

DBT-Bleed: Dual-Branch Temporal Modeling with Key-Frame Selection for Surgical Bleeding Detection

Sudhanshu Mishra^{1*}, Jialang Xu^{3*}, Jensen Ang⁴, Evangelos B. Mazomenos³, Beng Ti Ang⁴, and Yueming Jin^{1,2**}

¹ Department of Electrical and Computer Engineering, National University of Singapore, Singapore

² Department of Biomedical Engineering, National University of Singapore, Singapore

³ UCL Hawkes Institute, Department of Medical Physics and Biomedical Engineering, University College London, London, United Kingdom

⁴ Department of Neurosurgery, National Neuroscience Institute, Singapore
ymjin@nus.edu.sg

Abstract. Intraoperative Adverse Events (IAEs) detection is critical for improving surgical safety, with bleeding being among the most frequent events across many surgery types. Existing methods struggle to distinguish bleeding IAE from visually similar residual blood due to limited temporal reasoning. Moreover, modeling long surgical videos while preserving fine-grained temporal dynamics remains computationally challenging. We propose *DBT-Bleed*, a dual-branch multi-scale temporal modeling framework disentangling bleeding and normal representations using layer-wise temporal adapters for short- and long-term bleeding progression. To efficiently process long surgical videos without sacrificing fine-grained temporal information, we introduce HiRED, a Hierarchical Entropy-Driven frame selection strategy that retains temporally informative segments while removing redundancy. Experiments on the MultiBypass dataset demonstrate gains of 6.53% in F1, 5.62% in Recall and 9% in MCC values for bleeding IAE detection, consistently outperforming video-level baselines. Additionally, we evaluate cross-procedure generalization on a newly curated dataset from a different surgical procedure type, where DBT-Bleed demonstrates robust transferability by achieving gain of 6% in F1 and 8% in MCC under zero-shot setting. To support this evaluation, we introduce *EndoPit-IAE*, an Endonasal Pituitary Surgery dataset annotated for IAEs, representing the first IAE-annotated dataset in neurosurgery. Code is available at: <https://github.com/jinlab-imvr/DBT-Bleed>

Keywords: Intraoperative Adverse Events · Endonasal Pituitary Surgery · Gastric Bypass injury

* Equal contribution

** Corresponding author

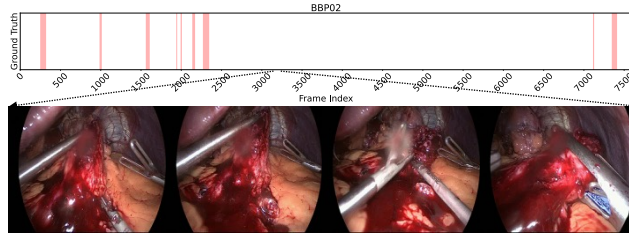


Fig. 1: Non-IAE residual blood visibly similar to bleeding IAEs.

1 Introduction

Intraoperative Adverse Events (IAEs) are rare but critical occurrences during surgery. They can lead to severe postoperative complications, including infection, organ dysfunction, or even mortality [25], and often disrupt surgical workflow by necessitating immediate corrective interventions. They not only jeopardize patient safety but also increase healthcare costs due to prolonged recovery times. IAEs contribute upto 400k deaths annually [8], with bleeding being the most common at 23% [25].

Therefore, it is crucial to timely detect bleeding to enable prompt intervention and improve surgical outcomes.

Despite promising efforts [21], video-level IAEs detection remains relatively underdeveloped. Methods focus on bleeding segmentation [9,11,26], blood source and flow detection [15,4,10,12,18,24]. However, detecting bleeding as an IAE differs from the tasks described above, because it requires distinguishing bleeding caused by a manual error from anticipated surgical bleeding.

The lack of IAEs-annotated public datasets has constrained research in this field, making the recent extension of MultiBypass dataset [7,6,14,13] for IAEs annotation [1,5] a vital advancement. The dataset naturally contains frames with visible blood that are labeled as “normal” (Fig. 1), as the blood may originate from clinically resolved events or intentional surgical preparation. Such bleeding constitutes ‘Residual-bleeding’ leading to ambiguous frame-level supervision, whereas ‘Bleeding-IAE’ is unintended surgical-site bleeding from mechanical or thermal incision. Horita et al. [3] applied YOLOv7 for frame-wise real-time bleeding detection without explicit temporal modeling. Bose et al. [1] incorporated short-range temporal context but focused on limited clip lengths, restricting multi-scale modeling of bleeding progression. Consequently, two key challenges remain: (1) the need for robust temporal modeling to distinguish bleeding IAE from visually similar residual blood, and (2) the difficulty of efficiently modeling long surgical video sequences without losing fine-grained temporal dynamics.

