Test-Time Adaptation for Rare Surgical Phase Recognition: Bridging the Coverage-Gap Paradox*

Ho-min Park[1**], Francesca Tozzi[2], Robbe De Muynck[2], Narim Kim[1], Niki Rashidian[3], Wouter Willaert[3], Wesley De Neve[1], and Joris Vankerschaver[1]

[1] Ghent University Global Campus, Incheon, South Korea
{homin.park,narim.kim,wesley.deneve,joris.vankerschaver}@ghent.ac.kr
[2] Ghent University, Ghent, Belgium
{francesca.tozzi,rodmuync.demuynck}@ugent.be
[3] Ghent University Hospital, Ghent, Belgium
{nikdokht.rashidian,wouter.willaert}@ugent.be

Abstract. Surgical phase recognition models struggle with *rare phases*—states under-represented in training annotations, occupying <5% of training frames. We identify the **coverage-gap paradox**: increasing rare-phase training videos fails to proportionally improve accuracy because prototype dilution averages out intra-class variation, a phenomenon we observe consistently across three representative classifier families (k-NN, linear probes, and prototypes). To bridge this gap, we propose a lightweight **test-time adaptation (TTA)** framework (20K parameters, <0.1 ms/frame) whose core mechanism is an *adaptive pseudo-label threshold* (q-th percentile) that avoids the empty-pool failure of fixed thresholds, combined with temporal smoothing for pseudo-label precision filtering and auto-guided annotation (B=5 frames). On **three datasets** spanning gastric bypass (MB140, 13 phases), cholecystectomy (Cholec80, 7 phases), and cataract surgery (Cat-101, 10 phases)—our method improves rare-phase accuracy by **+29.2/+31.5/+33.3 pp**, outperforming temporal convolutional network (TCN), gated recurrent unit (GRU), and TENT-style baselines.

Keywords: Coverage-gap paradox · Rare phases · Surgical phase recognition · Test-time adaptation

1 Introduction

Automated surgical phase recognition enables real-time decision support, operating-room scheduling, and skill assessment [16,18]. While temporal models [5,6,7] achieve high overall accuracy on common phases, their per-class performance on *rare phases*—states occupying <5% of training frames—remains poor, as we

* Code: https://github.com/powersimmani/tta-rare-surgical
** Corresponding author.

demonstrate in Sec. 4.2. We define a phase as *rare* when it has near-zero representation in labeled training data, whether due to clinical infrequency (e.g., closure of Petersen's space in gastric bypass, ∼0.5% of frames) or incomplete annotation of an otherwise routine step.

The fundamental challenge is *notably skewed class distributions at the training-set level*: in multi-centric datasets like MultiBypass140 [13], phases such as *closure_petersen_space* occupy <1% of total frames (∼0.3–0.7% per-video frame share at Inselspital Bern) yet appear in nearly every video at Strasbourg. Standard remedies (class-balanced loss [4], oversampling [2], long-tail learning [29]) assume sufficient examples exist and cannot compensate when entire classes have near-zero coverage.

The coverage-gap paradox. We empirically characterize a counter-intuitive finding: let $\mathrm{Acc}_{\mathcal{R}}(m)$ denote rare-phase accuracy (frame-level) when m training videos contain rare-phase labels. We observe $\mathrm{Acc}_{\mathcal{R}}(5) - \mathrm{Acc}_{\mathcal{R}}(1) < 5$ percentage points (pp) on MB140, while the gap to oracle (accuracy with full rare-phase labels) remains >28 pp at m=20. This paradox—more data fails to close the gap—arises from *prototype dilution* (Fig. 1A): averaging features from diverse surgical contexts moves centroids toward the global mean, narrowing rare-phase coverage while common phases dominate the feature space. We observe this consistently across three representative classifier families: on MB140 (m=1), k-nearest neighbor (k-NN) achieves 3.0–3.8% rare-phase accuracy, linear probes 6.2%, and prototypes 21.7%—all below 22% despite >63% overall accuracy.

