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.

SurgAnon: Out-of-Body Detection in Robotic Surgical Video via Gated Multi-Prompt Text Fusion with Temporal Modeling

Suyash Kumar[1,2⋆†], Sébastien Frey[1,3⋆], Ezem Sura Ekmekci[1], Duccio Lalli[1,5,6], Quentin Rudondy[3], Pierre Berthet-rayne[4], Herve Delingette[1], Francois Bremond[1], and Nicholas Ayache[1]

[1] Université Côte d'Azur, Inria, Sophia-Antipolis, France
[2] Indian Institute of Technology (BHU) Varanasi, India
[3] CHU Nice, Nice, France
[4] Carvolix, Nice, France
[5] EURECOM, Sophia Antipolis, France
[6] Politecnico di Torino, Turin, Italy
kumar.suyash2304@gmail.com (corresponding author), frey.s@chu-nice.fr

Abstract. Out-of-body (OoB) detection identifies frames in which the laparoscopic camera exits the patient. It is a prerequisite for privacy-preserving surgical video anonymisation and downstream AI tasks, yet no standardized evaluation protocol exists for the robotic domain under annotation scarcity. We address this gap with three contributions. First, a **lightweight timestamp annotation protocol** requiring only 15–20 minutes per video, enabling institutions to annotate robotic procedures without per-frame clinical review. Second, a **controlled comparative evaluation** of 17 da Vinci robotic procedures across 12 procedure types ($\approx$ 24h 26 min; 5 training, 12 test), on which OoBNet, IODA (zero-shot and supervised), and four ablated variants are all evaluated under *identical* training conditions. Third, **GatedTextFusion-LSTM**, a strong baseline that fuses SurgVLP visual embeddings with dynamically gated class-specific text prompts and models the sequence with a stateful LSTM, achieving $F1_{smooth} = \mathbf{0.9849}$ and $AP = \mathbf{0.9956}$ on the held-out test set, outperforming OoBNet supervised ($F1_{smooth} = 0.9390$, $AP = 0.9572$) and IODA supervised ($F1_{smooth} = 0.9663$, $AP = 0.9883$) with 24% fewer parameters compared to IODA. Notably, our model yields only 1 missed (false negative) OoB segment across the 12 test videos, and this residual error occurs at an IB↔OoB transition frame rather than mid-segment inside an out-of-body scene. A semi-supervised extension via LSTM self-training on 15 pseudo-labelled videos further improves AP to 0.9964, demonstrating that the evaluation protocol enables label-efficient research. On the gynaecological procedure, the model achieves perfect segment-level accuracy (zero FP/FN), addressing the heightened privacy sensitivity of such cases for women's health surgery.

⋆ Equal contribution.
† Corresponding author.

Keywords: Out-of-body detection · Surgical video evaluation · Annotation protocol · Vision-language models · Surgical Video Anonymisation

1 Introduction

Robotic minimally invasive surgery produces long endoscopic recordings that routinely contain out-of-body (OoB) segments. Detecting and excising these segments is a prerequisite for privacy-compliant data sharing and for downstream surgical AI tasks including phase recognition and skill assessment. Despite growing interest in OoB detection [1–3], existing methods were each developed and evaluated under different data regimes: OoBNet [1] was validated across multiple laparoscopic GI centres, while IODA [2] was trained on 100 laparoscopic procedures across 23 procedure types, requiring dense manual annotation by a medical expert across the full dataset. Critically, purely visual methods struggle with ambiguous frames like electrocautery smoke, blurry trocar transitions, and warm-toned exterior skin textures that share low-level statistics with intra-abdominal tissue, none of which can be resolved without semantic understanding of the surgical scene. Neither provides an annotation protocol suited to institutions that can realistically label only a handful of robotic videos, and neither has been validated specifically on da Vinci robotic surgery under annotation scarcity. In this work, we propose a new anonymisation evaluation protocol and detection method to address this gap. Using a segment-level **timestamp annotation protocol** requiring only 15–20 minutes per video, we annotate 17 da Vinci robotic procedures spanning 12 clinically diverse procedure types, establishing a **controlled comparison** where all methods train on the *same* 5 labelled videos and evaluate on the same 12 held-out procedures. Our method yields only 1 missed OoB segment across the 12 test videos, and this residual error occurs at an IB↔OoB transition frame rather than mid-segment inside an out-of-body scene.