Subsequently, (1) We propose DBT-Bleed, a dual-branch multi-scale temporal framework for surgical bleeding detection. A layer-wise Multi-scale Temporal Adapter(MTA) captures both short- and long-term bleeding progression to distinguish bleeding adverse events from residual blood. (2) We introduce Hierarchical Entropy-Driven (HiRED) frame selection to process long surgical

videos without sacrificing fine-grained dynamics. This strategy removes redundant segments while preserving temporally informative frames. (3) We demonstrate cross-procedure robustness of our method via zero-shot evaluation on a newly curated inhouse Endonasal Pituitary Surgery dataset. In this context, we introduce EndoPit-IAE, the first neurosurgical dataset annotated for IAEs. (4) We demonstrate state-of-the-art performance on both MultiBypass and EndoPit-IAE through extensive experiments and ablation studies.

2 Methods

2.1 Dataset

MultiBypass [1,5] is a multi-center dataset of laparoscopic bypass surgery videos comprising 140 full-length surgical procedures, with corresponding frame-level IAE annotations at 1 fps available at [1]. All training, validation, and testing experiments are conducted on MultiBypass.

To further assess generalization, we curate **EndoPit-IAE**, an in-house Endonasal Pituitary Surgery dataset with detailed IAEs annotations, representing—to the best of our knowledge—the first video dataset dedicated to neurosurgical IAEs detection. Annotations were performed at 1-second intervals by expert surgeons. Three annotators labeled the in-house dataset: one consultant with > 10 years of experience and two consultant-trained trainees. Trainee annotations were reviewed by the consultant, and discrepancies were resolved by joint re-review and consensus, with an average inter-rater discrepancy $< 0.5\%$ per video. Annotators estimated the flow rate of any bleeding present on the screen, and scored it using the VIBe (Validated Intraoperative Bleeding) [16] scale. EndoPit-IAE serves exclusively as an external zero-shot benchmark evaluating cross-procedure robustness. Frame-level supervision is unreliable due to persistent residual blood and temporally sparse events in long videos (Fig. 1), motivating video-level supervision by aggregating $N = 300$ frames with 100 frame overlap into videos (Table. 1).

Table 1: Dataset statistics

Dataset	Split	Non-bleed	Bleeding
MultiBypass	Train	1,586	600
	Val	416	173
	Test	884	267
EndoPit-IAE	Test	64	128

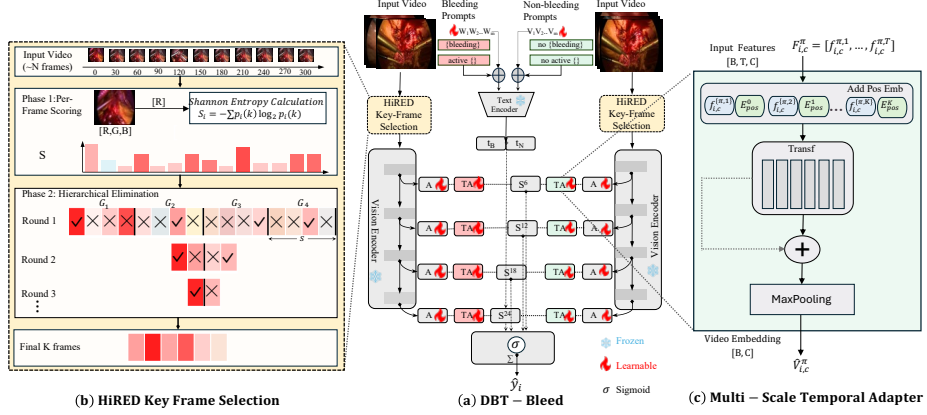


Fig. 2: Overview of the proposed (a) DBT-Bleed framework, including (b) HiRED key frame selection, and (c) dual-branch multi-scale temporal adapter.

