Distilling Temporal Coherence into 2D Networks for Transrectal Ultrasound Prostate Video Segmentation

Dong Yeong Kim*[1,2], JunGyu Lee*[3†], Jaewon Choi[2], June Young Seo[4], Myeongseop Kim[2], Jinwook Choi[1], Taek Min Kim[4‡], and Young-Gon Kim[2‡]

[1] Interdisciplinary Program in Bioengineering, Seoul National University
[2] Department of Transdisciplinary Medicine, Seoul National University Hospital
[3] Department of Artificial Intelligence, Yonsei University
[4] Department of Radiology, Seoul National University Hospital
steven6774@snu.ac.kr, jungyu.lee@yonsei.ac.kr

Abstract. Real-time video segmentation of the prostate in Transrectal Ultrasound (TRUS) is essential for image-guided interventions. While conventional 2D methods suffer from inter-frame inconsistencies by disregarding temporal context, 3D architectures incur prohibitive latency. To resolve this dilemma, we present a Temporally Consistent Learning Framework that distills temporal coherence into a 2D network during training, preserving single-frame inference efficiency. Our design is driven by a key clinical observation: the prostate exhibits geometric stability, whereas the surrounding acoustic environment fluctuates due to physiological motion and transducer pressure. Because conventional temporal constraints propagate erroneous gradients from these unstable regions, we introduce a Confidence-Weighted Temporal Consistency objective derived from optical flow warping residuals, selectively attenuating contributions from unreliable regions. Complementing this pixel-wise constraint, a Dual-scale Prototype Alignment Module enforces semantic coherence through contrastive optimization of local boundary and global semantic features. Furthermore, to eliminate the need for dense per-frame video annotations, we employ geometric equivariance-based pseudo-labeling with knowledge distillation from a pretrained teacher. Extensive experiments on SUN-SEG and our newly introduced TRUS-V benchmark (2,679 frames) demonstrate state-of-the-art accuracy and temporal consistency at real-time speed. Code and dataset are available at https://github.com/DYDevelop/DTC-TRUS.

Keywords: Video Prostate Segmentation · Temporal Consistency · Self-Supervised Learning · Prototype Alignment · Ultrasound.

1 Introduction

Prostate cancer remains a leading cause of male malignancy worldwide [21], necessitating reliable imaging for early diagnosis and intervention. Transrectal

* Equal contribution ‡ Corresponding Author † Work done prior to joining[3]

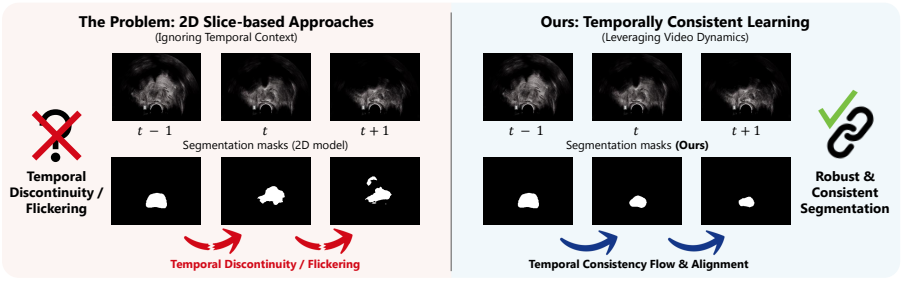

Fig. 1. Impact of the proposed consistency learning. Illustration of the limitations of 2D slice-based methods (left) versus our temporally consistent approach (right). Our method significantly reduces flickering artifacts and improves boundary robustness in challenging prostate ultrasound videos.

Ultrasound (TRUS) has established itself as the standard modality for procedures such as biopsy and brachytherapy, owing to its real-time capability, cost-effectiveness, and radiation-free nature [15]. In clinical workflows, accurate segmentation of the prostate boundary from TRUS videos is a prerequisite for precise volume estimation and treatment planning. However, relying on manual delineation is labor-intensive and prone to inter-observer variability. This challenge is further exacerbated by the inherent artifacts of TRUS, such as low contrast, severe speckle noise, and ambiguous boundaries, which complicate the development of robust real-time automatic segmentation systems [24].

