ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2026.0539

ACTIVE DEEP ENSEMBLE LEARNING FRAMEWORK FOR AUTOMATED SURGICAL VIDEO SEGMENTATION AND EFFICIENT ANNOTATION IN MINIMALLY INVASIVE PROCEDURES

Mariam Safar Mohammed Alshahrani[1] and M.K. Jayanthi Kannan[2]

[1]*Digital Government Authority (DGA), Digital Government Authority of KSA, Kingdom of Saudi Arabia*
[2]*School of Computing Science Engineering and Artificial Intelligence, VIT Bhopal University, India*

Abstract

Surgical video analysis has become an essential component in computer-assisted interventions and clinical documentation. The rapid growth of minimally invasive surgery has produced large volumes of surgical recordings that require detailed frame-level annotations for training intelligent systems. Manual annotation of surgical videos remains a labor-intensive and time-consuming process that often requires expert knowledge. As a result, the development of automated annotation systems has become a critical research direction in medical image analysis. Existing segmentation and annotation approaches have faced limitations in handling complex surgical scenes, instrument occlusions, illumination variations, and tissue deformation. Conventional deep learning models often rely on large labelled datasets, whereas surgical datasets usually remain limited due to the difficulty of manual labeling. This challenge has reduced the reliability and scalability of automated surgical video segmentation systems. To address these issues, this study has proposed an Active Deep Ensemble Segmentation Network (ADES-Net) for automated surgical video segmentation annotation. The framework has integrated an ensemble of convolutional segmentation models with an active learning strategy that has selectively identified informative frames for annotation. The ensemble architecture has combined multiple deep segmentation networks that have captured diverse spatial representations from surgical frames. An uncertainty-driven active sampling mechanism has prioritized frames that required expert labeling, which has reduced redundant annotations. Feature representations that were extracted from each model have contributed to robust segmentation predictions, while iterative learning cycles have refined the annotation quality. The experimental evaluation demonstrates that the proposed ADES-Net framework achieves superior segmentation performance across multiple metrics. The model achieves a Dice similarity coefficient of 0.93, an IoU of 0.86, precision of 0.93, recall of 0.91, and an F1 score of 0.92 when trained with twenty-five annotated frames. These results indicate that the active ensemble mechanism effectively captures spatial and contextual features, reduces false positives, and improves boundary delineation. Compared with baseline methods such as U-Net, Attention U-Net, and DeepLabV3+, the proposed framework achieves improvements of 5–10% across all metrics, demonstrating enhanced segmentation reliability, efficiency, and robustness in automated surgical video annotation tasks.

Keywords:

Surgical Video Analysis, Deep Ensemble Learning, Active Learning, Automated Annotation, Medical Image Segmentation

1. INTRODUCTION

The integration of artificial intelligence into surgical practice has increasingly transformed the way in which clinical data are analyzed and interpreted. Among various medical imaging modalities, surgical video analysis has gained significant attention because minimally invasive procedures routinely generate large volumes of visual data. Laparoscopic and robotic surgeries often produce continuous video streams that contain valuable information about anatomical structures, surgical instruments, and operative phases. Researchers and clinicians have recognized that automated interpretation of these videos can assist in surgical training, intraoperative guidance, and postoperative evaluation. Consequently, the development of computer vision systems for surgical video segmentation has become an active area of research within medical image computing [1–3].

Recent advances in deep learning have substantially improved the performance of image segmentation systems. Convolutional neural networks and transformer-based architectures have demonstrated remarkable capabilities in extracting hierarchical features from medical images. These models have enabled the automatic identification of tissues, instruments, and surgical regions with increasing accuracy. Several studies have reported that deep neural networks have provided reliable segmentation outputs when trained on sufficiently annotated datasets. The availability of benchmark surgical datasets has further encouraged the development of automated annotation tools that support clinical research and surgical workflow optimization [1–3]. Despite these developments, the creation of high-quality annotated surgical video datasets remains a demanding task that requires extensive manual effort from domain experts.