2.2 Problem Definition

Let $\mathcal{D} = \{(V_i, y_i)\}_{i=1}^M$ denote a set of surgical videos, where each video $V_i = \{x_{i,t}\}_{t=1}^{N_i}$ contains total N_i RGB frames and is associated with a binary video-level label $y_i \in \{0, 1\}$ indicating the absence or presence of bleeding IAE. We aim to learn a weakly supervised classifier f_θ that predicts a video-level probability $\hat{y}_i = f_\theta(V_i)$, with positive labels indicating at least one bleeding IAE in V_i .

2.3 DBT-Bleed

Overview Architecture We propose DBT-Bleed—a CLIP-based dual-branch framework with multi-scale temporal modeling, shown in Fig. 2, to effectively capture the dynamic characteristics of adverse events in complex surgical scenes. Given a long surgical video of N frames, HiRED selects the top K informative frames via Shannon entropy estimation followed by hierarchical pruning. The selected frames are processed by a frozen CLIP-image encoder augmented with trainable spatial adapters (A) and temporal adapters (TA) that explicitly model inter-frame dependencies, enabling both short- and long-range temporal reasoning across the video. To guide the dual-contrast alignment, we prepend learnable context vectors (V_i, W_i) to class-specific text prompts—“active {bleeding}”, “no active {bleeding}”, etc—and encoded through CLIP’s frozen text encoder, producing bleeding (t_B) and non-bleeding (t_N) text embeddings. Multi-scale visual-text similarity scores $S^6, S^{12}, S^{18}, S^{24}$ are aggregated and passed through a sigmoid function to yield the final video-level bleeding prediction $\hat{y}_i$.

We optimize a log-sigmoid dual-contrast loss that encourages each branch to align with its matched text prompt while repelling the opposite, summed over

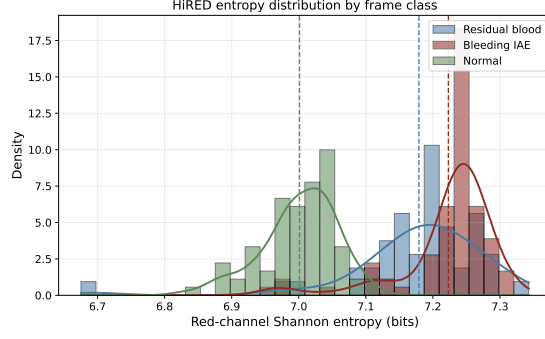


Fig. 3: Red-channel Shannon Entropy Distribution.

adapter layers $\pi \in \{6, 12, 18, 24\}$ and both branches $c \in \{b, n\}$:

$$\mathcal{L} = -\frac{1}{B} \sum_{i, \pi, c} \log \sigma \left(\ell_{i, c} \cdot (\hat{v}_{i, c}^\pi \cdot \hat{t}_c - \hat{v}_{i, c}^\pi \cdot \hat{t}_{\bar{c}}) \right)$$

where $\hat{v}_{i, c}^\pi$ and $\hat{t}_c$ are L2-normalized video and text embeddings, $\bar{c}$ denotes the opposite branch, and $\ell_{i, c} \in \{-1, +1\}$ is a signed ground-truth label.

HiRED: Hierarchical Entropy-Driven Key-Frame Selection From each temporal window of N frames, only $K \ll N$ key-frames are selected for the CLIP encoder. Uniform subsampling treats all frames as equally informative, often retaining redundant content (instrument occlusion, smoke, static tissue). HiRED addresses this via two-phase entropy-guided selection.

Phase 1: Per-Frame Scoring. Red channel of each frame is extracted, and the Shannon entropy [19] of the normalized 256-bin intensity histogram is computed as $S_i = -\sum_{k=0}^{255} p_i(k) \log_2 p_i(k)$, where $p_i(k)$ denotes the fraction of pixels with intensity k in frame i . High entropy indicates diverse red-channel intensities—a visual cue of bleeding. This yields a score vector $\mathbf{S} \in \mathbb{R}^N$.

