TCCNet: Temporally Consistent Context-Free Network for Semi-supervised Video Polyp Segmentation

Xiaotong Li[1] , **Jilan Xu**[1] , **Yuejie Zhang**[1,✉] , **Rui Feng**[1] , **Rui-Wei Zhao**[2] , **Tao Zhang**[3,✉] , **Xuequan Lu**[4] and **Shang Gao**[4]

[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University

[2]Academy for Engineering and Technology, Fudan University

[3]School of Information Management and Engineering, Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics

[4]School of Information Technology, Deakin University

{xtli20, jilanxu18, yjzhang, fengrui, rwzhao}@fudan.deu.cn, taozhang@mail.shufe.edu.cn, {xuequan.lu, shang.gao}@deakin.edu.au

Abstract

Automatic Video Polyp Segmentation (VPS) is highly valued for the early diagnosis of colorectal cancer. However, existing methods are limited in three respects: 1) most of them work on static images, while ignoring the temporal information in consecutive video frames; 2) all of them are fully supervised and easily overfit in presence of limited annotations; 3) the context of polyp (i.e., lumen, specularity and mucosa tissue) varies in an endoscopic clip, which may affect the predictions of adjacent frames. To resolve these challenges, we propose a novel Temporally Consistent Context-Free Network (TCCNet) for semi-supervised VPS. It contains a segmentation branch and a propagation branch with a co-training scheme to supervise the predictions of unlabeled image. To maintain the temporal consistency of predictions, we design a Sequence-Corrected Reverse Attention module and a Propagation-Corrected Reverse Attention module. A Context-Free Loss is also proposed to mitigate the impact of varying contexts. Extensive experiments show that even trained under 1/15 label ratio, TCCNet is comparable to the state-of-the-art fully supervised methods for VPS. Also, TCCNet surpasses existing semi-supervised methods for natural image and other medical image segmentation tasks.

1 Introduction

Automatic video endoscopic polyp segmentation is of great research value for the prevention of Colorectal Cancer (CRC). However, most previous methods are fully supervised and trained on static images. They have three major limitations. Firstly, recent methods like [Fan *et al.*, 2020; Kim *et al.*, 2021] only relied on static images to train and evaluate their models, while ignoring the temporal information in an endoscopic sequence. As shown in Fig. 1, the images are from the same endoscopic sequence and focus on the same polyp object. Their trajectories and appearance changes are temporally correlated. It is insufficient to only focus on independent static images for the Video Polyp Segmentation (VPS) task. Secondly, limited annotated data is the bottleneck for the VPS task. Due to the fuzzy boundary of polyp and its similarity to the background tissue, even skilled clinicians may fail to reach an agreement on the annotations for successive frames, as shown in the first row of Fig. 1. Thirdly, although an endoscopic video focuses on the same polyp tissue, the context of the polyp (i.e., lumen, specularity and mucosa tissue) changes due to angles of camera or lights, which may impact on the prediction results of adjacent frames. E.g., in the second row of Fig. 1, the contexts of the 2^{nd} and 3^{rd} frames are different from that of the 1^{st} one.

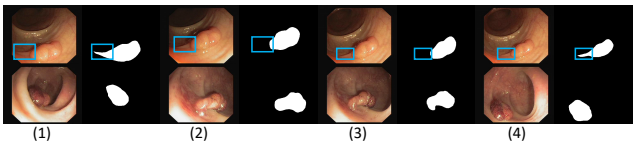

Figure 1: Clips from two endoscopic videos. In the first row, disagreements among successive annotations are bounded by blue squares. The same tissue in blue square is labeled as polyp in the 1^{st} frame, but labeled as background in the 2^{nd} and 3^{rd} frames.

To tackle the aforementioned limitations, we propose a novel Temporally Consistent Context-Free Network (TCCNet) for the semi-supervised VPS task, which contains a segmentation branch and a propagation branch. For *the first limitation*, we design a Sequence-Corrected Reverse Attention (SC-RA) module and a Propagation-Corrected Reverse Attention (PC-RA) module to fully exploit the temporal information and keep the prediction temporally consistent among successive frames. Reverse Attention [Fan *et al.*, 2020] is employed to refine the boundary of saliency map. To correct the map's prediction error, we introduce an error correction

mechanism. Positional information [Hu *et al.*, 2021] is obtained from the endoscopic sequence and effectively prevents a suspected area from being misclassified. For *the second limitation*, we design a co-training scheme to train the network in a semi-supervised manner. A sequence clip with one labeled reference frame and several unlabeled frames is fed into the two parallel branches. For the unlabeled frames, the outputs of one branch are supervised by the pseudo labels [Chen *et al.*, 2021] generated by the other branch. Different from existing semi-supervised methods, we focus on the temporal consistency between the ground truth of the reference frame and the pseudo labels of the unlabeled frames. For *the third limitation*, we propose a Context-Free Loss to mitigate the impact of varying contexts within successive frames.