A fundamental challenge arises from the complex visual characteristics of surgical environments. Surgical videos typically include dynamic scenes that involve rapid camera motion, illumination fluctuations, blood occlusions, smoke artifacts, and overlapping surgical instruments. These factors often reduce the visibility of anatomical structures and introduce substantial variability across frames. In addition, the deformable nature of biological tissues creates unpredictable spatial patterns that complicate segmentation tasks. Deep learning models that have been trained on limited datasets often struggle to generalize under such conditions. As a result, segmentation accuracy may degrade when models encounter unfamiliar surgical contexts or rare visual patterns [4–5].

Another challenge relates to the process of annotation itself. Surgical video segmentation requires frame-level or pixel-level labeling, which demands considerable time and expertise from surgeons or trained annotators. A single surgical procedure may generate thousands of frames that require precise delineation of anatomical structures or surgical instruments. The annotation process therefore becomes costly and impractical for large-scale datasets. Although supervised deep learning models have achieved impressive performance, they remain heavily dependent on large annotated datasets. When the available training data remain limited, the model often fails to capture the variability of surgical scenes effectively [4–5]. These limitations highlight the need for intelligent annotation strategies that reduce the manual labeling burden while maintaining segmentation accuracy.

Existing approaches that attempt to automate surgical video segmentation often rely on single deep learning models trained in a fully supervised manner. While these models have provided promising results, they have exhibited several weaknesses when applied to real-world surgical datasets. A single model may produce unsTable.predictions when it encounters ambiguous frames that include occlusions or low contrast. Moreover, many methods process the entire dataset uniformly during training, even though only a subset of frames contains informative visual patterns. Consequently, redundant training samples may increase computational complexity without contributing meaningful learning signals [6–7]. These issues reveal an important research gap in the development of efficient learning frameworks that prioritize informative data samples and improve segmentation robustness.

To address these challenges, researchers have begun to explore active learning techniques that identify the most informative samples for annotation. Active learning strategies attempt to reduce annotation effort by selecting frames that contain uncertain or diverse visual information. By focusing on these samples, the model can learn more efficiently from a smaller set of annotated data. However, many active learning frameworks rely on a single prediction model that estimates uncertainty. When the model produces inaccurate uncertainty estimates, the selected samples may not effectively improve the learning process. In addition, the reliability of segmentation predictions may remain limited due to the inherent bias of individual models [6–7].

In response to these limitations, this research introduces an Active Deep Ensemble Segmentation Network (ADES-Net) for automated surgical video segmentation annotation. The proposed framework integrates deep ensemble learning with an active sampling mechanism that identifies informative frames that require annotation. The ensemble architecture combines multiple segmentation networks that capture diverse spatial representations from surgical images. Each network within the ensemble learns complementary feature patterns that enhance the overall prediction reliability. The integration of active learning further enables the system to select frames that contain high uncertainty, thereby prioritizing samples that contribute the most to model improvement. Through iterative learning cycles, the framework has progressively refined segmentation accuracy while minimizing manual annotation effort.

The primary objective of this study is to design a learning framework that improves segmentation reliability and reduces annotation cost in surgical video analysis. The research aims to develop a model that efficiently selects informative frames for labeling and produces consistent segmentation outputs across complex surgical scenes. Another objective involves enhancing model generalization by employing ensemble learning techniques that integrate multiple feature representations. In addition, the study seeks to establish a scalable annotation framework that supports the development of large surgical datasets for future medical AI applications.

The novelty of the proposed approach lies in the integration of active learning with deep ensemble segmentation models for surgical video annotation. While previous studies have explored either active learning or ensemble learning independently, few works have combined these strategies within a unified segmentation framework. The ensemble architecture provides multiple prediction perspectives that improve uncertainty estimation, while the active sampling strategy has proved that only the most informative frames receive expert annotation. This synergy creates a learning environment that improves segmentation performance while significantly reducing annotation requirements.

