Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **10**, Number **2**, May **2025**

Mutual Consistency Dual-stream Tumor Segmentation Network Based on Semi-supervised Training

Yuhuan Zhang, Xiong Zhang, Jinjin Wang, Hong Shangguan, Xueying Cui
Xiao-Jia Wu, Ai-Ping Ning, An-Hong Wang

School of Electronic Information Engineering
Taiyuan University of Science and Technology, Taiyuan 030024, China
S202115110175@stu.tyust.edu.cn, zx@tyust.edu.cn, S202115110165@stu.tyust.edu.cn,
shangguan_hong@tyust.edu.cn, 2004003@tyust.edu.cn, 2005074@tyust.edu.cn,
2006052@tyust.edu.cn, 1997030@tyust.edu.cn

Yanyan Wang*

Shanxi Cancer Hospital
Cancer Hospital of the Chinese Academy of Medical Sciences Shanxi Hospital, Taiyuan 030024, China
13513611493@139.com

*Corresponding author: Yanyan Wang

Received March 9, 2024, revised August 8, 2024, accepted December 30, 2024.

Abstract. *During the prediction of pathological staging and the determination of appropriate treatment plans, accurate detection and segmentation of gastrointestinal stromal tumors (GISTs) and their surrounding tissues play a crucial role. Training deep neural network models for stomach tumor segmentation using limited annotated data poses challenges such as insufficient model training and inaccurate segmentation results. This paper proposes a semi-supervised training-based consistency dual-stream segmentation network. The aim is to leverage different types of data to enhance the accuracy of GIST segmentation. Firstly, a semi-supervised training strategy is employed to expand the amount of data used for training, thereby improving the accuracy of model training. Secondly, a dual-stream decoder is designed to decompose semantic information into edge flow and morphology flow, independently processing edge and morphology information. This enhances the effectiveness of feature extraction and improves tumor segmentation accuracy. Additionally, a mutual consistency loss function is introduced, reducing the differences between multiple outputs during training. This guides different sub-models to generate low-entropy predictions in challenging areas. This supervised approach enables the participation of unlabeled image samples in training. Compared to the current state-of-the-art methods, the proposed approach achieves a 2% to 5% performance improvement across five metrics: Dice coefficient, Intersection over Union (IoU), Precision, Jaccard index, and Recall.*
Keywords: gastrointestinal stromal tumors, semi-supervised training, encoder-decoder, mutual consistency loss function, segmentation

1. **Introduction.** Gastrointestinal stromal tumors (GISTs) have high incidence and mortality rates globally [1]. The occurrence rates in the stomach are approximately 60-65%, in the small intestine 20-25%, in the colon and rectum around 10%, and in the esophagus about 5% [2, 3]. Accurate detection and segmentation of tumors and their surrounding tissues are crucial for predicting pathological staging and determining treatment plans. In

clinical practice, traditional tumor region segmentation methods involve manual annotation and delineation by experienced doctors on abdominal CT images. However, with the increasing detection rate of tumors and the scarcity of medical resources in China, manual annotation cannot meet the demand. Scholars both domestically andinternationally are researching techniques for automatically and accurately segmenting gastrointestinal stromal tumors to address this need.

In recent years, with the development of deep learning, end-to-end trained fully supervised convolutional neural networks have been widely used in medical image target classification and detection. They extract rich features automatically from datasets. Currently, convolutional neural network frameworks for image target segmentation are mostly based on the U-Net [4] network's encoder-decoder structure. This network consists of an encoder and a decoder, which use skip connections to concatenate the two parts in order to fuse low-level positional information with deep semantic information. Inspired by classical ResNet [5], replaces ordinary convolution+ normalization+ReLU activation functions in the encoder and decoder with more powerful residual blocks. Medical images, compared to natural images, have simpler semantics, fixed structures, but lower contrast between different parts. This makes it challenging for U-Net and ResUnet to accurately segment regions with complex boundaries or small target areas, referred to in this paper as challenging regions. Improving the accuracy of segmentation in challenging regions is a major problem faced by current deep learning-based image segmentation networks. Addressing this issue, NU-Net [6] uses a deeper U-Net structure to extract deeper semantic structures and designs a multi-output U-Net module to refine the features extracted by the network. CA-Net [7] introduces a joint attention module composed of spatial attention, channel attention, and scale attention between the encoder and decoder to give higher weight to relevant features. This enhances network performance. MSCA-Net [8] building on CA-Net, adds global attention and short-term attention, and also employs multi-scale attention in skip connections to further improve network accuracy. These methods uniformly input image information such as shape and texture into the network to extract target features, but they face significant errors in the segmentation of target edges.

