StepAL: Step-aware Active Learning for Cataract Surgical Videos

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

[1] Johns Hopkins University, Baltimore, MD 21218, USA
[2] Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD
[3] Malone Center for Engineering in Healthcare, Johns Hopkins University
snisarg812@gmail.com

Abstract. Active learning (AL) can reduce annotation costs in surgical video analysis while maintaining model performance. However, traditional AL methods, developed for images or short video clips, are suboptimal for surgical step recognition due to inter-step dependencies within long, untrimmed surgical videos. These methods typically select individual frames or clips for labeling, which is ineffective for surgical videos where annotators require the context of the entire video for annotation. To address this, we propose StepAL, an active learning framework designed for full video selection in surgical step recognition. StepAL integrates a step-aware feature representation, which leverages pseudo-labels to capture the distribution of predicted steps within each video, with an entropy-weighted clustering strategy. This combination prioritizes videos that are both uncertain and exhibit diverse step compositions for annotation. Experiments on two cataract surgery datasets (Cataract-1k and Cataract-101) demonstrate that StepAL consistently outperforms existing active learning approaches, achieving higher accuracy in step recognition with fewer labeled videos. StepAL offers an effective approach for efficient surgical video analysis, reducing the annotation burden in developing computer-assisted surgical systems.

Keywords: Cataract surgery · Active Learning · Step Recognition.

1 Introduction

Automated surgical step recognition is critical for real-time surgical assistance [15], objective skill assessment [32], automated report generation [33], and improved training curricula [7]. Annotated videos are necessary to develop algorithms for automated surgical step recognition, but reliable annotations are expensive because they require significant effort by trained experts [24,25,23,26].

Techniques such as active learning (AL) [20,16,17] can address the challenge of limited annotations by iteratively selecting the most informative and

* Equal contribution

diverse surgical videos for annotation, minimizing labeling costs while maximizing model performance. Existing AL methods for recognition predominantly focus on image-level [22,13] or single-label short video clip classification [27,28]. Common strategies include uncertainty sampling [30,31,19], diversity sampling [22,2,18], heuristic approaches [6], and ensemble models [10].

While effective in their respective domains, existing AL techniques are not directly transferable to the complexities of long, multi-step surgical videos. A fundamental challenge is the granularity mismatch: frame- or clip-level selection conflicts with the practical need for complete, multi-step video annotation in surgical procedures. Due to the inherent sequential dependencies between surgical steps [14,32,5], individual clips often lack sufficient context for accurate labeling. As a result, partial video annotations are ineffective, as the entire video must be reviewed to ensure contextual accuracy. Standard AL methods typically operate directly on unlabeled training inputs, individual clips in the case of video recognition, which can lead to suboptimal performance for surgical videos that require step-level information embedded across sequential clips.

Standard AL strategies also often overlook the structural and temporal information implicitly available in pseudo-labels. While pseudo-labels may be imperfect, particularly in early AL cycles, they provide a valuable approximation of the step distribution within a video. This approximation offers a more informative selection signal than treating all clips equally, which is the implicit assumption in methods that rely solely on clip-level averaging. Furthermore, AL must account for both uncertainty and diversity, ensuring that the uncertainty-based selection process does not lead to redundant sample selection.

StepAL addresses these challenges with two key components. The Step-aware Feature Representation (SFR) captures inter-step dependencies by encoding the distribution of surgical steps within each video, leveraging pseudo-labels predicted by the step recognition model. This step-specific representation allows the selection process to effectively distinguish between different surgical videos based on their step composition. Complementing this, the Entropy-weighted Clustering (EWC) prioritizes videos exhibiting high overall uncertainty. Critically, EWC leverages the step-aware representation of SFR, ensuring that the selected videos are not only uncertain but also represent a diverse range of surgical step sequences. This combined approach ensures that StepAL focuses annotation efforts on videos that are both highly uncertain and representative of the diverse range of step sequences present in the dataset.

To the best of our knowledge, StepAL is the first AL framework specifically designed for video selection in the context of long, multi-step surgical video step recognition. It directly addresses the practical constraints and inherent sequential structure of surgical video data, surpassing the limitations of traditional frame- or clip-level AL approaches by jointly optimizing for uncertainty and diversity at the video level.

Our key contributions are:

- We propose a novel active learning framework, StepAL, tailored for surgical video step recognition, effectively reducing the high annotation costs associated with labeling hour-long, untrimmed videos while achieving performance comparable to training using full annotations.