Our approach. We shift adaptation to *test time* (Fig. 1B): (1) cosine similarity prediction, (2) confidence-weighted temporal smoothing, (3) *adaptive-threshold* pseudo-label prototype updating, (4) auto-guided frame selection, and (5) minimal human annotation (B=5 frames). Our approach requires *no source data access*—only compact prototypes (<20K params).

Contributions. (i) We identify and empirically characterize the coverage-gap paradox across three representative classifier families (k-NN, linear, prototype). (ii) We propose a lightweight TTA framework centered on an *adaptive pseudo-label threshold* that resolves the empty-pool failure of fixed thresholds, with temporal smoothing serving as a precision filter for downstream pseudo-label selection. (iii) We validate on three datasets spanning three surgery types with statistical significance tests, cross-center evaluation, and a 2×2 factorial analysis isolating each component's contribution.

2 Related Work

Surgical phase recognition. Methods evolved from HMMs [18] and CNNs [11,22] to temporal models (TCN [5], MS-TCN [1], Transformers [6,7]) and foundation-model features [8,17,28]. The review presented in [15] reports phase accuracy varying widely (57–91%), yet *none target rare phases*. Lavanchy et al. [13] demonstrated cross-institutional degradation on MultiBypass140; Yuan et al. [27] proposes few-shot TTA for cross-institutional workflow—complementary to our within-procedure rare-phase focus. Prototypical networks [21,24] suit data scarcity,

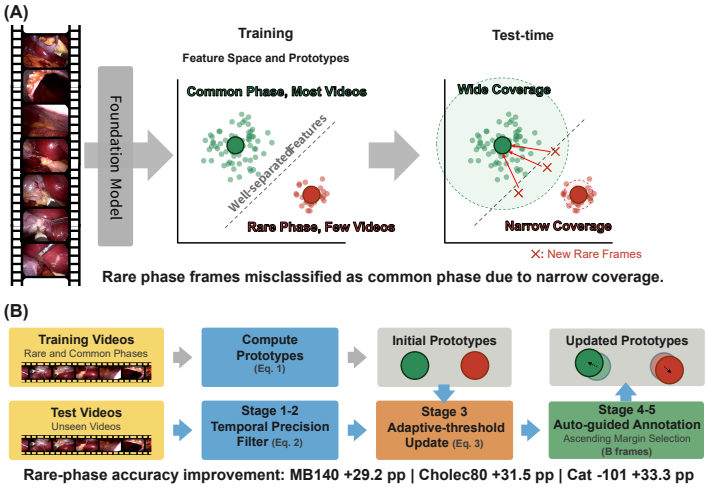

Fig. 1. (A) The coverage-gap paradox: common-phase prototypes trained on most videos achieve wide coverage in feature space, while rare-phase prototypes from few videos have narrow coverage, causing new rare frames (×) to be misclassified as common phases. (B) Our TTA framework (blue arrows = test-time path, grey = training): base prototypes computed from training data (Eq. 1) are adapted at test time through temporal precision filtering (Stage 1–2, Eq. 2), adaptive-threshold pseudo-label update (Stage 3, Eq. 3), and auto-guided annotation (Stage 4–5, Eq. 4, B=5 frames).

but in surgical video within-class variation causes *prototype dilution*—the coverage-gap paradox we formalize.

Class imbalance and test-time adaptation. Standard methods for overcoming imbalance [2,4,29] address frame-level imbalance but not the video-level coverage gap where entire classes are absent. Class-conditional self-training methods (CBST [30], SHOT [14]) use per-class thresholds for pseudo-label selection during retraining. Our method differs by targeting source-free TTA over *fixed prototypes* (no backbone retraining), with an explicit empty-pool guarantee via the per-class adaptive percentile. TTA adapts during inference without source data: TENT [23] minimizes entropy, SHOT [14] aligns distributions; medical TTA targets imaging domain shifts [10,12,26]. We differ by targeting surgical *temporal* data with lightweight prototype adaptation (<20K params) rather than deep parameter updates.

Active learning. Uncertainty selection [9,20] fails for rare classes due to poorly calibrated confidence; our auto-guided annotation pre-filters rare-phase candidates via TTA predictions before margin-based selection.