The contributions are three-fold: (i) We propose a novel TCCNet for the semi-supervised VPS. To the best of our knowledge, this is the first work to study the VPS task using semi-supervised learning. (ii) We design the SC-RA and PC-RA modules to keep predictions temporally consistent and introduce the Context-Free Loss to alleviate the impact of varying contexts. (iii) We conduct experiments on three VPS datasets. Results show that even trained under 1/15 label ratio, TCCNet is comparable to the state-of-the-art fully supervised methods for VPS. Specially, TCCNet exhibits obvious superiority over the existing semi-supervised methods for natural image and other medical image segmentation. [1]

2 Related Work

Polyp Segmentation. Deep Convolutional Neural Networks (DCNN) are widely used in polyp segmentation. U-Net++ [Zhou *et al.*, 2018] designed dense connections and enabled feature fusion of varying scales. ResUNet++ [Jha *et al.*, 2019] merged residual blocks, attention blocks, Atrous Spatial Pyramidal Pooling and Squeeze-and-Excitation (SE) blocks into a U-shaped architecture. PraNet [Fan *et al.*, 2020] fused features from varying scales with a Parallel Partial Decoder (PPD) [Chen *et al.*, 2018] and recovered the boundary cues with a Reverse Attention (RA) module. Inspired by the overall structure of PraNet, UACANet [Kim *et al.*, 2021] integrated the features from the foreground, background and uncertain area to refine the boundary cues. However, these methods only work on static images and cannot capture the temporal information in the endoscopic videos. PNSNet [Ji *et al.*, 2021] proposed a Normalized Self-attention (NS) block to dynamically update the receptive field of the network and obtain the temporal representation. Limited by the scale of video polyp datasets, PNSNet was pre-trained on static images and fine-tuned on video polyp images. Such a training strategy requires abundant annotations and does not suit the semi-supervised segmentation. By contrast, our semi-supervised framework fully exploits the temporal information. It keeps the prediction temporally consistent and breaks the bottleneck of limited annotations.

Semi-supervised Medical Image Segmentation. Semi-supervised Learning (SSL) improves the performance of medical image segmentation by utilizing a large set of unlabeled data. Adversarial learning [Zhang *et al.*, 2017] and consistency regularization [Tarvainen and Valpola, 2017] are the two most common semi-supervised methods. To avoid the inefficient perturbation, ICT [Verma *et al.*, 2019] introduced a consistency strategy which encouraged the consistency between the prediction of two unlabeled images' interpolation and the interpolation of those two images' predictions. UA-MT [Yu *et al.*, 2019] took the reliability of predictions into consideration and designed a consistency loss with the guidance of an estimated uncertainty. Apart from the input perturbation, feature perturbation [Ouali *et al.*, 2020] and model perturbation [Chen *et al.*, 2021] are also used for consistency regularization. URPC [Luo *et al.*, 2021] presented an uncertainty rectified pyramid consistency loss which encouraged the prediction at different scales to be consistent. In our framework, we introduce the model perturbation [Chen *et al.*, 2021] method into the VPS task and preserve the temporal consistency between the ground truth of the reference frame and the pseudo label of the unlabeled frame.

[1] Code is available at https://github.com/wener-yung/TCCNet

3 Methodology

3.1 Temporally Consistent Context-Free Network

Fig. 2(a) shows the overall architecture of the proposed TCCNet. It consists of a segmentation branch and a propagation branch. In the training phase, a training clip is parallelly fed into the two branches. The SC-RA module in the segmentation branch and the PC-RA module in the propagation branch are designed to correct the error of saliency maps from the previous layer and keep the temporal consistency of the predictions. The outputs of these two branches are supervised with a co-training scheme. From each input image, we synthesize two different images S_1 and S_2 with the predicted polyp locations, which are fed into the two branches again to obtain the global maps and calculate the CFLoss in Sec. 3.4.

The encoder structures of the two branches are the same. For a given endoscopic clip with T frames, $\{I_t\}_{t=1}^{T}$, five levels of features $\{F_t^l\}_{t=1,l=1}^{T,5} \in \mathbb{R}^{H^l \times W^l \times C}$ are extracted from the encoder. For fair comparison, we use the same backbone in PraNet [Fan *et al.*, 2020] as the encoder (i.e., Res2Net pre-trained on ImageNet). Following [Fan *et al.*, 2020], we integrate the high-level features to obtain the global features $\{F_t^6\}_{t=1}^{T}$ and the global maps $\{M_t^6\}_{t=1}^{T}$. To improve the quality of feature embedding and reduce computing resources, we update the encoder weights in the propagation branch as an Exponential Moving Average (EMA) [Tarvainen and Valpola, 2017] of the weights in the segmentation branch.

The decoder structures of these two branches are different, so are the prediction processes. Specifically, in the segmentation branch, taking the l^{th} layer as an example, $\{F_{s,t}^l\}_{t=1}^{T}$, $\{F_{s,t}^{l+1}\}_{t=1}^{T}$ and $\{M_{s,t}^{l+1}\}_{t=1}^{T}$ are sent to the SC-RA module to obtain $\{M_{s,t}^l\}_{t=1}^{T}$. The predictions of the T frames are calculated as a whole. While in the propagation branch, the predictions are calculated frame by frame. Taking the t^{th} frame as an example, the feature maps of the previous $t-1$ frames are stored as memory features in a memory pool. In the l^{th} layer, the memory features, $F_{p,t}^l$ and $M_{p,t}^{l+1}$ are sent to the PC-RA module, and $M_{p,t}^l$ is therefore obtained.

