This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Local Motion-Adaptive Temporal Smoothing for Real-Time Laparoscopic Video Segmentation

Minato Yamashita[1 [0000−0002−3826−0225]] and Masaaki Ito[1,2 [0000−0002−1101−2707]]

[1] Department for the Promotion of Medical Device Innovation, National Cancer Center Hospital East, Japan
[2] Department of Colorectal Surgery, National Cancer Center Hospital East, Japan
{miyamas3,maito}@east.ncc.go.jp

Abstract. Reliable real-time semantic segmentation of laparoscopic video is critical for intraoperative decision support. Nonetheless, frame-wise predictions are prone to noisy temporal flickering, while tissue deformation and scene motion require responsiveness to local changes. Although exponential moving average (EMA) filtering serves as a model-agnostic and lightweight post-processing strategy for mitigating such flickering, its fixed coefficient imposes a trade-off between temporal stability and motion responsiveness. To address this limitation, we propose a local motion-adaptive temporal smoothing method that replaces the fixed EMA coefficient with a spatially varying coefficient map derived from block-wise aggregation of sparse optical flow without requiring training or architectural modification. In this method, block-wise local motion estimated from sparse optical flow is converted to adaptive coefficients through a smoothstep mapping, and the resulting coefficients are applied online to segmentation probability maps. Experiments on the CholecSeg8k dataset show that the proposed method improves temporal stability in low-motion regions while preserving frame-wise accuracy in high-motion regions. Qualitative results demonstrate reduced flickering in stable regions without visible lag in dynamic regions. Runtime analysis shows a 90th-percentile latency of 9.41 ms, well below the 30-fps frame interval. These results support the potential of the proposed method for stabilizing real-time surgical video segmentation in intraoperative decision-support systems.

Keywords: Laparoscopic surgery · Semantic segmentation · Surgical video analysis · Temporal smoothing · Real-time inference.

1 Introduction

Semantic segmentation of laparoscopic video is a key task in surgical artificial intelligence [10, 26]. In particular, real-time pixel-level recognition of clinically relevant regions, including anatomical structures and Go/No-Go zones, is fundamental to safer decision-making in intraoperative navigation and robot-assisted surgery [16, 12, 21, 6, 24]. For these applications, segmentation must be accurate,

real-time, robust to noise, and temporally stable, because flickering masks can reduce usability and obscure safety-critical boundaries [7, 13]. However, improving temporal consistency is challenging in laparoscopic videos: noisy frame-to-frame variations should be reduced [18, 7, 13], while predictions should still follow genuine scene and appearance changes involving tissue deformation, camera motion, instrument motion, and blood [19, 11].

Recent studies have improved temporal consistency by incorporating temporal information directly into the segmentation architecture [20, 22, 7, 23, 13]. Although effective, such methods are difficult to apply to already optimized and validated systems because they are typically architecture-specific and require training. Considering the constraints of real-world deployment on resource-constrained edge platforms, temporal stabilization should instead be applicable to existing systems and add minimal computational overhead [3, 21].

This motivates post-processing methods that operate directly on segmentation outputs, such as inference-time smoothing of probability maps. Exponential moving average (EMA) filtering [9] is attractive because it is model-agnostic, requires only class probabilities, and incurs negligible overhead. However, standard EMA uses a single fixed coefficient, creating an inherent trade-off: stronger smoothing improves temporal stability, whereas weaker smoothing preserves responsiveness to motion. This trade-off is especially limiting when stable and dynamic regions coexist within the same frame. To address this trade-off, recent methods combine current and past predictions or features using motion compensation. ViperSAM [4] performs adaptive EMA in probability space by warping previous-frame probability maps, but relies on computationally costly dense optical flow. MCMA [17] applies a motion-corrected moving average in feature space with reduced overhead, yet requires integration into the model. These methods leave a gap for lightweight post-processing that adapts smoothing to local motion without dense optical flow or architectural modification.

