# CMSA-Net: Causal Multi-scale Aggregation with Adaptive Multi-source Reference for Video Polyp Segmentation

Tong Wang[1,2,3], Yaolei Qi[1,2], Siwen Wang[3], Imran Razzak[3], Guanyu Yang[1,2(✉)], and Yutong Xie[3(✉)]

[1] School of Computer Science and Engineering, Southeast University, China
[2] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
yang.list@seu.edu.cn
[3] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
yutong.xie678@gmail.com

**Abstract.** Video polyp segmentation (VPS) is an important task in computer-aided colonoscopy, as it helps doctors accurately locate and track polyps during examinations. However, VPS remains challenging because polyps often look similar to surrounding mucosa, leading to weak semantic discrimination. In addition, large changes in polyp position and scale across video frames make stable and accurate segmentation difficult. To address these challenges, we propose a robust VPS framework named CMSA-Net. The proposed network introduces a Causal Multi-scale Aggregation (CMA) module to effectively gather semantic information from multiple historical frames at different scales. By using causal attention, CMA ensures that temporal feature propagation follows strict time order, which helps reduce noise and improve feature reliability. Furthermore, we design a Dynamic Multi-source Reference (DMR) strategy that adaptively selects informative and reliable reference frames based on semantic separability and prediction confidence. This strategy provides strong multi-frame guidance while keeping the model efficient for real-time inference. Extensive experiments on the SUN-SEG dataset demonstrate that CMSA-Net achieves state-of-the-art performance, offering a favorable balance between segmentation accuracy and real-time clinical applicability. The code is available at https://github.com/wangtong627/CMSA-Net.

**Keywords:** Video polyp segmentation · Spatial-temporal attention · Semantic segmentation.

## 1 Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer-related death worldwide [1]. Since most CRC cases develop from colorectal polyps, early detection and accurate segmentation during colonoscopy are critically important.

✉ Corresponding authors.

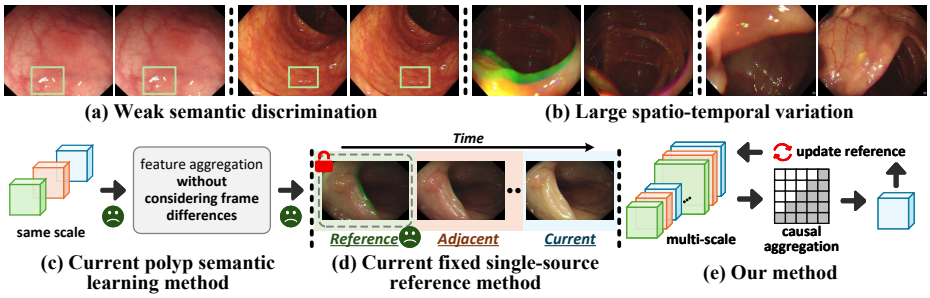


**Fig. 1. (a) Challenge 1:** Weak semantic discrimination, caused by the low contrast between polyps and background (case14_6 and case19). **(b) Challenge 2:** Large spatio-temporal variations across frames (case51_1 and case71_1). **(c) Limitation 1:** Existing feature fusion methods for polyp semantic learning , which ignore multi-scale information and intrinsic identity relationships. **(d) Limitation 2:** Current approaches adopt a fixed single-source reference, failing to capture dynamic and diverse spatio-temporal cues. **(e) Our Method**.

However, the miss rate for polyps during colonoscopy remains as high as 25%, posing a significant clinical challenge [15]. Therefore, developing effective video polyp segmentation (VPS) methods is of great clinical value for assisting real-time diagnosis and intervention.

Despite recent progress, accurate VPS still faces challenges (see Fig. 1(a-b)). **(1) Weak semantic discrimination:** polyps often exhibit low semantic contrast with surrounding mucosa, making discriminative semantic learning difficult. **(2) Large spatio-temporal variation:** irregular camera motion causes drastic changes in polyp scale and position across frames, disrupting temporal consistency. **(3) Real-time requirement:** clinical applications demand low-latency inference during procedures.

Early polyp segmentation methods [7,19,27,22,20,24,23,26,2] followed static image segmentation paradigms and showed limited performance in videos due to the lack of temporal modeling. More recently, VPS methods [13,14,3,12,4,17,28] introduce additional frames to assist current-frame segmentation. In particular, *reference frames* are used to provide semantic guidance, while *adjacent frames* are exploited to model short-term temporal consistency. Despite their utilization of temporal information, several limitations still remain: **(1) Limited spatio-temporal fusion (see Fig. 1(c)):** existing methods usually aggregate features at a single spatial scale, using Mamba-based modules [25,9,4,30] or attention mechanisms [13,14,12]. This design fails to fully exploit rich semantic cues across different scales in the reference frame and adjacent frames. Moreover, they often overlook the intrinsic identity relationships among the reference, adjacent and current frames, which can lead to feature contamination when large appearance variations occur between frames [8]. **(2) Single-source reference dependency (see Fig. 1(d)):** most approaches rely on a single and fixed reference source, which makes them less robust to large target variations. While

memory-based methods suffer from computational inefficiency due to redundant frames and may fail to retrieve the optimal reference information [29].