Additionally, existing convolutional neural networks require training with large datasets for end-to-end training to achieve optimal performance. However, a natural limitation in medical imaging is the difficulty in obtaining a large number of case samples for a specific disease. Training with a small number of samples can lead to overfitting of the network model, resulting in suboptimal outcomes. Scholars worldwide are dedicated to researching the application of semi-supervised learning in deep learning techniques. Semi-supervised learning, which requires only a small amount of labeled data and a large amount of unlabeled data for training, can yield satisfactory results. Unlabeled medical images are relatively easy to obtain. Currently, semi-supervised methods can be broadly categorized into two types. The first type is consistency models based on the smoothness assumption [9–11], where small perturbations in the input should not result in significant deviations in the corresponding output [12]. For instance, Ouali et al. [13] disrupts feature maps in the network through data augmentation while constraining the model output to remain unchanged, achieving consistency in segmentation results. Wang et al. [14] employs semantic direction in feature space for semantic data augmentation and then uses consistency constraints for semi-supervised learning. The second type is entropy minimization methods based on the clustering assumption [15–17], where each class should be compact and have low entropy. Kalluri et al. [18] proposed an entropy module to enable the model to generate low-entropy predictions in unlabeled sets. In the medical imaging field, CLCC [19] slices the original image and feeds the slices along with the original image into the network. Subsequently, the segmentation results from multiple slices are

combined to constrain the segmentation results of the original image, achieving semi-supervised learning. UCMT [20], based on a teacher-student network framework, has the teacher network generate high-quality predicted labels to optimize two student networks. Using cross supervision as a supervisory mechanism promotes mutual learning between the two students, enhancing their consistency.

Existing semi-supervised methods are mostly constrained by the performance of the supervised training network, such as the teacher-student framework. The quality of predictions generated by the teacher network directly influences the performance ceiling of the student network. Moreover, in most semi-supervised tasks, pseudo-labels generated from unlabeled data are not fully utilized [21]. In fact, deep learning models can produce pixel-level uncertain segmentation results. Through cross-learning with these uncertain pseudo-labels, unlabeled data can be effectively leveraged to optimize the network. This further enhances segmentation accuracy in challenging regions through semi-supervised training. Based on these considerations and the distribution characteristics of clinical data specific to gastrointestinal tumor segmentation,In this paper, we proposes a novel semi-supervised training network—the Mutual Consistency Dual-stream Network (MCDS-Net). Firstly, the paper improves the decoder by using the traditional U-Net decoder to handle the morphological information of the target and adds an edge flow decoder to separately process edge information. Secondly, a new gated convolutional layer is designed to extract edge information from the morphological flow, aiding more effective feature extraction for the edge flow. This enables the network to generate more accurate segmentation in challenging areas that are difficult to precisely delineate accurately. Finally, two decoder branches with different upsampling strategies are added. Mutual consistency constraints are applied to the outputs of multiple decoders. This minimizes differences between multiple outputs during training, compelling the model to generate consistent results in challenging regions. The main contributions of this paper can be summarized as follows:

(1) Introduction of a novel mutual consistency dual-stream network, where the morphological flow handling morphological information is retained in a U-Net decoder, and an edge flow decoder is added to parallelly process edge information. Spatial pyramid integration of both pieces of information enhances segmentation accuracy.

(2) Proposal of a novel gated convolutional layer that explicitly reinforces edge information. This layer utilizes deep semantic features to denoise the edge flow while enhancing edge features in deep semantic features. This allows the edge flow to complement fine edges that the morphological flow cannot generate.

(3) Design of a mutual consistency semi-supervised training scheme. Constraints on minimizing entropy are imposed by applying mutual consistency constraints to the outputs of multiple decoders. This achieves complementary features between multiple decoders, enhancing the network's segmentation capability in challenging regions.

2. **Methods.** The paper introduces a mutually consistent dual-stream tumor segmentation framework. It employs dual-stream decoders to extract morphological and edge features separately, achieving feature complementation. By utilizing pseudo-labels generated from unlabeled data and imposing constraints with entropy minimization, more attention is paid to challenging regions in different decoder outputs. The structure of the proposed mutual consistency dual-stream network is illustrated in Figure 1. Initially, the image is input into the encoder to obtain high-level semantic information, which is then fed into multiple dual-stream decoders using different upsampling strategies. To acquire more precise edge and morphological information, each dual-stream decoder processes the encoded image information through two independent edge and morphological flows to obtain edge and morphological information. In the edge flow, a newly designed gated

convolutional layer is used to enhance edge information for generating finer edges. Finally, the two types of information are fused through Atrous Spatial Pyramid Pooling (ASPP) to generate the ultimate segmentation result.