Our contributions. **(1) Timestamp annotation protocol and controlled evaluation:** a lightweight 15–20 min per-video labelling approach enabling 17 da Vinci procedures across 12 types to be annotated, with all methods trained and evaluated under identical conditions. **(2) GatedTextFusion-LSTM:** a strong baseline achieving $F1_{smooth} = 0.9849$ and $AP = 0.9956$, outperforming OoBNet ($F1_{smooth} = 0.9390$, $AP = 0.9572$) and IODA supervised ($F1_{smooth} = 0.9663$, $AP = 0.9883$), while using 24% fewer parameters than IODA and yielding only 1 missed OoB segment (false negative) across the 12 test videos, located at a transition frame rather than within an out-of-body scene. **(3) Semi-supervised extension:** LSTM self-training on 15 pseudo-labelled videos further improves AP to 0.9964, requiring no additional human annotation.

2 Related Work

Out-of-body detection. OoBNet [1] uses MobileNetV2 with a stateful LSTM, validated across multiple laparoscopic centres. IODA [2] employs AlexNet with

a stateless LSTM, trained on 100 laparoscopic videos across 23 procedure types. ROBAN [3] targets real-time live-streaming robotic surgery. Zohar et al. [5] apply semi-supervised re-annotation, Wang et al. [6] use self-supervised pre-training with minimal labels, and Leon Barbed et al. [7] use unsupervised descriptors for endoscopy segmentation. None exploit surgical language supervision.

Vision-language models and prompt engineering. SurgVLP [4] pre-trains a joint image-text encoder on surgical video, producing 768-dimensional embeddings where similar surgical concepts are geometrically close. CLIP [8] showed that a single text query can guide zero-shot classification; CoOp [9] and ProDA [10] learn prompt representations from data. Our approach instead maintains a fixed interpretable prompt bank with per-frame soft selection, preserving human readability.

3 Method

3.1 Problem Formulation

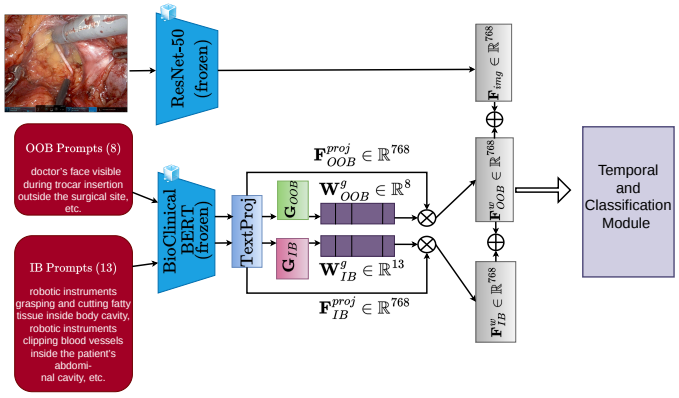
Given a robotic surgical video $\{x_t\}_{t=1}^{T}$, the goal is to predict a binary label $y_t \in \{0, 1\}$ per frame ($y_t = 1$ indicates OoB). Each frame is encoded by a frozen SurgVLP ResNet-50 backbone(layer 4 unfrozen in the case of GatedTextFusion MLP): $\mathbf{F}_{\text{img},t} = \text{SurgVLP}_{\text{vision}}(x_t) \in \mathbb{R}^{768}$. We process videos in non-overlapping windows of $L = 64$ frames, carrying the LSTM hidden state across windows. Fig. 1 gives an overview.