Recent advances in deep learning have achieved remarkable success in medical image segmentation [19]. However, most existing approaches for prostate segmentation treat video data as a set of independent 2D static images [12, 14, 5]. While computationally efficient, these 2D slice-based models inherently neglect temporal context, leading to unstable predictions across adjacent frames, commonly manifesting as "flickering" artifacts that are unacceptable for clinical applications requiring smooth tracking (Fig. 1). To capture spatiotemporal features, 3D CNNs and Recurrent Neural Networks have been employed [17, 26, 28, 18]. However, these methods incur high computational costs, rendering them impractical for intra-operative procedures where low-latency feedback is strictly mandated [24]. To balance accuracy and efficiency, a third line of work augments a 2D feature extractor with a dedicated temporal-interaction module [7, 27]. Nevertheless, such designs still introduce auxiliary temporal modules at inference and rely on temporally annotated video for supervision, leaving the trade-off between runtime and annotation cost unresolved.

In this work, we propose a **Temporally Consistent Learning Framework** that distills temporal coherence into a standard 2D network during training, preserving single-frame inference efficiency. A key observation motivating our design is that the prostate exhibits geometric stability, whereas the surrounding acoustic environment fluctuates substantially due to physiological motion and transducer pressure. Consequently, naive temporal constraints propagate erroneous gradients from unreliable background regions, degrading rather than improving consistency.

Specifically, we introduce two synergistic mechanisms. First, a **Confidence-Weighted Temporal Consistency** objective incorporates a weighting scheme derived from optical flow warping residuals. This mechanism selectively attenuates contributions from dynamically unstable regions, concentrating supervisory signals on the anatomically consistent prostate structure. Second, a **Dual-scale Prototype Alignment Module** enforces semantic coherence in the latent space through contrastive optimization of both local boundary features and global semantic representations across temporal neighbors, ensuring robustness against speckle noise and intensity variations.

To alleviate the burden of dense per-frame video annotation, we employ a self-supervised strategy leveraging geometric equivariance for pseudo-label generation. Additionally, a Student-Teacher architecture with knowledge distillation mitigates catastrophic forgetting of spatial priors learned from labeled static images. Furthermore, we introduce **TRUS-V**, a large-scale video benchmark comprising 2,679 annotated frames across axial and sagittal views from 10 patients, thereby addressing the scarcity of publicly available video datasets in this domain.
Our main contributions are summarized as follows:

- We propose a Temporally Consistent Learning Framework that distills temporal coherence into a 2D segmentation model during training, eliminating the need for heavy 3D computations or temporal-interaction modules at inference while achieving temporally stable predictions.
- We introduce Confidence-Weighted Temporal Consistency and Dual-scale Prototype Alignment, which synergistically suppress flickering artifacts and enhance boundary robustness by focusing on stable anatomical structures.
- We release TRUS-V, a large-scale benchmark comprising 2,679 densely annotated frames (Axial/Sagittal) from 10 patients. Extensive experiments on TRUS-V and SUN-SEG demonstrate state-of-the-art performance in both accuracy and temporal coherence.

2 Datasets

Newly Collected TRUS-V Benchmark. To address the lack of multi-view video datasets for prostate segmentation, we introduce TRUS-V, an in-house benchmark comprising 2,679 frames collected from 10 patients. Unlike existing image-based datasets, TRUS-V captures temporal dynamics through 20 continuous video sequences, consisting of paired Axial and Sagittal views for each patient. To ensure rigorous evaluation, we partitioned the dataset at the patient level, yielding 2,400 training and 279 testing frames. Given the prohibitive cost of dense manual annotation for video data, we established the ground truth using a semi-automated annotation strategy. First, we utilized a separate static dataset containing 2,140 Axial and 2,260 Sagittal images to train a 5-fold ensemble U-Net, which achieved a high mean Dice Similarity Coefficient (DSC) of 0.95 (Axial) and 0.93 (Sagittal). This model generated initial candidate masks for the TRUS-V video sequences. Subsequently, experienced radiologists visually