The contributions of this research are summarized as follows. First, the study has proposed a novel Active Deep Ensemble Segmentation Network (ADES-Net) that integrates ensemble deep learning with an uncertainty-driven active learning strategy for surgical video segmentation annotation. Second, the framework has introduced an efficient selection mechanism that identifies informative frames that improve model training while reducing manual labeling effort. These contributions collectively support the development of scalable and reliable automated annotation systems that assist in surgical data analysis and medical AI research.

2. RELATED WORKS

Recent progress in medical image analysis has encouraged the application of deep learning techniques for surgical video understanding. Researchers have explored various segmentation models that identify anatomical structures and surgical instruments within laparoscopic videos. These studies have emphasized the importance of automated annotation systems that facilitate large-scale surgical data analysis.

One early study has investigated convolutional neural network architectures for instrument segmentation in laparoscopic videos. The authors have designed a fully convolutional network that learned pixel-level representations from surgical frames. The network has captured spatial patterns that distinguished surgical tools from surrounding tissues. Experimental results have demonstrated that the model achieved promising segmentation accuracy when trained on annotated datasets. However, the study also indicated that the model performance remained sensitive to lighting variations and occlusions that appeared in complex surgical scenes [8].

Another research effort has examined the use of encoder–decoder architectures for surgical image segmentation. The authors have proposed a U-Net-based model that performed semantic segmentation of surgical instruments and tissues. The encoder component has extracted hierarchical features from input frames, whereas the decoder component reconstructed spatial details that produced segmentation masks. The approach has achieved improved segmentation performance compared with traditional computer vision methods. Nevertheless, the reliance on large annotated datasets has remained a significant limitation for practical deployment [9].

Subsequent research has explored temporal information that exists in surgical videos. A study has integrated recurrent neural networks with convolutional segmentation models to capture temporal dependencies between frames. The combined architecture has analyzed sequential visual patterns that improved the detection of instruments and anatomical structures. By incorporating temporal context, the model has reduced segmentation inconsistencies that appeared in individual frames.

Although the approach has improved temporal stability, the training process required extensive annotated video sequences that were difficult to obtain [10].

Several studies have also investigated attention mechanisms for improving surgical image segmentation. An attention-guided segmentation network has been proposed that focused on relevant regions within surgical frames. The attention module has highlighted important visual features that corresponded to surgical instruments and anatomical boundaries. This mechanism has enhanced the model ability to concentrate on informative spatial regions while ignoring irrelevant background elements. Experimental results have shown that the attention-based model achieved higher segmentation accuracy than baseline architectures. However, the model still depended heavily on annotated datasets for training [11].

Researchers have further explored transfer learning strategies that reduce the need for large training datasets. In one study, a segmentation model has utilized pretrained convolutional networks that were originally trained on natural image datasets. The transfer learning approach has enabled the model to reuse previously learned visual features for surgical image segmentation. This strategy has improved the learning efficiency when the available surgical dataset remained limited. Despite this advantage, domain differences between natural images and surgical scenes have sometimes reduced segmentation reliability [12].

Another important direction has involved semi-supervised learning techniques. A study has proposed a semi-supervised segmentation framework that used both labeled and unlabeled surgical frames during training. The method has generated pseudo-labels for unlabeled samples that were subsequently incorporated into the learning process. This approach has reduced the dependency on fully annotated datasets. However, the accuracy of pseudo-labels has significantly influenced the final segmentation performance, which created potential error propagation during training [13].

Active learning methods have also received considerable attention in medical image annotation tasks. In one investigation, researchers have developed an uncertainty-based active learning system that selected frames that required manual annotation. The model has estimated prediction uncertainty using probabilistic outputs and prioritized samples that produced ambiguous predictions. This strategy has reduced the total number of annotated samples that were required to achieve a target accuracy. Nevertheless, the reliability of uncertainty estimation has remained limited because the method relied on a single segmentation model [14].