3.2 Multi-Prompt Text Bank

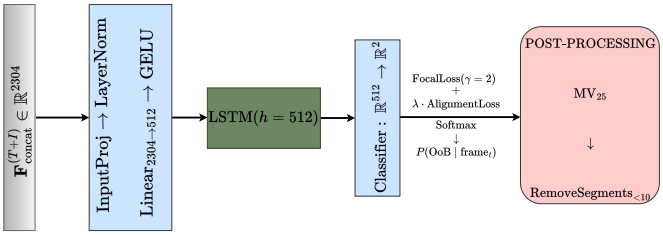
We define a curated bank of $K_{\text{IB}} = 13$ in-body and $K_{\text{OB}} = 8$ out-of-body prompts describing instrument-tissue interactions and extra-corporeal scenes, respectively, encoded once offline and L2-normalised. Table 1 shows representative examples.

Table 1. Representative prompt examples (2 of 13 IB; 2 of 8 OoB).

Class	Example prompt
In-body	Robotic instruments grasping and cutting fatty tissue inside body cavity
In-body	Placing surgical mesh inside the abdominal wall during hernia repair
Out-of-body	Doctor's face visible during trocar insertion outside the surgical site
Out-of-body	Camera outside the patient showing the operating room environment



(a) Gated Multi-Prompt Text Fusion



(b) Temporal and Classification Module

Fig. 1. Overview of **GatedTextFusion-LSTM**. **(a)** A frame is encoded by a frozen ResNet-50 backbone into $\mathbf{F}_{\text{img}} \in \mathbb{R}^{768}$. A bank of 13 IB and 8 OoB text prompts is projected via TextProj; gate networks $\mathbf{G}_{\text{IB}}$, $\mathbf{G}_{\text{OoB}}$ produce soft attention weights, yielding $\mathbf{F}^{\text{w}}_{\text{IB}}, \mathbf{F}^{\text{w}}_{\text{OoB}} \in \mathbb{R}^{768}$ concatenated into $\mathbf{F}^{(T+I)}_{\text{concat}} \in \mathbb{R}^{2304}$. **(b)** The representation passes through LayerNorm $\rightarrow$ $\text{Linear}_{2304\rightarrow512}$ $\rightarrow$ GELU, a 2-layer stateful LSTM ($h=$ 512), and a linear classifier. Post-processing: sliding majority vote (window 25) and short-segment suppression (<10 frames).

3.3 Gated Text Fusion

A shared trainable projection $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{768\times768}$ maps frozen text embeddings into the visual feature space: $\mathbf{T}'_{\text{IB}} = \mathbf{T}_{\text{IB}}\mathbf{W}^{\top}_{\text{proj}}$, $\mathbf{T}'_{\text{OB}} = \mathbf{T}_{\text{OB}}\mathbf{W}^{\top}_{\text{proj}}$. Two separate gate networks are used because IB and OoB scenes are visually disjoint and a single shared gate would introduce conflicting gradient signals. Per-frame soft attention weights over the prompt banks are conditioned on the visual feature; the joint representation concatenates the visual and both fused text embeddings:

$$\mathbf{w}^{\text{IB}}_t = \text{softmax}(\mathbf{W}_{\text{IB}}\mathbf{F}_{\text{img},t}), \quad \mathbf{F}^{\text{w}}_{\text{IB},t} = \mathbf{w}^{\text{IB}}_t \cdot \mathbf{T}'_{\text{IB}},$$
$$\mathbf{F}^{(T+I)}_{\text{concat},t} = [\mathbf{F}_{\text{img},t} \parallel \mathbf{F}^{\text{w}}_{\text{IB},t} \parallel \mathbf{F}^{\text{w}}_{\text{OB},t}] \in \mathbb{R}^{2304}, \tag{1}$$

and symmetrically for $\mathbf{W}_{\text{OB}} \in \mathbb{R}^{8\times768}$, preserving visual information while enriching it with adaptive semantic context from both classes.