In this work, we propose a model-agnostic, lightweight post-processing method for local motion-adaptive temporal smoothing in laparoscopic video segmentation. We estimate local motion from sparse optical flow, aggregate it over local blocks, and obtain a pixel-wise EMA coefficient map by upsampling block-wise coefficients. This allows stronger smoothing in locally stable regions and weaker smoothing in locally dynamic regions, thereby mitigating the stability–responsiveness trade-off of standard EMA without dense optical flow or architectural modification.

Our contributions are threefold. First, we formulate a local motion-adaptive temporal EMA with spatially varying smoothing coefficients. Second, we show that the method improves temporal stability in stable regions while maintaining responsiveness in dynamic regions. Third, we quantify its runtime overhead and demonstrate real-time compatibility with a 90th-percentile latency below 10 ms.

2 Method

2.1 Overview

Fig. 1 summarizes the proposed method. Given consecutive frames, we estimate block-wise local motion using sparse optical flow and convert the resulting block-wise motion estimates into a pixel-wise adaptive EMA coefficient map via a smoothstep mapping. This coefficient map controls the relative contribution of the current probability map P_t and the previous smoothed probability map $\hat{P}_{t-1}$ at each pixel. By adapting the smoothing strength spatially within each frame, the proposed filter suppresses flickering while preserving responsiveness.

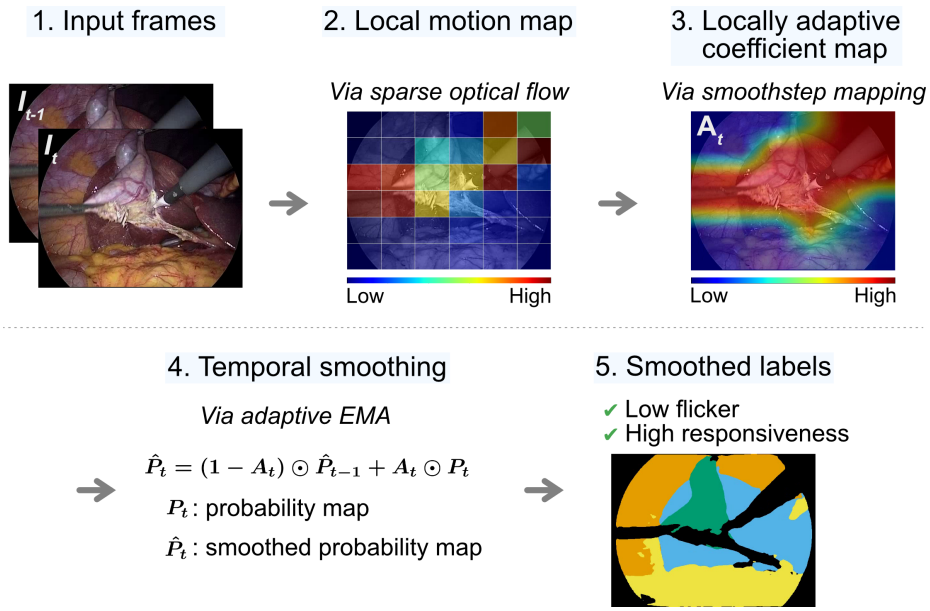


Fig. 1. Overview of proposed method.

2.2 Standard Fixed-Coefficient EMA Filtering

Let $P_t \in [0,1]^{C \times H \times W}$ denote the class probability map predicted for frame t, where C is the number of semantic classes, and let $\hat{P}_t$ denote its temporally smoothed counterpart.

Fixed-coefficient EMA updates the smoothed probability map with a fixed coefficient $\alpha \in (0,1]$ as

$$\hat{P}_t = (1-\alpha)\hat{P}_{t-1} + \alpha P_t.$$

A small α yields stronger temporal smoothing, whereas a large α makes the output more responsive to the current prediction. However, because the same coefficient is applied to all pixels regardless of local motion, fixed-coefficient EMA can lead to delayed responses in dynamic regions or insufficient stabilization in stable regions.

