance of our proposed AML-CT model. In this section, we provide an overview of the dataset used, elaborate on the implementation details, and describe the evaluation metrics employed in our experimental process. Subsequently, we present a comprehensive comparative study to highlight the effectiveness of our approach. This involves a performance comparison between AML-CT and recent popular semi-supervised medical image segmentation baselines. Furthermore, we conducted ablation studies to validate the effectiveness and necessity of each module proposed in AML-CT. Additionally, we carried out additional experiments to assess the rationale of our model. These experiments involved a comparison of the semi-supervised segmentation performance and complexity of our proposed model and its variations.

4.1. Datasets

To assess the performance of our proposed AML-CT model across various medical image segmentation tasks with objects of different sizes, we conducted extensive experiments on two available image datasets, namely ACDC and BraTs2019. Each 3D sample was split into a varying number of 2D CT images [38] (ranging from a few hundred to over a thousand), and the image sizes were standardized to 256 × 256. Approximately 70% of the samples in each dataset were selected as the training set, 10% as the validation set, and 20% as the test set. The detailed information for these datasets is provided below, with their statistical details shown in Table 1.

ACDC [39]: This dataset consists of publicly available segmented cardiac MRI images, comprising 200 annotated short-axis cardiac magnetic resonance images derived from 100 patients. The dataset is designed to serve both clinical and algorithmic research purposes, offering segmentation masks for the left ventricle (LV), myocardium (Myo), and right ventricle (RV). To accommodate the specific characteristics of large slice gaps, we chose to employ a 2D segmentation approach rather than a 3D segmentation method. In the preprocessing stage, we uniformly resized all slices to a resolution of 256 × 256 pixels and normalized the image intensity range to [0, 1]. To expand the size of the training dataset and mitigate overfitting, we applied standard data augmentation techniques, including random cropping of images, and random rotation and flipping within the range of [−25, 25]. During the inference stage, we generated results by segmenting predictions and then stacked these results into a three-dimensional volume. This processing approach significantly contributes to the enhancement of segmentation accuracy and effectiveness.

BraTs2019 [40]: The BraTs2019 dataset is a compilation of MRI images primarily employed for brain tumor segmentation. In our study, we exclusively utilized the HGG (High-Grade Glioma) data from this dataset, as the test set lacked labels, and the tumors present in HGG cases exhibit relatively distinct characteristics. The dataset encompasses a total of 259 cases, with each case comprising 155 MRI image slices, each sized at 240 × 240 pixels. For each case, four modalities are provided, namely T1, T2, Flair, and T1ce. In our study, we specifically selected the T1ce modality for segmenting the whole tumor (WT), tumor-enhancing (ET), and tumor core (TC) regions. It is worth noting that the BraTs2019 dataset poses two primary challenges for image segmentation: the intricate nature of the target (brain tumors) and the heterogeneity of their localization. The segmentation task in this dataset is particularly challenging due to the fact that the tumor-enhancing regions are quite small, occupying only a few hundred pixels within the image. The preprocessing methodology employed for this dataset aligns with that used for the ACDC dataset, with the additional step of preemptively excluding non-disease slices from each case.

4.2. Implementation details

Our experiments were conducted using the PyTorch framework[1] and executed on NVIDIA GeForce RTX 2080Ti and 4090 GPUs. We employed the same codebase as all baselines, including identical segmentation network architecture, training procedures, and evaluation methods. A fixed random seed was used to ensure consistent training

[1] https://pytorch.org/.

and evaluation results under the same hyperparameter settings. To ensure a fair comparison, comprehensive hyperparameter tuning was conducted for each model, and the best results obtained in each case are reported. The implementation details of the proposed AML-CT model are outlined as follows.

The U-Net architecture serves as the foundational segmentation backbone network for all approaches. The U-Net encoder employs four maximum pooling operations to downsample the original image resolution by a factor of 16. It consists of five layers, each comprising dual convolutional blocks with feature dimensions of 16, 32, 64, 128, and 256, respectively. Regarding the U-Net decoder, we utilize four transposed convolution operations to restore the original image resolution. In the semi-supervised scenario, we randomly sample both labeled and unlabeled data at two distinct scales (5% and 10%). Therefore, during the training process, there are two types of mini-batches for each of the two datasets. For the ACDC dataset, the numbers of labeled slices are 68 and 136, respectively. As for the BraTs2019 dataset, the numbers of labeled slices are 598 and 1197, respectively.

AML-CT and all baseline models were trained using the SGD optimizer with a weight decay parameter of 0.0001, momentum set to 0.9, and initialize the learning rate γ is 0.05, which decayed over the training epochs. Specifically, we employed polynomial learning rate decay, multiplying the learning rate by $\left(1-\frac{t}{T}\right)^{0.9}$, where t represents the current training iteration, T is the total number of training iterations which is set to 60k. The batch size for input data was set to 8, with half of it being labeled data. Finally, we set the stabilized value of the consistency loss weight λ in the model's loss function (i.e., Eq. (7)) to 0.1, and λ gradually increased from 0 to 0.1 according to Eq. (8) and then stabilized. To avoid overfitting and gradient vanishing, we augmented the training set with random cropping, flipping, and rotation methods, and applied Dropout and Batch Normalization during training.

Regarding the selection of the above hyperparameters, we employed a combination of grid search and empirical adjustment, taking into account the capacity and performance of the experimental equipment. Specifically, we search for the initial learning rate γ within the range [0.01, 0.025, 0.05, 0.075] and for the stable value of the consistency weight λ (Eq. (8)) within the range [0.05, 0.1, 0.25, 0.5]. For certain hyperparameters, such as weight decay and momentum, we made empirical adjustments based on previous experience and preliminary experiments [12,41]. For batch size, we chose the largest batch size allowed by the GPU capacity to improve the stability of model training.

4.3. Evaluation metrics

To evaluate the segmentation performance of our proposed method and some state-of-the-art baselines, two widely used segmentation evaluation metrics, the Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD95), were adopted. The formal definitions of DSC and HD95 are as follows:

$$\mathrm{DSC}(A,B)=\frac{2\times|A\cap B|}{|A|+|B|} \tag{11}$$

$$\mathrm{HD}_{95}(A,B)=95\text{th percentile of }\{d(a,b)\mid a\in A,b\in B\} \tag{12}$$

where we use A to represent the prediction, indicating the set of predicted segmented pixels, and B to represent the ground truth, indicating the set of true segmented pixels. $|A|$ and $|B|$ denote the cardinality of sets A and B, respectively, representing the number of elements (i.e., pixels) in each set. Furthermore, $|A|+|B|$ represents the cardinality of the intersection of sets A and B, which corresponds to the number of shared pixels between the prediction and the ground truth. In particular, we employ the DSC, commonly referred to as the F1-score [42], to quantitatively assess the similarity between these two sets. Additionally, we utilize the HD95 to measure the distance between the two segmentation results, taking into account shape disparities between the predicted and true segmentations. DSC is well-suited for evaluating pixel-level similarity in segmentation outcomes, while HD95 places greater emphasis on analyzing shape distinctions and the alignment of segmentation boundaries.