3.4 Alignment Loss

An alignment auxiliary loss minimises cosine distance between fused text embeddings and visual features ($\mathbf{F}_{\text{img},t}$ detached, updating only $\mathbf{W}_{\text{proj}}$ and gate networks), class-averaged to equalise gradient contributions from both classes:

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left[\frac{1}{|\mathcal{B}_0|}\sum_{t:\, y_t=0}\big(1-\cos(\mathbf{F}^{\text{w}}_{\text{IB},t}, \mathbf{F}_{\text{img},t})\big) + \frac{1}{|\mathcal{B}_1|}\sum_{t:\, y_t=1}\big(1-\cos(\mathbf{F}^{\text{w}}_{\text{OB},t}, \mathbf{F}_{\text{img},t})\big)\right]. \tag{2}$$

Total loss: $\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda\mathcal{L}_{\text{align}}$, $\lambda = 0.5$, focal $\gamma = 2$.

3.5 Temporal Modelling with Stateful LSTM

Per-frame representations are projected and processed by a 2-layer LSTM:

$$\begin{aligned}\mathbf{e}_t &= \text{GELU}(\text{Linear}_{2304\to512}(\text{LayerNorm}(\mathbf{F}^{(T+I)}_{\text{concat},t}))),\\ \hat{y}_t &= \text{Linear}(\text{LSTM}(\mathbf{e}_t, \mathbf{s}_{t-1})).\end{aligned} \tag{3}$$

Hidden state $\mathbf{s}_t$ is carried across windows within a video, reset at video boundaries; overlapping windows (stride 32) increase gradient updates during training (truncated BPTT).

3.6 Training Procedure

We train in two stages to separate per-frame text-visual alignment (Stage 1) from sequence-level temporal modelling (Stage 2), avoiding LSTM gradient interference with the projection layer before convergence. **Stage 1**: GatedTextFusion MLP, `layer4` partially unfrozen ($\text{lr} = 10^{-3}$, $\text{lr}_{\text{bb}} = 10^{-4}$, 15 epochs). **Stage 2**: LSTM head warm-started, backbone frozen ($\text{lr} = 10^{-3}$, 15 epochs). Post-processing: majority-vote filter ($w = 25$) and short-segment suppression (<10 frames), applied to all models.

3.7 Semi-Supervised Extension

We extend to unlabeled data via LSTM self-training: the supervised model generates confidence-filtered pseudo-labels on unannotated videos and fine-tunes on the combined pool. Concretely, the supervised model runs in inference mode, producing a video-level confidence score $\bar{c}_v = \frac{1}{T_v}\sum_{t=1}^{T_v}\max(p_t, 1-p_t)$, where p_t is the OoB probability at frame t. We select the $K = 15$ videos with highest $\bar{c}_v$ and fine-tune for 3 epochs on the combined pool of 5 labeled and 15 pseudo-labeled videos ($\text{lr} = 2\times10^{-4}$, $\lambda = 0.2$, backbone frozen).

4 Experiments

4.1 Dataset

We use a proprietary dataset of robotic minimally invasive surgical procedures (da Vinci system) spanning **12 procedure types**: inguinal hernia repair, radical prostatectomy, gynaecological procedures, fundoplication, radical cystectomy, partial nephrectomy, and others. The dataset comprises 17 fully annotated videos split into **5 training videos** ($\approx$ 7h 32 min) and **12 held-out test videos** ($\approx$ 16h 54 min), giving $\approx$ **24h 26 min** of annotated footage.

Ethics and annotation. Recordings were collected under an institutional data-use agreement with written informed consent, approved under institutional ethics review (approval number 2025-R004-33). A single annotator marked OoB segment start/end timestamps in one pass (15 to 20 min per video), requiring no per-frame labelling; frame-level labels are derived by converting timestamps to frame indices. This protocol is practical for institutions with limited annotation resources. All models are evaluated against the same timestamp-derived labels.

4.2 Models and Baselines

OoBNet [1]: MobileNetV2 + stateful LSTM (17.0M); no language guidance. Evaluated zero-shot and supervised.
IODA [2]: AlexNet + stateless LSTM on 32-frame windows (40.7M); no language supervision. Evaluated zero-shot and supervised. Neither IODA nor OoBNet was originally developed or validated on robotic surgical video: IODA was trained exclusively on laparoscopic procedures, and OoBNet was not validated on da Vinci robotic surgery.