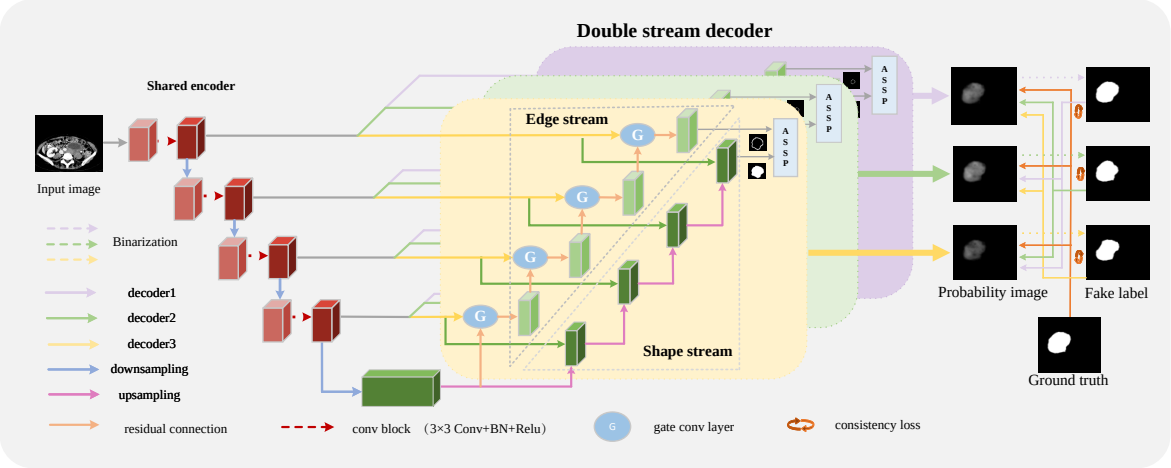

FIGURE 1. Mutual consistency dual-stream segmentation network.

2.1. **Mutual Consistency Dual-stream Segmentation Network.** Currently, medical image segmentation methods based on deep convolutional neural networks typically adopt encoder-decoder structures similar to U-Net. However, this structure lacks clear differentiation between morphological and edge information, as well lacks a loss function constraint for edge information. This results in poor accuracy in the segmentation results, especially in the edge regions. To address this issue, this paper proposes a mutually consistent dual-stream segmentation structure, as illustrated in Figure 1. The structure consists of a shared encoder and multiple dual-stream decoders that employ different upsampling strategies, including transpose convolution layers, linear interpolation layers, and nearest-neighbor interpolation layers. These are used to extract and fuse edge and morphological features, as depicted in the right half of Figure 1. Each decoder is composed of two streams and a dual-stream fusion module. The first stream is the morphological flow, composed of a standard U-Net decoder. The second stream is the edge flow, reinforced by a gated convolutional layer to enhance the boundary features of the tumor. It is constrained using manually annotated edge images, forcing the edge flow to handle only boundary-related information. Ultimately, the fusion of semantic features from the morphological flow and boundary features from the edge flow produces more accurate segmentation results. The following provides a detailed description of the dual-stream decoder and the gated convolutional layer in the framework.

2.1.1. *Dual-stream Decoder.* The module comprises a morphological flow, an edge flow, and a fusion module. The morphological flow takes the features $f \in \mathbb{R}^{C\times\frac{H}{m}\times\frac{W}{m}}$ encoded by the encoder as input and outputs Regular $\in \mathbb{R}^{C\times H\times W}$, and representing the height and width of the input image, and as the downsampling ratio when encoding with the encoder. The structure of ResNet-101 [22] is used as the feature extraction network for the morphological flow. Similarly, the edge flow takes the output f of the encoder as input, producing boundary information edge $\in \mathbb{R}^{H\times W}$. Binary cross-entropy loss is applied to supervise this output. In the fusion module, spatial pyramid pooling is utilized to merge the semantic features from the morphological flow with the boundary features output by the edge flow to retain multi-scale contextual information. The internal structure of

spatial pyramid pooling is shown in the blue part of Figure 1. First, the initial input image is processed using the Canny operator to extract edge features. Then, the extracted edge features are concatenated and convolved with the features in the edge flow. Afterward, they are concatenated with the features extracted by the morphological flow. Feature downsampling is achieved through convolution, batch normalization, and ReLU activation functions. Finally, using a spatial pyramid and convolution, the ultimate segmentation result is generated. The spatial pyramid structure is shown in Figure 2.

2.1.2. *Gated Convolutional Layer.* To more effectively extract the edge information of tumors, a new gated convolutional layer has been designed to assist in processing only edge-relevant information. This helps filter out other interfering information. Gated convolutional layers are applied at each level between the edge flows. Let m represent the number of positions, and $t \in 0, 1, ..., m$ is a running index, where r_t and s_t represent the intermediate representations processed by the gated convolutional layer for the corresponding morphological flow and edge flow. To apply the gated convolutional layer, an attention map $\alpha \in \mathbb{R}^{H \times W}$ is obtained by concatenating r_t and s_t, then normalized with a 1×1 convolution layer $C_{1\times1}$, and finally passed through a Sigmoid activation function:

$$\alpha_t = \sigma\left(C_{1\times1}\left(r_t \| s_t\right)\right) \tag{1}$$

The symbol represents the connection of feature maps. The gated convolutional layer first performs element-wise multiplication between the attention map and the edge flow feature. Subsequently, it adds the resulting product as residual information to the edge features. Finally, a learnable weight is applied to each channel of the output feature. For each pixel, represents the gated convolutional layer. Its computation process is as follows:

$$\begin{aligned} \hat{s}_t^{(i,j)} &= \left(s_t * \omega_t\right)_{(i,j)} \\ &= \left(\left(s_{t_{(i,j)}} \odot \alpha_{t_{(i,j)}}\right) + s_{t_{(i,j)}}\right)^T \omega_t \end{aligned} \tag{2}$$

The resulting is then passed to the next layer of the edge flow for further processing. Here, both attention map computation and gated convolution are differentiable, enabling end-to-end backpropagation. In the experiments, three gated convolutions were used, and connected to each layer of the edge flow.