2.3 Sparse Optical Flow-Based Local Motion Estimation

To make the EMA coefficient spatially adaptive to local motion, we estimate motion between consecutive frames. We first place a regular grid of sample points in the previous frame, denoted as $\mathcal{K}_{t-1} = \{x_i^{t-1}\}_{i=1}^{N}$.

These points are then tracked from I_{t-1} to I_t using pyramidal Lucas–Kanade sparse optical flow [15, 1]. For each successfully tracked point, the displacement vector is computed as $\mathbf{u}_i^t = x_i^t - x_i^{t-1}$. To obtain a local motion representation, the motion-estimation domain is divided into blocks $\{b_j\}_{j=1}^{B}$, and the local motion magnitude for each block is defined as

$$m_t(b_j) = Q_{0.9}\big(\|\mathbf{u}_i^t\|_2 \mid x_i^{t-1} \in b_j\big)\,,$$

where $Q_{0.9}$ denotes the 90th percentile over tracked points whose initial positions fall inside the block.

2.4 Smoothstep-Based Adaptive EMA Coefficient

The block-wise motion values are converted into block-wise EMA coefficients so that low-motion regions are smoothed more strongly, whereas high-motion regions emphasize the current prediction more.

For each block b_j, we first normalize the motion magnitude using two thresholds:

$$z_t(b_j) = \mathrm{clip}\left(\frac{m_t(b_j) - \tau_{\mathrm{low}}}{\tau_{\mathrm{high}} - \tau_{\mathrm{low}}}, 0, 1\right),$$

where τ_{low} and τ_{high} denote the lower and upper motion thresholds, respectively. We then define the block-wise adaptive coefficient using the smoothstep function:

$$a_t(b_j) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\left(3z_t(b_j)^2 - 2z_t(b_j)^3\right),$$

where $\alpha_{\min}$ and $\alpha_{\max}$ are the lower and upper bounds of the EMA coefficient, respectively.

When $m_t(b_j) \le \tau_{\mathrm{low}}$, the coefficient becomes $\alpha_{\min}$, resulting in strong temporal smoothing; when $m_t(b_j) \ge \tau_{\mathrm{high}}$, the coefficient becomes $\alpha_{\max}$, reducing the influence of past predictions and placing more weight on the current prediction. Between these thresholds, the smoothstep mapping provides a smooth transition, avoiding a hard switch between stable and dynamic regions.

The resulting block-wise coefficients $\{a_t(b_j)\}_{j=1}^{B}$ are bilinearly upsampled to the probability-map resolution to obtain the pixel-wise adaptive EMA coefficient map $A_t(x)$.

2.5 Local Motion-Adaptive EMA Filtering

The final temporally smoothed probability map is obtained by replacing the scalar coefficient in fixed-coefficient EMA with the pixel-wise coefficient map $A_t(x)$:

$$\hat{P}_t(c,x) = (1 - A_t(x))\hat{P}_{t-1}(c,x) + A_t(x)P_t(c,x),$$

for each class c and pixel location x. The same coefficient $A_t(x)$ is applied to all classes at a given pixel, preserving the class-probability structure while adapting the smoothing strength spatially.

For the first frame of a video sequence, the smoothed probability map is initialized as $\hat{P}_0 = P_0$.

The update requires no future frames, making the filter suitable for real-time inference.