Phase 2: Hierarchical Elimination. Given $\mathbf{S}$, frames are iteratively pruned over multiple rounds as shown in Fig 2(b). Let $\mathcal{C}^{(t)}$ denote the candidate set at round t , with $\mathcal{C}^{(0)} = \{1, \dots, N\}$. At each round, the candidates are partitioned into $G^{(t)}$ contiguous segments of roughly equal size (segment size s), where

$$G^{(t)} = \max \left(K, \left\lfloor \frac{|\mathcal{C}^{(t)}|}{s} \right\rfloor \right) \quad (1)$$

Within each segment $G_g^{(t)}$, only the highest-entropy frame ($c_g^{(t)}$) survives, and thus $\mathcal{C}^{(t+1)} = \{c_1^{(t)}, \dots, c_{G^{(t)}}^{(t)}\}$ for the next round. The process terminates when $|\mathcal{C}^{(t)}| = K$. Early rounds eliminate low-entropy frames globally while later rounds

make fine-grained distinctions within progressively smaller, higher-quality pools. Because segments are always contiguous, the final K key-frames are guaranteed to span the full temporal extent of the clip. During training, a stochastic variant (HiRED-J) injects diversity by selecting the second-best frame per segment with probability $\rho=0.3$; at test time, selection is deterministic.

Residual-blood segments are often static and redundant, yielding lower red-channel entropy. HiRED retains high-diversity frames more indicative of active bleeding IAEs despite overlap (Fig. 3), thereby suppressing residual-blood evidence within each 300-frame window and improving temporal discrimination.

Multi-scale Temporal Adapter (MTA) To capture fine-grained temporal reasoning, we introduce MTA, a lightweight temporal adapter operating across multiple semantic scales of the frozen CLIP image encoder as shown in Fig 2. Each adapter consists of 6 self-attention layers, 12 attention heads, and learnable frame positional embeddings. Intermediate representations are extracted at four depths $\pi \in \{6, 12, 18, 24\}$, spanning a hierarchy from low-level texture and color transitions to high-level semantic features encoding instrument-tissue interactions. At each scale, per-frame patch tokens produced by the dual spatial adapters are spatially mean-pooled into frame-level embeddings $\mathbf{F}_{i,c}^\pi = [\mathbf{f}_{i,c}^{\pi,1}, \dots, \mathbf{f}_{i,c}^{\pi,T}] \in \mathbb{R}^{T \times d}$. These are augmented with learnable temporal position embeddings and processed by a temporal transformer with a residual connection:

$$\hat{\mathbf{v}}_{i,c}^\pi = \text{norm} \left(\text{MaxPool} \left(\text{Transf}(\mathbf{F}_{i,c}^\pi + \mathbf{E}_{\text{pos}}) + \mathbf{F}_{i,c}^\pi \right) \right)$$

where $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{T \times d}$ are learnable temporal position embeddings, $\text{Transf}(\cdot)$ is a stack of multi-head self-attention layers with feed-forward networks performing global temporal reasoning, and the residual connection preserves pretrained CLIP representations by initializing the module as a near-identity mapping. This multi-scale design allows adapter layers to learn depth-specific temporal aggregation: shallow layers capture low-level dynamics (e.g., color changes), while deeper layers model higher-level semantics like bleeding progression.

3 Experimental Setup

Implementation We use CLIP ViT-L/14 architecture as backbone and 240-pixel input images. All backbone parameters are frozen during training; only the lightweight spatial adapter modules, temporal adapter modules & learnable text prompts are optimized. Training is performed with Adam optimizer with learning rate of $1 \times e^{-3}$, batch size 16 for 90 epochs.

Baselines We compare against four representative state-of-the-art methods: MadCLIP [17] and VadCLIP [22] (dual-branch CLIP-based anomaly detection), ActionCLIP [20] (temporal transformer over CLIP features), and SEDMamba [23] (a Mamba-based surgical error detection model). These baselines span both

CLIP-based and domain-specific paradigms relevant to our design. BetaMixture [1] is excluded due to the lack of publicly available code, preventing reproducible comparison.

Evaluation Metrics We adopt standard classification metrics, including F1 score and Recall for performance evaluation. We additionally report the Matthews Correlation Coefficient (MCC), a correlation-based metric between ground-truth and predicted labels, widely regarded as a robust performance measure for imbalanced binary classification [2]. MCC score ranges from -1 (complete disagreement) through 0 (random prediction) to $+1$ (perfect prediction), providing a balanced measure even under class imbalance.