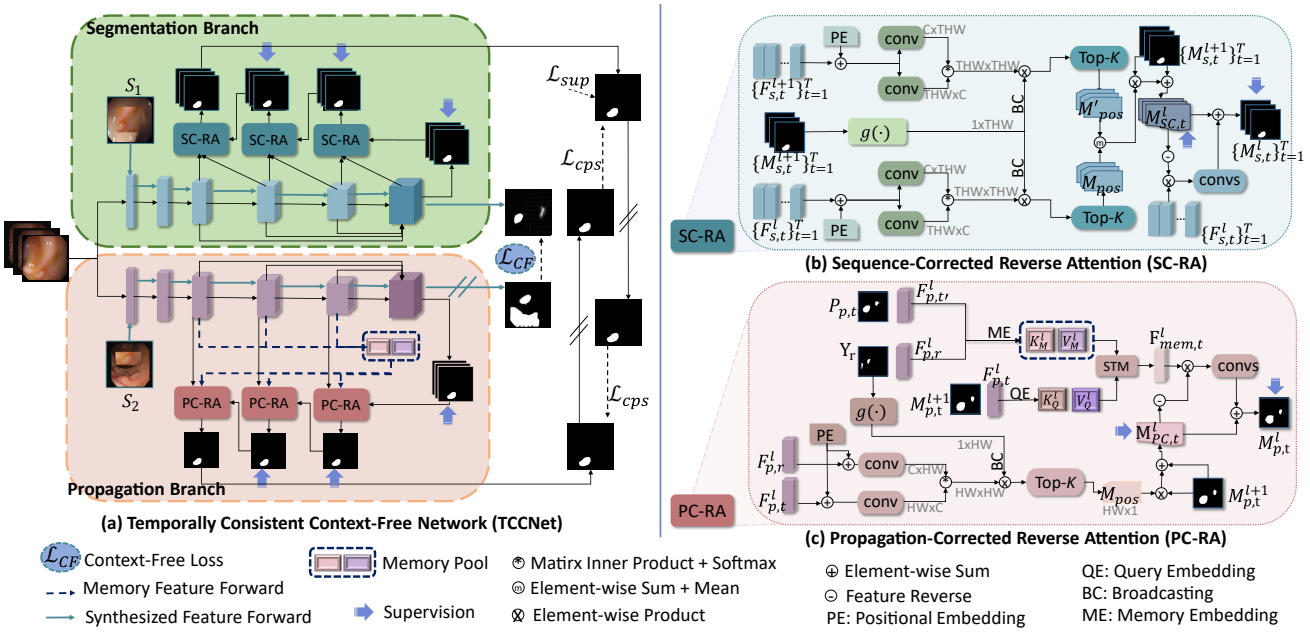

Figure 2: (a) An overall architecture of TCCNet. The approach to synthesizing S_1 and S_2 is described in Sec. 3.4. (b) The detail of SC-RA module. (c) The detail of PC-RA module. '//' on $\longrightarrow$ denotes stop-gradient.

TCCNet focuses on the VPS task with a semi-supervised method, where a training clip $\{I_t\}_{t=1}^{T}$ consists of one labeled frame and $T-1$ unlabeled frames. We denote the labeled frame as the reference frame (I_r, Y_r) and the unlabeled frames as $\mathcal{D}_u = \{I_t\}_{t=2}^{T}$. The training clip is sent to the two parallel branches. The predictions of segmentation branch are denoted by $\{P_{s,t}\}_{t=1}^{T}$ and the predictions of propagation branch are denoted by $\{P_{p,t}\}_{t=2}^{T}$. It should be highlighted that the segmentation branch predicts the entire clip and outputs T masks, while the propagation branch outputs $T-1$ masks of the unlabeled frames, excluding the mask of the reference frame.

The supervised loss on the reference frame is formulated using the Binary Cross Entropy Loss (BCELoss) and the Intersection over Union Loss (IoULoss). It is calculated as:

$$\mathcal{L}_{sup}(P_{s,r}, Y_r) = \mathcal{L}_{BCE}(P_{s,r}, Y_r) + \mathcal{L}_{IoU}(P_{s,r}, Y_r), \quad (1)$$

where $\mathcal{L}_{BCE}(\cdot)$ and $\mathcal{L}_{IoU}(\cdot)$ denote the BCELoss and IoULoss functions, respectively.

The unsupervised loss on the unlabeled frames is defined using the Cross Pseudo Supervision Loss (CPSLoss) [Chen *et al.*, 2021]. We get the pseudo segmentation labels $\{Y'_{s,t}\}_{t=2}^{T}$ and $\{Y'_{p,t}\}_{t=2}^{T}$ by thresholding the corresponding predictions. The CPSLoss is bidirectional. Specifically, the predictions $\{P_{s,t}\}_{t=2}^{T}$ of the segmentation branch are supervised by the pseudo labels $\{Y'_{p,t}\}_{t=2}^{T}$ generated from the propagation branch and vice versa. It is calculated as:

$$\mathcal{L}_{cps} = \frac{1}{T-1}\sum_{t=2}^{T}(\mathcal{L}_{BCE}(P_{s,t}, Y'_{p,t}) + \mathcal{L}_{BCE}(P_{p,t}, Y'_{s,t})). \quad (2)$$