- Specifically, we introduce step-aware feature representations to effectively capture inter-step dependencies in surgical videos and an entropy-weighted clustering strategy to jointly prioritize videos with high model uncertainty and diverse surgical step distributions.
- Through extensive experiments on two publicly available cataract surgery datasets, we demonstrate the effectiveness of StepAL in enabling efficient and accurate surgical video step recognition.

2 The StepAL Method

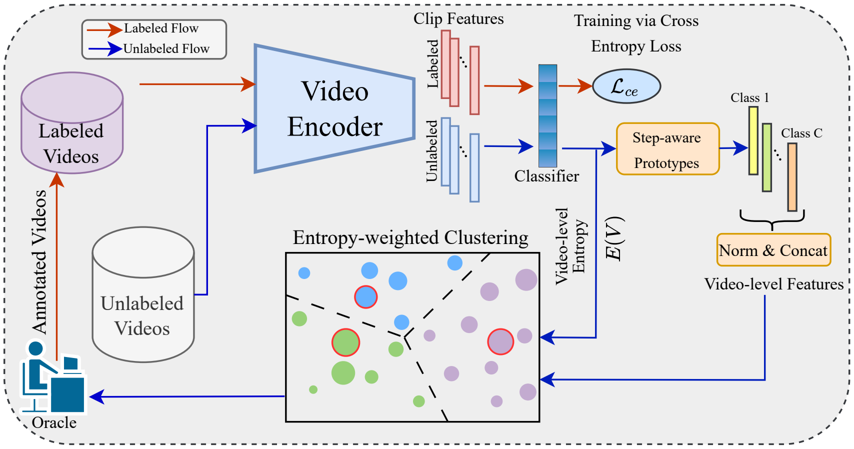


Fig. 1. Overview of our proposed *StepAL* framework. Given long, untrimmed surgical videos as unlabeled data, StepAL employs a hybrid AL approach that selects informative samples based on both uncertainty and representativeness. Step-aware representations are obtained by concatenating prototypes from clip-level features of different pseudo-labels. Video-level uncertainty is measured by averaging clip-level entropies. Finally, entropy-weighted clustering selects videos closest to cluster centers, striking a balance between the diversity and uncertainty of the selected videos for annotation.

We introduce StepAL, an AL framework designed for efficient step recognition in long, untrimmed surgical videos. Our framework addresses the core challenge of minimizing annotation costs while maximizing model performance in multi-step procedures. Let $\mathcal{D} = \{V_n\}_{n=1}^{N}$ denote a surgical video dataset comprising N videos. We partition $\mathcal{D}$ into a labeled set, $\mathcal{D}_L$, and an unlabeled set, $\mathcal{D}_U$. The AL process iteratively selects videos from $\mathcal{D}_U$ for annotation by an expert (e.g., a surgeon) and adds them to $\mathcal{D}_L$ to retrain the step recognition model.

Algorithm 1 StepAL: Active Learning for Multi-step Surgical Videos

```
Require: Dataset $\mathcal{D}$, initial labeled set $\mathcal{D}_L$, unlabeled set $\mathcal{D}_U$, total AL cycles $R$,
    budget per cycle $b$, number of classes $C$.
for cycle $r = 1$ to $R$ do
    Train classifier $F(\cdot;\theta)$ on $\mathcal{D}_L$.
    for each video $V \in \mathcal{D}_U$ do
        Infer clip-level logits $\ell_t \in \mathbb{R}^C$ and features $\phi_t \in \mathbb{R}^D$.
        Compute pseudo-labels $\hat{y}_t = \arg\max_c (\ell_t)_c$ for all clips $t$.
        Construct step-aware representation $z_V$ (Eq. 3).
        Compute video entropy $E(V)$ (Eq. 5).
    end for
    Perform Weighted KMeans on $\{z_V\}$ with weights $\{E(V)\}$ (Eq. 6).
    Select top-$b$ videos $\mathcal{Q} \subseteq \mathcal{D}_U$ nearest to each cluster center.
    Annotate all clips of videos in $\mathcal{Q}$; update $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{Q}$ and $\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{Q}$.
end for
return Final labeled set $\mathcal{D}_L$ and trained model $F(\cdot;\theta)$.
```