To address these issues, we propose the **C**ausal **M**ulti-scale **S**patio-temporal **A**ggregation **Net**work (***CMSA-Net***). CMSA-Net integrates causal multi-scale modeling with a dynamic multi-source reference strategy to enhance semantic perception of polyps with weak discriminability while maintaining real-time efficiency (see Fig. 1(e)). (1) We propose the **Causal Multi-scale Aggregation (CMA)** module for current-frame semantic learning. Unlike single-scale feature interactions, CMA enables the current frame to aggregate multi-scale semantic priors from reference and adjacent frames. By exploiting temporal information across different spatial scales, CMA improves the representation of weakly discriminative polyp features. Moreover, a causal attention mechanism is introduced to model the logical relationships among reference, adjacent, and current frames, encouraging temporally coherent feature propagation and reducing feature contamination under large inter-frame variations. (2) We propose a **Dynamic Multi-source Reference (DMR)** strategy to provide multiple meaningful semantic references. DMR adaptively updates the reference set by selecting reliable frames from the video sequence according to the current frame, ensuring semantic consistency while avoiding redundant computation. The resulting multi-source references provide stable and robust semantic cues, which further enhance the effectiveness of CMA and support real-time inference. We conducted extensive experiments on SUN-SEG dataset. Results show that CMSA-Net achieves state-of-the-art performance while maintaining real-time inference speeds. Our contributions are summarized as follows:

- We propose CMSA-Net for VPS through causal multi-scale modeling and dynamic multi-source references.
- We design a Causal Multi-scale Aggregation (CMA) module for causal multi-scale spatio-temporal aggregation to enhance current-frame semantics.
- We introduce a Dynamic Multi-source Reference (DMR) strategy that selects reliable reference frames for stable and efficient semantic guidance.
- Extensive experiments on the SUN-SEG dataset demonstrate that CMSA-Net achieves state-of-the-art performance with real-time inference speed.

## 2 Method

Fig. 2 shows our CMSA-Net. Given an input video clip with $T$ frames, we denote the clip as $\{I_t\}_{t=1}^{T}$. The first $R$ frames $\{I_1, \ldots, I_R\}$ are *reference frames* $I_{\text{ref}}$, where $R$ corresponds to the number of multi-source references; the last frame $I_T$ is the *current frame* $I_{\text{cur}}$; and the remaining frames $\{I_{R+1}, \ldots, I_{T-1}\}$ are *adjacent frames* $I_{\text{adj}}$. We extract multi-scale representations for all frames using a shared **Image Encoder** backbone with $S$ stages. We denote the feature at the $s$-th stage of the $r$-th reference frame as $\mathbf{F}^s_{\text{ref}_r}$, where $r \in \{1, \ldots, R\}$ indexes the reference source and $s \in \{0, \ldots, S\}$ indexes the backbone stage. Similarly, the $s$-th stage representation of the current frame is denoted as $\mathbf{F}^s_{\text{cur}}$, and that of the $k$-th adjacent frame is denoted as $\mathbf{F}^s_{\text{adj}_k}$ for $k \in \{R+1, \ldots, T-1\}$.

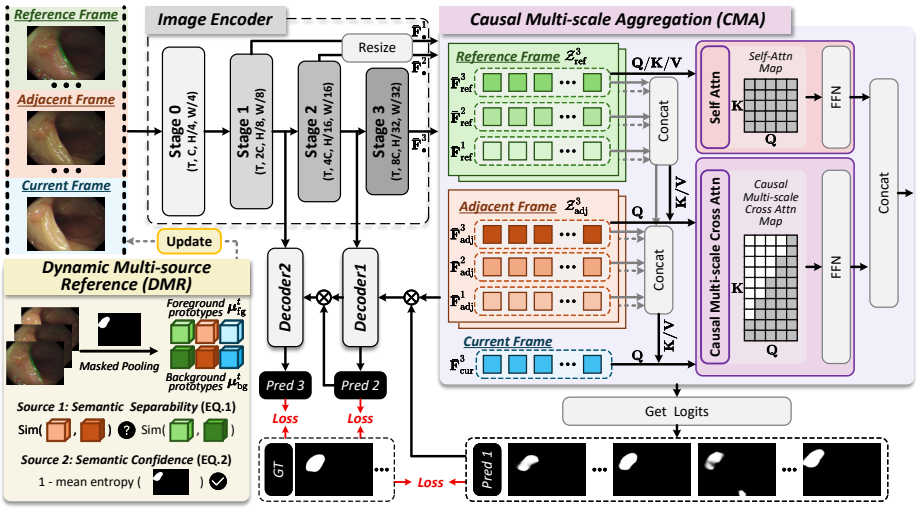


**Fig. 2.** Framework of the proposed CMSA-Net, including the CMA module (Sec. 2.1) with DMR strategy (Sec. 2.2).

For convenience, we concatenate the features of all reference frames, adjacent frames, and the current frame at the $s$-th stage to form a unified representation: $\mathbf{F}^s = \left[\mathbf{F}^s_{\mathrm{ref}_1}, \ldots, \mathbf{F}^s_{\mathrm{ref}_R}, \mathbf{F}^s_{\mathrm{adj}_{R+1}}, \ldots, \mathbf{F}^s_{\mathrm{adj}_{T-1}}, \mathbf{F}^s_{\mathrm{cur}}\right]$, where $[\cdot]$ denotes concatenation. Following [14], the input clip $\{I_t\}_{t=1}^{T}$ is stacked into $[T, C, H, W]$. The image encoder produces multi-scale features at 4 stages with decreasing spatial resolutions and increasing channel dimensions, namely, $\mathbf{F}^0 \in \mathbb{R}^{T\times 2C\times \frac{H}{4}\times\frac{W}{4}}$, $\mathbf{F}^1 \in \mathbb{R}^{T\times 4C\times \frac{H}{8}\times\frac{W}{8}}$, $\mathbf{F}^2 \in \mathbb{R}^{T\times 8C\times \frac{H}{16}\times\frac{W}{16}}$, and $\mathbf{F}^3 \in \mathbb{R}^{T\times 16C\times \frac{H}{32}\times\frac{W}{32}}$.