3.2 Sequence-Corrected Reverse Attention

The SC-RA module is described in Fig. 2(b). Taking the l^{th} layer as an example, the SC-RA module receives feature maps $\{F^l_{s,t}\}_{t=1}^{T}$, $\{F^{l+1}_{s,t}\}_{t=1}^{T}$ and segmentation maps $\{M^{l+1}_{s,t}\}_{t=1}^{T}$ to calculate the sequence-corrected position maps $\{M_{pos,t}\}_{t=1}^{T}$, which are used to correct $\{M^{l+1}_{s,t}\}_{t=1}^{T}$. $\{M_{pos,t}\}_{t=1}^{T}$ is the average of M_{pos} and M'_{pos}. Specifically, we use $\{F^l_{s,t}\}_{t=1}^{T}$ to compute the position map M_{pos} and $\{F^{l+1}_{s,t}\}_{t=1}^{T}$ to compute M'_{pos}. The calculation processes of M_{pos} and M'_{pos} are similar. Taking M_{pos} as an example, we complement the feature maps $\{F^l_{s,t}\}_{t=1}^{T}$ with 2D positional encoding [Carion *et al.*, 2020] and calculate the query vector Q and key vector K with the $1\times1\times1$ convolutional layer. We reshape Q and K, and apply matrix inner product to obtain the similarity map Sim.

The segmentation maps $\{M^{l+1}_{s,t}\}_{t=1}^{T}$ from the previous layer are used to reduce the response of background region. We get the local response by:

$$\{g^l\}_{t=1}^{T} = \{\exp(\sigma(\mathcal{U}(M^{l+1}_{s,t})))/e\}_{t=1}^{T}, \quad (3)$$

where σ is the sigmoid function; and $\mathcal{U}(\cdot)$ is the up-sampling operation. g is then reshaped to $g'^l \in \mathbb{R}^{1\times TH^lW^l}$.

We apply element-wise product between Sim and g' to suppress the response of non-polyp region. Then we select the top-K values on the key dimension and average them to get position map M^l_{pos}.

The sequence-corrected position maps $\{M_{pos,t}\}_{t=1}^{T}$ are used to compute the sequence-corrected segmentation maps $\{M^l_{SC,t}\}_{t=1}^{T}$ by:

$$\{M^l_{SC,t}\}_{t=1}^{T} = \{M_{pos,t} * \sigma(\mathcal{U}(M^{l+1}_{s,t})) + \sigma(\mathcal{U}(M^{l+1}_{s,t}))\}_{t=1}^{T}. \quad (4)$$

The current segmentation maps are then calculated using the Reverse Attention [Fan *et al.*, 2020],

$$\{M^l_{s,t}\}_{t=1}^{T} = \{convs(\ominus M_{SC,t} * F^l_{s,t}) + M^l_{SC,t}\}_{t=1}^{T}, \quad (5)$$

where $convs(\cdot)$ denotes the convolution layers.

We adopt supervision on the segmentation predictions from the middle layers and the sequence-corrected segmentation maps. The calculation is:

$$\begin{aligned}\mathcal{L}^s_{deep} = &\sum_{l=4}^{6}\mathcal{L}_{sup}(P^l_{s,r}, Y_r) + \sum_{l=3}^{5}\mathcal{L}_{sup}(P^l_{SC,r}, Y_r) \\ &+ \frac{1}{T-1}\sum_{t=2}^{T}(\sum_{l=4}^{6}\mathcal{L}_{BCE}(P^l_{s,t}, Y'_{p,t}) + \sum_{l=3}^{5}\mathcal{L}_{BCE}(P^l_{SC,t}, Y'_{p,t})),\end{aligned} \tag{6}$$

where $P_{SC} = \sigma(\mathcal{U}(M_{SC}))$ and $P_s = \sigma(\mathcal{U}(M_s))$.

3.3 Propagation-Corrected Reverse Attention

The PC-RA module is shown in Fig. 2(c). The Space-Time Memory (STM) [Oh *et al.*, 2019] module aims to utilize all the available temporal cues. For the t^{th} frame, the feature map $F^l_{p,t}$ along with the segmentation map $M^{l+1}_{p,t}$ are embedded as a pair of 2D key and value maps, which are defined as:

$$\begin{aligned}K^l_{Q,t} \in \mathbb{R}^{H^l\times W^l\times C/8} &= \phi_q(F^l_{p,t} + con_p(\sigma(\mathcal{U}(M^{l+1}_{p,t})))), \\ V^l_{Q,t} \in \mathbb{R}^{H^l\times W^l\times C/2} &= g_q(F^l_{p,t} + con_p(\sigma(\mathcal{U}(M^{l+1}_{p,t})))),\end{aligned} \tag{7}$$

where $\phi_q(\cdot)$ and $g_q(\cdot)$ are two parallel 3×3 convolutional layers, and $con_p(\cdot)$ is a 7×7 convolutional layer.

Each of the previous frame is independently embedded into a pair of key and value maps by following the above strategy. We concatenate them in the temporal dimension and store them in the memory pool. $K^l_M \in \mathbb{R}^{T'\times H^l\times W^l\times C/8}$ and $V^l_M \in \mathbb{R}^{T'\times H^l\times W^l\times C/2}$ are obtained, where T' is the number of memory frames. The memory embedding and the current feature embedding are fed into the STM module to obtain a memory map $F^l_{mem,t}$.