2.6 Implementation Details

The proposed method was applied to the segmentation model output at inference time as a post-processing step. For computational efficiency, sparse optical flow was implemented with OpenCV's `calcOpticalFlowPyrLK` function on grayscale frames after black-border cropping and resizing the frames to a width of 640 pixels for flow estimation while preserving the aspect ratio. Specifically, we (i) estimated sparse motion on a regular grid of points with a stride of 16 pixels and a 16-pixel margin, (ii) aggregated the motion over a 6×6 block grid, (iii) converted it into a block-wise adaptive EMA coefficient map, and (iv) upsampled the coefficient map to the probability-map resolution by bilinear interpolation before filtering. In the current implementation, the default adaptive-EMA parameters were set to $\alpha_{\min} = 0.1$, $\alpha_{\max} = 0.9$.

3 Experiments

3.1 Experimental Setup

Dataset. We used the CholecSeg8k dataset, which provides 8,080 laparoscopic cholecystectomy frames organized as short consecutive sequences with pixel-level anatomical labels from 17 videos, allowing us to evaluate temporal stability and responsiveness under realistic surgical scene motion [8]. We considered four foreground classes: liver, fat, abdominal wall, and gallbladder; all remaining labels were mapped to background. Input frames were originally provided at 854×480 and 25 fps. The dataset was split at the video level into training, validation, and test sets. Evaluation was performed on the held-out test set comprising three videos (video IDs 12, 24, and 35).

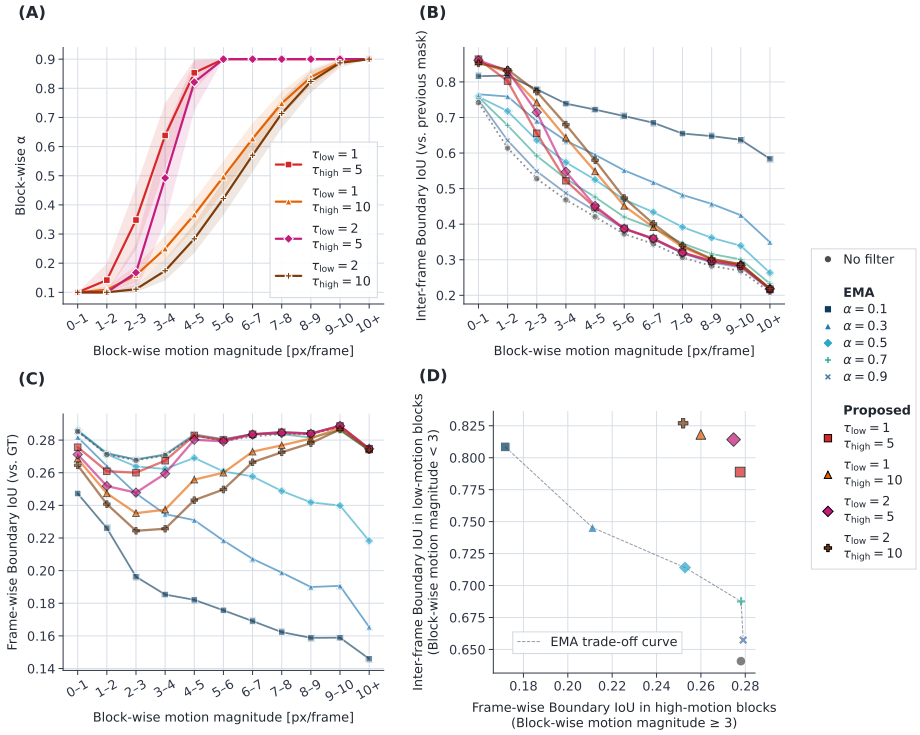


Fig. 2. Motion-stratified analysis of the proposed method compared with No filter and fixed-coefficient EMA baselines. Motion bins are defined using Farneback-derived block-wise motion. (A) Mean adaptive EMA coefficient with the 10th–90th percentile range. (B) Inter-frame Boundary IoU vs. previous masks. (C) Frame-wise Boundary IoU vs. ground truth. (D) Stability–responsiveness trade-off.