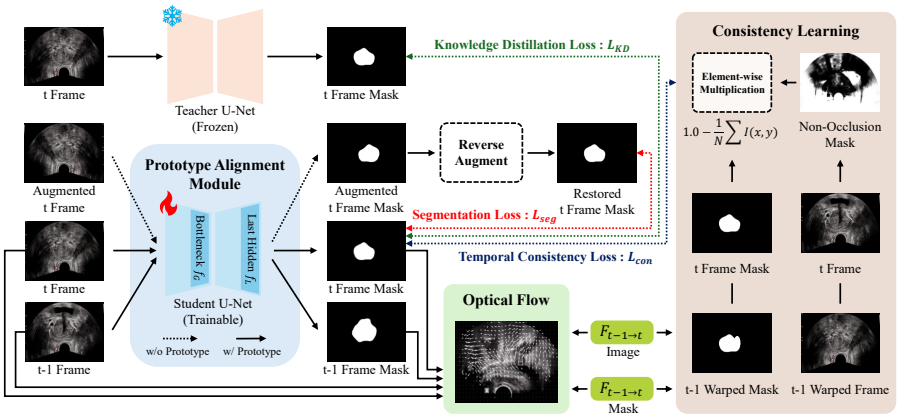

Fig. 2. Overview of the proposed framework. The network distills temporal coherence via Confidence-Weighted Temporal Consistency and Dual-scale Prototype Alignment. We employ a self-supervised Student-Teacher strategy to utilize unlabeled videos via pseudo-labeling and knowledge distillation. Crucially, only the efficient 2D Student is required during inference, ensuring real-time performance.

inspected and manually refined these masks frame-by-frame to correct inaccuracies. This human-in-the-loop process ensures the high reliability of our reference standard.

Public Dataset: SUN-SEG. To assess generalization beyond the ultrasound modality, we utilized SUN-SEG [11], a large-scale Video Polyp Segmentation benchmark (158,690 frames), characterized by diverse challenges such as fast motion and occlusion. We followed the dataset's standard training and evaluation protocols to validate that our Temporally Consistent Learning Framework is effective for general video segmentation tasks requiring temporal coherence.

3 Methods

3.1 Overview of the Framework

To achieve real-time video segmentation without the computational overhead of 3D networks, we propose a Temporally Consistent Learning Framework (see Fig. 2) that empowers a standard 2D backbone to capture video dynamics. Given an unlabeled video sequence $\mathcal{V} = \{I_1, \ldots, I_T\}$, our objective is to train a Student model $\mathcal{S}(\cdot; \theta_S)$ that generates accurate and consistent masks $P_t = \mathcal{S}(I_t)$ without frame-wise annotations. To achieve this, we employ a Student-Teacher architecture enforcing coherence at two synergistic levels: (1) pixel-level geometric alignment via optical flow to explicitly model tissue deformation, and (2) feature-level semantic alignment via prototypes to robustly suppress speckle noise. This bi-level strategy ensures robust temporal representation learning while maintaining high inference efficiency.

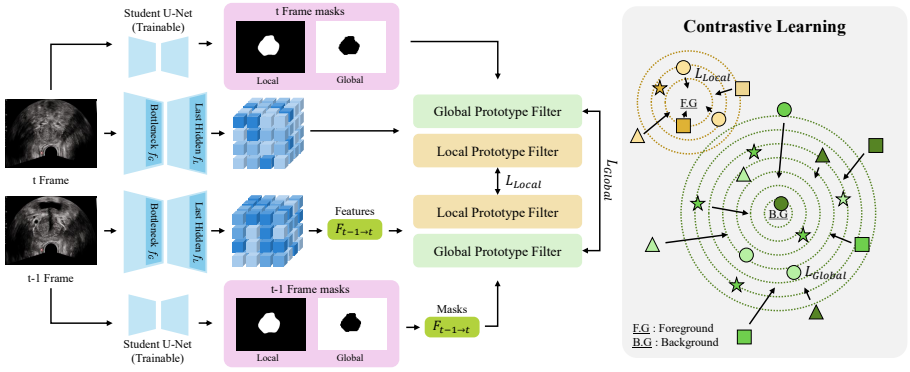

Fig. 3. Dual-scale Prototype Alignment Module. We extract global and local prototypes to enforce semantic consistency. A temporal alignment objective (Right) optimizes intra-class compactness by matching features across adjacent frames, ensuring robust target boundaries and scene stability against video artifacts.