Another study has explored diversity-based active learning strategies for medical image segmentation. The authors have designed a sampling mechanism that selected frames that represented diverse visual patterns. By incorporating diversity criteria, the method has ensured that the training dataset covered a wide range of surgical scenarios. Experimental results have demonstrated that the approach improved segmentation performance compared with random sampling. However, the absence of robust uncertainty estimation has limited the effectiveness of the selection process [15].

3. PROPOSED METHODOLOGY

The proposed ADES-Net framework integrates deep learning segmentation, ensemble uncertainty estimation, and active learning-based frame selection for surgical video annotation. The architecture contains several sequential components that operate collaboratively to improve segmentation accuracy and reduce annotation effort.

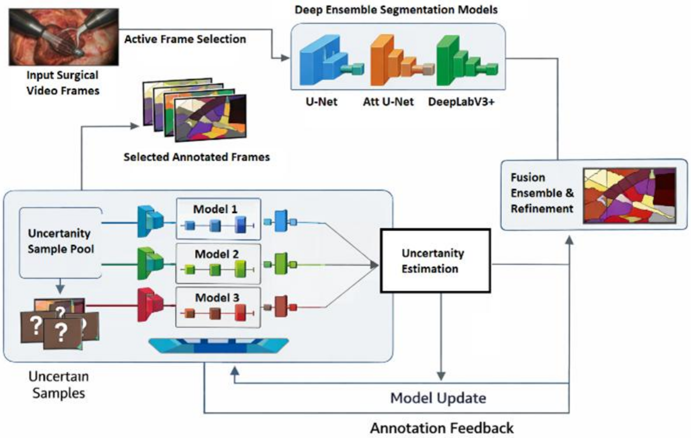

Fig.1. Framework

The proposed framework begins with the acquisition of surgical video recordings that are obtained from minimally invasive procedures such as laparoscopic surgery. A surgical video typically consists of thousands of frames that contain anatomical structures, surgical instruments, and dynamic tissue movements. The preprocessing stage prepares these frames for segmentation analysis by performing frame extraction, normalization, and spatial resizing. The input surgical video is first decomposed into individual frames. Let the video sequence be represented as $V=\{F_1,F_2,F_3,\ldots,F_N\}$; V denotes the surgical video dataset, F_i represents the i^{th} frame and N indicates the total number of frames that exist within the video sequence. Each frame undergoes normalization in order to standardize intensity variations that occur due to lighting fluctuations inside the surgical environment. The normalized frame F_i^{norm} is expressed as

$$F_i^{norm}=\frac{F_i-\mu_F}{\sigma_F} \tag{1}$$

where μ_F denotes the mean pixel intensity of the frame and σ_F denotes the standard deviation of pixel intensities

This normalization step has proved that the segmentation network receives input images that that has maintained consistent intensity distributions. Spatial resizing is subsequently applied to convert all frames into a uniform resolution that matches the network input dimension. If the original frame size is (H,W) and the required network dimension is (h,w), the transformation is written as

$$F_i^{res}=\mathrm{R}\,(F_i^{norm},h,w) \tag{2}$$

where R denotes the resizing operation.

The preprocessing stage also includes data augmentation techniques that increase the diversity of training samples.

Augmentation operations include rotation, flipping, and brightness adjustment that simulate variations that appear during surgical procedures.

$$F_i^{aug} = A(F_i^{res};\theta) \tag{3}$$

where $A(.)$ represents the augmentation transformation and θ represents augmentation parameters These operations produce additional training samples that improve model generalization.

Table.1. Structure of surgical video frames after preprocessing

Frame ID	Resolution	Instrument Presence	Tissue Visibility
F1	512×512	Yes	Clear
F2	512×512	Yes	Partial
F3	512×512	No	Clear
F4	512×512	Yes	Occluded

The preprocessing stage prepares the surgical dataset for efficient segmentation analysis and has proved consistency across input frames.