The formulation of the propagation-corrected map $M_{PC,t}$ is similar to that of the sequence-corrected map in Sec. 3.2. Similarity map Sim is calculated with the feature map of the t^{th} frame and that of the reference frame. The ground truth of the reference frame Y_r is used to calculate the local response by $g_r = \exp(Y_r)/e$. We then get the position map $M_{pos,t}$ by applying element-wise product between Sim and g_r, selecting top-K values on the dimension and reshaping operation.

The propagation-corrected segmentation map $M^l_{PC,t}$ is then calculated similarly to the sequence-corrected segmentation map in Eq. (4). We apply the Reverse Attention on $F^l_{mem,t}$ and $M^l_{PC,t}$ to get the segmentation map $M^l_{p,t}$ by Eq. (5). Similar to Sec. 3.2, we adopt the supervision upon the segmentation maps. The loss function is defined as:

$$\begin{aligned}\mathcal{L}^p_{deep} = &\frac{1}{T-1}\sum_{t=2}^{T}(\sum_{l=4}^{6}\mathcal{L}_{BCE}(\sigma(\mathcal{U}(M^l_{p,t})), Y'_{s,t}) \\ &+ \sum_{l=3}^{5}\mathcal{L}_{BCE}(\sigma(\mathcal{U}(M^l_{PC,t})), Y'_{s,t})).\end{aligned} \tag{8}$$

3.4 Context-Free Loss

In an endoscopic video, the context around one polyp may vary due to angles of camera or lights, which may affect the prediction results. For our semi-supervised VPS task, the network may overfit on the limited labeled data and context. Therefore, we propose a Context-Free Loss (CFloss) to reduce the dependency on varying contexts and improve the robustness of the network.

For a training clip, we obtain the locations of polyps by applying average, erosion and dilation operations on $\{P_{s,t}\}^T_{t=2}$ and $\{P_{p,t}\}^T_{t=2}$. For each frame in this sequence, two patches are randomly cropped with an overlapping region, which must contain the polyp tissue. To further increase the diversity of context, we randomly sample two frames from two different sequences as the background images and synthesize two images by overlaying the two patches onto the background images, thus obtaining two synthesized images $S_1$and S_2.

The synthesized images are fed into the segmentation branch and the propagation branch, and two global maps are obtained. Denoting the two segmentation maps of the overlapping region as $\Omega_{s,1}$ and $\Omega_{p,2}$, our CFLoss is bidirectional and formulated as:

$$\mathcal{L}_{CF} = \frac{1}{2}\sum_{i\in\Omega}(|\omega_{s,1,i} - \omega_{p,2,i}|^2 + |\omega_{s,2,i} - \omega_{p,1,i}|^2), \tag{9}$$

where $i \in \Omega$ indicates a pixel in the overlapping region.

3.5 Training Strategy

Pre-training on pseudo sequences. We generate a pseudo sequence dataset with static frames to pre-train the network, which is usually used in Video Object Segmentation (VOS) task [Oh *et al.*, 2019; Hu *et al.*, 2021]. For a training clip, the first frame is sampled from the labeled frames and the rest are generated by applying random affine transforms on the first frame, such as translation, zooming, cropping, flip and rotation. In the pre-training phase, only the labeled frames are used and our network is fully supervised.

Main-training on real sequences. In the main-training phase, our network is semi-supervised. For a training clip, the first frame is sampled from the labeled frames and works as the reference frame. The rest unlabeled frames are randomly selected from the same sequence in temporal order.

Total loss. The total loss of the network is formulated as:

$$\begin{aligned}\mathcal{L}_{total} = &\mathcal{L}_{sup} + \lambda_{cps}\mathcal{L}_{cps} + \lambda_s\mathcal{L}^s_{deep} \\ &+ \lambda_p\mathcal{L}^p_{deep} + \lambda_{cf}\mathcal{L}_{CF},\end{aligned} \tag{10}$$

where λ_{cps}, λ_s, λ_p and λ_{cf} are the hyper-parameters to balance the loss terms.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on three video-based polyp datasets: CVC-300 [Tajbakhsh *et al.*, 2016], CVC-612 [Bernal *et al.*, 2015] and ETIS [Silva *et al.*, 2014]. CVC-300/CVC-612/ETIS contains 300/612/196 images from 13/29/26 video clips. The image resolutions are 384×288, 500×574 and 1225×966 pixels, respectively. Following [Ji *et al.*, 2021], 60% images from CVC-612 and 60% images from