Table 2: Performance comparison on the MultiBypass dataset and zero-shot transfer evaluation on the EndoPit-IAE dataset ($p < 0.001$).

Method	MultiBypass			EndoPit-IAE (Zero-shot)		
	F1 $\uparrow$	Recall $\uparrow$	MCC $\uparrow$	F1 $\uparrow$	Recall $\uparrow$	MCC $\uparrow$
SEDMamba	45.13	46.82	0.28	70.49	67.19	0.19
VadCLIP	54.31	69.66	0.38	63.00	53.91	0.18
ActionCLIP	58.38	73.03	0.44	75.77	91.15	0.18
MadCLIP	58.22	69.66	0.44	77.00	88.48	0.27
DBT-Bleed (Ours)	64.91	78.65	0.53	83.00	89.53	0.35

4 Results and Discussions

Table 2 shows quantitative results on MultiBypass dataset. Our method achieves the best overall performance, surpassing the second-best ActionCLIP method by 6.53% gain in F1 score and 5.62% gain in Recall. DBT-Bleed improves MCC by 9%, indicating a clearer bleeding/non-bleeding decision boundary while accounting for all confusion-matrix entries under class imbalance. DBT-Bleed also improves AUPRC over MadCLIP (61.6 vs. 56.29). All improvements over the baselines are statistically significant (McNemar’s exact test, $p < 0.001$ for all pairwise comparisons).

We present ribbon visualizations in Fig. 4, enabling a direct comparison of prediction dynamics between our method and the baselines. Fig. 4 demonstrates that our method achieves superior bleeding IAE video-clip detection performance compared to all baselines, with improved sensitivity and specificity. We further provide qualitative comparisons in Fig. 5. Notably, our method successfully identifies several critical cases with high confidence where residual blood is present on the scene, highlighting its robustness in difficult scenarios.

We further evaluate cross-procedure generalization on the in-house EndoPit-IAE dataset in a zero-shot setting. As shown in Table 2, although EndoPit-IAE

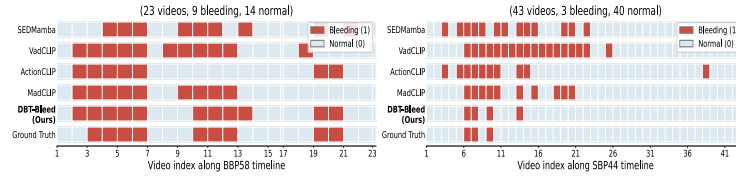


Fig. 4: Ribbon visualizations of model predictions on MultiBypass dataset.

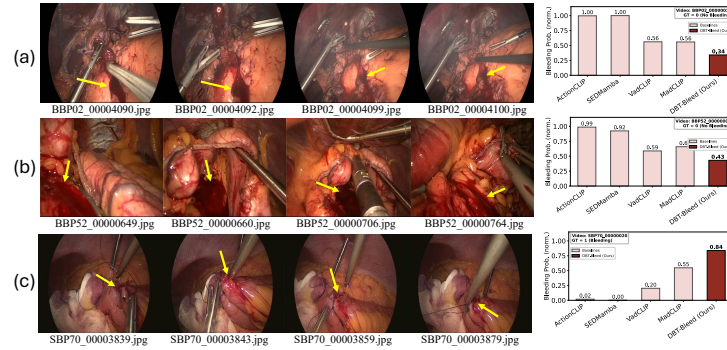


Fig. 5: Qualitative MultiBypass results: (a)-(b) show non-IAE residual blood; (c) subtle bleeding IAE. Our method yields accurate, confident predictions.

has substantial domain shift between surgery types against MultiBypass, the proposed DBT-Bleed still achieves the best zero-shot performance (6% gain in F1 and 8% gain in MCC score), showing the better generalization across different datasets. The higher zero-shot F1 and Recall score on EndoPit-IAE compared to supervised performance on MultiBypass is potentially accounted by its higher resolution and stronger color contrast; importantly, under the same zero-shot protocol, DBT-Bleed consistently outperforms all baselines.