3.2 Self-Supervised Equivariance and Distillation

To circumvent the prohibitive cost of dense per-frame video annotation, we adopt a self-supervised strategy anchored by geometric equivariance [1] and knowledge distillation [6]. Notably, the video adaptation stage requires no frame-wise video labels; spatial priors are instead inherited from a Teacher pre-trained solely on the separate static image dataset. First, relying on the hypothesis that segmentation should be equivariant to geometric transformations, we generate pseudo-labels $\hat{Y}_t$ by applying the inverse transformation to predictions of a flipped input $\mathcal{A}(I_t)$. The student minimizes $\mathcal{L}_{seg} = \mathcal{H}(P_t, \text{stop\_grad}(\hat{Y}_t))$, where $\mathcal{H}$ denotes the BCE-Dice loss and the stop-gradient prevents mode collapse [2]. Simultaneously, to mitigate catastrophic forgetting of spatial representations, we distill knowledge from a frozen static Teacher (sharing the identical architecture with the Student) using the ℓ_2-based objective $\mathcal{L}_{KD} = ||\sigma(Z_S/\tau) - \sigma(Z_T/\tau)||_2^2$, where σ represents the sigmoid function, τ is the temperature parameter, and Z_S and Z_T denote the student and teacher logits, respectively. This effectively regularizes training, ensuring the model retains robust anatomical priors while adapting to video dynamics.

3.3 Confidence-Weighted Temporal Consistency

To explicitly model the temporal continuity of the prostate anatomy, we introduce a motion-aware consistency loss. We estimate the optical flow field $\mathcal{F}_{t-1\to t}$ between adjacent frames I_{t-1} and I_t using a dense flow estimation algorithm (specifically, Farneback [4]). The predicted mask of the previous frame P_{t-1} is warped to the current coordinate system using a warping operation $\mathcal{W}$ [8]:

$$\tilde{P}_{t-1\to t} = \mathcal{W}(P_{t-1}, \mathcal{F}_{t-1\to t}) \tag{1}$$

We enforce the warped mask $\tilde{P}_{t-1\to t}$ to be consistent with the current prediction P_t. To handle occlusions or motion estimation errors, we compute a non-occlusion

confidence map $M_{noc} = \exp(-|I_t - \mathcal{W}(I_{t-1}, \mathcal{F}_{t-1\to t})|)$. Unlike standard weighted ℓ_1 losses, we formulate the objective to maximize the structural similarity in reliable regions [13]. The temporal consistency loss is defined as:

$$\mathcal{L}_{con} = 1 - \frac{1}{N}\sum_{p} M_{noc}(p) \cdot \Big(1 - |P_t(p) - \tilde{P}_{t-1\to t}(p)|\Big) \tag{2}$$

where N is the total number of pixels and p denotes the pixel index. This formulation minimizes the loss when the predictions are consistent ($|P_t - \tilde{P}_{t-1\to t}| \approx 0$) in regions with high confidence ($M_{noc} \approx 1$), while effectively ignoring gradients in occluded or unstable regions where $M_{noc} \approx 0$.

3.4 Dual-scale Prototype Alignment Module

Pixel-level consistency alone may be insufficient under severe speckle noise or acoustic shadows. Therefore, we propose a Dual-scale Prototype Alignment Module to enforce semantic consistency across different feature granularities. We extract prototypes from both Global (Bottleneck) and Local (Decoder) features of the Student network (see Fig. 3).

