A40 (Fri-PM)

MICCAI 2025

28th International Conference on Medical Image Computing and Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA

MAMBA-Based Weakly Supervised Medical Image Segmentation with Cross-Modal Textual Information

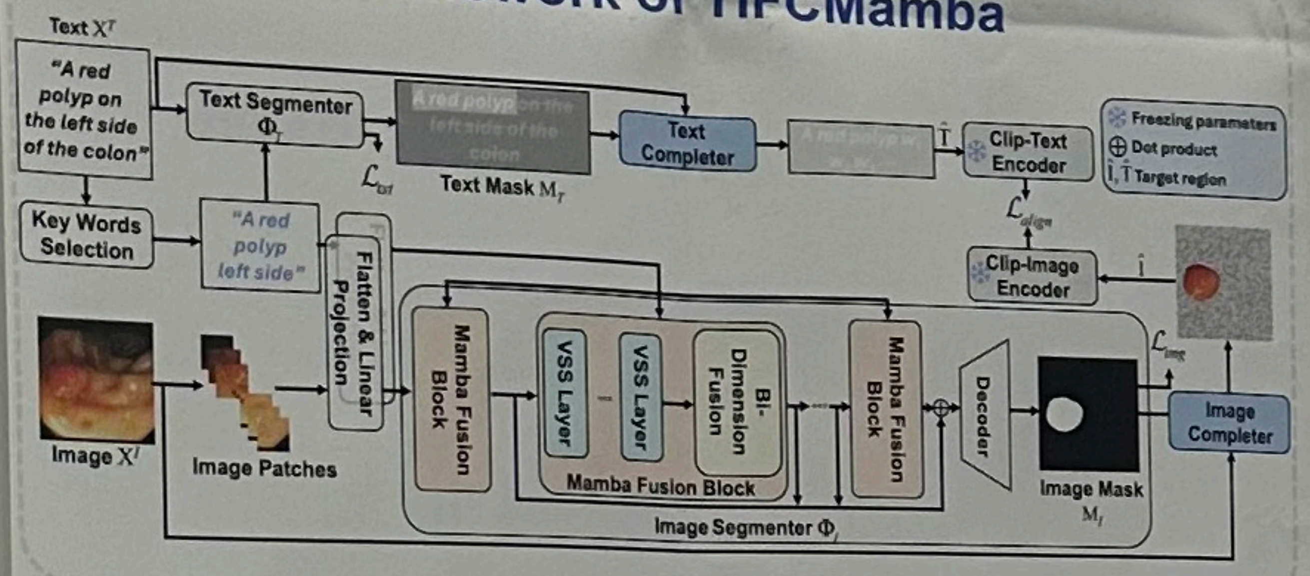
Zhen Pan, Wenhui Huang and Yuanjie Zheng

Shandong Normal University
School of Information Science and Engineering

Introduction

In medical image segmentation, obtaining pixel-level annotated data is costly. While semi-supervised and weakly-supervised methods reduce annotation dependence, they still require some pixel-level annotations. In contrast, leveraging textual descriptions corresponding to medical images as supervisory information for segmentation is more promising. Textual descriptions are easier to acquire, as users only need to provide location and appearance details of lesions. We present TIFCMamba, a Mamba-based architecture for text-image fusion segmentation. The framework processes images and texts in parallel to establish cross-modal correspondences, aligning CLIP-encoded features through contrastive learning. We propose Mamba Fusion module integrates text and image features through Bi-Dimension Fusion, enabling both intra-modal refinement and inter-modal interaction while preserving computational efficiency. Experiments on polyp and skin lesion datasets demonstrate competitive performance against fully supervised methods and state-of-the-art weakly-supervised approaches. Code and dataset will be available at https://github.com/silentyuchen/TIFCMamba.