4.4. Main results

In this subsection, we compare our proposed method with several recent state-of-the-art (SOTA) semi-supervised medical image segmentation approaches. These baseline methods are categorized into two groups based on their network structures. The selected baseline semi-supervised methods mainly fall into two categories. The first category consists of semi-supervised models based on CNN, including Mean Teacher (MT) [34], Entropy Minimization (EM) [43], Uncertainty Aware Mean Teacher (UAMT) [44], Cross Consistency Training (CCT) [46], Regularized Dropout (RD) [45], Cross Pseudo Supervision (CPS) [12]; the second category is based on Transformer methods including Cross Teaching between CNN and Transformer (CTCT) [13], When CNN meet with ViT (S4CVnet) [15]. In addition, we also compared our method with some fully supervised models. Due to the AML-CT model's primary segmentation network being based on the U-Net architecture, we initially compared it with a fully supervised U-Net model. Additionally, we employed the fully supervised approach using only 5% and 10% labeled data (referred to as Sup). Furthermore, since our model incorporates a transformer module and the introduction of Transformer models into segmentation models has become popular in recent years, we also compared it to some fully supervised segmentation networks that include Transformer models, including Swin-Unet and TransUNet models.

For a fair comparison, all implementations of these methods for the same task are consistent with our framework, and these methods are available online [41]. During validation and testing, all auxiliary training modules are discarded, and only the trained main segmentation network is used to generate the final predictions. Experiments are conducted using only 5% and 10% of the training images as labeled data. The comparative results on different datasets are presented below.

4.4.1. Results on ACDC

As shown in Table 2, it lists the segmentation results and averages for LV, Myo, and RV on the ACDC dataset for various baseline methods and our proposed AML-CT model. It is worth noting that among all baseline methods, the best baseline model (S4CVnet) outperforms other existing methods at the 5% and 10% labeled data settings. Initially, in comparison to the fully supervised approach (Sup), all semi-supervised segmentation methods demonstrated a certain level of performance improvement. This suggests that semi-supervised learning can leverage unlabeled data to improve segmentation performance, and introducing Transformer models into traditional CNN-based semi-supervised learning methods can further significantly enhance model segmentation performance. Secondly, compared to other semi-supervised methods, the co-training approach achieved superior segmentation performance, validating the effectiveness of co-training.

Meanwhile, compared to our proposed method, AML-CT achieves better performance in both DSC and HD95 evaluation metrics than all existing methods, especially with a significant improvement in the DSC metric compared to the best baseline (S4CVnet) at the 5% labeled data setting. It is noteworthy that at the 10% labeled data setting, our proposed method's performance is comparable to fully supervised U-net in both DSC (0.4%) and HD95 (0.509 mm). Additionally, it slightly outperforms Swin-Unet and TransUNet in image segmentation tasks across two datasets. This can be attributed to the fact that, in medical image segmentation, lesions typically reside in specific local regions, whereas Transformer-based models tend to focus on global features of the image, limiting their effectiveness in this domain. Our method, by integrating local and global features extracted by CNNs and Transformer networks, enhances the model's understanding of medical images, demonstrating the effectiveness of our approach.

Table 2
The results of our proposed AML-CT and the state-of-the-art semi-supervised medical image segmentation baselines on ACDC dataset with different ratio of labeled data, where the best and the second best results are bold and underlined, respectively.

Labeled	Method	LV		Myo		RV		Mean	
		DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
5%	Sup	0.4723	40.76	0.6576	8.737	0.7231	10.41	0.6176	19.97
	MT [34]	0.5073	30.94	0.6607	17.33	0.7409	22.16	0.6362	23.48
	EM [43]	0.5346	11.61	0.6694	6.491	0.7657	9.283	0.6566	9.127
	UAMT [44]	0.5077	33.16	0.6523	18.74	0.7158	25.38	0.6253	25.76
	RD [45]	0.5725	7.120	0.6764	6.261	0.7500	5.619	0.6663	6.333
	CCT [46]	0.5630	15.92	0.6352	15.81	0.7018	21.11	0.6333	17.61
	CPS [12]	0.5714	11.72	0.6692	5.374	0.7527	7.544	0.6644	8.215
	CTCT [13]	0.7091	3.114	0.7236	4.464	0.8005	4.810	0.7444	4.130
	S4CVnet [15]	0.7150	3.741	0.7288	**2.804**	0.7977	**2.624**	0.7472	**3.056**
	AML-CT (ours)	**0.8579**	**1.681**	**0.8104**	3.021	**0.9104**	5.505	**0.8595**	3.402
10%	Sup	0.8540	5.398	0.8238	11.10	0.8740	9.634	0.8506	8.712
	MT [34]	0.8420	2.690	0.8306	8.043	0.8845	11.53	0.8524	7.420
	EM [43]	0.8615	4.735	0.8389	7.818	0.8903	11.36	0.8635	7.971
	UAMT [44]	0.8253	13.52	0.8107	17.76	0.8726	18.53	0.8362	16.61
	RD [45]	0.8691	1.719	0.8382	3.953	0.8950	11.14	0.8674	5.604
	CCT [46]	0.8318	4.058	0.8135	11.18	0.8836	14.03	0.8451	9.756
	CPS [12]	0.8751	1.924	0.8529	7.235	0.9076	7.429	0.8785	5.529
	CTCT [13]	0.8586	3.797	0.8527	4.446	0.9044	7.721	0.8719	5.321
	S4CVnet [15]	0.8775	4.100	0.8564	2.570	0.9102	9.263	0.8813	5.311
	AML-CT (ours)	**0.9064**	**1.462**	**0.8820**	**1.500**	**0.9360**	**2.061**	**0.9081**	**1.674**
100%	U-net [27]	0.9023	1.311	0.8913	1.578	0.9427	3.663	0.9121	2.184
	Swin-Unet [14]	0.9025	1.219	0.8590	2.479	0.9215	3.116	0.8925	2.271
	TransUNet [23]	0.9054	2.498	0.8728	2.697	0.9309	3.318	0.9030	2.227

Table 3
The results of our proposed AML-CT and the state-of-the-art semi-supervised medical image segmentation baselines on BraTs2019 dataset with different ratio of labeled data, where the best and the second best results are bold and underlined, respectively.