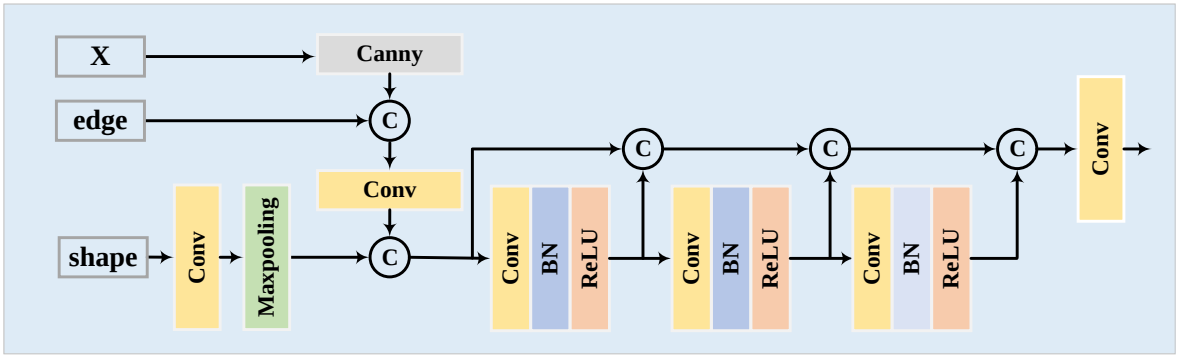

FIGURE 2. Space Pyramid Fusion Module

2.2. **Network Training.** Let $x \in X$ represent the input image, $(y_{\text{pred}} \mid x;\theta)$ the probability map generated by x, where θ represents the parameters of the backbone network f_θ. Let $y_l \in Y_l$ represent the given segmentation label, and the labeled and unlabeled datasets are denoted as $\boldsymbol{D}_L = \{x_l^i, y_l^i \mid i = 1, \cdots, N_l\}$ and $\boldsymbol{D}_U = \{x_u^i, y_u^i \mid i = 1, \cdots, N_u\}$ respectively.

The proposed mutual consistency dual-stream network can enhance the segmentation accuracy in edge regions by minimizing the model's uncertainty. The uncertainty estimation process is defined as:

$$\mu_x = D\left[p\left(y_{\text{pred}} \mid x;\theta_{\text{sub}}^I\right), \cdots, p\left(y_{\text{pred}} \mid x;\theta_{\text{sub}}^n\right)\right] \tag{3}$$

where μ_x is the pixel-level uncertainty, and D represents the statistical discrepancy among n outputs. Utilizing a shared encoder $f_{\theta_{sub}^i}, i \in 1,\cdot\cdot,n$, n sub-models are predefined before uncertainty estimation. Thus, the uncertainty μ_x of x input becomes:

$$f_{\theta_{sub}^i} = f_{\theta_e} \oplus f_{\theta_d^i}, i \in 1, \cdots, n \tag{4}$$

$$\mu_x = D\left[p\left(y_{\text{pred}} \mid x;\theta_{\text{sub}}^I\right), \cdots, p\left(y_{\text{pred}} \mid x;\theta_{\text{sub}}^n\right)\right] \tag{5}$$

The symbol $\oplus$ represents that sub-model $f_{\theta_{sub}^i}$ consists of a shared encoder f_{θ_e} and a decoder $f_{\theta_d^i}$. Each sub-model is a standard encoder-decoder architecture, where the encoder is a standard U-Net network, and the decoder structure is the dual-stream decoder mentioned earlier.

Given that consistency constraints and entropy minimization constraints can effectively utilize unlabeled data, a new mutual consistency training strategy is proposed, using these two constraints to train the mutual consistency dual-stream network.

Specifically, first, transform the output probability map $p\left(y_{\text{pred}} \mid x;\theta\right)$ into a pseudo-label $p^*\left(y_{\text{pred}} \mid x;\theta\right)$. Then, perform mutual learning between the probability output of one decoder and the pseudo-labels of other decoders [23]. This generates n disparate outputs, guiding the model to produce consistent results in challenging areas. The advantages of this design include: (1) enforcing consistency constraints to encourage all sub-models to produce invariant outputs, (2) learning the model to generate low-entropy results as a minimum entropy constraint under pseudo-label supervision, and (3) training the mutual consistency dual-stream network model in an end-to-end manner without multiple forward passes. Finally, the proposed mutual consistency dual-stream network model is trained using the weighted sum of the supervised loss and mutual consistency loss, as follows:

$$L_{mc} = \sum_{i,j=1\&i\neq j}^{n} D\left[p^*\left(y_{\text{pred}}^* \mid x;\theta_{\text{sub}}^i\right), p\left(y_{\text{pred}} \mid x;\theta_{\text{sub}}^j\right)\right] \tag{6}$$

$$Loss = \lambda \times \sum_{i=1}^{n} L_{\text{seg}}\left(p\left(y_{\text{pred}} \mid x;\theta_{\text{sub}}^j\right), y_l\right) + \beta \times L_{mc} \tag{7}$$

where L_{seg} is the commonly used Dice loss for segmentation tasks, D is the mean squared error (MSE) loss for paired inputs, λ and β are two hyperparameters used to balance the supervised loss and mutual consistency loss. Here, L_{mc} is applied to both labeled and unlabeled sets $\boldsymbol{D}_L$ and $\boldsymbol{D}_U$.