Framework of TIFCMamba



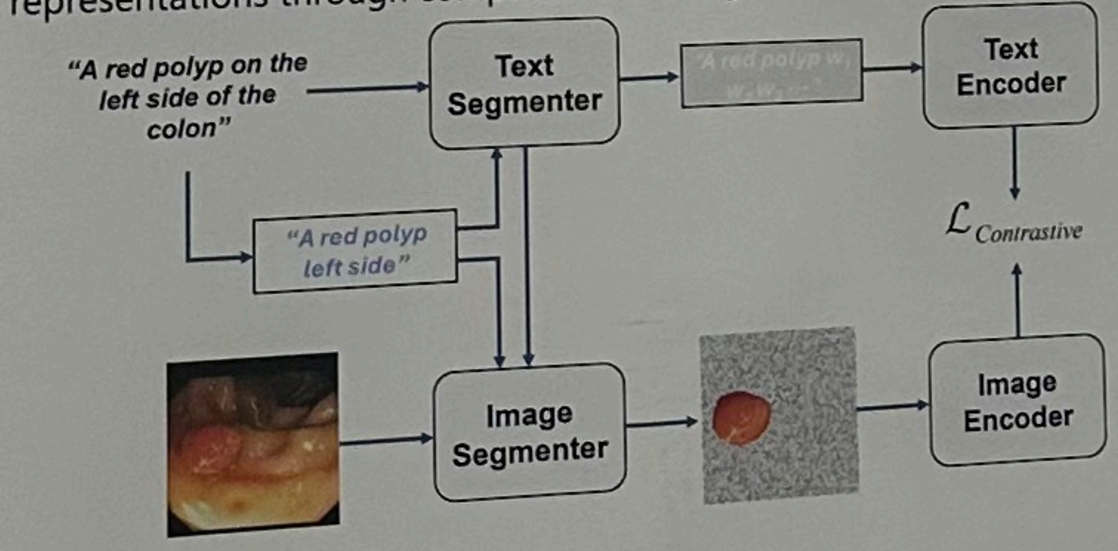
Main contributions:

- *Proposing the TIFCMamba framework.* A text-image fusion segmentation architecture based on Mamba has been designed to achieve text-supervised segmentation of medical images. This framework employs multimodal contrastive learning to reduce reliance on pixel-level annotations while circumventing the high computational costs of traditional Transformer models.
- *Designing the Mamba Fusion Module.* Bi-Dimension Fusion enables deep interaction between image and text features while maintaining computational efficiency, resolving Mamba's insufficient token interaction in multimodal feature fusion.
- *Introduction of Image-Text Mutual Alignment Mechanism.* Achieves precise alignment between local image regions and textual semantics during both training and testing phases, rectifying the inconsistency between global semantic alignment and local region alignment inherent in conventional text-supervised methods.