Ablation Studies We conduct experiments to select an appropriate video length. On MultiBypass, we observe that approximately 95% of continuous adverse events span fewer than 300 frames. Accordingly, we perform an ablation study over N in 16, 32, 64, 100, 300, 400, 500 frames per video. Peak performance occurs at $N = 300$ as shown in Table 3, which we adopt as the default setting. We hypothesize that this regime best leverages the proposed HiRED key-frame selection strategy by preserving sufficient temporal context. Furthermore, both the MTA and HiRED modules yield consistent improvements across all evaluation metrics as shown in Table 3.

Table 3: Ablation study. Left: video length N . Right: the proposed components.

$N \rightarrow$	16	32	64	100	300	400	500	Base	MTA	HiRED	F1	Recall	MCC
F1	52.42	54.51	55.10	56.72	64.91	58.70	59.17	✓	✗	✗	58.22	69.66	0.44
Recall	59.38	59.30	60.63	66.76	78.65	74.06	74.28	✓	✓	✗	60.15	74.91	0.47
MCC	0.48	0.50	0.49	0.50	0.53	0.41	0.42	✓	✓	✓	64.91	78.65	0.53

5 Conclusions

We present DBT-Bleed, a dual-branch multi-scale temporal framework to capture fine-grained short- and long-term bleeding progression while addressing the challenges posed by visually similar residual blood. The proposed HiRED frame selection strategy further enables efficient long-range modeling without compromising fine-grained temporal information. Our method improves F1 by 6.53% and MCC by 9% on the MultiBypass dataset. We also demonstrate zero-shot cross-procedure generalization of DBT-Bleed with 6% F1 and 8% MCC gain on a newly curated Endonasal Pituitary Surgery dataset, EndoPit-IAE, the first neurosurgical IAE-annotated dataset. While this study focuses on binary bleeding detection, future work will develop larger, more diverse datasets and extend DBT-Bleed toward multi-label, clinically representative IAEs modeling.

Acknowledgments. This work was supported by Ministry of Education Tier 2 grant, Singapore (T2EP20224-0028), NUS-UCL Research and Innovation Collaboration Fund 2025, and an EPSRC Standard Grant [EP/Z534754/1].

Disclosure of Interests. The authors declare no conflict of interests.

References

1. Bose, R., Nwoye, C.I., Lazo, J.F., Lavanchy, J.L., Padoy, N.: Feature mixing approach for detecting intraoperative adverse events in laparoscopic roux-en-y gastric bypass surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 178–188. Springer (2025)
2. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics **21**(1), 6 (2020)
3. Horita, K., Hida, K., Itatani, Y., Fujita, H., Hidaka, Y., Yamamoto, G., Ito, M., Obama, K.: Real-time detection of active bleeding in laparoscopic colectomy using artificial intelligence. Surgical Endoscopy **38**(6), 3461–3469 (2024)
4. Hua, S., Gao, J., Wang, Z., Yeerkenbieke, P., Li, J., Wang, J., He, G., Jiang, J., Lu, Y., Yu, Q., et al.: Automatic bleeding detection in laparoscopic surgery based on a faster region-based convolutional neural network. Annals of Translational Medicine **10**(10), 546 (2022)
5. Lavanchy, J.L., Alapatt, D., Sestini, L., Kraljević, M., Nett, P.C., Mutter, D., Müller-Stich, B.P., Padoy, N.: Analyzing the impact of surgical technique on intraoperative adverse events in laparoscopic roux-en-y gastric bypass surgery by video-based assessment. Surgical Endoscopy **39**(3), 2026–2036 (2025)