3 Method

3.1 Problem Formulation

Given training videos $\mathcal{V}_{\text{train}}$ with per-frame labels across C phase classes, let $\mathcal{R} \subset \{1, \ldots, C\}$ denote the set of rare classes and $\mathcal{D}_c$ the set of training frame indices with label c. We compute ℓ_2-normalized prototypes:

$$\mathbf{p}_c = \frac{\bar{\mathbf{x}}_c}{\|\bar{\mathbf{x}}_c\|}, \quad \bar{\mathbf{x}}_c = \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \mathbf{x}_i \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the foundation model feature vector for frame i. At test time, each frame t is assigned the class whose prototype has highest cosine similarity: $\hat{y}_t = \arg\max_c \cos(\mathbf{x}_t, \mathbf{p}_c)$.

3.2 Coverage-Gap Paradox

When only m videos contain rare phase c, increasing m should improve $\mathbf{p}_c$, but *prototype dilution* moves $\mathbf{p}_c$ toward the global mean, reducing discriminability. On MB140, rare-phase accuracy plateaus beyond m=2, and all three families (k-NN, linear, prototype) fail (<22% despite >63% overall).

3.3 TTA Framework

Stage 1–2: Temporal Precision Filter. Stage 1 computes cosine-similarity predictions $\hat{y}_t$ for all test frames. Stage 2 applies a confidence-weighted majority vote within sliding window w (w=15 frames at 1 fps, a 15-s window; Sec. 4.4) to suppress transient false positives, yielding smoothed predictions $\tilde{y}_t$ and improving pseudo-label pool precision (27%→44% on MB140):

$$\tilde{y}_t = \arg\max_c \sum_{j=t-\lfloor w/2 \rfloor}^{t+\lfloor w/2 \rfloor} \mathbf{1}[\hat{y}_j = c] \cdot \text{conf}_j, \quad \text{conf}_j = \max_c \cos(\mathbf{x}_j, \mathbf{p}_c) \tag{2}$$

Stage 3: Adaptive-Threshold Pseudo-Label Update (core mechanism). For rare class $c \in \mathcal{R}$, update via exponential moving average (EMA):

$$\mathbf{p}'_c = \text{normalize}\big(\alpha \cdot \mathbf{p}_c + (1{-}\alpha) \cdot \bar{\mathbf{x}}_c^{\text{pseudo}}\big) \tag{3}$$

where α=0.5, $\bar{\mathbf{x}}_c^{\text{pseudo}} = \text{mean}(\{\mathbf{x}_t : \tilde{y}_t{=}c \wedge \text{conf}_t{\geq}\tau_c\})$. Critically, τ_c is the q-th percentile of confidence scores for class c (default q=75), not a fixed threshold. A 2×2 factorial confirms this design is the dominant accuracy driver (+22 pp), while fixed thresholds (p=0.95) yield empty pseudo-label pools for rare phases (Sec. 4.2).

Stage 4–5: Auto-Guided Annotation. Pre-filter rare candidates ($\mathcal{C} = \{t : \tilde{y}_t \in \mathcal{R}\}$), select B frames with minimum decision margin:

$$s_t = \max_c \cos(\mathbf{x}_t, \mathbf{p}'_c) - \max_{c' \neq \hat{c}} \cos(\mathbf{x}_t, \mathbf{p}'_{c'}), \quad \hat{c} = \arg\max_c \cos(\mathbf{x}_t, \mathbf{p}'_c) \tag{4}$$

Algorithm 1 Test-Time Adaptation for Rare Phase Recognition

Require: Test video $\{\mathbf{x}_t\}_{t=1}^{T}$, prototypes $\{\mathbf{p}_c\}_{c=1}^{C}$, rare classes $\mathcal{R}$, window w, percentile q, EMA weight α, budget B
Ensure: Adapted prototypes $\{\mathbf{p}'_c\}$