Methods

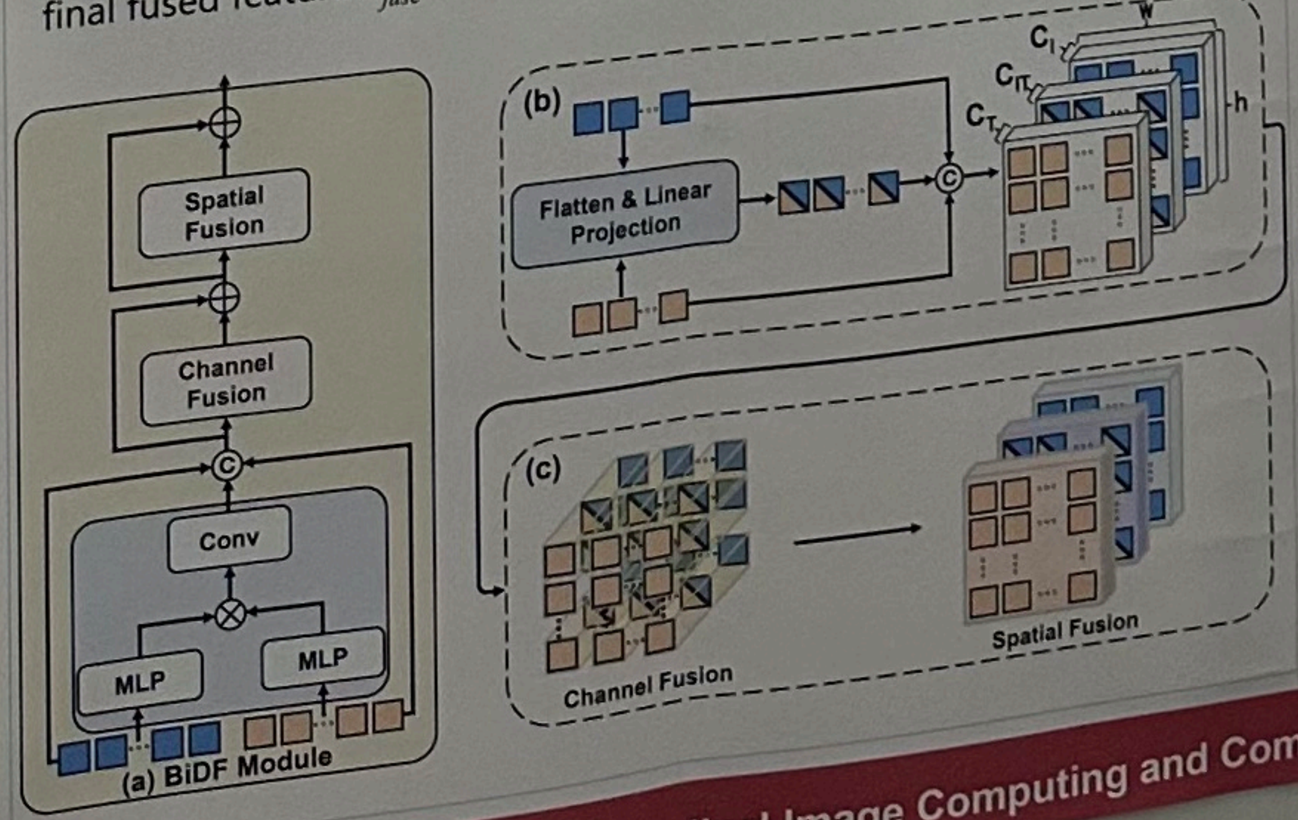
Overall Workflow

For each image-text pair (X_i^I, X_i^T), the keyword selector extracts the keyword W_T (e.g., "A red polyp left side") from X_i^T. The image segmenter Φ_I obtains the image feature e^I through X_i^I and W_T at the last layer of the Decoder, and dot product it with the text feature e^T output from the CLIP text encoder E_T to get the image mask M_I; whereas the text segmenter Φ_I extracts the keyword W_T (e.g., "a red polyp left side"). Φ_T processes X_i^T and W_T to generate the text mask M_T. Cropping X_I with M_I and randomly filling the background gives the mask image $\hat{I}$, and similarly M_T constructs the complete text $\hat{T}$. Finally, the image encoder E_I and the text encoder E_T of CLIP extract features from $\hat{I}$ and $\hat{T}$ respectively, and align their representations through comparative learning.



BiDF module

The BiDF module fuses cross-modal features by sequentially arranging image, image-text, and text features into a feature tensor, and performing fusion along both the spatial and channel dimensions through two SSM modules. In the first stage, text features are expanded and fused with image features to allow each image patch to incorporate textual information, and then concatenated along the channel dimension. In the second stage, the concatenated features undergo spatial fusion via a 2D selective scan and channel fusion via a 1D selective scan, producing the final fused feature F_{fuse}.



Experiments

Quantitative Analysis of TIFCMamba

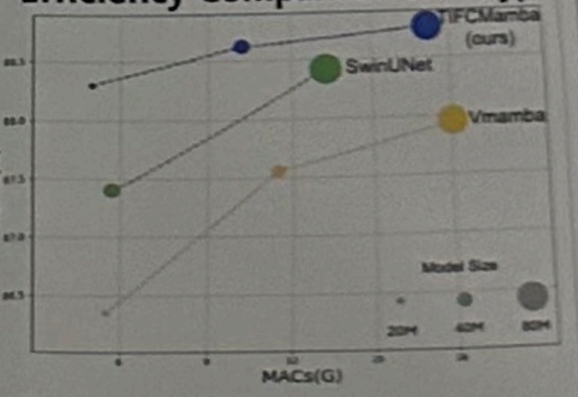
Method	ClinicDB		ColonDB		LaribPolypDB		ISIC2017	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
ResUNet (2020)	81.33	77.40	83.62	75.78	79.88	76.47	83.52	78.06
SwinUnet-T (2021)	86.64	82.24	85.90	83.76	88.43	80.05	85.47	81.35
WeakPolyp (2023)	84.30	81.56	86.67	79.89	82.79	80.60	85.41	82.67
TCL (2023)	84.35	80.89	85.02	81.67	85.58	78.32	85.46	81.90
SimSeg (2023)	85.17	80.38	84.92	80.16	85.60	79.73	87.28	82.49
SimTxtSeg (2024)	86.38	81.72	85.18	80.95	86.43	80.30	86.51	80.94
CoDe (2024)	86.98	82.45	86.58	81.45	87.35	81.07	87.09	83.51
XCoOp (2024)	86.55	82.43	85.73	80.18	88.31	80.61	86.36	81.52
TIFCMamba-T	87.50	81.53	87.38	80.93	87.67	81.09	87.20	83.13
TIFCMamba-S	88.07	83.93	87.67	81.45	87.81	81.77	87.79	83.59
TIFCMamba-B	88.24	84.22	87.74	82.56	88.92	82.43	87.95	83.76

- Our model demonstrates improvements in mDice and mIoU by **+1.26%** and **+1.77%, +1.16%** and **+0.89%**, **+0.61%** and **+1.36%**, and **+0.67%** and **+0.25%** on four datasets, respectively.

Ablation Study of Fusion Mode

Fusion Mode		Polyp	ISIC2017
Spatial	Channel	mDice	mDice
×	×	63.57	58.39
✓	×	72.36	70.95
×	✓	79.86	77.49
✓	✓	88.24	87.95

Efficiency Comparison on Polyp



Qualitative Analysis of TIFCMamba