6. Lavanchy, J.L., Gonzalez, C., Kassem, H., Nett, P.C., Mutter, D., Padoy, N.: Proposal and multicentric validation of a laparoscopic roux-en-y gastric bypass surgery ontology. *Surgical endoscopy* **37**(3), 2070–2077 (2023)
7. Lavanchy, J.L., Ramesh, S., Dall’Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N.: Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. *International journal of computer assisted radiology and surgery* **19**(11), 2249–2257 (2024)
8. Makary, M.A., Daniel, M.: Medical error—the third leading cause of death in the us. *Bmj* **353** (2016)
9. Miao, A.J., Lin, S., Lu, J., Richter, F., Ostrander, B., Funk, E.K., Orosco, R.K., Yip, M.C.: Hemoset: The first blood segmentation dataset for automation of hemostasis management. In: 2024 International Symposium on Medical Robotics (ISMR). pp. 1–7. IEEE (2024)
10. Pei, J., Zhou, Z., Guo, D., Li, Z., Qin, J., Du, B., Heng, P.A.: Synergistic bleeding region and point detection in laparoscopic surgical videos. *arXiv preprint arXiv:2503.22174* (2025)
11. Rabbani, N., Seve, C., Bourdel, N., Bartoli, A.: Video-based computer-aided laparoscopic bleeding management: a space-time memory neural network with positional encoding and adversarial domain adaptation. In: *Medical Imaging with Deep Learning* (2022)
12. Rahbar, M.D., Reisner, L., Ying, H., Pandya, A.: An entropy-based approach to detect and localize intraoperative bleeding during minimally invasive surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery* **16**(6), 1–9 (2020)
13. Ramesh, S., Dall’Alba, D., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Fiorini, P., Padoy, N.: Weakly supervised temporal convolutional networks for fine-grained surgical activity recognition. *IEEE Transactions on Medical Imaging* **42**(9), 2592–2602 (2023)
14. Ramesh, S., Dall’Alba, D., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Fiorini, P., Padoy, N.: Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *International journal of computer assisted radiology and surgery* **16**(7), 1111–1119 (2021)
15. Richter, F., Shen, S., Liu, F., Huang, J., Funk, E.K., Orosco, R.K., Yip, M.C.: Autonomous robotic suction to clear the surgical field for hemostasis using image-based blood flow detection. *IEEE Robotics and Automation Letters* **6**(2), 1383–1390 (2021)
16. Sciubba, D.M., Khanna, N., Pennington, Z., Singh, R.K.: Vibe scale: validation of the intraoperative bleeding severity scale by spine surgeons. *International Journal of Spine Surgery* **16**(4), 740 (2022)
17. Shiri, M., Beyan, C., Murino, V.: Madclip: few-shot medical anomaly detection with clip. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 416–426. Springer (2025)
18. Sogabe, M., Ishikawa, K., Takamatsu, T., Takeuchi, K., Kanno, T., Fujimoto, K., Miyazaki, T., Kawase, T., Sato, T., Kawashima, K.: Bleeding alert map (bam): The identification method of the bleeding source in real organs using datasets made on mimicking organs. *Array* **19**, 100308 (2023)
19. Spadaro, G., Renzulli, R., Bragagnolo, A., Giraldo, J.H., Fiandrotti, A., Grangetto, M., Tartaglione, E.: Shannon strikes again! entropy-based pruning in deep neural

- networks for transfer learning under extreme memory and computation budgets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1518–1522 (2023)
20. Wang, M., Xing, J., Mei, J., Liu, Y., Jiang, Y.: Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE transactions on neural networks and learning systems* (2023)
 21. Wei, H., Rudzicz, F., Fleet, D., Grantcharov, T., Taati, B.: Intraoperative adverse event detection in laparoscopic surgery: Stabilized multi-stage temporal convolutional network with focal-uncertainty loss. In: *Machine Learning for Healthcare Conference*. pp. 283–307. PMLR (2021)
 22. Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., Zhang, Y.: Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 6074–6082 (2024)
 23. Xu, J., Sirajudeen, N., Boal, M., Francis, N., Stoyanov, D., Mazomenos, E.B.: Sedmamba: Enhancing selective state space modelling with bottleneck mechanism and fine-to-coarse temporal fusion for efficient error detection in robot-assisted surgery. *IEEE Robotics and Automation Letters* **10**(1), 232–239 (2024)
 24. Xu, M., Zhou, R., Wang, A., Lyu, C., Li, Z., Zhong, N., Ren, H.: Bleedorigin: Dynamic bleeding source localization in endoscopic submucosal dissection via dual-stage detection and tracking. *arXiv preprint arXiv:2507.15094* (2025)
 25. Zegers, M., de Bruijne, M.C., de Keizer, B., Merten, H., Groenewegen, P.P., van der Wal, G., Wagner, C.: The incidence, root-causes, and outcomes of adverse events in surgical units: implication for potential prevention strategies. *Patient safety in surgery* **5**(1), 13 (2011)
 26. Zou, S., Li, J., Ji, W., Huang, J., Wang, K., Dan, G., Si, W., Pan, Y.: Surgical scene segmentation using a spike-driven video transformer with real-time potential. *arXiv preprint arXiv:2512.21284* (2025)