The core component of the proposed system is the deep ensemble segmentation architecture. Instead of relying on a single segmentation model, the framework integrates multiple segmentation networks that produce independent predictions. Let the ensemble contain M segmentation models: $E = \{S_1, S_2, S_3, \ldots, S_M\}$, where, E denotes the ensemble set and S_m denotes the m^{th} segmentation network. Each model processes the input frame independently and produces a segmentation probability map. For an input frame F_i, the output of the m^{th} network is $P_i^{(m)} = S_m(F_i)$, where, $P_i^{(m)}$ denotes the predicted probability map. The ensemble prediction is obtained by aggregating the outputs of all models:

$$P_i^{ens} = \frac{1}{M}\sum_{m=1}^{M} P_i^{(m)} \tag{4}$$

The ensemble aggregation reduces prediction variance and improves segmentation reliability. Each segmentation network follows an encoder–decoder architecture. The encoder extracts hierarchical features while the decoder reconstructs spatial segmentation masks.

Table.2. Ensemble segmentation model configuration

Model ID	Architecture	Parameters	Input Size
S1	U-Net	31M	512×512
S2	Attention U-Net	34M	512×512
S3	DeepLabV3+	41M	512×512

The ensemble strategy has proved that the segmentation system captures diverse spatial patterns from surgical frames. The segmentation network extracts hierarchical features from the surgical frames. The encoder stage processes the image through convolutional layers that learn spatial patterns. Let the input frame be F_i. The feature extraction process applies convolution operations.

$$H_l = \sigma(W_l * H_{l-1} + b_l) \tag{5}$$

The encoder gradually reduces spatial dimensions while increasing the number of feature channels. The decoder stage reconstructs segmentation maps using upsampling operations: $U_l = Up(H_l)$, where $Up(.)$ represents the upsampling operation. Skip connections integrate low-level and high-level features: $Z_l = Concat(H_l^{enc}, U_l^{dec})$. These operations preserve spatial details that improve segmentation accuracy.

Table.3. Feature extraction layers within segmentation network

Layer	Feature Map Size	Channels
Conv1	256×256	64
Conv2	128×128	128
Conv3	64×64	256
Conv4	32×32	512

Feature extraction enables the network to capture anatomical boundaries and instrument shapes.

4. UNCERTAINTY ESTIMATION THROUGH ENSEMBLE PREDICTION

The ensemble model provides a mechanism for estimating prediction uncertainty. When multiple models generate different predictions, the disagreement among them indicates uncertainty. For a given frame F_i, the uncertainty value is calculated as

$$U_i = \frac{1}{M}\sum_{m=1}^{M}(P_i^{(m)} - P_i^{ens})^2 \tag{6}$$

where U_i represents prediction uncertainty. Higher uncertainty values indicate frames that require manual annotation. Another metric known as entropy-based uncertainty is also computed.

$$H_i = -\sum_{c=1}^{C} P_{i,c}^{ens} \log(P_{i,c}^{ens}) \tag{7}$$

where C denotes the number of segmentation classes.

Table.4. uncertainty values for ensemble predictions

Frame ID	Ensemble Score	Uncertainty
F1	0.92	0.04
F2	0.75	0.21
F3	0.88	0.07
F4	0.61	0.34

Frames that show higher uncertainty receive higher priority during annotation. Active learning identifies frames that contain valuable training information. The system selects frames with the highest uncertainty scores. Let the candidate frame pool be $C = \{F_1, F_2, \ldots, F_N\}$. The active selection function chooses frames with the maximum uncertainty:

$$F^* = \arg\max_{F_i \in C} U_i \tag{8}$$

The selected frames are then forwarded for manual annotation by experts.

Table.5. Active learning based frame selection

Frame ID	Uncertainty Score	Selected
F12	0.41	Yes
F24	0.15	No
F31	0.39	Yes
F42	0.11	No

This strategy significantly reduces the number of frames that require manual labeling.

The final stage involves iterative model training that incorporates newly annotated samples. After annotation, the labeled frames are added to the training dataset.