3. **Results.**

3.1. **Dataset.** The Liver Tumor Segmentation Challenge (LiTS) dataset [24] and a self-constructed Gastrointestinal Stromal Tumors Segmentation (GISTS) dataset were utilized in this study. The LiTS dataset comprises contrast-enhanced computed tomography (CT) images during the portal venous phase of the abdomen, along with pixel-level annotations for liver and liver tumor boundaries. Each CT volume corresponds to an individual patient. The original dataset consists of 130 cases and 70 cases. To create a new 2D dataset, relevant liver slices were extracted only from publicly available training cases. For each lesion in the training set, slices with at least one tumor and a minimum size of 60 pixels were labeled. The remaining slices were discarded. This resulted in 4,302 affected slices from 112 cases. Ultimately, the input images fed into the network were resized to 512x512.

The GISTS dataset, obtained from Shanxi Cancer Hospital, was collected using Discovery CT750 HD and Siemens Force CT scanners (tube current: 100mAs, tube voltage: 120kV) with a slice thickness of 5mm. Similar to the LiTS dataset, the original data were filtered to create a dataset for network training. The initial dataset comprises 5,019 slices from 74 cases of gastrointestinal stromal tumors. Images containing tumors with a minimum diameter greater than 20 pixels were selected as lesion images, and the rest were discarded. The final training data includes 774 lesion images.

3.2. **Experimental Environment and Parameter Settings.** During the training process, the Adam optimizer with a momentum of 0.5 and weight decay of 0.0001 was employed for 30,000 iterations. All experiments were conducted using PyTorch 1.8 on a 24GB NVIDIA RTX 3090 GP. The initial learning rate was set to 0.0002, and decayed by 0.1 every 100 iterations.

Two typical semi-supervised experimental setups were conducted, utilizing either 30% or 50% of labeled data along with the remaining unlabeled data for training. To assess the performance of the proposed network, it was compared with several state-of-the-art methods, including Unet, ResNuet, NU-Net, CA-Net, and MSCA-Net.

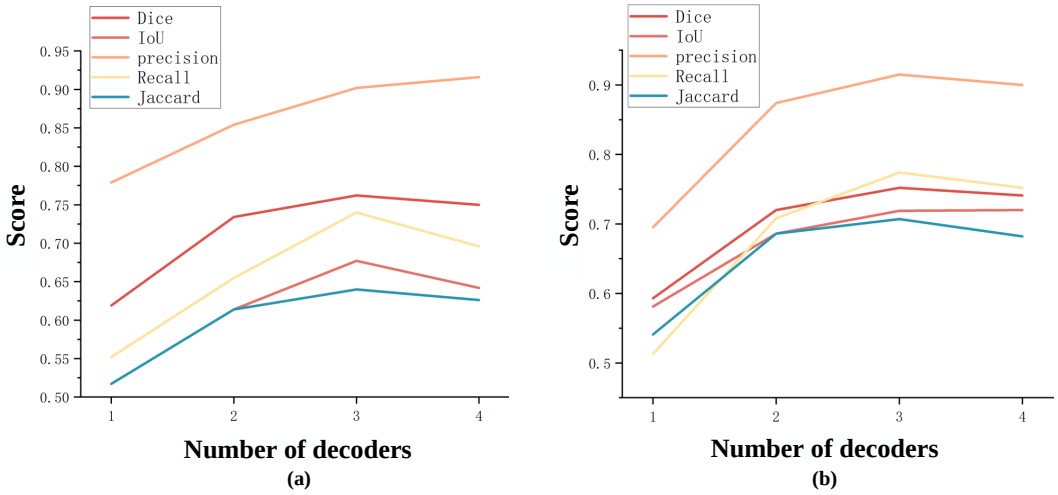

FIGURE 3. The impact of the number of decoders on the final segmentation ((a) is the LiTS dataset, (b) is the GISTS dataset)

3.3. **Evaluating Indicator.** To evaluate the performance of the proposed multi-task semi-supervised medical image segmentation, five widely used evaluation metrics were selected, including Dice coefficient (Dice), Intersection over Union (IoU), Jaccard index

TABLE 1. The segmentation performance of different algorithms using 30% and 50% annotated datasets on the GISTS dataset

	Dice	Iou	Precicion	Recall	Jc	Hd95
Regular	0.739±0.023	0.608±0.025	0.870±0.045	0.673±0.053	0.614±0.0272	8.687±8.839
Edge	0.755±0.028	0.624±0.053	0.895±0.043	0.698±0.056	0.637±0.056	7.686±7.135

(Jaccard), Precision, and Recall. These metrics are widely recognized as effectively reflecting the model's performance in image segmentation tasks. The symbol ↑ indicates that a higher value is better.