Segmentation Model and Training. As the base segmentation model, we used an FPN segmentation decoder [14] with a pretrained EfficientNetV2-S encoder [25]. The segmentation model input resolution was 512×288. We trained the model with AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-4}, batch size 16) using mixed precision. The loss function was an equally weighted sum of cross-entropy and Dice loss. Training was run for up to 30 epochs with early stopping based on validation mean intersection over union (mIoU), and the checkpoint with the best validation mIoU was used for evaluation.

Runtime Environment. All experiments were conducted on a laptop with an Intel Core i9-12900H CPU, 32 GB RAM, and an NVIDIA RTX 3080 Ti GPU. Inference used a TensorRT-RTX engine (version 1.3) quantized to FP16.

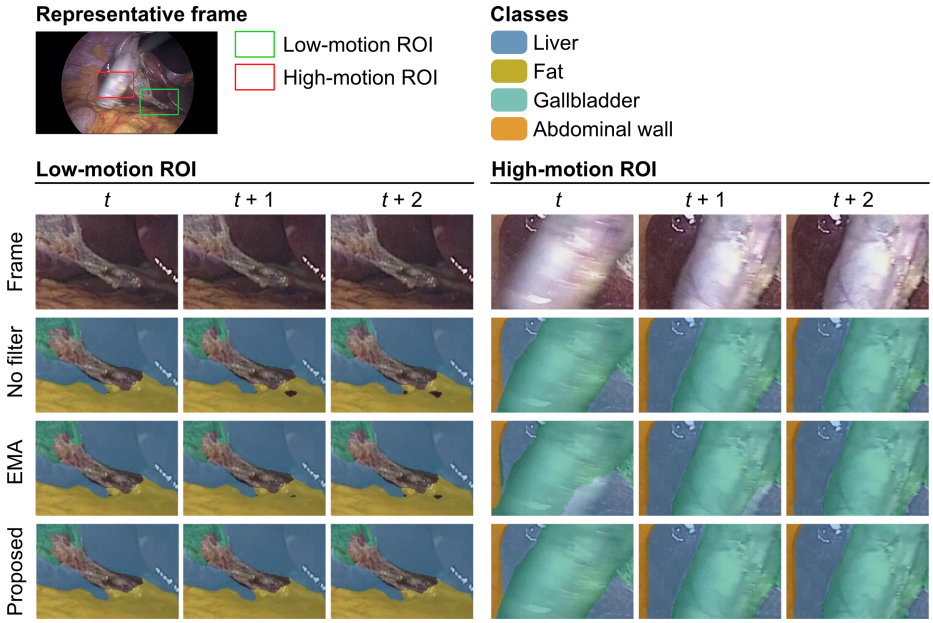


Fig. 3. Qualitative comparison in low- and high-motion ROIs across three consecutive frames. Fixed EMA: $\alpha = 0.5$; proposed: $\tau_{\text{low}} = 2$, $\tau_{\text{high}} = 5$.

3.2 Motion-Stratified Stability–Responsiveness Analysis

As a reference measure of local motion, we used block-wise motion derived from Farneback dense optical flow [5], defined as the 90th percentile of pixel-wise motion magnitudes within each block. The sparse-flow estimates used by the proposed method correlated strongly with this reference measure (Spearman's $\rho = 0.780$). The following analyses use these Farneback-derived motion bins. The number of blocks in each motion bin is provided in Supplementary Fig. S1.

We evaluated four representative parameter settings using all combinations of $\tau_{\text{low}} \in \{1.0, 2.0\}$ and $\tau_{\text{high}} \in \{5.0, 10.0\}$ pixels/frame. We compared them with the unfiltered prediction (No filter) and fixed-coefficient EMA baselines with $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. As a stability metric, we used inter-frame Boundary IoU, computed using Boundary IoU [2] between current- and previous-frame predicted masks across foreground classes. Because this metric can also penalize true motion, we interpret it mainly within low-motion bins, where temporal changes are expected to be small. As a responsiveness metric, we used frame-wise Boundary IoU against the ground truth labels. Corresponding analyses using mIoU are provided in Supplementary Fig. S2.