Four ablations: **Visual-Only MLP** (SurgVLP→MLP; no text), **GatedTextFusion MLP** (gated fusion + MLP; no LSTM), **Visual-Only LSTM** (SurgVLP→LSTM; no text), **GatedTextFusion LSTM** (full model; proposed). Our model has 31.1M inference-active parameters (ResNet-50 + task head; the BioClinicalBERT text encoder runs once offline to encode the 21 prompts and is not called at inference), 24% fewer than IODA's 40.7M, despite outperforming it on every metric.

4.3 Training Details

All SurgVLP-based models share an optimiser (AdamW, weight decay 10^{-4}), loss (focal $\gamma = 2$, to address the 9:1 IB/OoB class imbalance), and schedule (cosine annealing, 15 epochs). All videos are pre-extracted at 5 fps for all models. GatedTextFusion variants use $K_{\text{IB}} = 13$, $K_{\text{OB}} = 8$ prompts with $\lambda = 0.5$ alignment weight. LSTM models use hidden dimension 512, 2 layers, truncated BPTT ($L = 64$, stride 32). GatedTextFusion-LSTM uses two-stage training (§3.6); all others use single-stage with $\text{lr} = 10^{-3}$. Training is on a single NVIDIA Tesla V100 GPU. Baselines re-implemented in PyTorch. *OoBNet*: TensorFlow inference-only

release; training re-implemented in PyTorch from scratch per paper; 15 epochs without early stopping. *IODA*: focal loss ($\gamma = 2$) replaces original unweighted loss to address 9:1 IB/OoB imbalance; pretrained weights loaded with `strict=False` for zero-shot. Post-processing (majority-vote $w = 25$, min-segment 10) applied to all models. All other hyperparameters match respective papers exactly.

Evaluation protocol. All models use their last-epoch checkpoint. We report $\mathrm{F1}_{\mathrm{raw}}$, $\mathrm{F1}_{\mathrm{smooth}}$ (majority-vote smoothing, $w = 25$, min-segment 10), AP (threshold-free, post-processing-independent), MCC, Jaccard, and total FP/FN segment counts across the 12 test videos (defined in Table 2 footnote). All threshold-dependent metrics are computed at $\tau = 0.5$.

5 Results

5.1 Main Results

Table 2 reports all models, all trained on the same 5 labelled videos and evaluated on the same 12-video test set. GatedTextFusion-LSTM achieves $\mathrm{F1}_{\mathrm{smooth}} = 0.9849$, outperforming IODA supervised ($\mathrm{F1}_{\mathrm{smooth}} = 0.9663$) and OoBNet ($\mathrm{F1}_{\mathrm{smooth}} = 0.9390$). Both baselines drop in zero-shot transfer, confirming a domain gap between laparoscopic and robotic video. AP is threshold-free and unaffected by smoothing asymmetry: our AP = 0.9956 exceeds IODA supervised (0.9883) by +0.0073 and OoBNet supervised (0.9572) by +0.0384. Semi-supervised self-training further improves AP to 0.9964. Ablations confirm synergy: LSTM alone (AP = 0.9868) and text fusion alone (AP = 0.9923) each underperform the full combination. The FP/FN columns reveal segment-level failure modes: IODA zero-shot achieves few false positives (FP = 4) but a high false negative count (FN = 14); GatedTextFusion-LSTM achieves the best supervised balance (FP = 19, FN = 1) across the 12 test videos. This single FN occurs at a transition frame, not inside a real out-of-body segment, so no sensitive content is left unredacted. Performance remained consistent across Full-HD and SD-PAL acquisition settings.