Overall Pipeline. Algorithm 1 provides an overview of the StepAL procedure. The process begins by training a step recognition model, $F(\cdot;\theta)$, on the available labeled data, $\mathcal{D}_L$. For each unlabeled video, $V \in \mathcal{D}_U$, we compute clip-level logits, $\ell_t \in \mathbb{R}^C$, and extract corresponding feature embeddings, $\phi_t \in \mathbb{R}^D$, where C represents the number of distinct surgical steps and D denotes the dimensionality of the feature space. Pseudo-labels, $\hat{y}_t = \arg\max_c(\ell_t)_c$, are generated for each clip based on the model's predictions. These pseudo-labels are then used to construct a step-aware feature representation, $z_V \in \mathbb{R}^{C\times D}$ (detailed in Sec. 2.1), which captures the distribution of predicted steps within the video. Concurrently, a video-level entropy, $E(V)$, is computed by averaging the clip-level probability distributions and calculating the resulting entropy (Sec. 2.2). This entropy serves as a measure of the model's overall uncertainty for the given video. The core of the active learning selection strategy lies in applying weighted KMeans clustering to the set of step-aware feature representations, $\{z_V\}$, using the corresponding video entropies, $\{E(V)\}$, as sample weights. This strategically biases the clustering towards videos exhibiting higher uncertainty. A predefined budget, b, determines the number of videos selected; specifically, those closest to the cluster centers are chosen for full annotation. These newly annotated videos are then incorporated into the labeled set, $\mathcal{D}_L$, and the unlabeled set, $\mathcal{D}_U$, is updated accordingly. The entire process is repeated for a predetermined number of active learning cycles, R.

2.1 Step-aware Feature Representation

Surgical procedures inherently consist of a sequence of distinct steps, each possessing unique visual and temporal characteristics. Traditional approaches that rely on simple feature averaging across all clips within a video discard this crucial step-specific information. To mitigate this limitation, we introduce a step-aware feature representation. This representation organizes clip-level embeddings according to their predicted surgical steps, thereby preserving the subtle, yet significant, differences between the various steps of a surgical procedure. This preservation of step-specific information is critical for enabling a more diverse and informative selection of videos during the active learning process.

For each unlabeled video, $V \in \mathcal{D}_U$, composed of T clips $\{x_1, x_2, \ldots, x_T\}$, we extract clip-level features, $\phi_t \in \mathbb{R}^D$, and generate corresponding pseudo-labels, $\hat{y}_t \in \{1, \ldots, C\}$. We define the set $I_V^{(c)}$ as the indices of clips predicted to belong to step c:

$$I_V^{(c)} \;=\; \bigl\{\, t \mid \hat{y}_t = c \bigr\}. \tag{1}$$

For each surgical step c, a step-specific feature, $f_V^{(c)}$, is computed as follows:

$$f_V^{(c)} \;=\; \begin{cases} \dfrac{1}{|I_V^{(c)}|} \displaystyle\sum_{t \in I_V^{(c)}} \phi_t, & \text{if } I_V^{(c)} \neq \varnothing, \\ f_{\text{average}}(V), & \text{otherwise}, \end{cases} \tag{2}$$

where $f_{\text{average}}(V)$ represents the global average of ϕ_t across all clips in video V. This ensures that all surgical steps, even those not predicted in a particular video, are represented in the final feature vector. Each $f_V^{(c)}$ is then ℓ_2-normalized, and these normalized vectors are concatenated to form the final step-aware representation, z_V:

$$z_V \;=\; \Bigl[\tilde{f}_V^{(1)} \parallel \tilde{f}_V^{(2)} \parallel \ldots \parallel \tilde{f}_V^{(C)}\Bigr], \quad \tilde{f}_V^{(c)} \;=\; \frac{f_V^{(c)}}{\bigl\| f_V^{(c)} \bigr\|_2 + \epsilon}. \tag{3}$$

The resulting step-aware representation, $z_V \in \mathbb{R}^{C \times D}$, effectively encodes the distribution of predicted steps within each video. By maintaining distinct feature representations for each predicted step, z_V captures the inherent compositional diversity of multi-step surgical procedures, a crucial factor for effective active learning.

