Lightweight Temporal Models for Bolus Segmentation in VFSS

Yuanhan Chen
Affiliations, University of Toronto

Motivation

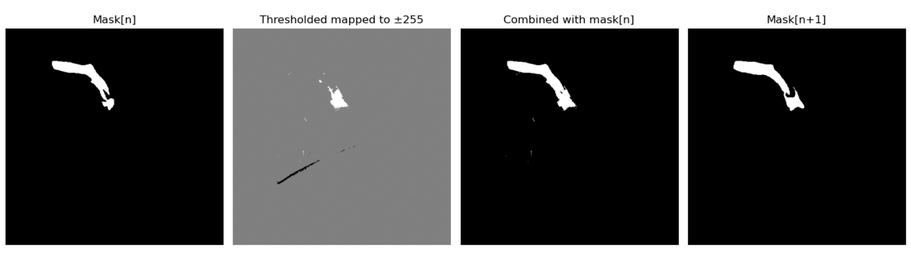
Accurate bolus segmentation in Videofluoroscopic Swallow Studies (VFSS) is essential for analyzing swallowing safety, penetration/aspiration risk, and residue. However, the task is difficult because VFSS frames suffer from:

- low contrast and overlapping anatomical structures,
- motion blur and translucent appearance of tissues,

Key observation from exploratory analysis:

Even a simple frame-to-frame difference contains strong temporal cues about the moving bolus. When thresholded and combined with the previous mask, we obtain a shape that closely resembles the next ground-truth mask. This motivated all later designs involving:

- temporal fusion,
- clean-frame reconstruction, and
- lightweight spatiotemporal heads.



Related Work

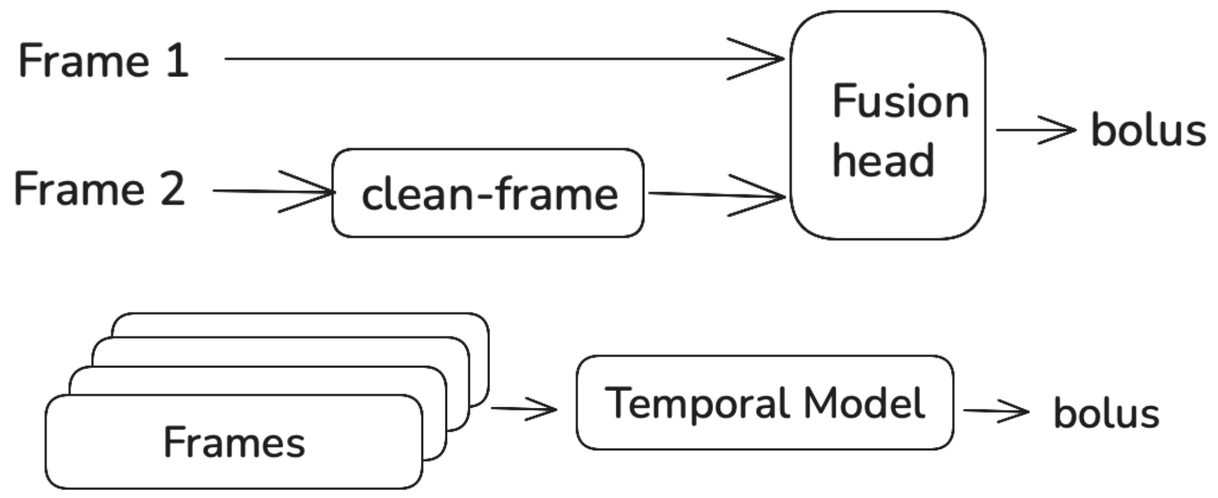
- Single-frame models: U-Net and UNet++ are common for VFSS segmentation. Lightweight variants (e.g., MobileNetV2-UNet) reach near 0.8 Dice, improve speed but still use >6M parameters.
- Temporal transformer methods: Video-TransUNet and Video-SwinUNet reach Dice ≈0.87–0.89, but rely on heavy transformer backbones, making real-time use impractical.
- Preprocessing-based methods: PECI-Net shows that contrast enhancement and edge cues help in low-contrast VFSS settings.
- Gap: Prior work focuses either on heavy high-accuracy models or single-frame lightweight baselines.
- Lightweight temporal or two-stage designs for bolus segmentation remain underexplored.

References

[1] Li, W., Mao, S., Mahoney, A., Coyle, J., Sejdić, E. Deep Learning Models for Bolus Segmentation in Videofluoroscopic Swallow Studies, Journal of Real-Time Image Processing, 2024
[2] Fakhry, A., et al. Deep Learning for Video Fluoroscopic Swallowing Study Analysis, IEEE Access, 2025
[3] Kim, I. J., et al. PECI-Net: Preprocessing Ensemble and Contrast Integration for Bolus Segmentation in VFSS, arXiv, 2024

New Technique

- We develop lightweight (~1M parameter) segmentation models suitable for efficient VFSS analysis on resource-limited systems.
- Prior VFSS methods (e.g., Video-SwinUNet) achieve high Dice (≈0.87–0.89) but rely on heavy transformer backbones. Even “small” MobileNetV2-UNet variants still exceed 6M parameters.
- Our approach focuses on efficient spatial–temporal fusion using:
- Two-stage pipeline: clean-frame reconstruction → lightweight fusion head.
- Lite temporal model: a compact Temporal Context Module (TCM).



Experimental Results

Lightweight temporal or two-stage models achieve Dice ≈0.6 with significantly fewer parameters than state-of-the-art models.
Near MobileNet result and 6x smaller.

Model	Clean Stage	Fusion Head	Seg (IoU / Dice / AUC)	Clean L1 / PSNR
Two-UNet Pipeline	SmallUNet (masked L1)	SmallUNet	**0.519 / 0.628 / 0.995**	
ViT Two-Stage	ViT encoder–decoder (masked L1)	Windowed attn + conv	**0.486 / 0.609 / 0.988**	**0.0077 / 33.37 dB**
Cross-Attention Two-Stage	ViT decoder (L1)	Cross-frame attention	**0.448 / 0.578 / 0.989**	**0.0112 / 31.36 dB**
UNet++ Two-Stage	TinyUNet (L1 + SSIM)	UNet++	**0.471 / 0.602 / 0.990**	**0.0056 / 40.23 dB**
Masked-AE Two-Stage	Conv autoencoder (masked L1)	Channel + spatial attention	**0.468 / 0.595 / 0.985**	**0.0059 / 38.15 dB**
Masked-UNet Two-Stage	TinyUNet (masked L1)	Channel + spatial attention	**0.435 / 0.561 / 0.977**	**0.0055 / 41.18 dB**

Model	Architecture	Seg (IoU / Dice / AUC)
TCM-UNet	Temporal Context Module UNet	**0.515 / 0.644 / 0.996**
UNet Baseline	MobileNetV2 encoder + unet	**0.542 / 0.669 / 0.995**