for $t = 1$ **to** T **do**
  $\text{sim}_{t,c} \leftarrow \cos(\mathbf{x}_t, \mathbf{p}_c)$ for all c; $\hat{y}_t \leftarrow \arg\max_c \text{sim}_{t,c}$; $\text{conf}_t \leftarrow \max_c \text{sim}_{t,c}$
end for
for $t = 1$ **to** T **do**
  $\tilde{y}_t \leftarrow \arg\max_c \sum_{j=\max(1,t-w/2)}^{\min(T,t+w/2)} \mathbf{1}[\hat{y}_j{=}c] \cdot \text{conf}_j$
end for
for each $c \in \mathcal{R}$ **do**
  $\tau_c \leftarrow \text{percentile}_q(\{\text{conf}_t : \tilde{y}_t = c\})$ ▷ adaptive, not fixed
  $\mathcal{P}_c \leftarrow \{\mathbf{x}_t : \tilde{y}_t{=}c \wedge \text{conf}_t \geq \tau_c\}$ ▷ pseudo-label pool
  if $|\mathcal{P}_c| > 0$ **then**
    $\bar{\mathbf{x}}_c^{\text{pseudo}} \leftarrow \text{normalize}(\text{mean}(\mathcal{P}_c))$
    $\mathbf{p}'_c \leftarrow \text{normalize}(\alpha \cdot \mathbf{p}_c + (1{-}\alpha) \cdot \bar{\mathbf{x}}_c^{\text{pseudo}})$
  end if
end for
$\mathcal{C} \leftarrow \{t : \tilde{y}_t \in \mathcal{R}\}$ ▷ rare-phase candidates
for $b = 1$ **to** B **do**
  $t^* \leftarrow \arg\min_{t \in \mathcal{C}} (\max_c \cos(\mathbf{x}_t, \mathbf{p}'_c) - \max_{c' \neq \hat{c}} \cos(\mathbf{x}_t, \mathbf{p}'_{c'}))$
  Query label y^* for $\mathbf{x}_{t^*}$; update $\mathbf{p}'_{y^*} \leftarrow \text{normalize}(\alpha \cdot \mathbf{p}'_{y^*} + (1{-}\alpha) \cdot \mathbf{x}_{t^*})$
  $\mathcal{C} \leftarrow \mathcal{C} \setminus \{t^*\}$
end for

A human annotator labels $B{=}5$ frames (oracle in experiments); prototypes are updated accordingly (Algorithm 1).

Multi-prototype bank. For $C \geq 10$, we use $K{=}3$ sub-prototypes per class via k-means; $K{=}1$ for $C \leq 7$ (over-partitioning creates boundary confusion with few classes).

Temporal vs. Combined. We report two operating points: *Temporal* (Stage 1–3) is fully source-free and unsupervised; *Combined* (Stage 1–5) additionally consumes $B{=}5$ human-annotated frames, so its gains partly reflect this minimal supervision, whereas Temporal rows are strictly source-free.

4 Experiments

4.1 Datasets and Setup

MultiBypass140 (MB140) [13]: 140 gastric bypass videos (70 Bern + 70 Strasbourg), 13 phases. Rare: *closure_petersen_space*, *closure_mesenteric_defect*. Features: LemonFM [3] (1536d). 5-fold cross-validation on Bern, cross-center to Strasbourg.

Cholec80 [22]: 80 cholecystectomy videos, 7 phases. 40/8/32 split (standard [22]). Features: DINOv2 ViT-L/14 (1024d). We target *CalotTriangleDissection*: present in every case but short (<5% of frames), it suits annotation-scarcity simulation via label masking (below).

Cataract-101 (Cat-101) [19]: 101 cataract surgery videos, 10 phases. 70/11/20 split (following [19]). Features: LemonFM (1536d). Rare: *CapsulePolishing* (4.2%), *Incision* (4.8%).

Rarity simulation. Training restricted to m *videos* containing rare-phase labels (m=1 for MB140/Cat-101, m=5 for Cholec80). For Cholec80, CalotTriangleDissection appears in nearly all videos, so removing videos would confound annotation scarcity with overall data loss. Instead, we use *label masking*: all 40 training videos are retained for common-phase prototype computation, but rare-phase labels are masked in all but m videos—simulating the practical scenario where a phase exists but lacks annotations.

