StepAL: Step-aware Active Learning for Cataract Surgical Videos

Nisarg A. Shah*[1], Bardia Safaie*[1], Shameema Sikder[2,3], S. Swaroop Vedula[3], Vishal M. Patel[1]

Johns Hopkins University[1], Wilmer Eye Institute[2], Malone Center for Engineering in Healthcare[3]

Introduction

Automated surgical step recognition is critical for real-time surgical assistance, skill assessment, and automated reporting. Developing these systems requires vast amounts of expertly annotated video data, a prohibitively expensive process.

Traditional Active Learning (AL) methods are suboptimal, suffering from a granularity mismatch by selecting individual frames/clips, which is ineffective as surgeons require full video context for accurate labeling. This work addresses the need for an AL strategy designed for full video selection.

Contributions

We propose StepAL, a novel active learning framework that effectively reduces high annotation costs. Our primary contributions are twofold:

- First, a Step-aware Feature Representation (SFR) that captures inter-step dependencies by modeling the unique distribution of surgical steps within each video.
- Second, an Entropy-weighted Clustering (EWC) strategy that jointly prioritizes videos with high model uncertainty and diverse step compositions.

Methodology

StepAL strategically selects the most informative full videos for annotation by jointly modeling data diversity and model uncertainty. The pipeline begins with an initial model to generate pseudo-labels for unlabeled videos. These are used to compute our novel representations, which guide the selection process.

A. Step-aware Feature Representation (SFR)

Captures the unique composition of each surgery by creating a rich feature vector based on the distribution of predicted steps (pseudo-labels).

B. Entropy-weighted Clustering (EWC)

Prioritizes uncertain videos by using video-level entropy as a weight during clustering, ensuring selected videos are both diverse and challenging for the model.

$$E(V) = \frac{1}{T}\sum_{t=1}^{T} H(p_t) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C} p_t^{(c)} \log(p_t^{(c)} + \epsilon).$$

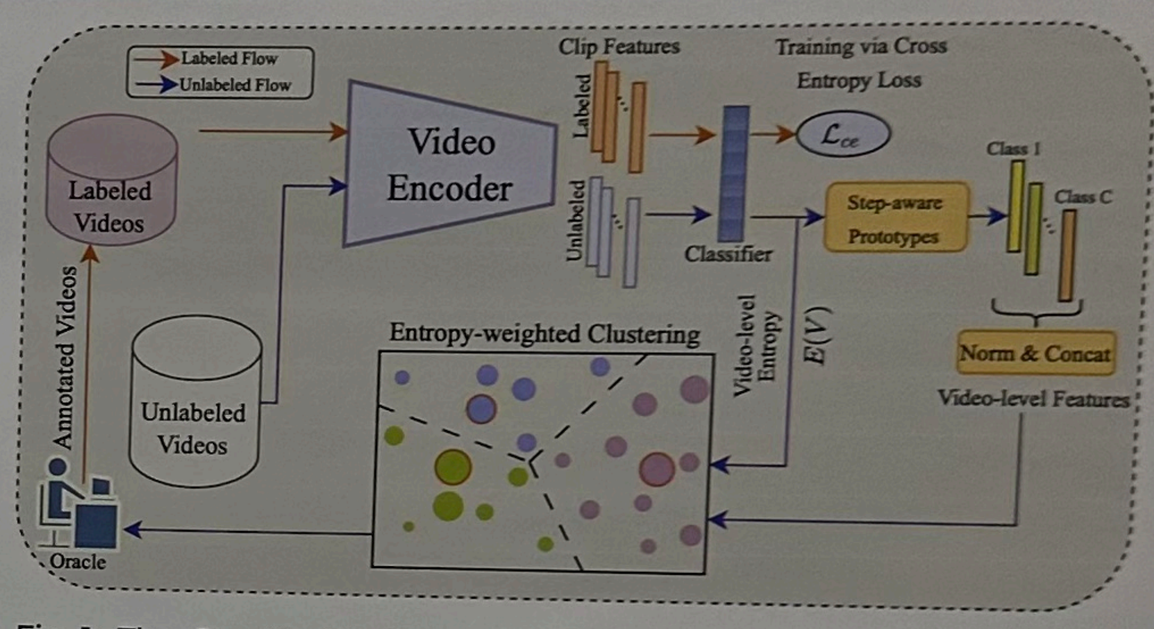


Fig 1: The StepAL Pipeline. Our method computes a step-aware representation and video-level entropy to select the most valuable videos for expert annotation via Entropy-weighted Clustering.

Quantitative Evaluation

Table 1. Performance metrics for two cataract surgery datasets (Cataract-1k and Cataract-101) after the first AL cycle (R=1). StepAL shows substantial improvements, highlighting its efficiency in identifying informative videos early.