Labeled	Method	WT		TC		ET		Mean	
		DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
5%	Sup	0.5929	14.28	0.4757	28.05	0.7314	9.793	0.5983	17.37
	MT [34]	0.6003	14.22	0.4765	32.15	0.7431	10.05	0.6067	18.81
	EM [43]	0.5862	16.73	0.4724	27.99	0.7386	12.86	0.5991	19.19
	UAMT [44]	0.5930	16.92	0.4751	35.54	0.7371	13.44	0.6018	21.97
	RD [45]	0.5874	13.32	0.4751	28.36	0.7266	12.49	0.5964	17.72
	CCT [46]	0.6050	17.02	0.4902	22.12	0.7469	15.02	0.6140	18.05
	CPS [12]	0.5971	18.18	0.4715	34.01	0.7451	19.11	0.6046	23.77
	CTCT [13]	0.6281	16.43	0.5029	29.68	0.7681	13.13	0.6330	19.75
	S4CVnet [15]	0.6247	17.54	0.5085	24.81	0.7843	15.95	0.6392	19.43
	AML-CT (ours)	**0.6277**	**12.53**	**0.5133**	**21.26**	**0.7904**	**6.247**	**0.6438**	**13.35**
10%	Sup	0.6143	10.76	0.4991	18.32	0.7649	6.269	0.6261	11.79
	MT [34]	0.6263	11.44	0.5091	28.76	0.7680	15.41	0.6344	18.53
	EM [43]	0.6144	14.23	0.5053	29.83	0.7637	15.12	0.6278	19.73
	UAMT [44]	0.6232	12.63	0.5066	27.43	0.7759	11.71	0.6353	17.25
	RD [45]	0.6283	13.81	0.5072	25.99	0.7760	14.57	0.6372	18.12
	CCT [46]	0.6255	13.80	0.5136	22.91	0.7714	9.454	0.6368	15.39
	CPS [12]	0.6288	10.52	0.5041	21.04	0.7756	14.49	0.6362	15.35
	CTCT [13]	0.6225	11.45	0.5006	26.60	0.7726	10.34	0.6319	16.13
	S4CVnet [15]	0.6447	10.64	0.5360	18.90	0.7959	7.373	0.6571	12.31
	AML-CT (ours)	**0.6578**	**8.369**	**0.5384**	**4.44**	**0.7989**	**5.406**	**0.6651**	**9.406**
100%	U-net [27]	0.6683	8.017	0.5791	17.08	0.8098	7.262	0.6857	10.79
	Swin-Unet [14]	0.6500	6.410	0.5741	3.664	0.8128	3.664	0.6790	7.500
	TransUNet [23]	0.6326	10.84	0.5856	11.35	0.7811	4.040	0.6664	8.743

4.4.2. Results on BraTs2019

We further evaluated our proposed method on the BraTs 2019 dataset. As shown in Table 3 which lists the results of all methods on the BraTs2019 dataset. It can be observed that our model achieves better or comparable performance on the two evaluation metrics at 5% and 10% labeled data settings compared to all existing methods. Similar to ACDC, the CTCT and S4CVnet models outperform other existing methods in both annotation data settings, further confirming the improvement in semi-supervised medical image segmentation performance with the introduction of Transformer models. In particular, in the case of 10% annotated data, the S4CVnet model has a DSC metric of 65.71%, surpassing other existing methods, while our method shows an improvement of 0.80% in DSC, achieving better performance. When the labeled data is set to 5%, our method's segmentation performance remains slightly better than all existing methods, demonstrating the efficiency of our approach. Compared to fully supervised methods, in the 10% labeled data setting, our method performs slightly lower than U-Net in terms of DSC but achieves comparable segmentation performance to Transformer-based models such as Swin-Unet and TransUNet, demonstrating the effectiveness of our approach.

4.4.3. Visualization comparison of different methods

Finally, we performed visualizations and comparisons on the test sets of the ACDC and BraTs2019 datasets using models trained with 10% labeled data. As shown in Fig. 3, it shows the visualized segmentation results of AML-CT and the baselines on two datasets.

Specifically, the results of ACDC dataset (at the first four rows) show that: (i) the segmentation results of traditional semi-supervised

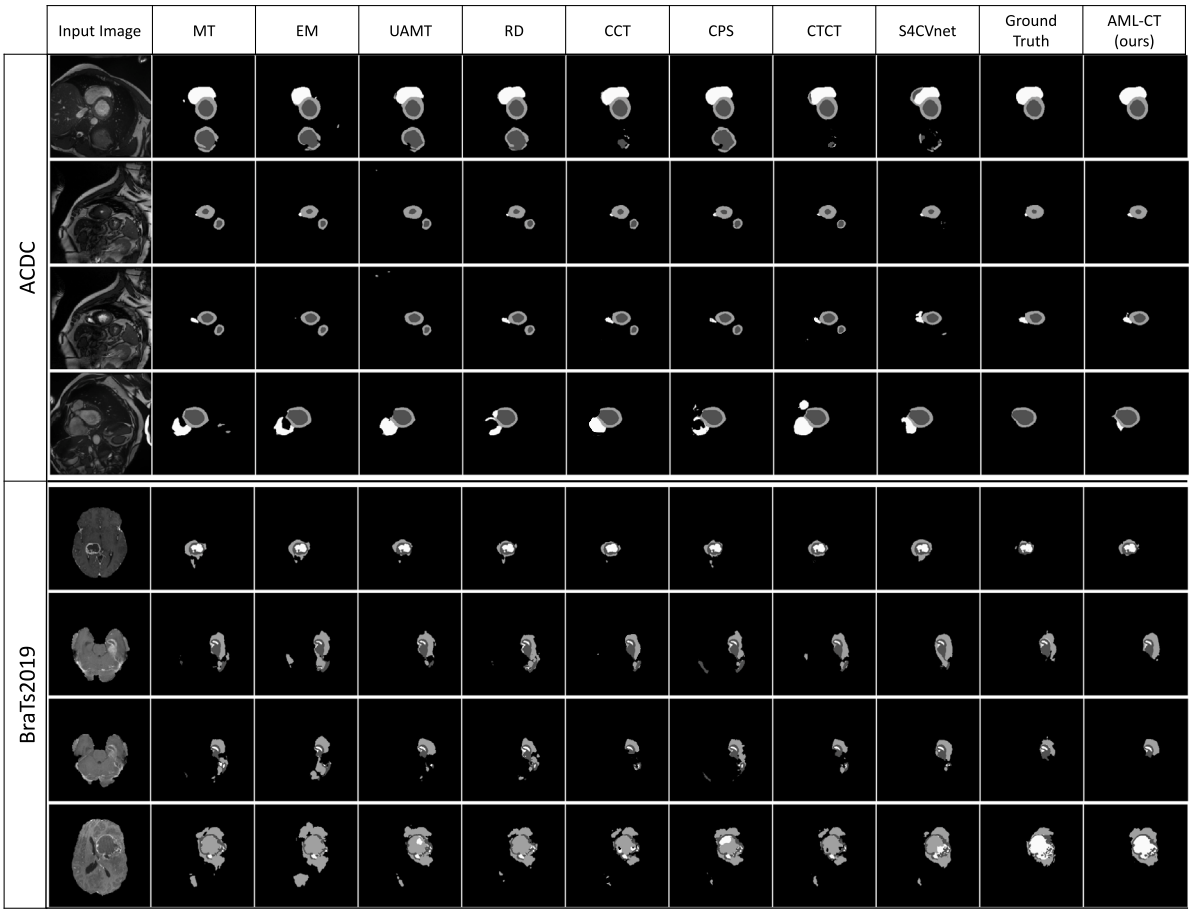


Fig. 3. Examples of visualized segmentation results of our proposed AML-CT and the baselines on the ACDC and BraTs2019 datasets, where each segmentation model was trained with 10% labeled data.