### 2.1 Causal Multi-scale Aggregation (CMA) Module

CMA enhances the discriminability of current-frame features by aggregating multi-scale spatio-temporal information. For a target backbone stage $s$, the module treats features from all available backbone stages $u \in \{1, \ldots, S\}$ as complementary semantic scales to capture rich contextual priors. Specifically, for any frame $I_\bullet$ in the clip, where the placeholder $\bullet \in \{\mathrm{ref}, \mathrm{adj}, \mathrm{cur}\}$ denotes the frame type, we denote its original feature at the $u$-th stage as $\mathbf{F}^u_\bullet \in \mathbb{R}^{C_u\times H_u\times W_u}$. To enable cross-scale interaction, all source features are aligned to the spatial resolution $(H_s, W_s)$ and unified to the channel dimension $C$ of the target stage $s$ via: $\bar{\mathbf{F}}^{u\to s}_\bullet = \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\times3}(\mathrm{Resize}_{H_s\times W_s}(\mathbf{F}^u_\bullet))) \in \mathbb{R}^{C\times H_s\times W_s}$. These aligned multi-scale features are then concatenated channel-wise to form a per-frame multi-scale token set: $\mathcal{Z}^s_\bullet = [\bar{\mathbf{F}}^{1\to s}_\bullet, \ldots, \bar{\mathbf{F}}^{S\to s}_\bullet] \in \mathbb{R}^{(S\cdot C)\times H_s\times W_s}$.

To leverage temporal priors without future leakage, CMA employs causal attention, restricting features at time $t$ to attend only to reference features and past frames, and accordingly constructs frame-dependent query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$). **(1) For a reference feature** $\bar{\mathbf{F}}^s_{\mathrm{ref}_r}$ $(r \in \{1, \ldots, R\})$, self-attention is applied with $\mathbf{Q}_{\mathrm{ref}_r} = W_Q \cdot \mathcal{Z}^s_{\mathrm{ref}_r}$, $\mathbf{K}_{\mathrm{ref}_r} = W_K \cdot \mathcal{Z}^s_{\mathrm{ref}_r}$, and $\mathbf{V}_{\mathrm{ref}_r} = W_V \cdot \mathcal{Z}^s_{\mathrm{ref}_r}$.

**(2) For an adjacent feature** $\bar{\mathbf{F}}^s_{\mathrm{adj}_t}$ $(t \in \{R{+}1, \ldots, T{-}1\})$, the query is generated from its current-scale feature $\mathbf{Q}_t = W_Q \cdot \bar{\mathbf{F}}^s_{\mathrm{adj}_t}$, while keys and values are constructed by concatenating multi-scale token sets of all reference frames and adjacent frames up to time $t$: $\mathbf{K}_t = W_K \cdot [\, \mathcal{Z}^s_{\mathrm{ref}_1}, \ldots, \mathcal{Z}^s_{\mathrm{ref}_R}, \mathcal{Z}^s_{\mathrm{adj}_{R+1}}, \ldots, \mathcal{Z}^s_{\mathrm{adj}_t} \,]$ and $\mathbf{V}_t = W_V \cdot [\, \mathcal{Z}^s_{\mathrm{ref}_1}, \ldots, \mathcal{Z}^s_{\mathrm{ref}_R}, \mathcal{Z}^s_{\mathrm{adj}_{R+1}}, \ldots, \mathcal{Z}^s_{\mathrm{adj}_t} \,]$. **(3) For the current feature** $\bar{\mathbf{F}}^s_{\mathrm{cur}}$ $(t = T)$, the query is generated from its high-level representation $\mathbf{Q}_{\mathrm{cur}} = W_Q \cdot \bar{\mathbf{F}}^s_{\mathrm{cur}}$, and keys and values are formed by concatenating multi-scale token sets of all available frames: $\mathbf{K}_{\mathrm{cur}} = W_K \cdot [\, \mathcal{Z}^s_{\mathrm{ref}_1}, \ldots, \mathcal{Z}^s_{\mathrm{ref}_R}, \mathcal{Z}^s_{\mathrm{adj}_{R+1}}, \ldots, \mathcal{Z}^s_{\mathrm{adj}_{T-1}}, \mathcal{Z}^s_{\mathrm{cur}} \,]$ and $\mathbf{V}_{\mathrm{cur}} = W_V \cdot [\, \mathcal{Z}^s_{\mathrm{ref}_1}, \ldots, \mathcal{Z}^s_{\mathrm{ref}_R}, \mathcal{Z}^s_{\mathrm{adj}_{R+1}}, \ldots, \mathcal{Z}^s_{\mathrm{adj}_{T-1}}, \mathcal{Z}^s_{\mathrm{cur}} \,]$. All attention operations strictly follow the causal constraint, ensuring that no feature at time $t$ attends to features from future frames. The causal multi-scale aggregation is $\mathcal{A}^s_t = \mathrm{Softmax}\Big(\frac{\mathbf{Q}_t \cdot \mathbf{K}_t^\top}{\sqrt{d}}\Big)\mathbf{V}_t$, where $d$ denotes the feature dimension. The aggregated output is fused with the original feature through residual learning: $\hat{\mathbf{F}}^s_t = \bar{\mathbf{F}}^s_t + \mathcal{A}^s_t, \quad \tilde{\mathbf{F}}^s_t = \hat{\mathbf{F}}^s_t + \mathtt{FFN}\big(\mathtt{LN}(\hat{\mathbf{F}}^s_t)\big)$. CMA aggregates multi-scale temporal context to enhance polyp representations.