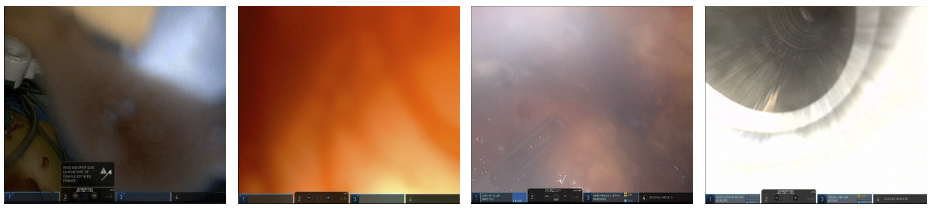
5.2 Qualitative Analysis

Fig. 2 contrasts our model against the baselines (OoBNet, IODA), with the ground-truth class and which method is correct varying across panels. These are illustrative examples, not an exhaustive account of every error. (a) Out-of-body OR view: **Ours** correct, both baselines misclassify as in-body. (b) In-body, heavily blurred near-field occlusion: both baselines correct, **Ours** produces a false positive. (c) In-body frame with electrocautery smoke: **Ours** correct, both baselines misclassify as OoB. (d) Out-of-body trocar-valve transition frame: a baseline is correct, **Ours** misclassifies as in-body, and it occurs at a transition frame, not inside a real out-of-body scene.

Table 2. Comparison on 12 held-out robotic procedures. AP is threshold-free and post-processing-independent. FP/FN: total segment counts across the 12 test videos (post-smoothing). †Pretrained laparoscopic weights, no robotic fine-tuning. ‡Original TF weights, no robotic fine-tuning.

Model	$F1_{raw}\uparrow$	$F1_{smooth}\uparrow$	AP↑	MCC↑	Jac↑	FP↓	FN↓
OoBNet [1] (zero-shot)‡	0.8545	0.8671	0.9282	0.8517	0.7821	114	10
OoBNet [1] (supervised)	0.9317	0.9390	0.9572	0.9247	0.8883	43	1
IODA [2] (zero-shot)†	0.8427	0.8436	0.9767	0.8336	0.7516	4	14
IODA [2] (supervised)	0.9601	0.9663	0.9883	0.9603	0.9367	8	6
Visual-Only MLP	0.9685	0.9796	0.9902	0.9744	0.9607	41	2
Visual-Only LSTM	0.9691	0.9808	0.9868	0.9762	0.9629	35	2
GatedTextFusion MLP	0.9694	0.9791	0.9923	0.9738	0.9598	46	1
GatedTextFusion LSTM	0.9742	**0.9849**	0.9956	**0.9809**	**0.9704**	19	1
+ semi-sup.	**0.9793**	0.9840	**0.9964**	0.9800	0.9688	18	1

$F1_{raw}$: no post-processing. $F1_{smooth}$: majority-vote ($w = 25$, min-seg 10). AP is threshold-free and post-processing-independent; primary comparison metric. All threshold-dependent metrics were computed at $\tau = 0.5$. FP: total spurious OoB segments across the 12 test videos (data loss). FN: total missed OoB segments across the 12 test videos (privacy risk). Both computed post-smoothing. Smoothing is integral to SurgAnon but not to baseline design.



(a) TP (b) FP (c) TN (d) FN

Fig. 2. Qualitative comparison between **Ours** and the baselines (OoBNet, IODA), relative to **Ours**: (a) is a True Positive, (b) is a False Positive, (c) is a True Negative, and (d) is a False Negative.

6 Discussion and Analysis

Where baselines fail and why language helps. Purely visual methods struggle where surface appearance is ambiguous or shifts across domains; our learned gates leverage semantic language cues to resolve these cases and consistently outperform the baselines. The frozen SurgVLP backbone (AP = 0.9902) surpasses IODA supervised (AP = 0.9883) with only 5.99M trainable task-head parameters vs. IODA's 40.7M total; gated fusion adds +0.0021 AP and, with LSTM, yields a 4.8× larger F1 gain, confirming temporal modelling and language guidance are synergistic. Gate weights averaged across training frames suggest coherent selection: IB by grasping/cutting (0.260) and vessel clipping (0.255); OoB by trocar insertion (0.308) and face visibility (0.255); cross-class similarity drops to −0.88. Despite training on only five procedures, performance remained strong

across twelve held-out procedure types, suggesting that the learned representation generalises across diverse robotic surgeries within our dataset, though limited to one institution and platform. The F1 drop ($0.9849 \rightarrow 0.9840$) in semi-supervised training reflects noisier pseudo-label boundaries, while AP improves ($0.9956 \rightarrow 0.9964$). GatedTextFusion-LSTM has just 1 missed OoB segment across the 12 test videos, and it occurs at a transition frame, not inside a real out-of-body scene.