segmentation network based on CNNs are obvious incorrect and sometimes even misidentification; (ii) the segmentation results of S4CVnet are relatively better for segmenting the region of the heart, but their segmentation results are still inaccurate sometimes; and (iii) the segmentation performance of our AML-CT is much better than the baselines, its segmentation results for the small tumor objects are very close to ground truths. Similarly, we have the following observations for the BraTs2019 dataset. (i) The semi-supervised segmentation network based on CNNs methods can neither correctly recognize and segment the tumor in the slices; (ii) The S4CVnet model are better but their performances in segmenting the edge areas of the tumor are not satisfactory; and (iii) the segmentation results of the proposed AML-CT is better than the baselines. Furthermore, the segmentation results of AML-CT exhibit the highest similarity when compared to the ground truth in this two datasets. Therefore, these visualized examples greatly demonstrate again that by the various mechanisms in the model, AML-CT remedies the drawbacks of the existing deep segmentation models, and achieves much better performances in medical image segmentation tasks, especially for small objects.

4.5. Ablation study

To validate the effectiveness of our semi-supervised segmentation method, we conducted ablation experiments on two datasets. The results are shown in Tables 4 and 5. In these tables, CPS* represents the cross pseudo-supervision framework, Modules indicates the ablation study on the two proposed modules, and AML denotes the study of different branch combinations with our two proposed modules. Specifically, the models for ablation experiments and comparisons are as follows: **(i) CPS* w/ C × C** is a model that utilizes two convolution-based segmentation networks. During the training process, they generate pseudo labels for each other and use these pseudo labels for cross supervision; **(ii) CPS* w/ T × T** is a model that replaces the two convolutional networks in the CPS framework with networks built using our proposed ViT+ module, and utilizes these two Transformer networks for cross supervision; **(iii) CPS* w/ C × T** is a model that replace the one of CNNs network in the CPS model with a branch based on our proposed ViT+ module. It should be noted that our experimental model was conducted under the conditions of two branches based on CNNs and Transformers with a pyramid structure, thus, this model is the basic model for ablation experiments (denoted sup); **(iv) sup w/ CBC-Fusion** is an intermediate model that uses only the CBC-Fusion module on top of the supervised model (denoted sup+CBC); **(v) sup w/ Tb-net** is an intermediate model that uses only the TB-net structure on top of the supervised (sup) model (denoted sup+TB); **(vi) AML w/C×C** is a model that replaces the Transformer branch in the auxiliary module with a convolution-based network branch in our proposed model; **(vii) AML w/ C×T** is a model that replaces the CNNs branch in the auxiliary module with a Transformer-based network branch in our proposed model; and **(viii) AML-CT (ours)** is the final model constructed by incorporating all the proposed modules.

In this ablation study, we conducted comprehensive testing and comparisons on the ACDC and BraTs2019 datasets with 10% labeled data. The goal was to analyze the contributions of each module to the experimental results, with evaluation metrics including DSC (Dice Similarity Coefficient) and HD95 (95% Hausdorff Distance). The average results of the ablation experiment are shown in Tables 4 and 5.

Table 4
Ablation study on the ACDC dataset with 10% labeled data was conducted using DSC and HD95 as evaluation metrics.

Methods		LV		Myo		RV		Mean	
		DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
CPS*	C × C	0.8751	1.924	0.8529	7.235	0.9076	7.429	0.8785	5.529
	T × T	0.8333	12.35	0.8262	6.890	0.9053	9.090	0.8549	9.651
	C × T (sup)	0.8584	3.299	0.8524	9.378	0.9143	13.11	0.8750	8.596
Modules	sup+CBC	0.8926	2.264	0.8626	3.240	0.9139	5.089	0.8897	3.531
	sup+Tb	0.8573	4.681	0.8621	3.111	0.9197	5.318	0.8797	4.370
AML	C × C	0.8968	1.628	0.8749	1.918	0.9311	1.573	0.9009	1.706
	T × T	0.8813	1.674	0.8784	1.879	0.9294	1.689	0.8963	1.747
	C × T (ours)	0.9064	1.462	0.8820	1.500	0.9360	2.061	0.9081	1.674

Table 5
Ablation study on the BraTs2019 dataset with 10% labeled data was conducted using DSC and HD95 as evaluation metrics.

Methods		WT		TC		ET		Mean	
		DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
CPS*	C × C	0.6288	10.52	0.5041	21.04	0.7756	14.49	0.6362	15.35
	T × T	0.5900	15.65	0.4659	18.98	0.7367	16.55	0.5975	17.06
	C × T (sup)	0.5935	15.06	0.4986	17.39	0.7413	17.15	0.6112	16.53
Modules	sup+CBC	0.6191	9.242	0.5355	17.88	0.7731	11.43	0.6426	12.85
	sup+Tb	0.6203	10.44	0.5344	19.78	0.7691	10.38	0.412	13.53
AML	C × C	0.6480	8.664	0.5268	13.98	0.7925	6.543	0.6558	9.730
	T × T	0.6508	9.198	0.5312	14.93	0.7984	8.510	0.6601	10.88
	C × T (ours)	0.6578	8.369	0.5384	14.44	0.7989	5.406	0.6651	9.406

First, we removed one branch at a time under the three-branch network condition and compared the performance changes after removing each branch to verify the contribution of each branch to the overall model performance. Removing one branch from the three-branch network results in the cross-pseudo-supervision framework. Under this framework, we tested the performance of different branch combinations. The table shows that (i) using only the convolutional branch (denoted as CPS*+ C × C) and using both the convolutional and Transformer branches (denoted as CPS*+ C × T) yield similar performance; (ii) both configurations outperform using only the Transformer branch (denoted as CPS*+ T × T). This is because in medical image segmentation, lesions are usually located in specific local areas. Therefore, Transformer-based models have limitations in the field of medical image segmentation. Convolutional networks are very effective in focusing on and extracting these local features. Hence, in a semi-supervised medical image segmentation framework, it is necessary to incorporate a convolution-based Unet network while integrating Transformers to extract global features, enabling the model to better understand medical images. Then, as observed in Table, the segmentation performance of models sup+CBC and sup+TB consistently outperforms the use of the sup model. This is because the CBC-Fusion module introduces cross-branch connections and fusion between the two branches, allowing each model to access feature information extracted by the other branch. By combining this information, the models can better comprehend the input images. The TB-net incorporates a U-net model with strong representation capabilities on top of the sup model, enabling the aggregation of information extracted from the other two branches for improved image representation. Thus, this confirms the effectiveness of the proposed CBC-Fusion and TB-net modules in medical image segmentation tasks.