Self-Training baseline. TENT-style entropy minimization adapted to a prototype classifier: prototypes are updated with frames whose softmax-normalized similarity exceeds a fixed threshold (p=0.95). For rare classes, this fixed threshold typically yields an empty pseudo-label pool—the failure mode our adaptive-percentile design (Sec. 3, Stage 3) resolves by construction.

Evaluation. Overall accuracy and rare-phase mean accuracy, reported as mean±std over 5 random seeds. The seed determines which video(s) form the rare-phase source, substantially affecting accuracy (std ±15.4–16.5%); deployment would use all rare-phase videos, reducing this variance. Statistical significance: paired Wilcoxon signed-rank tests [25] over per-video scores (n=70 MB140 across 5 folds, n=160 Cholec80 = 32×5 seeds, n=100 Cat-101 = 20×5 seeds; $^{***}p<0.001$, $^{**}p<0.01$). *Caveat*: Cholec80 and Cat-101 pairs are pseudo-replicated across seeds (effective independent n=32 and n=20 respectively); reported p-values may be liberal.

4.2 Main Results

Table 1 presents the core comparison. **Temporal** (unsupervised, Stage 1–3) improves rare-phase accuracy on all datasets: +25.7 pp on MB140, +31.5 pp on Cholec80, +18.8 pp on Cat-101. A 2×2 factorial (5-fold×5-seed, Bern, K=1; MB140 rare-phase acc.) gives: (a) fixed, no-temporal 21.7% (=Base); (b) adaptive, no-temporal **43.7%**; (c) fixed, temporal 21.1%; (d) adaptive, temporal 40.5%. The *adaptive threshold* is the dominant driver (Δ=+22.0 pp on MB140, +15.4 pp on Cholec80). Temporal smoothing costs −3.2 pp in isolation [(d)<(b)] but raises pool precision (27%→44%), aiding Combined's oracle selection. With fixed p=0.95, zero pseudo-labels pass the threshold on MB140; the adaptive percentile resolves this by construction. **Combined** (K=3): +29.2 pp MB140 ($p<0.01$), +33.3 pp Cat-101. On Cat-101, Combined's rare-phase gain costs −10.8 pp overall—a structural limitation of cosine-similarity classifiers in high-C settings (Sec. 5). (The factorial's single-center K=1 Base of 21.7% differs from Table 1's K=3 Base of 25.9% due to independent rare-video draws.)

Causal variant. A past-only window on Cholec80 matches the bidirectional result (71.2% vs. 70.8%, w=15), enabling real-time deployment without future frames.

Table 1. Main TTA results. Rare = rare-phase mean accuracy (%), Ov = overall (%). Best non-oracle **bolded**; K=1 for Cholec80, K=3 for MB140/Cat-101. Temporal = Stage 1–3 (unsupervised); Combined = Stage 1–5 (+ B=5 oracle frames). Cholec80 uses label masking (all 40 training videos retained), under which "Combined"/"Oracle" are semantically circular (oracle annotation re-introduces masked labels), so we report only Temporal; MB140 and Cat-101 use video removal and are unaffected.

	MB140 (m=1)		**Cholec80** (m=5)		**Cat-101** (m=1)	
Method	Rare	Ov	Rare	Ov	Rare	Ov
Base	25.9	63.6	29.6	41.3	12.0	24.4
Temporal	51.6***	62.0	**61.1*****	48.6	30.8***	15.9
Self-Training	25.9	63.6	49.8***	45.8	12.0	24.4
Combined	**55.1****	65.2	–	–	**45.3*****	13.6
Oracle	85.3	63.5	–	–	98.7	10.7

$^{***}p<0.001$, $^{**}p<0.01$ (Wilcoxon vs. Base). Self-Training (fixed p=0.95) = Base on MB140/Cat-101 (zero pseudo-labels pass); on Cholec80 it admits some (+20.2 pp) vs. Temporal's +31.5 pp. Cat-101 Combined/Oracle degrade overall accuracy (−10.8/−13.7 pp; Sec. 5). We use K=1 for Cholec80 since K=3 over-partitions a 7-class space.