2.2 Entropy-weighted Clustering

In conjunction with the step-aware feature representation, we employ a strategy to prioritize videos with high model uncertainty. For each unlabeled video, V, we quantify this uncertainty using a video-level entropy measure. Given the logit vector, $\ell_t \in \mathbb{R}^C$, for clip t, and its corresponding softmax probabilities, $p_t = \text{softmax}(\ell_t)$, we first compute the clip-level entropy:

$$H(p_t) = -\sum_{c=1}^{C} p_t^{(c)} \log(p_t^{(c)} + \epsilon). \tag{4}$$

Then, the video-level entropy, $E(V)$, is calculated by averaging the clip-level entropies across all T clips in the video:

$$E(V) = \frac{1}{T} \sum_{t=1}^{T} H(p_t) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{c=1}^{C} p_t^{(c)} \log(p_t^{(c)} + \epsilon). \tag{5}$$

High values of $E(V)$ indicate that the model is uncertain about the step assignments within the video, often due to ambiguous surgical steps or complex transitions. Annotating such videos is expected to yield significant improvements in model performance.

To achieve a balance between uncertainty and diversity, we utilize weighted KMeans clustering. This technique operates on the step-aware representations, $\{z_V\}$, while incorporating the video entropies, $\{E(V)\}$, as sample weights. Let $\{c_k\}_{k=1}^{b} \subset \mathbb{R}^{C \times D}$ denote the cluster centers, and $\alpha(V)$ represent the cluster assignment for video V. The weighted KMeans objective function is:

$$\min_{\{c_k\}} \sum_{V \in \mathcal{D}_U} E(V) \left\| z_V - c_{\alpha(V)} \right\|^2. \tag{6}$$

This formulation biases the clustering process towards videos with higher entropy values, effectively prioritizing the selection of uncertain samples. Following the clustering process, we select up to b videos – those closest to each cluster center in the step-aware feature space – for full annotation. This selected set of videos represents a balance: high uncertainty (due to the entropy weighting) and diverse step compositions (due to the clustering on the step-aware representation). The annotated videos are then added to the labeled set, $\mathcal{D}_L$, driving the iterative learning process.

The combination of step-aware feature representations and entropy-weighted clustering enables StepAL to efficiently identify videos that are both challenging for the current model and representative of the wide variety of surgical procedures. This targeted approach to active learning minimizes the annotation effort while maximizing the information gain, ultimately leading to improved accuracy in surgical step recognition.

3 Experiments and Results

Datasets: We evaluate StepAL on two publicly available cataract surgery video datasets: Cataract-1k [9] and Cataract-101 [21]. The labeled subset of Cataract-1k provided by the authors includes 56 videos with a resolution of 1024x768 at 30 fps, annotated with 13 surgical steps. For this dataset, we use 25 videos for training, 7 for validation, and 24 for testing. Cataract-101 comprises 101 videos with a resolution of 720x540 at 25 fps, annotated with 10 surgical steps. We follow the standard split of 50 training, 10 validation, and 40 testing videos. Following prior work [8,29], all videos are subsampled to 1 fps and resized to 250x250. Model performance is evaluated using frame-wise accuracy, precision, recall, and Jaccard Index.

Implementation Details: We use a Video Vision Transformer (VideoViT) base model (‘VideoViT-B/16‘ architecture [1], pre-trained on Kinetics-400 [11]), with 16 frames sampled per video (16 x 224 x 224 input clips after resizing). The model produces 768-dimensional feature embeddings (ϕ_t) for each clip.

The active learning process starts with an initial labeled set, $\mathcal{D}_L$, containing 10% of the training videos. The step recognition model, $F(\cdot;\theta)$, is fine-tuned on $\mathcal{D}_L$, then iteratively updated following Algorithm 1. Each cycle ($R = 4$ total) selects a new batch from the unlabeled set, $\mathcal{D}_U$, using the Step-aware Feature Representation (Sec. 2.1) and Entropy-weighted Clustering (Sec. 2.2), adding 10% of the total training data to $\mathcal{D}_L$. This evaluates model performance up to 50% labeled data utilization. A model trained on the *complete* training set achieves mean accuracies of 92.01% (Cataract-1k) and 89.47% (Cataract-101), serving as an oracle performance reference.

Training uses a batch size of 14 on a single NVIDIA A100 GPU, employing the Adam optimizer [12] (learning rate = 1e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 5e-4) and Cross-Entropy Loss.