7 Conclusion

GatedTextFusion-LSTM achieves $\mathrm{F1}_{\mathrm{smooth}} = 0.9849$ and $\mathrm{AP} = 0.9956$ on 17 da Vinci procedures, outperforming OoBNet and IODA with 24% fewer parameters compared to IODA. It yields only 1 missed OoB segment across the 12 test videos, and this residual false negative occurs at a trocar transition rather than within an out-of-body scene, consistent with OoB gate weights prioritising face visibility (0.255); semi-supervised self-training further pushes AP to 0.9964. Due to patient privacy, the underlying video dataset cannot be released, consistent with the field norm followed by OoBNet and IODA. Code and trained model weights are available at `https://github.com/suyashkumar234/SurgAnon`.

Acknowledgments. This work was supported by the 3IA Côte d'Azur Investments under Grant ANR-23-IACL-0001.

Disclosure of Interests. The authors declare no competing interests.

References

1. Lavanchy, J.L., Vardazaryan, A., Mascagni, P., Mutter, D., Padoy, N.: Preserving privacy in surgical video analysis using a deep learning classifier to identify out-of-body scenes in endoscopic videos. Sci. Rep. **13**, 9235 (2023). https://doi.org/10.1038/s41598-023-36453-1
2. Schulze, A., Tran, D., Daum, M.T.J., Kisilenko, A., Maier-Hein, L., Speidel, S., Distler, M., Weitz, J., Müller-Stich, B.P., Bodenstedt, S., Wagner, M.: Ensuring privacy protection in the era of big laparoscopic video data: development and validation of an inside outside discrimination algorithm (IODA). Surg. Endosc. **37**, 6153–6162 (2023). https://doi.org/10.1007/s00464-023-10078-x
3. De Backer, P., Simoens, J., Mestdagh, K., Hofman, J., Eckhoff, J.A., Jobczyk, M., Van Eetvelde, E., D'Hondt, M., Moschovas, M.C., Patel, V., Van Praet, C., Fuchs, H.F., Debbaut, C., Decaestecker, K., Mottrie, A.: Privacy-proof live surgery streaming: Development and validation of a low-cost, real-time robotic surgery anonymization algorithm. Ann. Surg. **280**(1), 13–20 (2024). https://doi.org/10.1097/SLA.0000000000006245
4. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. Med. Image Anal. (2025). https://doi.org/10.1016/j.media.2025.103467

5. Zohar, M., Bar, O., Neimark, D., Hager, G.D., Asselmann, D.: Accurate detection of out of body segments in surgical video using semi-supervised learning. In: Proc. MIDL 2020, PMLR, pp. 1–14 (2020)
6. Wang, Z., Liu, X., Perreault, C., Jarc, A.: Automatic detection of out-of-body frames in surgical videos for privacy protection using self-supervised learning and minimal labels. arXiv preprint arXiv:2303.18106 (2023)
7. Leon Barbed, O., Azagra, P., Plo, J., Murillo, A.C.: Automated overview of complete endoscopies with unsupervised learned descriptors. Int. J. Comput. Assist. Radiol. Surg. **21**, 617–624 (2026). https://doi.org/10.1007/s11548-025-03502-1
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. ICML 2021, PMLR, vol. 139, pp. 8748–8763 (2021)
9. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vis. **130**, 2337–2348 (2022). https://doi.org/10.1007/s11263-022-01653-1
10. Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: Proc. CVPR 2022, pp. 5206–5215 (2022). https://doi.org/10.1109/CVPR52688.2022.00513