4.3 Comparison with Temporal Baselines

Table 2 compares against temporal post-processing on the same features. Non-adaptive baselines improve ≤8.4 pp; TENT-style [23] achieves +19.0–21.1 pp. Our Temporal reaches +16.3/+26.1 pp (MB140/Cholec80); TENT outperforms ours on MB140 (+19.0 vs. +16.3 pp) but trails on Cholec80 by 5.0 pp, using ∼100× more parameters. Table 2 uses video removal for Cholec80 (vs. Table 1's label masking), so Combined is reportable here. Its weak result (Table 2, +0.3 pp) reflects K=3 over-partitioning a 7-class space—*sub-prototype fragmentation*, a mechanism distinct from the Cat-101 overall-accuracy drop (Sec. 5), which stems from frozen-encoder/high-C boundary shift; the recommended Cholec80 configuration is K=1 (Temporal). Full end-to-end architectures (MS-TCN [1], Trans-SVNet [7]) lie outside our source-free regime; our claim is only that *temporal post-processing alone* does not bridge the coverage gap, as Table 2 evidences.

4.4 Ablation and Additional Analysis

Ablation. Sensitivity to key hyperparameters (MB140, m=1, single fold/seed): w=15 optimal; B=5 sufficient (33.6%; B=10: 41.8%); K=3 outperforms K=1 (42.7% vs. 33.6%). Both q and α are robust ($q \in \{50, 75, 90\}$, $\alpha \in \{0.3, 0.5, 0.7\}$, <3 pp variation). **Auto-guided hit rates** (fraction of selected frames truly rare-phase): Cholec80 37.6%, Cat-101 10.2%, MB140 3.14%. Ascending-margin selection beats random by 2.7×/2.3× on Cholec80/Cat-101; on MB140 the margin is small (1.05×) due to extreme rarity (<3% of frames).

Cross-center transfer. Table 3 (MB140 Bern↔Strasbourg, 5-fold; Temporal q=75, K=3): TTA improves in all four conditions, largest at m=1 (+37.7 pp,

Table 2. Temporal baseline comparison. Δ: rare-phase improvement (pp) over Base.

		Δ Rare Mean (pp)	
Method	**Ann.**	**MB140**	**Cholec80**
Majority Vote	—	+0.2	+6.1
TCN Post-hoc	—	−0.1	+5.5
GRU Smooth	—	−0.4	+8.4
TENT-style	—	+19.0	+21.1
Ours (Temporal)	—	+16.3	**+26.1**
Ours (Combined)	B=5	**+25.7**	+0.3

Base Rare: MB140 20.5%, Cholec80 44.7% (video removal). MB140 Base differs from Table 1 due to independent rare-video draws (±15.4%). TENT: entropy minimization over prototype logits.

Table 3. Cross-center transfer (MB140). Rare = rare-phase mean accuracy (%). Temporal = our unsupervised TTA.

		Base	**Temporal**	**Adapted**	**Oracle**
Bern→Stras	m=all	40.0	64.3	55.3	91.9
	m=1	8.9	**46.6**	42.6	86.8
Stras→Bern	m=all	26.8	58.5	50.4	92.1
	m=1	28.2	**57.0**	50.5	92.6

Temporal uses adaptive threshold (q=75). Adapted uses fixed p=0.95 + B=5 oracle frames; the fixed threshold limits pseudo-label generation (cf. factorial, Sec. 4.2).

Bern→Stras); even with all training data it adds +24–32 pp, confirming center-specific prototype dilution.

5 Discussion

Cross-center and self-targeting behavior. Per-video analysis (n=70) shows TTA is self-targeting: rare-phase videos gain +24.4 pp vs. +0.0 pp for non-rare ones, explaining the cross-center gains (Table 3) without target labels.

Structural trade-off. On MB140, Temporal costs only −1.6 pp overall and Combined improves it (+1.6 pp; Table 1). The larger Cat-101 drop (−10.8/−13.7 pp under Combined/Oracle) traces to the frozen-encoder representational ceiling under the high-C=10, source-free setting (prototype-anchoring is a scoped future direction; see Limitations).

6 Conclusion

We identified the coverage-gap paradox and proposed a lightweight TTA framework (20K params, 0.08 ms/frame on CPU) whose adaptive-threshold update

yields +22 pp in factorial analysis, with temporal filtering and auto-guided annotation (B=5) adding gains. Across three datasets we show +29–33 pp rare-phase improvement over temporal baselines; a causal variant retains >99% performance for real-time use.