Dataset	Metric	Random	Margin[3]	Entropy[30]	Coreset[22]	CoreGCN[4]	Ours
Cataract-1k	Accuracy	0.5795	0.6245	*0.6703*	0.6245	0.6679	**0.7169** (+4.66%)
	Precision	0.5074	0.5299	0.5706	0.5299	*0.5868*	**0.6485** (+6.17%)
	Recall	0.4691	0.5008	*0.5277*	0.5008	0.5242	**0.5785** (+5.08%)
	Jaccard	0.3028	0.3420	0.3801	0.3420	*0.3844*	**0.4308** (+4.64%)
Cataract-101	Accuracy	0.7859	*0.7893*	0.7589	0.7613	0.7700	**0.8016** (+1.23%)
	Precision	0.6937	*0.7495*	0.7002	0.7100	0.7300	**0.7635** (+1.40%)
	Recall	0.6791	*0.7314*	0.6891	0.7040	0.7054	**0.7333** (+0.19%)
	Jaccard	0.5404	*0.5877*	0.5376	0.5411	0.5495	**0.5977** (+1.00%)

Key Finding: Dramatic Gains

With only **20%** of data labeled on the complex Cataract-1k dataset, StepAL achieves:

✔ +4.66% in Accuracy

✔ +4.64% in Jaccard Index (vs. next best method)

Table 2. Comparison of step recognition accuracy in our ablation study, which validates that the full StepAL model outperforms variants lacking either the SFR or EWC components.

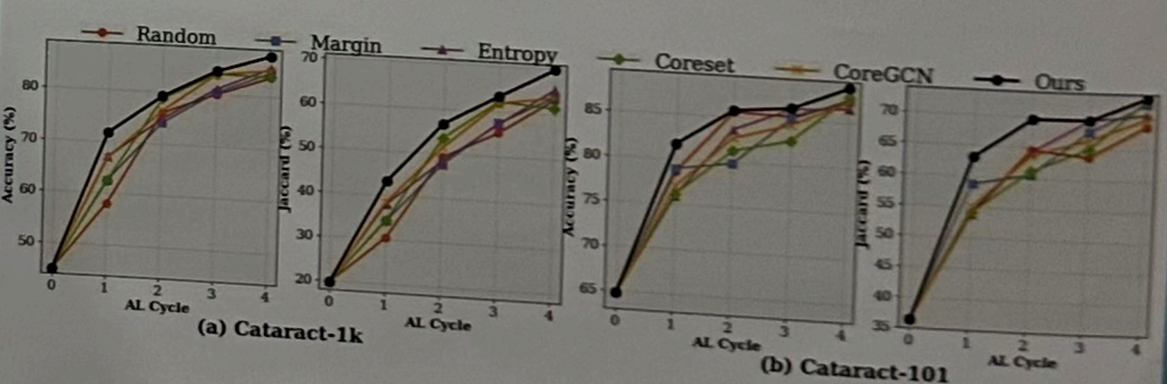
Metric	Random	Entropy	KMeans	ME-KMeans	EWC	Ours
Accuracy	0.5795	0.6703	0.6245	0.6807	0.6408	**0.7169**
Precision	0.5074	0.5706	0.5299	0.6157	0.5366	**0.6485**
Recall	0.4691	0.5277	0.5008	0.5317	0.5123	**0.5785**
Jaccard	0.3028	0.3801	0.3420	0.3941	0.3491	**0.4308**

Ablation Insights

- Representation is Crucial, as methods using naive feature averaging (like KMeans) underperform by obscuring vital step details. Our SFR is key to capturing this diversity
- Uncertainty Matters: while entropy-based selection is effective, our EWC method combines uncertainty with our representation to focus annotation effort where it is needed most, leading to superior performance

Performance Analysis

Fig 2: Comparison of quantitative performance across 5 Active Learning Cycles (R=0 to 4) on the Cataract-1k and Cataract-101 datasets. The plots show StepAL consistently outperforming baselines, achieving a higher performance ceiling with fewer labeled videos.



Conclusion

StepAL provides a practical and effective solution to the surgical data bottleneck. By designing a method for intelligent, full-video selection, we enable more efficient development of clinical AI tools. This work accelerates research in real-time surgical guidance, automated reporting, and objective skill assessment, bringing next-generation surgical AI closer to clinical reality.

Acknowledgements

This research was supported by a grant from the National Institutes of Health, USA: R01EY033065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Also, we would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources, where some of these analyses were conducted, especially DISCOVERY HPC. Their commitment to advancing research has been invaluable in the successful completion of this study.