Finally, when both modules were included (denoted as AML), we varied the branch combinations within the auxiliary module to verify the impact of different branch combinations on model performance. The results show that (i) using only the convolutional branch (denoted as AML+ C × C) and using only the Transformer branch (denoted as AML+ T×T) yield similar segmentation performance; (ii) the combination of both the convolutional and Transformer branches (denoted as AML+ C×T) outperforms other configurations; (iii) AML+ C×T shows significant improvement over sup+CBC and sup+Tb. Furthermore, from the results, it can be observed that the AML-CT model, which incorporates both the CBC-Fusion and TB-net modules, achieves more significant performance improvements compared to using only one of these modules. This is because the CBC-Fusion and TB-net modules address different aspects of the problem to enhance the segmentation performance of the U-Net model. The CBC-Fusion module is employed to extract feature information from input images, effectively combining features extracted from different perspectives. This mutual comparison and correction contribute to the improvement of the model's segmentation performance. On the other hand, TB-net is used to aggregate mixed information extracted from the CBC-Fusion module for better feature representation. This demonstrates that simultaneously incorporating both the CBC-Fusion and TB-net modules into the dual-teacher model is a reasonable approach for achieving more accurate semi-supervised image segmentation results.

Therefore, the above observations demonstrate that the proposed three advanced modules are all effective and essential for AML-CT to achieve the superior semi-supervised medical image segmentation performances.

Table 6
The number of parameters when gradually increase the number of ViT+ in the decoder of the Transformer branch, where the parameter count (para.) is measured in millions.

Model	Para./m	Model	Para./m
e = 0	3.61	d = 0	50.4
e = 1	30.8	d = 1	56.1
e = 2	38.2	d = 2	62.6
e = 3	44.6	d = 3	69.9
e = 4	50.4	d = 4	95.2

4.6. Additional experiments

4.6.1. Effect of the placement of ViT+ modules

Furthermore, we conducted additional validation experiments on the number of ViT+ modules in the Transformer branch of the AML-CT model. We introduced the Vision Transformer model only in the Encoder part of the pathway, while in the Decoder part, a structure similar to CNN, with convolutional layers, was employed. We designed the model in this way because we believe that in the Encoder–Decoder structure, the role of the Encoder is to gradually reduce the size of the input image and increase the receptive field through step-by-step

Table 7
The results of our proposed AML-CT on the ACDC dataset, gradually increasing the number of ViT+ modules in the Transformer branch when using 10% labeled data, where t represents the number of ViT+ modules introduced from the encoder to the decoder. The best results in the table are highlighted in bold.

Model	LV		Myo		RV		Mean	
	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
t = 0	0.8873	1.480	0.8665	6.169	0.9216	5.413	0.8918	4.354
t = 1	0.8937	1.463	0.8714	4.115	0.9284	3.379	0.8978	2.985
t = 2	0.8914	1.821	0.8767	2.301	0.9321	2.369	0.9001	2.163
t = 3	0.9006	1.377	0.8758	1.533	0.9336	2.657	0.9033	1.856
t = 4	**0.9064**	**1.462**	**0.8820**	1.500	0.9360	2.061	**0.9081**	1.674
t = 5	0.8934	1.641	0.8776	1.955	0.9324	2.353	0.9012	1.983
t = 6	0.9001	1.588	0.8813	**1.097**	**0.9402**	**1.081**	0.9072	**1.255**
t = 7	0.8989	1.576	0.8747	1.231	0.9302	2.093	0.9013	1.634
t = 8	0.8983	1.464	0.8742	2.232	0.9315	2.790	0.9014	2.162

Table 8
The results of our proposed AML-CT on the BraTs2019 dataset, gradually increasing the number of ViT+ modules in the Transformer branch when using 10% labeled data, where t represents the number of ViT+ modules introduced from the encoder to the decoder. The best results in the table are highlighted in bold.

Model	WT		TC		ET		Mean	
	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
t = 0	0.6416	8.794	0.5399	15.53	0.7941	6.363	0.6585	10.23
t = 1	0.6485	10.30	0.5210	15.70	0.7911	11.29	0.6535	12.43
t = 2	0.6529	8.317	0.5311	19.44	0.8002	5.811	0.6614	11.19
t = 3	0.6557	9.854	0.5264	13.61	0.7983	8.497	0.6602	10.65
t = 4	**0.6578**	8.369	0.5384	14.44	0.7989	5.406	0.6651	9.406
t = 5	0.6478	7.880	0.5379	15.87	0.7944	**4.231**	0.6600	9.329
t = 6	0.6482	**7.571**	**0.5475**	**13.53**	0.8024	4.353	0.6660	**8.488**
t = 7	0.6505	9.472	0.5362	15.27	0.8019	5.666	0.6629	10.23
t = 8	0.6559	8.433	0.5415	14.36	**0.8053**	5.076	**0.6675**	9.289

operations, allowing the model to extract more information about the features of the input image. In the Decoder part, the low-level features generated by the Encoder are combined with the features extracted by the Encoder to gradually reconstruct and restore the original image size. Therefore, there is no need to further extract attention features in the Decoder.

To validate the above hypothesis, in this section, we conducted relevant experiments on the ACDC and BraTs2019 datasets. The experiment involved gradually changing the number of introduced ViT+ modules in the Transformer branch on the basis of the proposed model. We compared the segmentation performance and the number of model parameters as the number of ViT+ modules in the decoder of the Transformer branch (denoted as d) varied from 0 to 4. Additionally, to further validate the effectiveness of the ViT+ modules we adopted, we also changed the number of ViT+ modules introduced in the encoder of this branch, comparing the segmentation performance and the number of model parameters as the number of ViT+ modules in the encoder of the Transformer branch (denoted as e) varied from 0 to 4. The baseline for the experiment was our proposed method when $e = 4$ or $d = 0$. The experimental results are shown in Tables 7 and 8. In order to compare the results of each group of experiments intuitively and obtain experimental conclusions, we display the experimental results of the two datasets in Fig. 4.

The experimental results shown in Tables 7 and 8. The experimental results from the table show that as t increases from 0 to 4, when introducing more ViT+ modules in the encoder of the Transformer branch, the segmentation results are improved. However, as t continues to increase, i.e., introducing more ViT+ modules in the decoder of that branch, there is no improvement in the final segmentation results. Specifically, on the ACDC dataset, our method (with $t = 4$) achieved the best results. As t increased from 0 to 4, our method showed a 1.63% improvement in average DSC compared to the worst result (with $t = 0$). However, as t continued to increase to 8, the segmentation results fluctuated but remained at a similar level. Similarly, on the BraTs2019 dataset, as t increased from 0 to 4, our method showed a 1.16% improvement in average DSC compared to the worst result (with $t = 1$). When $t = 8$, the model achieved the best segmentation results, but we can observe that the segmentation results between $t = 4$ and $t = 8$ are at a similar level.