Prototype Extraction. Let $f_t \in \{f_t^{loc}, f_t^{glob}\}$ be the local or global features. A prototype $v_{c,t}$ for class $c \in \{c_f, c_b\}$ is computed via masked average pooling [22]:

$$v_{c,t} = \frac{\sum_{x,y} \mathbb{1}_c(x,y) \cdot f_t(x,y)}{\sum_{x,y} \mathbb{1}_c(x,y) + \epsilon} \tag{3}$$

where $\mathbb{1}_c(\cdot)$ is a binary indicator isolating the foreground (c_f) or background (c_b) via an adaptive threshold.

Dual-scale Contrastive Alignment. To explicitly model multi-scale dynamics, we align c_f prototypes on local features (preserving boundaries) and c_b prototypes on global features (stabilizing scenes). We warp previous local features f_{t-1}^{loc} to t ($\tilde{v}^{loc}$) and directly compare spatially sparse global ones, whose high-level semantics are naturally motion-invariant—to maximize intra-class cosine similarity [25]. To handle scale variations dynamically, we cross-weight using background (w_{c_b}) and foreground (w_{c_f}) mask fractions:

$$\mathcal{L}_{proto} = w_{c_b}\big(1 - \cos(\tilde{v}_{c_f,t-1}^{loc}, v_{c_f,t}^{loc})\big) + w_{c_f}\big(1 - \cos(v_{c_b,t-1}^{glob}, v_{c_b,t}^{glob})\big) \tag{4}$$

This elegantly balances alignment; smaller targets (low w_{c_f}) receive higher local boundary weights (w_{c_b}), enforcing robust intra-class compactness.

3.5 Total Objective Function

The final objective function is a weighted sum of the aforementioned losses:

$$\mathcal{L}_{total} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{KD}\mathcal{L}_{KD} + \lambda_{con}\mathcal{L}_{con} + \lambda_{proto}\mathcal{L}_{proto} \tag{5}$$

where λ_{seg}, λ_{KD}, λ_{con}, and λ_{proto} are balancing coefficients. During inference, only the Student 2D network is employed, maintaining the computational efficiency of standard 2D models without calculating optical flow or teacher inference.

4 Experiments

4.1 Implementation Details

Our framework is backbone-agnostic; we instantiate it with U-Net++ [32] on TRUS-V and ACSNet [29] on SUN-SEG, all trained on a single NVIDIA A100 40GB. The static Teacher is pre-trained on a separate static image dataset and frozen, after which the Student is adapted on unlabeled videos with AdamW (learning rate 1×10^{-6}, batch size 8) for one epoch. We set λ_{seg}=λ_{KD}=λ_{proto}=1.0, and λ_{con}=0.25.

Table 1. Quantitative comparison on SUN-SEG dataset. The best and second-best results are highlighted in **bold** and underlined, respectively.

	Method	SUN-SEG-Easy (Unseen)						SUN-SEG-Hard (Unseen)					
		$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	$F_\beta^{mn} \uparrow$	Dice ↑	Sen ↑	$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	$F_\beta^{mn} \uparrow$	Dice ↑	Sen ↑
Image Models	U-Net [19]	0.669	0.677	0.459	0.528	0.530	0.420	0.670	0.679	0.457	0.527	0.542	0.429
	U-Net^{++} [32]	0.684	0.687	0.491	0.553	0.559	0.457	0.685	0.697	0.480	0.544	0.554	0.467
	ACSNet [29]	0.782	0.779	0.642	0.688	0.713	0.601	0.793	0.787	0.636	0.684	0.708	0.618
	PraNet [3]	0.733	0.753	0.572	0.632	0.621	0.524	0.717	0.735	0.544	0.607	0.598	0.512
	ColonSegNet [9]	0.776	0.768	0.621	0.666	0.683	0.594	0.789	0.770	0.636	0.662	0.685	0.599
	MedNeXt [20]	0.703	0.719	0.553	0.598	0.591	0.506	0.699	0.701	0.522	0.575	0.611	0.508
	MSRF-Net [23]	0.794	0.780	0.652	0.693	0.701	0.600	0.774	0.762	0.637	0.658	0.701	0.605
	TransNetR [10]	0.768	0.768	0.648	0.694	0.642	0.592	0.766	0.772	0.629	0.673	0.628	0.591
	Ours	0.816	**0.882**	**0.738**	**0.784**	0.746	0.719	**0.816**	**0.878**	**0.719**	**0.758**	**0.737**	**0.741**
Video Models	SSTAN [30]	0.805	0.838	0.691	0.745	0.726	0.662	0.801	0.733	0.682	0.734	0.718	0.676
	FLA-Net [16]	0.722	0.697	0.547	0.597	0.636	0.506	0.721	0.701	0.540	0.592	0.628	0.522
	DALA [31]	**0.837**	0.854	0.722	0.772	**0.768**	**0.721**	0.808	0.840	0.676	0.735	0.733	0.669