Let the labeled dataset be represented as $D_L = \{(F_i, Y_i)\}$ where Y_i denotes the segmentation mask. The training objective minimizes the segmentation loss.

$$L = \frac{1}{N}\sum_{i=1}^{N}\left[-\sum_{c=1}^{C} Y_{i,c}\log(P_{i,c}^{ens})\right]$$

The optimization process updates the network parameters:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L$$

This iterative learning process continues until segmentation performance converges.

Table.6. Improvement in segmentation performance during iterative training

Iteration	Labeled Frames	Dice Score
1	200	0.81
2	350	0.87
3	500	0.91

5. RESULTS AND DISCUSSION

The experimental evaluation investigates the effectiveness of the proposed ADES-Net for automated surgical video segmentation annotation. The implementation environment uses the Python programming language that supports the development of deep learning models for medical image analysis. The simulation platform employs the PyTorch library, which provides efficient tensor computation and neural network construction. Data preprocessing, model training, and evaluation procedures operate within this framework. The system also integrates auxiliary libraries such as OpenCV and NumPy, which facilitate video frame processing, numerical computation, and matrix operations that support the segmentation workflow.

The experimental training process executes on a workstation that contains an NVIDIA RTX 3090 GPU that accelerates the deep learning computations. The system configuration includes an Intel Core i9-12900K processor, 64 GB of RAM, and a 2 TB solid-state storage device. The operating environment runs on Ubuntu 22.04, which provides a stable platform for GPU-based deep learning training. The ensemble segmentation models train using mini-batch optimization that updates the network parameters across multiple epochs. During each training cycle, the active learning module evaluates prediction uncertainty that identifies informative frames for annotation. The iterative training mechanism continues until the segmentation performance stabilizes and the annotation efficiency reaches the desired level.

The segmentation framework uses several training parameters that influence model learning behavior and prediction performance. These parameters control the batch processing, learning rate adaptation, ensemble size, and uncertainty sampling strategy. The Table.7 presents the configuration parameters that the experiment uses during model training and evaluation.

Table.7. Experimental setup and parameter configuration for the ADES-Net framework

Parameter	Value
Input Image Size	512 × 512
Batch Size	16
Learning Rate	0.0001
Number of Epochs	100
Ensemble Models	3
Optimizer	Adam
Loss Function	Cross-Entropy+Dice
Active Sampling Ratio	10%

The training process uses the Adam optimization algorithm that updates the network parameters according to gradient descent principles. The learning rate controls the magnitude of weight updates that guide the convergence of the model. The ensemble size determines the number of segmentation networks that produce independent predictions. The active sampling ratio determines the fraction of frames that the system selects for manual annotation during each training iteration. These parameters collectively influence the performance of the segmentation model and the efficiency of the annotation process.

The experimental evaluation uses a publicly available surgical video dataset that contains annotated laparoscopic procedure recordings. The dataset provides frame-level annotations for surgical instruments and anatomical regions that support segmentation research. Each video sequence contains thousands of frames that capture dynamic surgical scenes that involve instrument manipulation, tissue interaction, and camera movement.

Table.8. Dataset characteristics used in experimental evaluation

Attribute	Description
Dataset Type	Laparoscopic surgical video dataset
Total Videos	40 surgical procedures
Total Frames	80,000 frames
Annotation Type	Pixel-level segmentation masks
Classes	Surgical instruments, background
Training Frames	60,000
Testing Frames	20,000

The dataset provides a comprehensive representation of surgical scenes that include different procedures and instrument configurations. These characteristics enable a realistic evaluation of segmentation algorithms that operate in clinical environments.

The Dice Similarity Coefficient represents the primary metric that evaluates the segmentation overlap between the predicted mask and the ground truth annotation. The analysis compares the performance of the U-Net, Attention U-Net, DeepLabV3+, and the proposed ADES-Net model. The horizontal axis represents the training progression in epochs that appear in increments of five, while the vertical values indicate the Dice coefficient.