### 2.2 Dynamic Multi-source Reference (DMR) Strategy

While CMA aggregates multi-scale spatio-temporal semantics, its effectiveness relies on high-quality references. Existing VPS methods depend on fixed references [4,13] or memory-based designs [12,29], which are unreliable over time, computationally expensive, and inflexible. We propose a DMR strategy that adaptively maintains a compact set of references during inference, leveraging two complementary sources: semantic separability and semantic confidence.

**Semantic Separability–based reference update.** Given the CMA output feature $\tilde{\mathbf{F}}^s_t$ at stage $s$, we first obtain the prediction logits by a $1 \times 1$ convolution, and the corresponding probability map $p_t \in [0,1]^{H \times W}$ after sigmoid activation. The probability map is used to perform masked pooling on $\tilde{\mathbf{F}}^s_t$ to extract foreground and background prototypes: $\boldsymbol{\mu}^t_{\mathrm{fg}} = \frac{\sum_i p_t(i)\,\mathbf{f}_t(i)}{\sum_i p_t(i)}, \quad \boldsymbol{\mu}^t_{\mathrm{bg}} = \frac{\sum_i (1-p_t(i))\,\mathbf{f}_t(i)}{\sum_i (1-p_t(i))}$, where $\mathbf{f}_t(i)$ denotes the feature vector at spatial location $i$. For a candidate frame $t$, we evaluate its semantic quality by foreground–background semantic separability and temporal consistency:

$$s^t_{\mathrm{sep}} = 1 - \cos\big(\boldsymbol{\mu}^t_{\mathrm{fg}}, \boldsymbol{\mu}^t_{\mathrm{bg}}\big)\,, s^t_{\mathrm{cons}} = \cos\big(\boldsymbol{\mu}^t_{\mathrm{fg}}, \boldsymbol{\mu}^{\mathrm{cur}}_{\mathrm{fg}}\big)\,, \mathrm{Score}^t_{\mathrm{sem}} = s^t_{\mathrm{sep}} + s^t_{\mathrm{cons}}. \quad (1)$$

The semantic reference $\tilde{\mathbf{F}}^{s,\mathrm{sem}}_{\mathrm{ref}}$ is updated only when a higher $\mathrm{Score}_{\mathrm{sem}}$ is achieved, with a cooldown interval to prevent drift.

**Semantic Confidence–based reference update.** In parallel, we maintain a confidence reference $\tilde{\mathbf{F}}^{s,\mathrm{conf}}_{\mathrm{ref}}$ to capture frames with reliable predictions. The semantic confidence of a candidate frame $t$ is quantified by an entropy-based determinacy measure:

$$c^t = 1 - \frac{1}{|\Omega|}\sum_{i \in \Omega} H(p_t(i))\,, \quad \mathrm{Score}^t_{\mathrm{conf}} = c^t + s^t_{\mathrm{cons}}, \quad (2)$$

where $H(p) = -p \log p - (1-p)\log(1-p)$. The confidence reference $\tilde{\mathbf{F}}^{s,\mathrm{conf}}_{\mathrm{ref}}$ is updated only when a higher $\mathrm{Score}_{\mathrm{conf}}$ is achieved, with a cooldown interval.

### 2.3 Learning Process and Implementation Details

**Learning Process and Loss Function.** During training, we supervise predictions at multiple stages using the decoder [13] to facilitate stable optimization. The CMA output feature $\tilde{\mathbf{F}}_t^s$ is first used to generate a coarse prediction $\text{pred}_1$. The CMA feature is then fused with the stage-2 feature and fed into **Decoder1**, yielding $\text{pred}_2$. Subsequently, the output features are combined with the stage-1 feature and passed to **Decoder2** to produce the final prediction $\text{pred}_3$. All predictions are resized to the ground-truth resolution. We employ Dice loss ($\mathcal{L}_{\text{Dice}}$), weighted IoU loss ($\mathcal{L}_{\text{wIoU}}$), and weighted BCE loss ($\mathcal{L}_{\text{wBCE}}$) jointly. The segmentation loss is defined as $\mathcal{L}_{\text{seg}}(m) = \mathcal{L}_{\text{Dice}}(m, g) + \mathcal{L}_{\text{wIoU}}(m, g) + \mathcal{L}_{\text{wBCE}}(m, g)$, and the total training loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}}(\text{pred}_1) + \mathcal{L}_{\text{seg}}(\text{pred}_2) + \mathcal{L}_{\text{seg}}(\text{pred}_3)$, where $m$ and $g$ denote the predicted masks and the corresponding ground-truth.

**Implementation Details.** We implement CMSA-Net using PyTorch and train it on a single NVIDIA RTX 3090 GPU. Res2Net-50 [10] and PVTv2-B2 [21] are adopted as backbone networks, both initialized with ImageNet pre-trained weights. All input frames are resized to $352 \times 352$, and the batch size is set to 4. The model is optimized using AdamW with a weight decay of $1 \times 10^{-4}$ and an initial learning rate of $1 \times 10^{-4}$, and trained for 30 epochs. Each input clip contains 6 frames, including reference, adjacent, and current frames. In implementation, CMA is applied to the stage-3 feature for aggregating high-level temporal semantics of the current frames, while multi-stage features are still utilized for the reference and adjacent frames. All intermediate features are projected to a unified channel dimension $C = 32$, and the CMA module is implemented with 4 attention heads. For the DMR strategy, the cooldown interval is set to 5 for semantic reference updates and 1 for confidence reference updates.