3.4. **Ablation Experiment.** To increase model diversity, the Mutual Consistency dual-stream Network model employs transpose convolution layers, linear interpolation layers, and nearest interpolation layers to construct three different decoders. These three decoders produce different outputs during processing, but when using mutual consistency constraints, the three different decoders often generate similar outputs. This characteristic helps reduce fuzzy predictions and decrease model uncertainty, enabling the Mutual Consistency dual-stream Network model to achieve better results. As the three decoders can produce similar results, to reduce the inference cost, the original encoder-decoder architecture, i.e., shared encoder and the first decoder are chosen during testing. Additionally, the number of decoders, denoted as n, is adjustable, and its impact is demonstrated through multiple experimental indicators. Figure 3 (a) and (b) show the influence of different numbers of decoders on performance. When the number of decoders is less than 3, increasing the number of decoders improves performance. However, as the number of decoders increases, the performance improvement gradually decreases, and in some metrics (such as recall), there is a performance decline. This is because too many decoders may lead to overfitting of the network to the data during supervised training. Therefore, to balance effectiveness and efficiency, this paper sets n to 3.

Table 1 illustrates the impact of boundary loss on tumor segmentation performance. "Regular" refers to the baseline network trained only through morphological flow without introducing edge flow. In contrast, "Edge" indicates the simultaneous use of both edge flow and morphological flow. The positive effect of boundary loss is clearly seen across various evaluation metrics. Compared to the baseline network, introducing edge flow to independently process boundary information and morphological information enhances the model's performance. This results in an improvement of around 2% in various metrics.

3.5. **Visual Effect and Quantitative Analysis.** Firstly, the proposed method was validated on the gastrointestinal tumor segmentation dataset. Table 2 displays the performance metrics of different algorithms for segmentation using 30% and 50% labeled datasets. The results indicate that the proposed algorithm exhibits the best performance on datasets with all labeling proportions. Figure 4 visualizes the segmentation results using 30% labeled data. Resunet and Unet algorithms show the poorest segmentation performance on the Gastrointestinal Stromal Tumors Segmentation (GISTS) dataset. They have numerous inaccurately segmented lesion areas, such as difficulty distinguishing areas with similar appearances to the lesion region. In contrast, although the UCMT and CLCC semi-supervised methods are relatively better than the aforementioned two methods, there is still a 2% to 5% gap in various metrics compared to the Mutual Consistency Dual-stream Network. Compared to other methods, the Mutual Consistency Dual-stream Network produces segmentation results that are more similar to manual segmentation. Furthermore, the design of the dual-stream decoder in this paper allows the edge flow to better handle edge information. This enables the final segmentation result

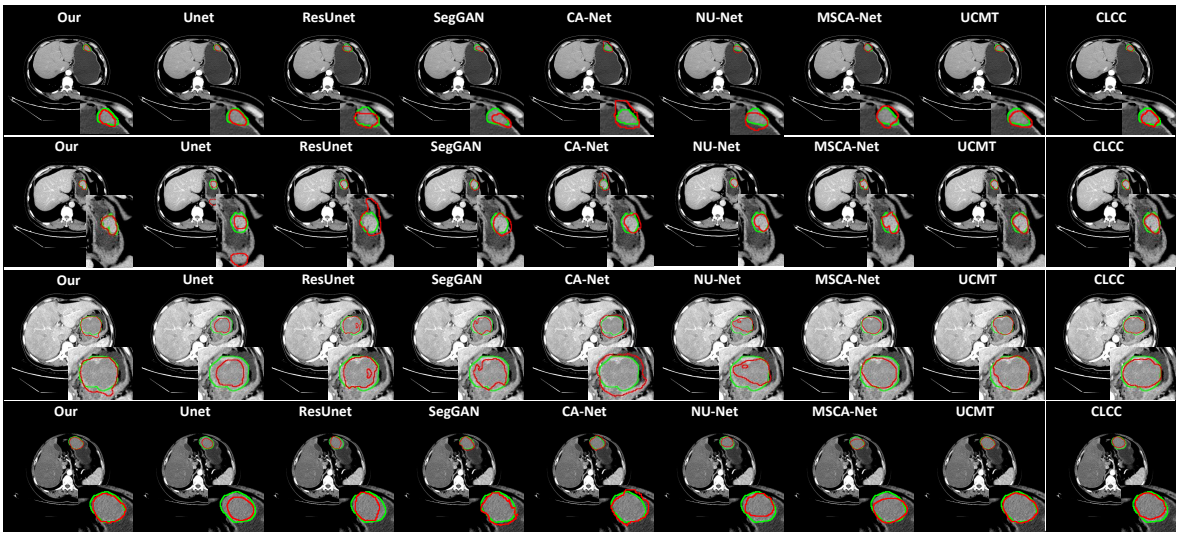

FIGURE 4. Visualization results of segmentation using different algorithms on the GISTS datase

to effectively separate background and target information. This means that even regions with appearances similar to the target can be well distinguished. As shown in the second row of Figure 4, lesions at the edges of organs, where other methods fail to accurately differentiate between the organ wall and the lesion edge, resulting in erroneous boundary predictions. The quantitative analysis results in Table 2 complement the visual results in Figure 4.