Fig. 2A shows that the adaptive EMA coefficient increases across the Farneback-derived motion bins, indicating that the sparse-flow-based coefficient map follows the intended motion-dependent behavior. In low-motion bins, the coefficient remains close to $\alpha_{\min}$, favoring stronger smoothing; as motion increases, it

smoothly rises toward $\alpha_{\max}$, placing more weight on the current prediction. The thresholds τ_{low} and τ_{high} determine the motion range over which this transition occurs. Lower τ values make the filter respond more quickly to motion, whereas higher τ values preserve stronger smoothing over a wider motion range.

In low-motion bins, the proposed method achieves inter-frame Boundary IoU close to fixed-coefficient EMA with $\alpha = 0.1$ and higher than larger-α baselines, indicating stronger flicker suppression in stable regions (Fig. 2B). As motion increases, its frame-wise Boundary IoU approaches that of No filter, whereas fixed-coefficient EMA with smaller coefficients exhibits a substantially larger degradation, indicating difficulty in following rapid motion (Fig. 2C). As expected from the behavior in Fig. 2A, lower τ values favor faster responsiveness to motion, whereas higher τ values favor stronger smoothing. The trade-off plot in Fig. 2D further shows that fixed-coefficient EMA exhibits a clear stability–responsiveness trade-off, whereas the proposed method mitigates it by maintaining high stability in low-motion regions (< 3 pixels/frame) while preserving responsiveness in high-motion regions (≥ 3 pixels/frame). Corresponding mIoU-based analyses show similar overall trends (Supplementary Fig. S2).

3.3 Qualitative Comparison

We further examined two ROIs, one with low motion and one with high motion, across three consecutive frames (Fig. 3). Fixed-coefficient EMA with $\alpha = 0.5$ slightly suppresses flicker within the fat label in the low-motion ROI, but introduces visible lag around the moving gallbladder Boundary in the high-motion ROI. In contrast, the proposed method provides stronger flicker suppression in the low-motion ROI while preserving rapid Boundary updates in the high-motion ROI, consistent with the motion-stratified quantitative analysis above.

3.4 Runtime Analysis

Table 1 compares runtime excluding disk I/O, measured from an input frame to the final mask label transferred back to the CPU. Runtime was measured over 1,830 frames after a 10-frame warm-up. The proposed method achieved a p90 latency of 9.41 ms, compared with 6.63 ms for No filter and 6.40 ms for fixed-coefficient EMA. Thus, the proposed filter added only 2.78 ms at p90 over No filter while remaining well below the 30-fps frame interval (33.33 ms).

Table 1. Runtime comparison of temporal smoothing methods.

Filter	**Mean ± SD (ms)**	**p50 (ms)**	**p90 (ms)**
No filter	5.89 ± 1.18	5.62	6.63
Fixed-coefficient EMA	5.84 ± 1.03	5.64	6.40
Proposed	7.41 ± 1.27	6.95	9.41

4 Limitations

The proposed method locally switches between stability and responsiveness according to motion magnitude, rather than achieving both simultaneously within the same moving region. Large motion shifts the filter toward weaker smoothing to preserve responsiveness. This can be triggered by anatomical or instrument motion, camera movement, or visual artifacts such as smoke or lens washing, potentially limiting flicker suppression in affected regions. Future work should investigate methods that maintain temporal stability in moving regions without sacrificing responsiveness.

A second limitation is that the motion thresholds τ_{low} and τ_{high} are defined in pixels per frame and must be specified in advance. Their optimal values may therefore depend on image resolution, camera distance, digital zoom, and acquisition conditions. Although the thresholds can be manually adjusted in real time, their automatic adaptation would be preferable for wider applicability across varying imaging conditions. Future work will focus on maintaining temporal stability in moving regions without sacrificing responsiveness and on automatically adapting motion thresholds across varying imaging conditions.