4.2 Comparisons with State-of-the-art Methods

Performance on Public Benchmark. We evaluate our framework against SOTA methods on the SUN-SEG dataset [11]. Following the standard protocol, we report Structure (S_α), Enhanced-alignment (E_ϕ^{mn}), weighted F-measures (F_β^w, F_β^{mn}), Dice, and Sensitivity (Sen). For fair comparison, all 2D models follow standard frame-by-frame inference. As Table 1 shows, our method not only significantly outperforms 2D baselines but also rivals or surpasses heavy video-based models. This confirms our strategy effectively transfers video dynamics to a lightweight 2D backbone, achieving real-time inference (89.95 FPS via ACSNet [29]) without heavy 3D computations.

Performance on Clinical TRUS-V. We further validate our method on the in-house TRUS-V benchmark (Table 2). Interestingly, we observed that existing video segmentation models (SSTAN [30], FLA-Net [16], DALA [31]), which generally excel in colonoscopy video scenes, struggle to outperform strong 2D baselines in the ultrasound domain. This performance degradation is likely attributed to the severe speckle noise and ambiguous boundaries inherent in TRUS, which disrupt standard temporal attention mechanisms. In contrast, our framework effectively handles these domain-specific challenges, achieving state-of-the-art performance across metrics at a real-time 127.97 FPS (U-Net^{++} [32]). These quantitative gains are corroborated by the visual comparisons in Fig. 4; while competing

Table 2. Quantitative comparison on the TRUS-V dataset.

	Method	TRUS-V					
		$S_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^w \uparrow$	$F_\beta^{mn} \uparrow$	Dice ↑	Sen ↑
Image Models	U-Net	0.964	0.985	0.808	0.814	0.817	0.829
	U-Net^{++}	**0.967**	**0.988**	0.821	0.820	0.820	0.827
	ACSNet	0.965	0.983	0.802	0.827	0.820	0.813
	PraNet	0.932	0.934	0.751	0.753	0.749	0.746
	ColonSegNet	0.959	0.981	0.806	0.810	0.808	0.812
	MedNeXt	0.886	0.924	0.771	0.799	0.795	0.832
	MSRF-Net	0.965	0.987	0.821	0.816	0.821	0.838
	TransNetR	0.964	0.982	0.815	0.817	0.816	0.821
	Ours	**0.967**	**0.988**	**0.830**	**0.828**	**0.829**	**0.839**
Video Models	SSTAN	0.963	0.980	0.816	0.814	0.819	0.837
	FLA-Net	0.963	0.978	0.789	0.817	0.811	0.809
	DALA	0.908	0.931	0.539	0.738	0.729	0.746

Table 3. Ablation study on the SUN-SEG-Easy (Unseen) dataset to investigate the effectiveness of each proposed component. Proto. refers to the prototype, and △ indicates partial (single-scale) usage.