Figure 5 and Figure 6 show the qualitative and quantitative analysis of the proposed method on the LiTS dataset. For liver tumor segmentation tasks with fuzzy boundaries, as shown in the fourth row of Figure 5, this paper uses mutual consistency loss constraints and supplements each other with uncertain segmentation results generated by multiple decoders to obtain more accurate segmentation results

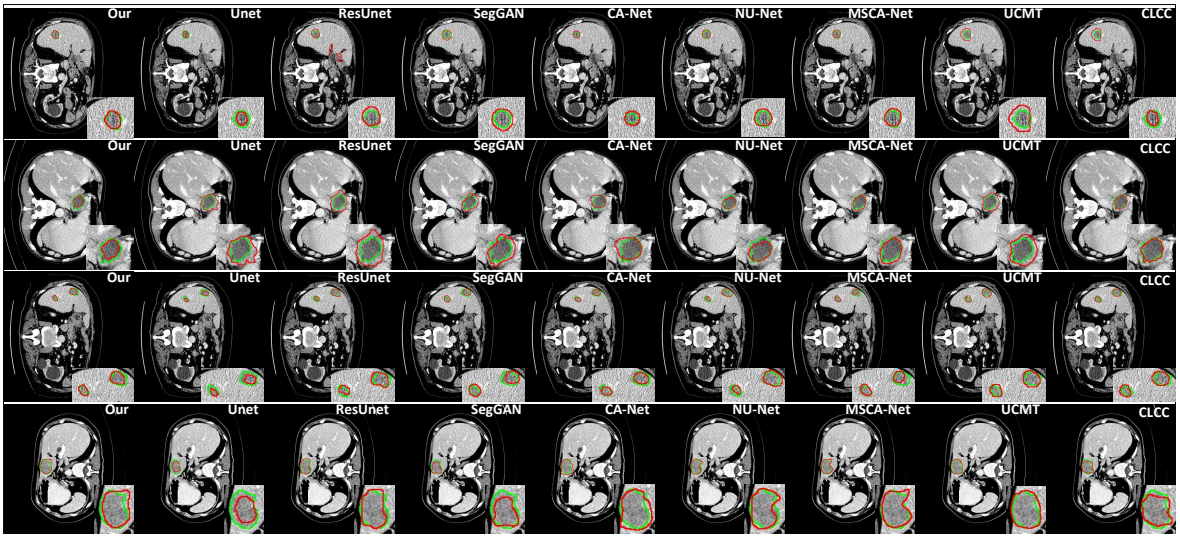

FIGURE 5. Visualization results of segmentation using different algorithms on the GISTS datase

4. **Conclusion.** The paper proposes a mutual consistency dual-stream network structure for semi-supervised medical image segmentation. The model fully leverages the diversity

TABLE 2. Segmentation performance of different algorithms using 30% and 50% annotated datasets on the GISTS dataset

Ratio	Methods	Dice↑	IoU↑	Precision↑	Jaccard↑	Recall↑
30%	Unet(baseline)	0.632±0.272	0.587±0.037	0.649±0.026	0.552±0.037	0.728±0.056
	ResUnet	0.643±0.066	0.626±0.063	0.702±0.014	0.561±0.063	0.686±0.077
	SegAN(2017)	0.672±0.033	0.590±0.027	0.682±0.037	0.528±0.027	0.716±0.047
	CA-Net(2020)	0.738±0.118	0.639±0.090	0.697±0.016	0.545±0.090	0.673±0.118
	NU-Net(2023)	0.681±0.069	0.573±0.052	0.601±0.083	0.516±0.052	0.604±0.078
	MSCA-Net(2023)	0.710±0.183	0.646±0.151	0.705±0.228	0.612±0.151	0.579±0.166
	UCMT(2023)	0.720±0.043	0.664±0.038	0.808±0.041	0.624±0.038	0.737±0.059
	CLCC(2022)	0.746±0.092	0.653±0.080	0.852±0.100	0.610±0.047	0.725±0.106
	ours	0.762±0.025	0.667±0.042	0.902±0.030	0.640±0.035	0.740±0.086
50%	UNet(baseline)	0.734±0.040	0.631±0.043	0.769±0.049	0.651±0.043	0.843±0.053
	ResUnet	0.726±0.066	0.641±0.063	0.754±0.033	0.641±0.063	0.788±0.077
	SegAN2017	0.733±0.107	0.655±0.100	0.809±0.059	0.611±0.061	0.724±0.135
	CA-Net2020	0.753±0.075	0.655±0.098	0.816±0.095	0.613±0.058	0.806±0.086
	NU-Net2023	0.742±0.080	0.656±0.070	0.796±0.087	0.646±0.070	0.814±0.095
	MSCA-Net2023	0.762±0.073	0.670±0.086	0.819±0.119	0.663±0.050	0.836±0.094
	UCMT	0.728±0.038	0.675±0.034	0.832±0.038	0.690±0.034	0.842±0.046
	CLCC	0.763±0.105	0.653±0.094	0.813±0.104	0.690±0.046	0.833±0.120
	ours	0.783±0.026	0.701±0.026	0.959±0.083	0.715±0.026	0.850±0.093