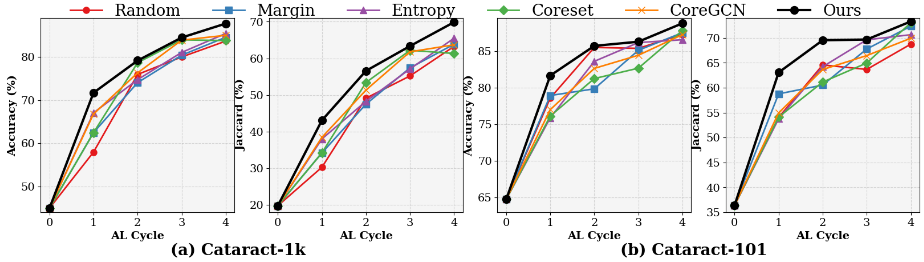
Dataset	**Metric**	**Random**	**Margin**[3]	**Entropy**[30]	**Coreset**[22]	**CoreGCN**[4]	**Ours**
Cataract-1k	Accuracy	0.5795	0.6245	*0.6703*	0.6245	0.6679	**0.7169** **(+4.66%)**
	Precision	0.5074	0.5299	0.5706	0.5299	*0.5868*	**0.6485** **(+6.17%)**
	Recall	0.4691	0.5008	*0.5277*	0.5008	0.5242	**0.5785** **(+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%)**

Table 1. Performance Metrics for Two Cataract Surgery Datasets, Cataract-1k [9] and the Cataract-101 dataset [21] for R = 1. Values in the green indicate the absolute percentage increase (from the next best result).

Results: Table 1 compares StepAL to state-of-the-art AL methods: Random, Margin [3], Entropy [30], Coreset [22], and CoreGCN [4]. Coreset focuses on diversity in feature space; CoreGCN uses a graph convolutional network. StepAL, however, uniquely integrates both uncertainty and diversity via its step-aware feature representation (Sec. 2.1) and entropy-weighted clustering (Sec. 2.2).

At AL cycle $R = 1$, StepAL outperforms all competing methods on both datasets. On Cataract-1k, StepAL achieves substantial improvements over the next best performing method, including a 4.66% increase in accuracy and a 4.64% increase in Jaccard index. Gains are also observed on Cataract-101, where StepAL surpasses the next best method by 1.23% in accuracy and 1.00% in Jaccard index, reflecting the dataset's inherent simplicity.

Figure 2 shows StepAL's consistent advantage across all active learning cycles. Its ability to achieve higher accuracy and Jaccard index from the initial stages highlights its effectiveness in rapidly identifying the most informative videos, crucial when annotation resources are limited. On Cataract-101, performance converges at later cycles ($R = 4$) as methods approach the oracle accuracy of 89.47%; yet, StepAL maintains a performance edge throughout.



(a) Cataract-1k

(b) Cataract-101

Fig. 2. Comparison of quantitative performance across 5 Active Learning Cycles (R = 0 to 4). (a) Results on Cataract-1k dataset and (b) Results on Cataract-101 dataset.

Ablations: Our ablation study on Cataract-1k, summarized in Table 2, illustrates the effectiveness of StepAL's components. The *Random* baseline serves as a benchmark, with the *Entropy* method improving accuracy by 15.7% over Random by selecting videos based on average clip-level entropy, demonstrating the value of incorporating uncertainty into active learning. In contrast, KMeans underperforms Entropy by 7.3% due to its reliance on averaged clip features, which obscure essential details. However, *ME-KMeans* (Maximum Entropy KMeans), which also uses averaged features but selects the most uncertain video in each cluster, surpasses both Entropy and KMeans, showing the importance of combining diversity with uncertainty for effective selection..

EWC shows only marginal improvement over KMeans and still lags behind ME-KMeans, emphasizing that the feature representation is a critical component. In stark contrast, **Ours** (StepAL) integrates step-aware feature representation (Sec. 2.1) and EWC (Sec. 2.2), outperforming all other methods by significant margins. StepAL not only improves accuracy by 5.3% over ME-KMeans but also enhances precision, demonstrating its robustness through consistent performance improvements across all metrics, effectively capturing step-level diversity and prioritizing overall video uncertainty for more effective video selection.

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**

Table 2. Performance Metrics for Ablation using the Cataract-1k[9] dataset.

4 Conclusion

In this paper, we present StepAL, a novel AL framework designed for the selection of videos to improve surgical step recognition. Unlike traditional AL methods that focus on individual frames or clips, StepAL selects entire videos for annotation, aligning better with real-world surgical workflows. Our approach combines step-aware feature representation, which captures fine-grained step-level information using pseudo-labels, with entropy-weighted clustering. This method prioritizes videos that are both highly uncertain and diverse in their step composition for further labeling. Experiments on two cataract surgery datasets demonstrate that StepAL consistently outperforms existing active learning approaches, achieving higher accuracy in step recognition with fewer labeled videos.