## 3 Experiments

### 3.1 Dataset and Metrics

**Dataset.** We evaluate on SUN-SEG [14], the largest VPS dataset. Following [14,4], the training set contains 112 video clips with 19,544 frames, while the test set is divided into four subsets: SUN-SEG-Easy-Seen (33 clips/4,719 frames), Easy-Unseen (86 clips/12,351 frames), Hard-Seen (17 clips/3,882 frames), and Hard-Unseen (37 clips/8,640 frames). Easy/Hard denote segmentation difficulty, while Seen/Unseen indicate whether test videos overlap with the training videos.
**Metrics.** We employ six metrics: Dice, IoU, MAE, structure-measure ($S_\alpha$) [5], enhanced-alignment measure ($E_\phi^{mn}$) [6] and weighted F-measure ($F_\beta^w$) [18].

### 3.2 Comparisons with State-of-the-art Methods

We compare CMSA-Net with 12 state-of-the-art methods: (1) natural video segmentation (NVS), including COSNet [16], PCSA [11], and 2/3D [19]; (2) image-based polyp segmentation (IPS), such as PraNet [7], ACSNet [27], SANet [22], and SEPNet [20]; and (3) video-based polyp segmentation (VPS), including PNS [13], PNS+ [14], MAST [3], SALI [12], and STDDNet [4].

**Table 1.** Quantitative comparison on SUN-SEG-Easy dataset. The best is **bold**.

| Method | Publication | Backbone | SUN-SEG-Easy-Seen (%) | | | | | | SUN-SEG-Easy-Unseen (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $S_\alpha \uparrow$ | $E_\phi^{mn} \uparrow$ | $F_\beta^w \uparrow$ | Dice↑ | IoU↑ | MAE↓ | $S_\alpha \uparrow$ | $E_\phi^{mn} \uparrow$ | $F_\beta^w \uparrow$ | Dice↑ | IoU↑ | MAE↓ |
| COSNet [16] | TPAMI'19 | - | 84.5 | 83.6 | 72.7 | 73.0 | 64.8 | 3.4 | 65.4 | 60.0 | 43.1 | 42.3 | 34.2 | 7.3 |
| PCSA [11] | AAAI'20 | - | 85.2 | 83.5 | 68.1 | 70.9 | 60.4 | 3.9 | 68.0 | 66.0 | 45.1 | 45.0 | 35.3 | 7.8 |
| 2/3D [19] | MICCAI'20 | - | 89.5 | 90.9 | 81.9 | 82.9 | 75.6 | 2.1 | 78.6 | 77.7 | 65.2 | 65.6 | 57.0 | 4.4 |
| PraNet [7] | MICCAI'20 | Res2Net-50 | 91.8 | 94.2 | 87.7 | 88.3 | 82.5 | 2.0 | 78.1 | 78.8 | 66.3 | 66.5 | 58.2 | 5.2 |
| ACSNet [27] | MICCAI'20 | ResNet-34 | 92.0 | 94.2 | 87.4 | 88.2 | 82.8 | 1.7 | 77.2 | 76.6 | 63.0 | 63.8 | 56.4 | 4.6 |
| SANet [22] | MICCAI'21 | Res2Net-50 | 91.6 | 93.3 | 86.6 | 87.2 | 82.0 | 1.8 | 75.0 | 72.8 | 59.0 | 59.3 | 52.4 | 5.2 |
| SEPNet [20] | TCSVT'24 | PVTv2-B2 | 93.1 | 96.2 | 88.3 | 89.6 | 83.4 | 1.7 | 82.9 | 88.3 | 73.5 | 75.1 | 66.6 | 4.2 |
| PNS [13] | MICCAI'21 | Res2Net-50 | 90.6 | 91.0 | 83.6 | 84.1 | 78.3 | 2.0 | 76.7 | 74.4 | 61.6 | 61.8 | 54.5 | 4.8 |
| PNS+ [14] | MIR'22 | Res2Net-50 | 91.7 | 92.5 | 84.8 | 85.5 | 78.7 | 2.1 | 80.6 | 79.8 | 67.6 | 67.8 | 59.1 | 4.4 |
| MAST [3] | Arxiv'24 | PVTv2-B2 | 92.5 | 96.2 | 87.8 | 89.3 | 82.7 | 1.6 | 83.2 | 89.4 | 74.9 | 77.0 | 67.4 | 3.7 |
| SALI [12] | MICCAI'24 | PVTv2-B2 | 90.2 | 93.2 | 84.9 | 85.8 | 78.9 | 2.4 | 73.1 | 75.2 | 58.7 | 59.2 | 50.2 | 6.3 |
| SALI [12] | MICCAI'24 | PVTv2-B5 | 90.7 | 93.7 | 85.1 | 86.2 | 79.6 | 2.2 | 77.1 | 82.1 | 64.6 | 65.6 | 56.8 | 5.5 |
| STDDNet [4] | ICCV'25 | Res2Net-50 | 93.5 | 96.0 | 89.7 | 90.5 | 85.0 | 1.5 | 81.7 | 83.0 | 72.1 | 72.4 | 64.3 | 3.7 |
| STDDNet [4] | ICCV'25 | PVTv2-B2 | 94.1 | 96.9 | 90.5 | 91.5 | 86.1 | 1.4 | 86.0 | **90.3** | 78.6 | 80.1 | 72.4 | 3.4 |
| Ours | - | Res2Net-50 | 94.5 | 97.3 | 90.5 | 91.9 | 86.5 | 1.2 | 84.4 | 90.1 | 75.3 | 77.5 | 69.2 | 3.5 |
| Ours | - | PVTv2-B2 | **95.1** | **97.5** | **91.6** | **92.6** | **87.6** | **1.1** | **86.7** | **90.3** | **79.3** | **80.3** | **72.6** | **2.9** |