5 Conclusion

We presented a lightweight, model-agnostic post-processing method for laparoscopic video segmentation. By deriving spatially varying EMA coefficients from sparse optical flow, the method provides stronger smoothing in low-motion regions while preserving responsiveness in high-motion regions. Experiments on the CholecSeg8k dataset showed that the proposed method alleviates the stability–responsiveness trade-off of fixed-coefficient EMA while maintaining a p90 latency below 10 ms. These results suggest that local motion-adaptive smoothing is a practical strategy for improving the temporal stability of real-time surgical video segmentation. Future work will focus on improving robustness to global motion and visual artifacts and on automatically adapting motion thresholds across different imaging conditions.

Acknowledgments. This research was supported by AMED under Grant Numbers JP25he2302012 and JP26ck0106105. ChatGPT and Codex were used to assist with language editing and the development and refinement of research code. The authors reviewed all outputs and take full responsibility for the final content.

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

References

1. Bouguet, J.P.: Implementation of the lucas kanade feature tracker. description of the algorithm; technical report. Intel Corporation Microprocessor Research Lab: Santa Clara, CA, USA (1999)

2. Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary iou: Improving object-centric image segmentation evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15334–15342 (2021). https://doi.org/10.1109/CVPR46437.2021.01508
3. Choksi, S., Szot, S., Zang, C., Yarali, K., Cao, Y., Ahmad, F., Xiang, Z., Bitner, D.P., Kostic, Z., Filicori, F.: Bringing artificial intelligence to the operating room: edge computing for real-time surgical phase recognition. Surg. Endosc. **37**(11), 8778–8784 (Nov 2023). https://doi.org/10.1007/s00464-023-10322-4
4. Deka, D.J.: Inference-time temporal probability smoothing for stable video segmentation with SAM2 under weak prompts (2026). https://doi.org/10.48550/arXiv.2604.17115
5. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) Image Analysis. pp. 363–370. Springer Berlin Heidelberg, Berlin, Heidelberg (2003). https://doi.org/10.1007/3-540-45103-X_50
6. Gon Park, S., Park, J., Rock Choi, H., Ho Lee, J., Tae Cho, S., Goo Lee, Y., Ahn, H., Pak, S.: Deep learning model for real-time semantic segmentation during intraoperative robotic prostatectomy. Eur. Urol. Open Sci. **62**, 47–53 (Apr 2024). https://doi.org/10.1016/j.euros.2024.02.005
7. Grammatikopoulou, M., Sanchez-Matilla, R., Bragman, F., Owen, D., Culshaw, L., Kerr, K., Stoyanov, D., Luengo, I.: A spatio-temporal network for video semantic segmentation in surgical videos. International Journal of Computer Assisted Radiology and Surgery **19**(2), 375–382 (2024). https://doi.org/10.1007/s11548-023-02971-6
8. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80 (2020), https://arxiv.org/abs/2012.12453
9. Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D.: Forecasting with exponential smoothing. Springer Series in Statistics, Springer, Berlin, Germany (Dec 2008). https://doi.org/10.1007/978-3-540-71918-2
10. Kamtam, D.N., Shrager, J.B., Malla, S.D., Lin, N., Cardona, J.J., Kim, J.J., Hu, C.: Deep learning approaches to surgical video segmentation and object detection: A scoping review. Computers in Biology and Medicine **194**, 110482 (2025). https://doi.org/10.1016/j.compbiomed.2025.110482
11. Khatibi, T., Sepehri, M.M., Shadpour, P., Zegordi, S.H.: Applications of image processing in laparoscopic surgeries: An overview. In: Computer Vision, pp. 1518–1544. IGI Global (2018). https://doi.org/10.4018/978-1-5225-5204-8.ch063
12. Laplante, S., Namazi, B., Kiani, P., Hashimoto, D.A., Alseidi, A., Pasten, M., Brunt, L.M., Gill, S., Davis, B., Bloom, M., Pernar, L., Okrainec, A., Madani, A.: Validation of an artificial intelligence platform for the guidance of safe laparoscopic cholecystectomy. Surg. Endosc. **37**(3), 2260–2268 (Mar 2023). https://doi.org/10.1007/s00464-022-09439-9
13. Lee, M., Bragman, F.J.S., Sanchez-Matilla, R., Luengo, I., Stoyanov, D.: Spatial-temporal nas for fast surgical segmentation. In: 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024. BMVA (2024), https://papers.bmvc2024.org/0042.pdf
14. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017). https://doi.org/10.1109/CVPR.2017.106

15. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence. vol. 2, pp. 674–679 (1981), https://www.ijcai.org/Proceedings/81-2/Papers/017.pdf
16. Madani, A., Namazi, B., Altieri, M.S., Hashimoto, D.A., Rivera, A.M., Pucher, P.H., Navarrete-Welton, A., Sankaranarayanan, G., Brunt, L.M., Okrainec, A., Alseidi, A., Schwaitzberg, S.D., De, S., Walsh, C.M., Fried, G.M., Vassiliou, M.C.: Artificial intelligence for intraoperative guidance: Using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Annals of Surgery **276**(2), 363–369 (2022). https://doi.org/10.1097/SLA.0000000000004594
17. Mendel, R., Rueckert, T., Wilhelm, D., Rueckert, D., Palm, C.: Motion-corrected moving average: Including post-hoc temporal information for improved video segmentation (2024). https://doi.org/10.48550/arXiv.2403.03120
18. Miksik, O., Munoz, D., Bagnell, J.A., Hebert, M.: Efficient temporal consistency for streaming video scene analysis. In: 2013 IEEE International Conference on Robotics and Automation. pp. 133–139 (2013). https://doi.org/10.1109/ICRA.2013.6630567
19. Münzer, B., Schoeffmann, K., Böszörmenyi, L.: Content-based processing and analysis of endoscopic images and videos: A survey. Multimed. Tools Appl. **77**(1), 1323–1362 (Jan 2018). https://doi.org/10.1007/s11042-016-4219-z
20. Nilsson, D., Sminchisescu, C.: Semantic video segmentation by gated recurrent flow propagation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6819–6828 (2018). https://doi.org/10.1109/CVPR.2018.00713
21. Protserov, S., Hunter, J., Zhang, H., Mashouri, P., Masino, C., Brudno, M., Madani, A.: Development, deployment and scaling of operating room-ready artificial intelligence for real-time surgical decision support. npj Digital Medicine **7**, 231 (2024). https://doi.org/10.1038/s41746-024-01225-2
22. Rebol, M., Knöbelreiter, P.: Frame-to-frame consistent semantic segmentation (2020). https://doi.org/10.3217/978-3-85125-752-6-18
23. Rueckert, T., Rueckert, D., Palm, C.: Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art. Comput. Biol. Med. **169**, 107929 (Feb 2024). https://doi.org/10.1016/j.compbiomed.2024.107929
24. Suzuki, Y., Kitaguchi, D., Kosugi, N., Hasegawa, H., Takeshita, N., Kinugasa, Y., Ito, M.: Artificial intelligence-guided total mesorectal excision: Development of a deep learning model to identify the pelvic fascial plane anatomy. Ann. Gastroenterol. Surg. **9**(6), 1225–1232 (Nov 2025). https://doi.org/10.1002/ags3.70064
25. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10096–10106. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/tan21a.html
26. Zhou, R., Wang, D., Zhang, H., Zhu, Y., Zhang, L., Chen, T., Liao, W., Ye, Z.: Vision techniques for anatomical structures in laparoscopic surgery: a comprehensive review. Frontiers in Surgery **12**, 1557153 (2025). https://doi.org/10.3389/fsurg.2025.1557153