It is important to note that as the number of Transformer modules in the channel decoder increases, the corresponding model parameters also increase, and the number of parameters in the ViT model is proportional to the size of the input image. Since the image size gradually increases after upsampling in the decoder, the model's parameter count continues to grow. Specifically, with the increase of t, the increment in parameter values between different models also increases. The parameter values for each experimental model are shown in Table 6. From the results, we can conclude that with the increase of t from 4 to 8, the segmentation model's performance did not significantly improve, but the number of model parameters increased significantly. To better illustrate the results, we show the DSC results for each model on the quantitative dataset and the number of model parameters in Fig. 4. From the figure, it can be observed that as t increases, the complexity of the model grows exponentially, while the segmentation performance does not show a significant increase. The experimental results indicate that increasing the number of ViT+ modules in the Transformer branch's decoder significantly increases the computational burden of the model without achieving better efficiency. Therefore, this also confirms the rationality of our approach. From the above results table, it can be observed that even when $t = 0$, meaning no Transformer modules are introduced in the entire model, decent segmentation results can still be obtained. This is because our proposed cross-branch feature cross-fusion module and three-branch network module are still included in the model. Since these two modules play a crucial role in improving segmentation performance, it confirms the effectiveness of the proposed modules.

4.6.2. Effect of decoder structure

In our proposed AML-CT model, the CBC-Fusion module, used to extract jointly learned features, has two outputs from two different segmentation networks, i.e., CNN and Transformer. However, when constructing the model, the decoder structures of both pathways are the same, utilizing convolutional structures. Through the experiments mentioned earlier, we successfully demonstrated that adding transformer modules to the decoder does not further improve the model's performance. Therefore, using only convolutional structures in the decoder is highly efficient. Consequently, in our model, the decoders of both the CNN and Transformer pathways have the same structure. The reason for designing the CBC-Fusion module with two outputs is to ensure smooth flow of feature information within the module without compression. To verify this, we conducted a corresponding experiment. In the control experiment, we fused the results of the encoders (smallest image size and maximum channel number) from the two segmentation networks within the CBC-Fusion module. Subsequently, starting from the fused result, we proceeded with the decoding process. The decoding operations were the same as before, and other model settings remained constant. We refer to this model as AML-F. The model diagram of the AML-F model is shown in Fig. 6, in this model, the fusion module provides only one output by merging the decoders. For unlabeled data, this module supervises the output against the main segmentation network's output. For labeled data, each part of the output is supervised against the corresponding true label, completing the semi-supervised task.

Furthermore, on the basis of the AML-F model, we also verified the effect of the Transformer module in the decoder. Therefore, we conducted experiments similar to those in Section 4.6.1, progressively increasing the number of Transformer modules in the decoder. We compared the results of each group of experiments to further validate the aforementioned experimental results. The results are shown in Table 9.

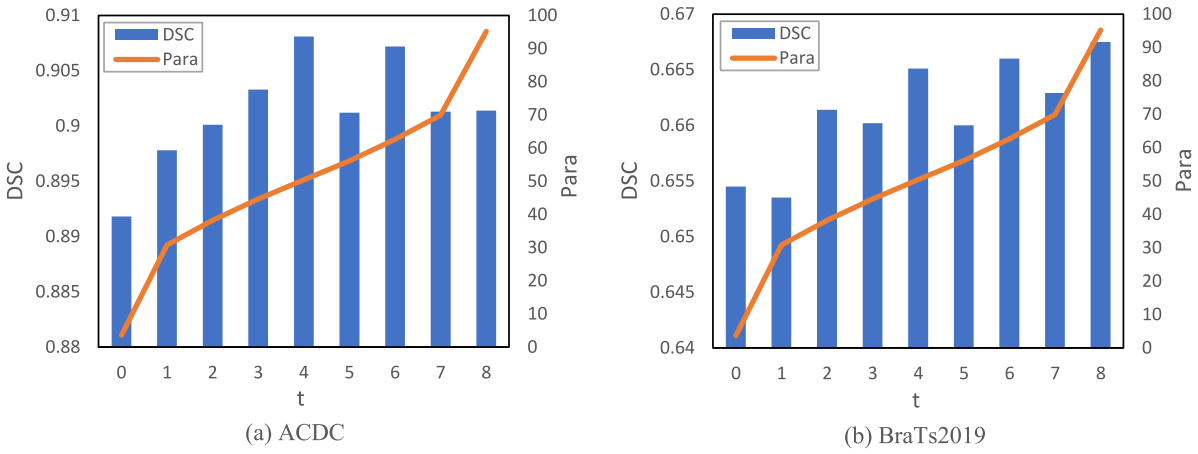


Fig. 4. The comparison of average segmentation results (DSC) and the model parameters of the AML-CT model and its variations. Where t represents the number of ViT+ modules introduced in the Transformer branch, where $t = 4$ corresponds to our model.

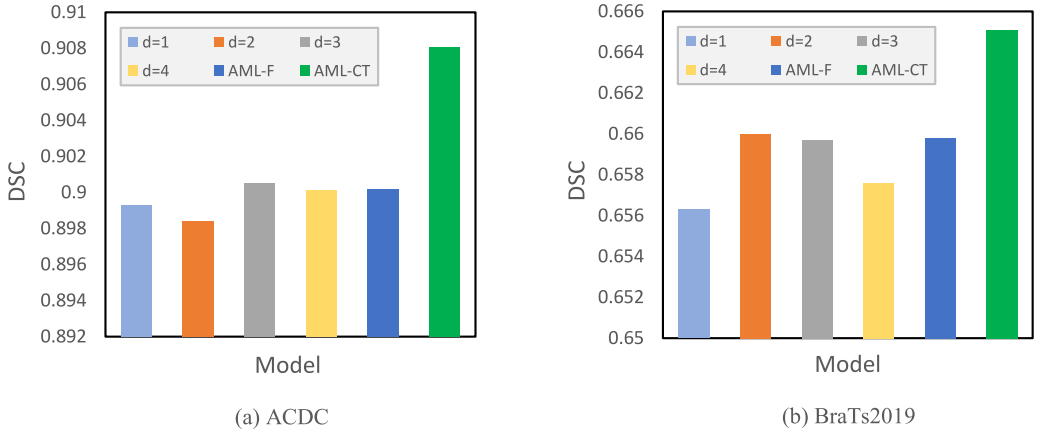


Fig. 5. Comparison of the average segmentation results (DSC) between the AML-F model and its variations with AML-CT, where d represents the number of ViT+ modules introduced in the decoder of the AML-F model.

Table 9
The results of AML-F on ACDC and BraTs2019 datasets with different organ of segmenting objects at 10% labeled data setting, where the best results are bold. Where d represents the number of ViT+ modules introduced in the decoder of the Transformer branch in the AML-F model. The Category.1, 2, and 3 represent three categories, where LV, Myo, and RV in ACDC; WT, TC, and ET in BraTs2019.