**Table 2.** Quantitative comparison on SUN-SEG-Hard dataset. The best is **bold**.

| Method | Publication | Backbone | SUN-SEG-Hard-Seen (%) | | | | | | SUN-SEG-Hard-Unseen (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $S_\alpha \uparrow$ | $E_\phi^{mn} \uparrow$ | $F_\beta^w \uparrow$ | Dice↑ | IoU↑ | MAE↓ | $S_\alpha \uparrow$ | $E_\phi^{mn} \uparrow$ | $F_\beta^w \uparrow$ | Dice↑ | IoU↑ | MAE↓ |
| COSNet [16] | TPAMI'19 | - | 78.5 | 77.2 | 62.6 | 63.3 | 54.1 | 4.6 | 67.0 | 62.7 | 44.3 | 43.8 | 35.3 | 7.0 |
| PCSA [11] | AAAI'20 | - | 77.2 | 75.9 | 56.6 | 58.5 | 47.9 | 5.7 | 68.2 | 66.0 | 44.2 | 45.0 | 35.1 | 8.0 |
| 2/3D [19] | MICCAI'20 | - | 84.9 | 86.9 | 75.3 | 76.4 | 67.1 | 3.5 | 78.6 | 77.5 | 63.4 | 64.4 | 55.8 | 4.4 |
| PraNet [7] | MICCAI'20 | Res2Net-50 | 88.4 | 91.9 | 83.1 | 83.9 | 76.6 | 3.1 | 78.7 | 80.2 | 66.7 | 67.5 | 58.7 | 5.3 |
| ACSNet [27] | MICCAI'20 | ResNet-34 | 87.2 | 91.0 | 80.6 | 82.0 | 74.8 | 3.6 | 76.2 | 77.6 | 61.0 | 62.4 | 54.7 | 5.3 |
| SANet [22] | MICCAI'21 | Res2Net-50 | 87.4 | 90.5 | 81.0 | 82.0 | 74.8 | 3.3 | 75.3 | 73.6 | 59.0 | 59.5 | 52.7 | 5.5 |
| SEPNet [20] | TCSVT'24 | PVTv2-B2 | 89.4 | 94.0 | 83.5 | 85.7 | 77.6 | 3.4 | 84.7 | 89.5 | 74.5 | 77.4 | 68.4 | 3.9 |
| PNS [13] | MICCAI'21 | Res2Net-50 | 87.0 | 89.2 | 78.7 | 79.6 | 72.1 | 3.3 | 76.7 | 75.5 | 60.9 | 61.5 | 53.9 | 5.0 |
| PNS+ [14] | MIR'22 | Res2Net-50 | 88.7 | 90.2 | 80.6 | 81.3 | 72.8 | 3.0 | 79.8 | 79.3 | 65.4 | 66.1 | 57.1 | 5.0 |
| MAST [3] | Arxiv'24 | PVTv2-B2 | 89.2 | 94.2 | 83.2 | 85.3 | 76.7 | 2.6 | 85.6 | 91.3 | 77.2 | 79.9 | 70.8 | 3.1 |
| SALI [12] | MICCAI'24 | PVTv2-B2 | 86.8 | 90.9 | 79.9 | 81.0 | 72.6 | 3.4 | 72.8 | 75.9 | 56.9 | 57.9 | 48.7 | 6.8 |
| SALI [12] | MICCAI'24 | PVTv2-B5 | 86.6 | 91.0 | 79.7 | 81.0 | 72.9 | 3.8 | 76.5 | 81.3 | 62.0 | 63.6 | 54.7 | 5.7 |
| STDDNet [4] | ICCV'25 | Res2Net-50 | 91.3 | 95.2 | 86.9 | 88.1 | 81.0 | 2.3 | 83.4 | 85.6 | 74.1 | 75.0 | 67.3 | 3.7 |
| STDDNet [4] | ICCV'25 | PVTv2-B2 | 91.1 | 95.0 | 86.0 | 87.8 | 80.6 | 2.8 | 86.3 | 90.2 | 78.1 | 80.2 | 72.2 | 3.5 |
| Ours | - | Res2Net-50 | **92.7** | **96.1** | **87.1** | **89.8** | **83.0** | **1.8** | 85.1 | 89.8 | 75.0 | 78.0 | 69.5 | 3.6 |
| Ours | - | PVTv2-B2 | 92.3 | 95.6 | **87.1** | 88.9 | 82.1 | 1.9 | **87.3** | **91.0** | **79.6** | **81.3** | **73.7** | **2.9** |

**Quantitative Comparisons.** Tab. 1 and Tab. 2 present the quantitative comparison results. [1] Overall, CMSA-Net achieves the best performance across both seen and unseen subsets under different difficulty levels. On the Easy setting, CMSA-Net consistently outperforms existing methods on both Seen and Unseen splits, achieving the highest Dice scores and the lowest MAE. On the more challenging Hard setting, our method shows clearer advantages, especially on unseen videos, where it surpasses the strongest baselines by up to **1.7%** Dice on Hard-Seen and **1.1%** Dice on Hard-Unseen, demonstrating superior robustness and generalization ability. These gains validate the effectiveness of jointly modeling CMA and DMR, where CMA enhances discriminative representations.

**Qualitative Comparisons.** In Fig. 3, we visualize the segmentation results of different methods on challenging cases. Our method achieved the best results.