	Polyp Segmentation Method	mDice ↑	mIoU ↑	wFm ↑	Sm ↑	Em ↑	MAE ↓
CVC-300-TV	U-Net++	0.743	0.649	0.733	0.860	0.868	0.018
	ResUNet++	0.473	0.386	0.459	0.718	0.678	0.034
	PraNet	0.826	0.734	0.811	0.904	0.927	0.013
	PNSNet	0.813	0.747	0.685	0.866	0.877	0.056
	UACANet	0.829	0.734	0.818	0.905	0.935	**0.012**
	Ours (1/15 labeled)	0.824	0.730	0.812	0.906	0.925	0.013
	Ours (1/2 labeled)	**0.848**	**0.760**	**0.832**	**0.920**	**0.943**	0.013
CVC-612-V	U-Net++	0.674	0.582	0.642	0.809	0.818	0.031
	ResUNet++	0.512	0.400	0.441	0.703	0.720	0.061
	Pranet	0.846	0.767	0.838	0.909	0.934	0.014
	PNSNet	0.844	0.764	0.830	0.912	0.933	0.012
	UACANet	0.840	0.765	0.830	0.909	0.929	0.012
	Ours (1/15 labeled)	0.854	0.777	0.844	0.916	**0.943**	0.012
	Ours (1/2 labeled)	**0.854**	**0.780**	**0.846**	**0.918**	0.938	**0.011**
CVC-612-T	U-Net++	0.739	0.655	0.734	0.834	0.831	0.056
	ResUNet++	0.591	0.494	0.559	0.734	0.766	0.074
	Pranet	0.837	0.762	0.829	0.895	0.900	0.039
	PNSNet	0.820	0.747	0.817	0.895	0.888	0.043
	UACANet	0.826	0.754	0.822	0.890	0.891	0.041
	Ours (1/15 labeled)	0.827	0.752	0.823	0.892	0.893	0.041
	Ours (1/2 labeled)	**0.843**	**0.772**	**0.838**	**0.902**	**0.906**	**0.038**
ETIS	U-Net++	0.309	0.251	0.293	0.624	0.571	0.045
	ResUNet++	0.136	0.106	0.134	0.527	0.463	0.057
	Pranet	0.585	0.509	0.560	0.778	0.759	**0.020**
	PNSNet	0.500	0.424	0.492	0.735	0.699	0.027
	UACANet	0.547	0.480	0.550	0.764	0.710	0.021
	Ours (1/15 labeled)	0.618	0.537	0.587	0.798	0.786	0.023
	Ours (1/2 labeled)	**0.641**	**0.550**	**0.597**	**0.806**	**0.813**	0.026

Table 1: Comparison with polyp segmentation methods. TCCNet achieves 65fps inference speed. Under the same settings, PNSNet reaches 69fps and UACANet reaches 74fps. In the inference phase, the segmentation branch is used for prediction with parameter size of 24.9M, which is less than UACANet (26.9M) and PNSNet (27.0M). ↑ denotes the higher the better and ↓ denotes the lower the better.

	Semi-supervised Method (1/15 labeled)	mDice ↑	mIoU ↑	wFm ↑	Sm ↑	Em ↑	MAE ↓
CVC-300-TV	UA-MT	0.803	0.703	0.761	0.897	0.917	0.016
	ICT	0.809	0.712	0.771	0.903	0.919	0.015
	CCT	0.813	0.713	0.773	0.897	**0.928**	0.015
	URPC	0.806	0.711	0.785	0.895	0.917	0.015
	CPS	0.802	0.706	0.782	0.894	0.911	0.016
	Ours	**0.824**	**0.730**	**0.812**	**0.906**	0.925	**0.013**
CVC-612-V	UA-MT	0.829	0.746	0.796	0.895	0.926	0.013
	ICT	0.827	0.744	0.802	0.894	0.928	0.014
	CCT	0.847	0.770	0.821	0.904	0.937	0.012
	URPC	0.837	0.758	0.814	0.900	0.934	0.012
	CPS	0.841	0.765	0.831	0.912	0.928	0.013
	Ours	**0.854**	**0.777**	**0.844**	**0.916**	**0.943**	**0.012**
CVC-612-T	UA-MT	0.818	0.741	0.809	0.889	0.888	0.043
	ICT	0.826	0.749	0.819	**0.893**	0.894	0.041
	CCT	0.815	0.736	0.805	0.884	0.886	0.044
	URPC	0.820	0.744	0.808	0.882	0.888	0.043
	CPS	0.814	0.740	0.809	0.887	0.883	0.043
	Ours	**0.827**	**0.752**	**0.823**	0.892	**0.893**	**0.041**
ETIS	UA-MT	0.588	0.485	0.490	0.768	0.762	0.037
	ICT	0.602	0.502	0.517	0.779	0.780	0.031
	CCT	0.608	0.512	0.515	0.779	0.780	0.024
	URPC	0.614	0.529	0.564	0.788	0.782	0.029
	CPS	0.605	0.515	0.550	0.781	0.779	0.027
	Ours	**0.618**	**0.537**	**0.587**	**0.798**	**0.786**	**0.023**

Table 2: Comparison with semi-supervised methods.

CVC-300 are used for training. For the semi-supervised training, we label an image every 15 frames for each sequence. That is, the label ratio is 1/15 in each sequence. We test the performance of TCCNet on the test datasets in [Ji *et al.*, 2021], including CVC-300-TV, CVC-612-V and CVC-612-T. Moreover, to verify the generalization ability of our TCCNet, we evaluate our network on ETIS, of which the image domain is unseen in the training set.

Evaluation metrics. We employ the same evaluation metrics in [Fan *et al.*, 2020], including mean Dice (mDice), mean IoU (mIoU), weighted F_β measure (wFm), S-measure (Sm), Enhanced-alignment measure (Em) and Mean Absolution Error (MAE).