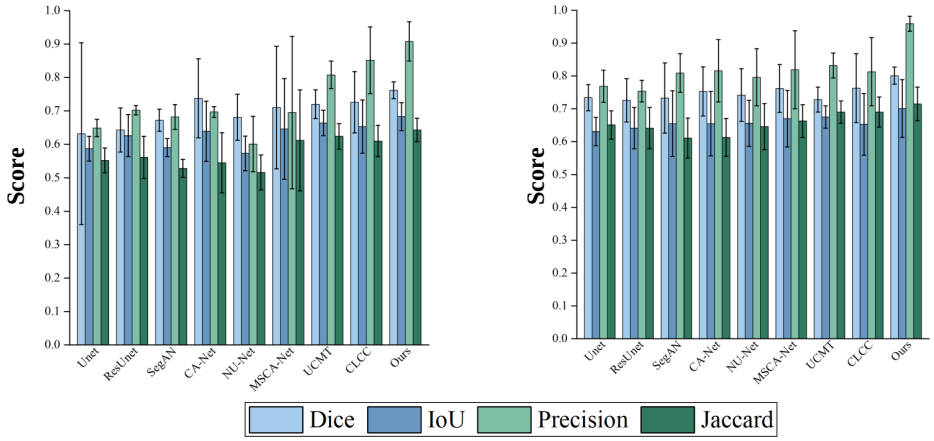

FIGURE 6. Visualization results of segmentation using different algorithms on the GISTS datase

of segmentation results in challenging regions from different sub-models, implementing a semi-supervised training mechanism through cross constraints. This allows for the effective utilization of training data, achieving superior segmentation results for lesion areas. In terms of model design, a dual-stream decoder structure is employed to independently process shape and edge information. This aids in generating finer predictions around the boundaries of lesion targets. Additionally, a loss function with mutual consistency constraints is designed. By associating the probability outputs of the three decoders with the pseudo-labels they generate, different sub-models are encouraged to produce low-entropy predictions in challenging regions. This further improves prediction accuracy. Experiments indicate that compared to eight existing models, the mutual consistency dual-stream network achieves a 2% to 5% improvement across five performance metrics

on two medical datasets. Using only 30% annotated data on a self-constructed dataset, the values for Dice, IoU, Precision, Jaccard, and Recall are 0.762, 0.667, 0.902, 0.640, and 0.740, respectively. Future research could further explore the network's multitasking capabilities, enabling it to simultaneously output segmentation and classification results for tumors. This would achieve automation in medical diagnosis.

Acknowledgment. This work is supported by the National Youth Science Foundation 62001321, National Natural Science Foundation of China 82171883, Shanxi Province Basic Research Program Project 202103021224265 and Shanxi Provincial Natural Science Foundation 20230302121144

REFERENCES

[1] T. Nishida, J. Y. Blay, S. Hirota, Y. Kitagawa, and Y. K. Kang, "The standard diagnosis, treatment, and follow-up of gastrointestinal stromal tumors based on guidelines," *Gastric Cancer*, vol. 19, no. 1, pp. 3–14, 2015.

[2] O. M. J. A. V. A. V. R. Soreide, KjetilSandvik, "Global epidemiology of gastrointestinal stromal tumours (gist): A systematic review of population-based cohort studies," *Cancer Epidemiology*, vol. 40, pp. 39–46, 2016.

[3] M. Mehren, R. Randall, R. Benjamin, S. G. Boles, M. Bui, E. Casper, E. Conrad, T. Delaney, K. Ganjoo, and S. George, "Gastrointestinal stromal tumors, version 2.2014: Featured updates to the nccn guidelines," *Journal of The National Comprehensive Cancer Network*, vol. 12, pp. 853–862, 2014.

[4] O. Ronneberger, P. Fischer, and T. 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.

[5] F. I. D. A, F. W. B, P. C. A, and C. W. C, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data - sciencedirect," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[6] G. Chen, L. Li, J. Zhang, and Y. Dai, "Rethinking the unpretentious u-net for medical ultrasound image segmentation," *Pattern Recognition*, vol. 142, p. 109728, 2022.

[7] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 699–711, 2020.

[8] Y. Sun, D. Dai, Q. Zhang, Y. Wang, S. Xu, and C. Lian, "Msca-net: Multi-scale contextual attention network for skin lesion segmentation," *Pattern Recognition*, vol. 139, p. 109524, 2023.

[9] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8801–8809, 2021.

[10] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 318–329.

[11] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*, pp. 605–613, 2019.

[12] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[13] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12 674–12 684, 2020.

[14] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, and Y. Wang, "Tripled-uncertainty guided mean teacher model for semi-supervised 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 II 24. Springer, 2021, pp. 450–460.

[15] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[16] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 557–11 568.

[17] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv preprint arXiv:2101.06329*, 2021.

[18] T. Kalluri, G. Varma, M. Chandraker, and C. Jawahar, "Universal semi-supervised semantic segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Cision*, pp. 5259–5270, 2019.

[19] X. Zhao, C. Fang, D.-J. Fan, X. Lin, F. Gao, and G. Li, "Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation," *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2022.

[20] Z. Shen, P. Cao, H. Yang, X. Liu, J. Yang, and O. R. Zaiane, "Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation," *arXiv preprint arXiv:2301.04465*, 2023.

[21] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *International Journal of Computer Vision*, vol. 130, no. 6, pp. 1526–1565, 2022.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[23] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4320–4328.

[24] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.