Disclosure of Interests

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

Acknowledgements.

This work 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.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
2. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671 (2019)
3. Balcan, M.F., Broder, A., Zhang, T.: Margin based active learning. In: International Conference on Computational Learning Theory. pp. 35–50. Springer (2007)
4. Caramalau, R., Bhattarai, B., Kim, T.K.: Sequential graph convolutional network for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9583–9592 (2021)
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 2020. pp. 343–352. Springer (2020)
6. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine learning **28**, 133–168 (1997)
7. Funke, I., Mees, S.T., Weitz, J., Speidel, S.: Video-based surgical skill assessment using 3d convolutional neural networks. IJCARS (2019)
8. 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 2021. pp. 593–603. Springer (2021)

9. Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S., Putzgruber-Adamitsch, D., Zinkernagel, M., Wolf, S., Schoeffmann, K., Sznitman, R.: Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. Scientific data **11**(1), 373 (2024)
10. Hino, H., Eguchi, S.: Active learning by query by committee with robust divergences. Information Geometry **6**(1), 81–106 (2023)
11. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Ma, S., Du, H., Curran, K.M., Lawlor, A., Dong, R.: Adaptive curriculum query strategy for active learning in medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 48–57. Springer (2024)
14. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**(9), 691–696 (2017)
15. Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies **28**(2), 82–90 (2019)
16. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
17. Safaei, B., Patel, V.M.: Active learning for vision language models. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 4902–4912 (February 2025)
18. Safaei, B., Siddiqui, F., Xu, J., Patel, V.M., Lo, S.Y.: Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14247–14256 (2025)
19. Safaei, B., Vibashan, V., de Melo, C.M., Patel, V.M.: Entropic open-set active learning. Proceedings of the AAAI Conference on Artificial Intelligence **38**(5), 4686–4694 (2024)
20. Safaei, B., VS, V., Patel, V.M.: Certainty and uncertainty guided active domain adaptation. arXiv preprint arXiv:2505.19421 (2025)
21. Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M.J., Putzgruber, D.: Cataract-101: video dataset of 101 cataract surgeries. In: Proceedings of the 9th ACM Multimedia Systems Conference. p. 421–425. MMSys '18, Association for Computing Machinery, New York, NY, USA (2018), https://doi.org/10.1145/3204949.3208137
22. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
23. Shah, N.A., Bandara, C., Skider, S., Vedula, S.S., Patel, V.M.: CSMAE: Cataract surgical masked autoencoder (MAE) based pre-training. In: Proceedings of the International Symposium on Biomedical Imaging (ISBI) (2025)
24. Shah, N.A., Sikder, S., Vedula, S.S., Patel, V.M.: Glsformer: Gated-long, short sequence transformer for step recognition in surgical videos. In: MICCAI (2023)
25. Shah, N.A., Sikder, S., Vedula, S.S., Patel, V.M.: Step detection in cataract surgery videos. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)

26. Shah, N.A., Xia, M., Vijay, S., Sikder, S., Vedula, S.S., Patel, V.M.: A vision foundation model for cataract surgery using joint-embedding predictive architecture. In: Medical Imaging with Deep Learning (2025)
27. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5972–5981 (2019)
28. Taketsugu, H., Ukita, N.: Active transfer learning for efficient video-specific human pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1880–1890 (2024)
29. 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 (2016)
30. Wang, D., Shang, Y.: A new active labeling method for deep learning. In: 2014 International joint conference on neural networks (IJCNN). pp. 112–119. IEEE (2014)
31. Wu, J., Chen, J., Huang, D.: Entropy-based active learning for object detection with progressive diversity constraint. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9397–9406 (2022)
32. Yu, F., Croso, G.S., Kim, T.S., Song, Z., Parker, F., Hager, G.D., Reiter, A., Vedula, S.S., Ali, H., Sikder, S.: Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. JAMA network open **2**(4), e191860–e191860 (2019)
33. Zisimopoulos, O., Flouty, E., Stoyanov, D., et al.: Deepphase: surgical phase recognition in cataracts videos. In: MICCAI 2018. pp. 265–272. Springer (2018)