Implementation details. The clip of input sequence T is set to 3 and the batch size is set to 2. In the pre-training phase, the network is trained for 200 epochs with Adam optimizer and a learning rate of 10^{-4}. In the main-training phase, the network is trained for 40 epochs with Adam optimizer. The initial learning rate is set to 10^{-4} with polynomial decay [Kim *et al.*, 2021]. All images are resized to 352×352 and normalized into $[-0.5, 0.5]$. Random data augmentation is performed. In the testing phase of TCCNet, we only use the output from the segmentation branch. C, K, λ_{cps}, λ_s, λ_p and λ_{cf} are empirically set to 32, 8, 1, 1, 1 and 2, respectively.

4.2 Comparison with State-of-the-art Methods

Comparison with methods for fully supervised polyp segmentation. We compare our semi-supervised TCCNet with the state-of-the-art (SOTA) fully supervised polyp segmentation models, including U-Net++ [Zhou *et al.*, 2018], ResUNet++ [Jha *et al.*, 2019], PraNet [Fan *et al.*, 2020], UACANet [Kim *et al.*, 2021] and PNSNet [Ji *et al.*, 2021]. We re-train the SOTA models with the released code under their default settings for further epochs to ensure fairness. The comparison results are shown in Table 1. It can be observed that our semi-supervised model trained under 1/15 label ratio is comparable to the fully supervised models. Under 1/2 label ratio, our TCCNet outperforms the fully supervised models in almost all cases. All the SOTA models perform poorly on the EITS dataset, since the image domain of ETIS is unseen in the training set and it is easy for the models to overfit on the CVC-300 and CVC-612 datasets. Thanks to the consistency regularization, our network under 1/15 label ratio gains 3.3% mDice improvement over the best SOTA fully supervised polyp segmentation model PraNet.

Comparison with methods for semi-supervised segmentation. We compare our method with the existing semi-supervised models, including UA-MT [Yu *et al.*, 2019] and URPC [Luo *et al.*, 2021] for other medical image segmentation tasks and ICT [Verma *et al.*, 2019], CCT [Ouali *et al.*, 2020] and CPS [Chen *et al.*, 2021] for natural image segmentation task. For fair comparison, we change the backbones of these SOTA networks to our segmentation network in the segmentation branch and train the networks with the same data augmentation and training strategy. The comparison results are listed in Table 2. It can be seen that our method achieves superior performance to other methods in almost all cases.

4.3 Ablation Studies

Ablation studies for different modules. The ablation results of different modules on the CVC-300-VT and CVC-612-V datasets are shown in Table 3. From the first four rows, we can observe that the proposed modules, SC-RA module, PC-RA module and CFLoss gain improvement over Baseline (the first row). The combination of SC-RA module and PC-RA module introduces the sequence information into the two branches and further improves the performance, as shown in

Modules			CVC-300-VT						CVC-612-V					
SC-RA	PC-RA	CFLoss	mDice ↑	mIoU ↑	wFm ↑	Sm ↑	mEm ↑	MAE ↓	mDice ↑	mIoU ↑	wFm ↑	Sm ↑	mEm ↑	MAE ↓
			0.791	0.695	0.746	0.890	0.900	0.017	0.803	0.718	0.773	0.891	0.904	0.016
√			0.809	0.716	0.801	0.893	0.918	0.014	0.836	0.762	0.825	0.908	0.927	0.012
	√		0.811	0.716	0.798	0.897	0.915	0.015	0.828	0.749	0.815	0.903	0.919	0.015
		√	0.809	0.712	0.798	0.895	**0.927**	0.018	0.817	0.734	0.720	0.891	0.914	0.019
√	√		0.814	0.717	0.800	0.895	0.926	0.014	0.842	0.758	0.824	0.908	0.938	0.014
√	√	√	**0.824**	**0.730**	**0.812**	**0.906**	0.925	**0.013**	**0.854**	**0.777**	**0.844**	**0.916**	**0.943**	**0.012**

Table 3: Ablation results on the CVC-300-VT and CVC-612-V datasets for SC-RA module, PC-RA module and CFLoss.

Dataset	Method	mDice ↑	mIoU ↑
CVC-300-VT	w/o. pre-train	0.800	0.703
	w. pre-train	**0.824**	**0.730**
CVC-612-V	w/o. pre-train	0.826	0.744
	w. pre-train	**0.854**	**0.777**

Table 4: Impact of our pre-training strategy.

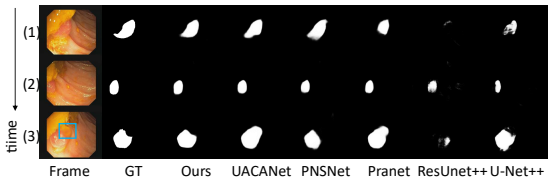

Figure 3: Qualitative results of temporal consistency. The three frames are from a testing clip in temporal order.

the fifth row. The full model (the last row) achieves better results, which is attributed to the context-free constraint.

Ablation studies for training strategy. Table 4 lists the effect of our pre-training strategy. It can be seen that the pre-training strategy brings 2.4%/2.8% of mDice and 2.7%/3.3% of mIoU increase on CVC-300-VT/CVC-612-V, respectively.

4.4 Qualitative Results

Temporal consistency of predictions. The quantitative comparison results between our model and other fully supervised VPS models are shown in Fig. 3. It can be observed that our model well maintains the temporal consistency of the predictions for a testing clip. Specifically, for the third frame, other models are likely to be influenced by the optical artifact region (blue square), leading to the misclassification as polyp tissue. Our model can learn the position and texture of polyps from previous frames, thus substantially alleviating false positives of segmentation.