Method	L_{seg}	L_{KD}	L_{proto}	L_{con}	$S_\alpha \uparrow$	**Dice ↑**
Baseline	✓	-	-	-	0.274	0.171
w/o Temporal	✓	✓	-	-	0.793	0.701
w/o Consistency	✓	✓	✓	-	0.813	0.735
w/o Proto.	✓	✓	-	✓	0.810	0.722
w/o Global Proto.	✓	✓	△	✓	0.808	0.721
w/o Local Proto.	✓	✓	△	✓	0.808	0.722
Proposed	✓	✓	✓	✓	**0.816**	**0.746**

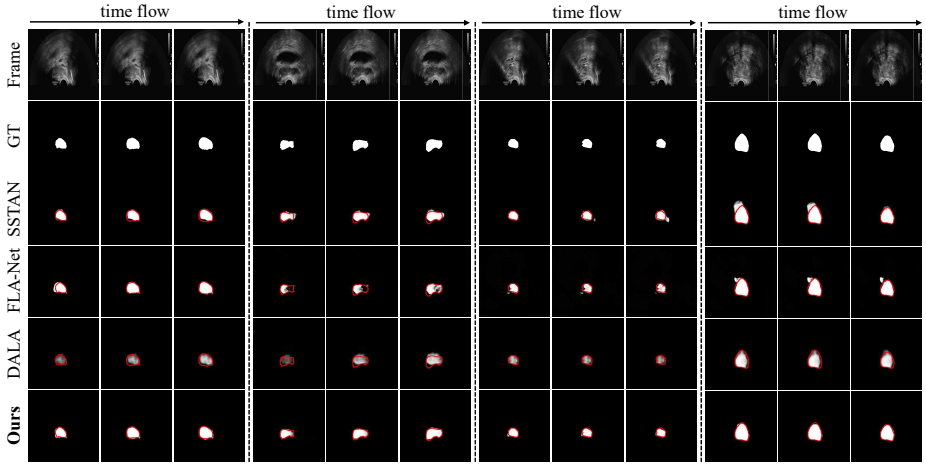

Fig. 4. Visual comparison on TRUS-V. Our method maintains robust boundaries aligned with the ground truth (red contours) across consecutive frames even under acoustic shadows, whereas competitors suffer from under-segmentation.

methods often produce intermittent false negatives under acoustic shadows or rapid probe motion, our model maintains smooth and anatomically plausible boundaries. This confirms that our temporal coherence distillation successfully learns robust representations that are resilient to signal dropouts.

4.3 Ablation Studies

As shown in Table 3, $\mathcal{L}_{KD}$ stabilizes the noisy baseline (Dice 0.171→0.701) but lacks temporal coherence. Individually, the pixel-level $\mathcal{L}_{con}$ and the dual-scale $\mathcal{L}_{proto}$ raise Dice to 0.722 and 0.735, respectively. Here, △ denotes single-scale usage: local-only in "w/o Global Proto." and global-only in "w/o Local Proto." These fail to improve over $\mathcal{L}_{con}$ (0.721/0.722 vs. 0.722), as local prototypes are noise-sensitive while global ones are too coarse for boundaries. Their area-adaptive combination yields the best performance (S_α 0.816, Dice 0.746).

5 Conclusion

In this paper, we presented a novel Temporally Consistent Learning Framework for robust prostate video segmentation. By successfully distilling temporal coherence through motion-aware consistency and prototype alignment, our approach effectively transfers temporal knowledge to a lightweight 2D backbone, solving the flickering problem inherent in slice-based models. Crucially, our self-supervised consistency strategy, empowered by a Student-Teacher architecture, enables the effective utilization of unlabeled video data while mitigating catastrophic forgetting. We also released TRUS-V, a comprehensive multi-view benchmark, to facilitate future research in this domain. Extensive experiments on both TRUS-V and SUN-SEG datasets demonstrated that our method achieves state-of-the-art performance, offering a promising solution for real-time, temporally stable image-guided prostate intervention. Future work will focus on extending this framework to 3D volume reconstruction and real-time biopsy guidance systems.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (Grant No. RS-2025-00553835 and RS-2024-00343630), and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No. RS-2025-02307233).
Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., De Bruijne, M.: Semi-supervised medical image segmentation via learning consistency under transformations. In: International conference on medical image computing and computer-assisted intervention. pp. 810–818. Springer (2019)
2. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
3. 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)
4. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. pp. 363–370. Springer (2003)
5. Ghavami, N., Hu, Y., Bonmati, E., Rodell, R., Gibson, E., Moore, C., Barratt, D.: Integration of spatial information in convolutional neural networks for automatic segmentation of intraoperative transrectal ultrasound images. Journal of Medical Imaging **6**(1), 011003 (2019)
6. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