Limitations. (i) On Cat-101, Combined updating trades −10.8 pp overall for the rare-phase gain (Sec. 5); prototype-anchoring is left as future work. (ii) Pool precision is low (11–25% on MB140) and convergence bounds remain open; the adaptive percentile prioritizes coverage over precision, recovered downstream by temporal smoothing and prototype geometry. (iii) The ablation (Sec. 4.4) uses a single fold/seed (per-fold values may diverge from multi-seed means by up to 12 pp); the headline factorial (Sec. 4.2) uses 5 seeds×5 folds (+22.0 pp adaptive-vs-fixed on MB140 Bern, K=1).

Acknowledgments. Ho-min Park and Joris Vankerschaver were supported by a UGent BOF grant (grant number BOF/STA/202109/039).

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

References

1. Abu Farha, Y., Gall, J.: MS-TCN: Multi-stage temporal convolutional network for action segmentation. In: CVPR. pp. 3575–3584 (2019)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
3. Che, C., Wang, C., Vercauteren, T., Tsoka, S., Garcia-Peraza-Herrera, L.C.: LEMON: A large endoscopic MONocular dataset and foundation model for perception in surgical settings. In: CVPR (2026)
4. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR. pp. 9268–9277 (2019)
5. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. In: MICCAI. pp. 343–352. Springer (2020)
6. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Opera: Attention-regularized transformers for surgical phase recognition. In: MICCAI. pp. 604–614. Springer (2021)
7. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: MICCAI. pp. 593–603. Springer (2021)
8. Hirsch, R., Caron, M., Cohen, R., et al.: Self-supervised learning for endoscopic video analysis. In: MICCAI. pp. 569–579. Springer (2023)
9. Holzinger, A.: Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Informatics **3**(2), 119–131 (2016)
10. Hu, M., Song, T., Gu, Y., et al.: Fully test-time adaptation for image segmentation. In: MICCAI. pp. 251–260. Springer (2021)
11. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network. IEEE Transactions on Medical Imaging **37**(5), 1114–1126 (2018)

12. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis **68**, 101907 (2021)
13. 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**, 2249–2257 (2024)
14. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In: ICML. pp. 6028–6039 (2020)
15. Liao, W., Zhu, Y., Zhang, H., Wang, D., Zhang, L., Chen, T., Zhou, R., Ye, Z.: Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: a systematic review. Frontiers in Surgery **12**, 1551838 (2025)
16. Maier-Hein, L., Eisenmann, M., et al.: Surgical data science–from concepts toward clinical translation. Medical Image Analysis **76**, 102306 (2022)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
18. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical Image Analysis **16**(3), 632–641 (2012)
19. Schoeffmann, K., Taschwer, M., Sarny, S., Munzer, B., Primus, M.J., Putzgruber, D.: Cataract-101: video dataset of 101 cataract surgeries. In: ACM Multimedia Systems Conference (MMSys). pp. 421–425. ACM (2018)
20. Settles, B.: Active learning. Morgan & Claypool Publishers (2012)
21. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS. pp. 4077–4087 (2017)
22. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging **36**(1), 86–97 (2017)
23. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
24. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys **53**(3), 1–34 (2020)
25. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945)
26. Wu, J., Liu, X., Wang, G., Zhang, S.: SicTTA: Single image continual test time adaptation for medical image segmentation. Medical Image Analysis **108**, 103859 (2025)
27. Yuan, K., Chen, T., Li, S., Lavanchy, J.L., Heiliger, C., Ozsoy, E., Huang, R., Bai, L., Navab, N., Srivastav, V., Ren, H., Padoy, N.: Recognizing Surgical Phases Anywhere: Few-Shot Test-time Adaptation and Task-graph Guided Refinement. In: MICCAI. Springer (2025)
28. 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. Medical Image Analysis **105**, 103644 (2025)

29. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10795–10816 (2023)
30. Zou, Y., Yu, Z., Vijaya Kumar, B.V.K., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018)