Analysis on various label ratios. We visualize the improvement over the supervised baseline under different label ratios, as shown in Fig. 4. The time intervals for annotations are set to 50, 15, 5, 2 and 1. Our network consistently surpasses Baseline. To be specific, the mDice gains of our network over Baseline are 18.1%/8.6%, 9.1%/6.5%, 3.1%/2.5%, 2.6%/1.7% and 2.1%/1.4% on CVC-300-VT/CVC-612-V.

Qualitative results of the error correction mechanism for segmentation map. Fig. 5(a) shows the qualitative results of the error correction mechanism for segmentation map. Using SC-RA module as an example, a suspected region is misclassified as the polyp tissue in the global segmentation map

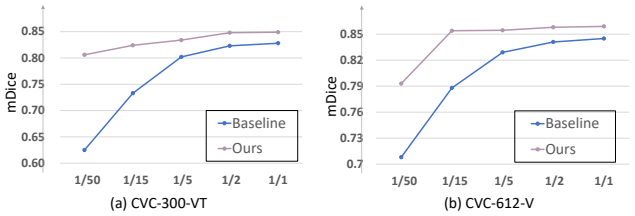

Figure 4: Results under different label ratios on the CVC-300-VT and CVC-612-V datasets. The time intervals for annotations are 50, 15, 5, 2 and 1.

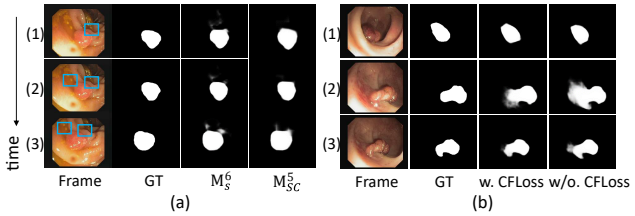

Figure 5: (a) Qualitative results of the error correction mechanism for segmentation map in the segmentation branch. The suspected region is bounded by blue squares. (b) Qualitative results of CFLoss.

M_s^6 but later corrected to M_{SC}^5 by the SC-RA module.

Qualitative results of the proposed CFLoss. Fig. 5(b) illustrates the qualitative results of the proposed CFLoss. Given the three consecutive frames, the context of polyp in the 2^{nd} and 3^{rd} frames are quite different from that in the 1^{st} one. As a result, the network without CFLoss misclassifies the background region as polyp tissue for the second and third frames. It can be observed that our network effectively reduces such false positives with the aid of CFLoss.

5 Conclusion

In this paper, we introduced a novel semi-supervised video polyp segmentation network, i.e., TCCNet. It consists of a segmentation branch and a propagation branch, which are trained with a co-training scheme. A SC-RA module and a PC-RA module are designed to keep the predictions temporally consistent for consecutive frames. A CFLoss is proposed to reduce the impact of varying contexts on the predictions of adjacent frames. Experiments demonstrate that our TCCNet gains better results than the SOTA methods on both fully supervised polyp segmentation and semi-supervised medical segmentation. In the future, we would like to extend our method to other challenging tasks, such as instrument segmentation from robotic surgical videos.

Acknowledgments

This work is supported by National Science and Technology Innovation 2030 – Major Project (No. 2021ZD0114001; No. 2021ZD0114000), National Natural Science Foundation of China (No. 61976057; No. 62172101), the Science and Technology Commission of Shanghai Municipality (No. 21511101000; No. 20511101203; No. 20511101403), the Science and Technology Major Project of Commission of Science and Technology of Shanghai (No. 2021SHZDZX0103), Shanghai Natural Science Foundation (No. 19ZR1417200). Yuejie Zhang and Tao Zhang are corresponding authors.

References

[Bernal *et al.*, 2015] Jorge Bernal, Francisco Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez de Miguel, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Medical Imaging Graph.*, pages 99–111, 2015.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[Chen *et al.*, 2018] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, pages 234–250, 2018.

[Chen *et al.*, 2021] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021.

[Fan *et al.*, 2020] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273, 2020.

[Hu *et al.*, 2021] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *CVPR*, pages 4144–4154, 2021.

[Jha *et al.*, 2019] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *ISM*, pages 225–2255, 2019.

[Ji *et al.*, 2021] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, pages 142–152, 2021.

[Kim *et al.*, 2021] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *ACM Multimedia*, pages 2167–2175, 2021.

[Luo *et al.*, 2021] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *MICCAI*, pages 318–329, 2021.

[Oh *et al.*, 2019] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019.

[Ouali *et al.*, 2020] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020.

[Silva *et al.*, 2014] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Comput. Assist. Radiol. Surg.*, pages 283–293, 2014.

[Tajbakhsh *et al.*, 2016] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *TMI*, pages 630–644, 2016.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, pages 1195–1204, 2017.

[Verma *et al.*, 2019] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *IJCAI*, pages 3635–3641, 2019.

[Yu *et al.*, 2019] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, pages 605–613, 2019.

[Zhang *et al.*, 2017] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *IJCAI*, pages 408–416, 2017.

[Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *MICCAI*, pages 3–11, 2018.