7. 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)
8. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
9. Jha, D., Ali, S., Tomar, N.K., Johansen, H.D., Johansen, D., Rittscher, J., Riegler, M.A., Halvorsen, P.: Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. Ieee Access **9**, 40496–40510 (2021)
10. Jha, D., Tomar, N.K., Sharma, V., Bagci, U.: Transnetr: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In: Medical Imaging with Deep Learning. pp. 1372–1384. PMLR (2024)
11. 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)
12. Karimi, D., Zeng, Q., Mathur, P., Avinash, A., Mahdavi, S.S., Spadinger, I., Abolmaesumi, P., Salcudean, S.E.: Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. Medical image analysis **57**, 186–196 (2019)
13. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018)
14. Lei, Y., Tian, S., He, X., Wang, T., Wang, B., Patel, P., Jani, A.B., Mao, H., Curran, W.J., Liu, T., et al.: Ultrasound prostate segmentation based on multidirectional deeply supervised v-net. Medical physics **46**(7), 3194–3206 (2019)
15. Lei, Y., Wang, T., Wang, B., He, X., Tian, S., Jani, A.B., Mao, H., Curran, W.J., Patel, P., Liu, T., et al.: Ultrasound prostate segmentation based on 3d v-net with deep supervision. Medical Imaging 2019: Ultrasonic Imaging and Tomography **10955**, 198–204 (2019)
16. Lin, J., Dai, Q., Zhu, L., Fu, H., Wang, Q., Li, W., Rao, W., Huang, X., Wang, L.: Shifting more attention to breast lesion segmentation in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 497–507. Springer (2023)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
18. Orlando, N., Gillies, D.J., Gyacskov, I., Romagnoli, C., D'Souza, D., Fenster, A.: Automatic prostate segmentation using deep learning on clinically diverse 3d transrectal ultrasound images. Medical physics **47**(6), 2413–2426 (2020)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. pp. 234–241. Springer (2015)
20. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
21. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: a cancer journal for clinicians **73**(1), 17–48 (2023)
22. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)

23. Srivastava, A., Jha, D., Chanda, S., Pal, U., Johansen, H.D., Johansen, D., Riegler, M.A., Ali, S., Halvorsen, P.: Msrf-net: A multi-scale residual fusion network for biomedical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(5), 2252–2263 (2021)
24. Wang, H., Wu, H., Wang, Z., Yue, P., Ni, D., Heng, P.A., Wang, Y.: A review of image processing methods in prostate ultrasound. arXiv preprint arXiv:2407.00678 (2024)
25. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206 (2019)
26. Wang, Y., Dou, H., Hu, X., Zhu, L., Yang, X., Xu, M., Qin, J., Heng, P.A., Wang, T., Ni, D.: Deep attentive features for prostate segmentation in 3d transrectal ultrasound. IEEE transactions on medical imaging **38**(12), 2768–2778 (2019)
27. Xu, H., Yang, Y., Aviles-Rivero, A.I., Yang, G., Qin, J., Zhu, L.: Lgrnet: Local-global reciprocal network for uterine fibroid segmentation in ultrasound videos. arXiv preprint arXiv:2407.05703 (2024)
28. Yang, X., Yu, L., Wu, L., Wang, Y., Ni, D., Qin, J., Heng, P.A.: Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
29. 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)
30. Zhao, X., Wu, Z., Tan, S., Fan, D.J., Li, Z., Wan, X., Li, G.: Semi-supervised spatial temporal attention network for video polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 456–466. Springer (2022)
31. Zhao, Y., Wang, X., Yin, J.: Efficient video polyp segmentation by deformable alignment and local attention. IEEE Journal of Biomedical and Health Informatics (2025)
32. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)