[1] For fair comparison, all methods were retrained on the SUN-SEG dataset. As the official SALI only provides a PVTv2-B5 version, we retrained it with the PVTv2-B2.

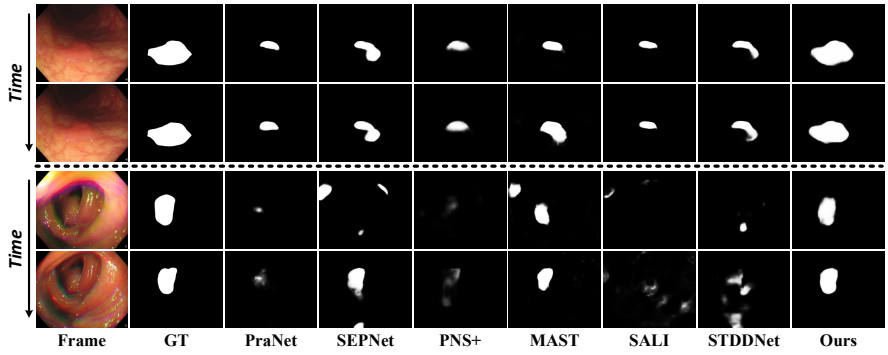


**Fig. 3.** Qualitative Comparisons. Upper (case91_2): a sequence of consecutive low-contrast frames; Lower (case32_4): significant variations between adjacent frames.

**Table 3.** Comparison of model parameters and inference time (clip = 6).

| Efficiency | PraNet | ACSNet | SEPNet | PNS+ | SALI | STDDNet (R) | STDDNet (P) | Ours (R) | Ours (P) |
|---|---|---|---|---|---|---|---|---|---|
| GFLOPs ↓ | 78.89 | 130.52 | 75.12 | 68.58 | 86.58 | 75.17 | 64.18 | 73.36 | 59.42 |
| Param. (M) ↓ | 30.50 | 29.45 | 25.96 | 9.8 | 82.73 | 38.16 | 29.67 | 29.61 | 25.79 |
| FPS ↑ | 42 | 21 | 33 | 64 | 10 | 45 | 36 | 47 | 38 |

**Table 4.** Ablation study of key components. The best is **bold**.

| Methods | SUN-SEG-Easy-Seen | | | | | SUN-SEG-Easy-Unseen | | | | | SUN-SEG-Hard-Seen | | | | | SUN-SEG-Hard-Unseen | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_\alpha$ | $E_\phi^{mn}$ | $F_\beta^w$ | Dice | IoU | $S_\alpha$ | $E_\phi^{mn}$ | $F_\beta^w$ | Dice | IoU | $S_\alpha$ | $E_\phi^{mn}$ | $F_\beta^w$ | Dice | IoU | $S_\alpha$ | $E_\phi^{mn}$ | $F_\beta^w$ | Dice | IoU |
| CMSA-Net | **95.1** | **97.5** | **91.6** | **92.6** | **87.6** | **86.7** | **90.3** | **79.3** | **80.3** | **72.6** | **92.3** | **95.6** | **87.1** | **88.9** | **82.1** | **87.3** | **91.0** | **79.6** | **81.3** | **73.7** |
| w/o CMA | 92.4 | 94.8 | 87.2 | 88.5 | 82.1 | 78.1 | 79.5 | 64.4 | 65.2 | 56.5 | 88.5 | 92.1 | 81.3 | 83.1 | 74.9 | 76.4 | 78.2 | 61.8 | 62.9 | 54.4 |
| w/o DMR | 93.8 | 96.2 | 89.1 | 90.3 | 84.7 | 80.2 | 81.6 | 67.5 | 68.8 | 60.3 | 90.4 | 94.2 | 84.8 | 86.6 | 79.1 | 79.1 | 80.2 | 65.3 | 67.0 | 58.4 |
| w/o CMA+DMR | 90.6 | 92.9 | 84.0 | 85.3 | 79.0 | 72.5 | 72.1 | 55.1 | 55.8 | 47.6 | 85.8 | 89.5 | 77.7 | 79.2 | 71.0 | 71.6 | 71.5 | 53.4 | 54.8 | 46.3 |
| w/o Multi-scale | 93.2 | 95.8 | 88.6 | 89.7 | 83.9 | 83.5 | 86.9 | 75.1 | 76.2 | 67.4 | 89.8 | 93.5 | 84.2 | 86.1 | 78.4 | 82.9 | 87.5 | 74.1 | 75.8 | 67.1 |
| w/o Causal Attn | 94.2 | 96.9 | 90.1 | 91.3 | 85.6 | 84.5 | 88.2 | 76.8 | 78.1 | 69.8 | 91.2 | 95.2 | 85.9 | 87.7 | 80.5 | 85.3 | 89.4 | 76.7 | 78.5 | 70.2 |
| w/o Multi-source | 94.0 | 96.7 | 89.9 | 91.1 | 85.5 | 86.7 | 90.7 | 79.7 | 80.9 | 73.0 | 91.0 | 95.0 | 85.6 | 87.5 | 80.3 | 86.9 | 89.9 | 78.4 | 80.4 | 72.6 |

**Model Complexity.** In Tab. 3, while improving segmentation performance, our method also meets the requirements for real-time inference.