Dataset	Model	Category.1		Category.2		Category.3		Mean	
		DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
ACDC	d = 1	0.8955	1.557	0.8786	1.450	0.9239	2.156	0.8993	1.721
	d = 2	0.8933	1.530	0.8768	2.855	0.9253	3.338	0.8984	2.574
	d = 3	0.8972	1.318	0.8748	1.141	0.9295	2.035	0.9005	1.498
	d = 4	0.8988	1.533	0.8759	1.200	0.9258	4.662	0.9001	2.465
	AML-F	0.8978	1.770	0.8765	1.677	0.9265	1.871	0.9002	1.772
	AML-CT	0.9064	1.462	0.8820	1.500	0.9360	2.061	0.9081	1.674
BraTs2019	d = 1	0.6481	7.739	0.5380	16.81	0.7827	4.276	0.6563	9.608
	d = 2	0.6472	7.783	0.5384	12.81	0.7943	6.664	0.6600	9.087
	d = 3	0.6461	8.021	0.5396	14.63	0.7935	7.963	0.6597	10.20
	d = 4	0.6474	8.412	0.5354	13.84	0.7900	6.070	0.6576	9.442
	AML-F	0.6493	8.004	0.5367	12.98	0.7933	4.620	0.6598	8.536
	AML-CT	0.6578	8.369	0.5384	14.44	0.7989	5.406	0.6651	9.406

The average segmentation results (DSC) of the AML-F model and its variations are shown in Fig. 5. From the experimental results, it can be observed that reducing the number of outputs in the CBC-Fusion module through feature fusion slightly decreases the segmentation performance on both datasets, but the difference is not significant. This is because, even though the AML-F model has only one output in the auxiliary module, the introduced Cross-Branch Cross-Fusion and Three-Branch Network modules are still present in the model. Since these two modules play a crucial role in improving segmentation performance, using only one decoder in the CBC-Fusion module has an impact on segmentation performance, but not too much. At the same time, it also demonstrates the importance of independent decoding processes in the two branches for the transfer of features. Additionally, we conducted experiments similar to Section 4.6.1, where we progressively increased the number of ViT+ modules in the decoder of the transformer branch in the AML-F model and compared the final segmentation results.

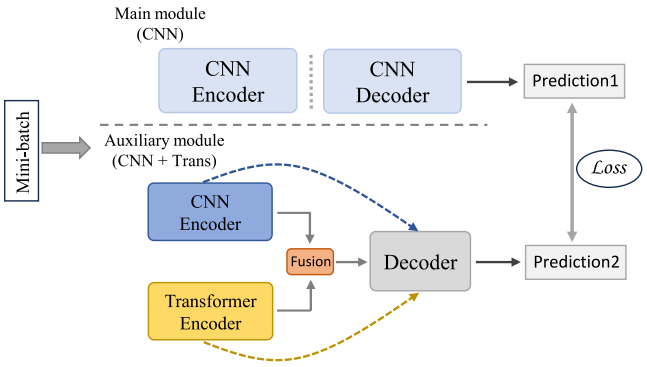


Fig. 6. The model diagram of the **AML-F** model, which is a variant of the AML-CT model. In the auxiliary module, the two different segmentation networks share one decoding network, and the auxiliary module has only one output, mutually supervised with the output of the main module. The rest of the model is the same as AML-CT.

The experimental results show that introducing ViT+ modules in the decoder of the transformer branch does not effectively improve segmentation metrics compared to AML-F, further confirming that continuing to introduce ViT+ modules in the decoder of the branch does not help enhance the model's performance.

5. Discussion and future work

5.1. Social impact for proposed model

The intelligent segmentation model proposed in this paper can be applied to various medical diagnostic scenarios, offering rapid, high-precision results to assist clinicians in making informed decisions. By providing objective reference data, the model significantly reduces diagnosis time and frees up physicians to focus on more complex cases. For example, in brain tumor screening, physicians often need to manually annotate potential tumor regions from a large set of brain medical images. Due to the irregular or unclear shape and size of the tumor, different physicians may interpret the same image differently, resulting in inconsistent annotations of tumor boundaries. This subjectivity not only affects the accuracy of diagnosis but also increases the risk of misdiagnosis or missed diagnoses. Moreover, manual annotation is time-consuming, particularly when handling large volumes of imaging data, which in turn reduces diagnostic efficiency. By applying the intelligent segmentation model proposed in this paper to such medical scenarios, the model can quickly identify and accurately segment lesions in the images, providing objective references for physicians. This assists them in more precise disease assessment, surgical planning, and treatment decision-making. Consequently, the model alleviates diagnostic workloads, saving time for both doctors and patients, and plays a crucial role in reducing the strain on medical resources.

However, when training segmentation models using traditional fully supervised methods, large amounts of labeled data are typically required to achieve high accuracy. Obtaining large-scale, high-quality labeled data is a time-consuming and resource-intensive process. In this paper, we propose a semi-supervised segmentation method that requires only a small amount of labeled data to achieve segmentation performance comparable to fully supervised methods. This greatly reduces the training cost of intelligent medical image segmentation models while improving data efficiency.

5.2. Limitations and future work

Despite the promising results achieved by the deep mutual learning framework that combines CNNs and Transformer networks, several limitations remain to be addressed. First, during the feature fusion process, the model may struggle to effectively emphasize key feature locations. Although the network integrates both local and global features through CNNs and Transformers, certain critical spatial or contextual information may not receive sufficient attention. To address this limitation, a potential future research direction may be to incorporate advanced attention mechanisms at the feature fusion stage to improve the model's ability in focusing on important regions within medical images, ultimately enhancing segmentation accuracy [47,48], particularly in complex cases with fine details.

Second, the current method of information transfer between the main module and the auxiliary module is governed by consistency loss. While this ensures some degree of information alignment, it may not fully capture the nuanced interactions required for optimal segmentation performance. Therefore, in the future, exploring novel loss functions specifically designed to better facilitate the flow of information between the auxiliary and main networks could be a potential research direction to further improve the model's performances, especially when balancing supervised and unsupervised data during the training process [49].

6. Conclusion

While automated medical image segmentation using deep learning has demonstrated remarkable success, it is still constrained by the demand for extensive and detailed annotations when applied in clinical settings. Semi-supervised learning, which leverages a limited set of labeled data and a substantial pool of unlabeled data, offers a promising solution to this challenge. In this research, we introduce the AML-CT model (CNN and Transformer co-learning) for semi-supervised medical image segmentation. This deep learning model effectively utilizes a small amount of labeled data alongside a vast collection of unlabeled data to achieve precise segmentation of medical images. The model features auxiliary modules that incorporate both CNN and Transformer components to extract comprehensive feature information from input medical images. The co-learning process between these two branches is facilitated through cross-branch feature cross-fusion mechanisms within the modules. Ultimately, the primary segmentation network optimizes the loss function through back-propagation, enabling the aggregation of jointly extracted feature information from medical images. This significantly enhances the segmentation performance of the U-net pathway.

We conducted a series of experiments on two publicly available datasets, and the results unequivocally demonstrate that our proposed method achieves segmentation performance on par with fully supervised methods even when employing a small fraction of labeled data. Furthermore, our approach surpasses the performance of recent state-of-the-art semi-supervised segmentation methods. In addition, we carried out experiments to validate the thoughtful design of our proposed model, and the results provide compelling evidence that our method significantly enhances segmentation performance while concurrently reducing computational overhead. These findings underscore the potential of our approach to substantially streamline clinical practices, thereby reducing the temporal and financial burdens on healthcare professionals and patients.