### 3.3 Ablation Study

We conduct ablation studies on SUN-SEG to evaluate the effectiveness of each component, with results summarized in Tab. 4. **(1) CMA and DMR.** Removing CMA or DMR consistently degrades performance, with more severe drops on Hard and Unseen subsets. On SUN-SEG-Hard-Unseen, the Dice score decreases from 81.3% to 62.9% without CMA and to 67.0% without DMR. Removing both modules causes a drastic collapse to 54.8% Dice, indicating that causal temporal aggregation and dynamic reference modeling are essential and complementary. **(2) Multi-scale and Causal Design in CMA.** Using a single-scale interaction or replacing causal attention with standard cross-attention leads to clear performance degradation, demonstrating the importance of multi-scale aggregation and causal modeling for robust temporal representation. **(3) Multi-source**

**Reference.** Using only a single reference also results in inferior performance, validating the benefit of maintaining multiple complementary reference sources.

## 4 Conclusion

We presented CMSA-Net, a causal multi-scale spatio-temporal framework for VPS. By integrating CMA and DMR, CMSA-Net enhances semantic discrimination under challenging conditions. Extensive experiments on SUN-SEG demonstrate consistent superiority over state-of-the-art methods on both seen and unseen subsets, particularly in hard scenarios, while maintaining real-time efficiency, highlighting its practical value for clinical applications.

**Acknowledgments.** Tong Wang, Yaolei Qi, and Guanyu Yang are supported by the National Natural Science Foundation of China (T2541064, 82441021) and the Jiangsu Province Cutting-edge Technology R&D Program (BF2025628). This work is also supported by the Big Data Computing Center of Southeast University.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Biller, L.H., Schrag, D.: Diagnosis and treatment of metastatic colorectal cancer: a review. Jama **325**(7), 669–685 (2021)
2. Chai, J., Luo, Z., Gao, J., Dai, L., Lai, Y., Li, S.: Querynet: A unified framework for accurate polyp segmentation and detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 544–554. Springer (2024)
3. Chen, G., Yang, J., Pu, X., Ji, G.P., Xiong, H., Pan, Y., Cui, H., Xia, Y.: Mast: Video polyp segmentation with a mixture-attention siamese transformer. arXiv preprint arXiv:2401.12439 (2024)
4. Chen, G., Wu, H., Qin, J.: Stddnet: Harnessing mamba for video polyp segmentation via spatial-aligned temporal modeling and discriminative dynamic representation learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 21364–21373 (2025)
5. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4548–4557 (2017)
6. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. Scientia Sinica Informationis **6**(6) (2021)
7. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 263–273. Springer (2020)
8. Feng, Y., Gao, S., Yan, F., Song, Y., Hong, L., Hu, J., Zhang, W.: Scoring, remember, and reference: Catching camouflaged objects in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 13043–13052 (2025)

9. Fu, R., Hu, S., Zheng, X., Tang, C., Liu, X.: Polymamba: Spatial-prior guided mamba for polyp segmentation with high-frequency enhancement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 455–465. Springer (2025)
10. Gao, S., Cheng, M.M., Zhao, K., Zhang, X., Yang, M.H., Torr, P.H.S.: Res2net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 652–662 (2019)
11. Gu, Y., Wang, L., Wang, Z., Liu, Y., Cheng, M.M., Lu, S.P.: Pyramid constrained self-attention network for fast video salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 10869–10876 (2020)
12. Hu, Q., Yi, Z., Zhou, Y., Peng, F., Liu, M., Li, Q., Wang, Z.: Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–541. Springer (2024)
13. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2021)
14. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)
15. Jiang, W., Xin, L., Zhu, S., Liu, Z., Wu, J., Ji, F., Yu, C., Shen, Z.: Risk factors related to polyp miss rate of short-term repeated colonoscopy. Digestive diseases and sciences **68**(5), 2040–2049 (2023)
16. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3623–3632 (2019)
17. Lu, Y., Yang, Y., Xing, Z., Wang, Q., Zhu, L.: Diff-vps: Video polyp segmentation via a multi-task diffusion network with adversarial temporal reasoning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 165–175. Springer (2024)
18. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2014)
19. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 295–305. Springer (2020)
20. Wang, T., Qi, X., Yang, G.: Polyp segmentation via semantic enhanced perceptual network. IEEE Transactions on Circuits and Systems for Video Technology (2024)
21. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**, 415 – 424 (2021)
22. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 699–708. Springer (2021)
23. Wu, H., Zhao, Z.: Epsegnet: Lightweight semantic recalibration and assembly for efficient polyp segmentation. IEEE Transactions on Neural Networks and Learning Systems (2025)

24. Wu, H., Zhao, Z., Zhong, J., Wang, W., Wen, Z., Qin, J.: Polypseg+: A lightweight context-aware network for real-time polyp segmentation. IEEE Transactions on Cybernetics **53**(4), 2610–2621 (2022)
25. Xu, Z., Tang, F., Chen, Z., Zhou, Z., Wu, W., Yang, Y., Liang, Y., Jiang, J., Cai, X., Su, J.: Polyp-mamba: Polyp segmentation with visual mamba. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 510–521. Springer (2024)
26. Ye, J., Su, Y., Wu, Y., He, J., Zhuang, B., Chen, Z., Cai, J.: Fpn-in-fpn: A nested multi-scale aggregation network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 199–208. Springer (2025)
27. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–262. Springer (2020)
28. Zhao, Y., Wang, X., Yin, J.: Efficient video polyp segmentation by deformable alignment and local attention. IEEE Journal of Biomedical and Health Informatics (2025)
29. Zhou, J., Pang, Z., Wang, Y.X.: Rmem: Restricted memory banks improve video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18602–18611 (2024)
30. Zhu, Y., Lv, H., Chen, G., Zhang, Z., Jiang, H., Xia, Y.: Hybrid graph mamba: Unlocking non-euclidean potential for accurate polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 277–286. Springer (2025)