CRediT authorship contribution statement

Zhenghua Xu: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Hening Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Runhe Yang:** Writing – review & editing, Methodology. **Yuchen Yang:** Writing – review & editing, Visualization. **Weipeng Liu:** Supervision, Funding acquisition. **Thomas Lukasiewicz:** Writing – review & editing, Supervision, Funding acquisition.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grants 62276089 and 62027813, by the Natural Science Foundation of Hebei Province, China, under the grant F2024202064, by the Natural Science Foundation of Tianjin, China, under the grant 24JCJQJC00200, by the Ministry of Human Resources and Social Security, China, under the grant RSTH-2023-135-1, by the S&T Program of Hebei under the grant 225676163GH, and by the AXA Research Fund.

Data availability

The data used are publicly available.

References

[1] Miao Yu, Miaomiao Guo, Shuai Zhang, Yuefu Zhan, Mingkang Zhao, Thomas Lukasiewicz, Zhenghua Xu, RIRGAN: An end-to-end lightweight multi-task learning method for brain MRI super-resolution and denoising, Comput. Biol. Med. 167 (2023) 107632.

[2] Di Yuan, Yunxin Liu, Zhenghua Xu, Yuefu Zhan, Junyang Chen, Thomas Lukasiewicz, Painless and accurate medical image analysis using deep reinforcement learning with task-oriented homogenized automatic pre-processing, Comput. Biol. Med. 153 (2023) 106487.

[3] Zhenghua Xu, Jiaqi Tang, Chang Qi, Dan Yao, Caihua Liu, Yuefu Zhan, Thomas Lukasiewicz, Cross-domain attention-guided generative data augmentation for medical image analysis with limited data, Comput. Biol. Med. 168 (2024) 107744.

[4] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).

[5] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, Sebastian Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.

[6] Zhenghua Xu, Biao Tian, Shijie Liu, Xiangtao Wang, Di Yuan, Junhua Gu, Junyang Chen, Thomas Lukasiewicz, Victor C.M. Leung, Collaborative attention guided multi-scale feature fusion network for medical image segmentation, IEEE Trans. Netw. Sci. Eng. 11 (2) (2024) 1857–1871.

[7] Jiaojiao Zhang, Shuo Zhang, Xiaoqian Shen, Thomas Lukasiewicz, Zhenghua Xu, Multi-ConDoS: Multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation, IEEE Trans. Med. Imaging 43 (1) (2024) 76–95.

[8] Jizong Peng, Guillermo Estrada, Marco Pedersoli, Christian Desrosiers, Deep co-training for semi-supervised image segmentation, Pattern Recognit. 107 (2020) 107269.

[9] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, Ming-Hsuan Yang, Adversarial learning for semi-supervised semantic segmentation, 2018, arXiv preprint arXiv:1802.07934.

[10] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, Ender Konukoglu, Contrastive learning of global and local features for medical image segmentation with limited annotations, Adv. Neural Inf. Process. Syst. 33 (2020) 12546–12558.

[11] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, Zhenghua Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, Med. Image Anal. 83 (2023) 102656.

[12] Xiaokang Chen, Yuhui Yuan, Gang Zeng, Jingdong Wang, Semi-supervised semantic segmentation with cross pseudo supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2613–2622.

[13] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, Shaoting Zhang, Semi-supervised medical image segmentation via cross teaching between cnn and transformer, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 820–833.

[14] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[15] Ziyang Wang, Tianze Li, Jian-Qing Zheng, Baoru Huang, When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 424–441.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[17] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[18] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[19] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, Shi-Min Hu, Segnext: Rethinking convolutional attention design for semantic segmentation, Adv. Neural Inf. Process. Syst. 35 (2022) 1140–1156.

[20] Seonkyeong Seong, Jaewan Choi, Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates, Remote Sens. 13 (16) (2021) 3087.

[21] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[22] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[23] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, Yuyin Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[24] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al., Segvit: Semantic segmentation with plain vision transformers, Adv. Neural Inf. Process. Syst. 35 (2022) 4971–4982.

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[26] Yundong Zhang, Huiye Liu, Qiang Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 14–24.

[27] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[28] Zhenghua Xu, Shijie Liu, Di Yuan, Lei Wang, Junyang Chen, Thomas Lukasiewicz, Zhigang Fu, Rui Zhang, ω-Net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution, Neurocomputing 500 (2022) 177–190.

[29] Zhenghua Xu, Yunxin Liu, Gang Xu, Thomas Lukasiewicz, Self-supervised medical image segmentation using deep reinforced adaptive masking, IEEE Trans. Med. Imaging Early Access (2024) 1–14.

[30] Di Yuan, Zhenghua Xu, Biao Tian, Hening Wang, Yuefu Zhan, Thomas Lukasiewicz, μ-Net: Medical image segmentation using efficient and effective deep supervision, Comput. Biol. Med. 160 (2023) 106963.

[31] Jonathan Long, Evan Shelhamer, Trevor Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[32] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, Jingdong Wang, Hrformer: High-resolution transformer for dense prediction, 2021, arXiv preprint arXiv:2110.09408.

[33] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

[34] Antti Tarvainen, Harri Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst. 30 (2017).

[35] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, Alan Yuille, Deep co-training for semi-supervised image recognition, in: Proceedings of the European Conference on Computer Vision, Eccv, 2018, pp. 135–152.

[36] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, Masashi Sugiyama, Co-teaching: Robust training of deep neural networks with extremely noisy labels, Adv. Neural Inf. Process. Syst. 31 (2018).

[37] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, Andrew Rabinovich, Training deep neural networks on noisy labels with bootstrapping, 2014, arXiv preprint arXiv:1412.6596.

[38] Qihang Yu, Yingda Xia, Lingxi Xie, Elliot K Fishman, Alan L Yuille, Thickened 2D networks for efficient 3D medical image segmentation, 2019, arXiv preprint arXiv:1904.01150.

[39] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11) (2018) 2514–2525.

[40] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[41] Xiangde Luo, SSL4mis, 2020, https://github.com/HiLab-git/SSL4MIS.

[42] Nancy Chinchor, Beth M. Sundheim, MUC-5 evaluation metrics, in: Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993, 1993.

[43] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, Patrick Pérez, Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2517–2526.

[44] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, Pheng-Ann Heng, Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer, 2019, pp. 605–613.

[45] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al., R-drop: Regularized dropout for neural networks, Adv. Neural Inf. Process. Syst. 34 (2021) 10890–10905.

[46] Yassine Ouali, Céline Hudelot, Myriam Tami, Semi-supervised semantic segmentation with cross-consistency training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12674–12684.

[47] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.

[48] Zhenghua Xu, Wenting Xu, Ruizhi Wang, Junyang Chen, Chang Qi, Thomas Lukasiewicz, Hybrid reinforced medical report generation with M-linear attention and repetition penalty, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–15, Early Access.

[49] Maxim Berman, Amal Rannen Triki, Matthew B. Blaschko, The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4413–4421.