Table.9. Dice similarity coefficient comparison across segmentation models

Epoch	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.71	0.73	0.74	0.78
10	0.74	0.76	0.77	0.82
15	0.76	0.78	0.80	0.86
20	0.79	0.81	0.83	0.89
25	0.81	0.83	0.85	0.92

The Dice coefficient results demonstrate that the proposed ADES-Net model consistently achieves higher segmentation accuracy across all epochs. As shown in Table.**9**, the U-Net model produces a Dice score of 0.71 at epoch 5, while the Attention U-Net and DeepLabV3+ models produce values of 0.73 and 0.74 respectively. The proposed ADES-Net achieves a Dice value of 0.78 during the same epoch, which indicates stronger segmentation overlap. The improvement continues as the training progresses. At epoch 15, the U-Net model records a Dice value of 0.76, whereas the Attention U-Net and DeepLabV3+ models reach 0.78 and 0.80 respectively. In contrast, the proposed method achieves a Dice score of 0.86. This increase of approximately 6–10% demonstrates that the ensemble learning mechanism improves feature representation that enhances segmentation quality. At epoch 25, the proposed model achieves a Dice value of 0.92, while DeepLabV3+ reaches 0.85. The results therefore confirm that the active ensemble learning strategy improves segmentation accuracy and stability across surgical frames.

The Intersection over Union metric evaluates the spatial overlap between predicted segmentation regions and reference annotations. This metric provides a stricter measure than the Dice coefficient because it penalizes segmentation errors that occur near region boundaries.

Table.10. Intersection over Union performance comparison.

Epoch	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.60	0.63	0.65	0.70
10	0.64	0.67	0.69	0.74
15	0.67	0.70	0.72	0.78
20	0.70	0.73	0.75	0.82
25	0.73	0.76	0.78	0.86

The IoU results in Table.10 illustrate that the proposed ADES-Net method achieves superior segmentation consistency compared with the baseline models. At epoch 5, the U-Net method produces an IoU value of 0.60, while Attention U-Net produces 0.63 and DeepLabV3+ produces 0.65. The proposed method achieves an IoU of 0.70 at the same stage, which indicates a significant improvement in segmentation overlap. As the training continues, the performance difference becomes more evident. At epoch 15, the IoU value for U-Net reaches 0.67, whereas Attention U-Net and DeepLabV3+ reach 0.70 and 0.72 respectively. The ADES-Net model produces an IoU value of 0.78 at this stage, which represents an improvement of approximately 6% relative to DeepLabV3+. By epoch 25, the proposed model reaches an IoU score of 0.86, whereas DeepLabV3+ reaches 0.78. These results indicate that the ensemble segmentation mechanism improves boundary detection that enhances the segmentation quality in surgical frames.

Precision measures the proportion of correctly identified segmentation pixels relative to all predicted positive pixels. This metric reflects the ability of the segmentation model to avoid false positive predictions.

Table.11. Precision comparison across segmentation models

Epoch	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.72	0.74	0.75	0.79
10	0.75	0.77	0.79	0.83
15	0.77	0.80	0.82	0.87
20	0.80	0.83	0.85	0.90
25	0.82	0.85	0.87	0.93

The precision analysis presented in Table.**11** indicates that the proposed ADES-Net model produces fewer false positive predictions than the baseline methods. At epoch 5, the U-Net model achieves a precision value of 0.72, while the Attention U-Net and DeepLabV3+ models achieve values of 0.74 and 0.75 respectively. The proposed method achieves a precision score of 0.79 during the same epoch. As the training continues, the precision values increase gradually across all models. At epoch 15, the U-Net model records a precision value of 0.77, whereas the Attention U-Net and DeepLabV3+ models reach 0.80 and 0.82 respectively. The proposed model achieves a precision value of 0.87, which indicates that the ensemble mechanism effectively reduces incorrect predictions that occur near the segmentation boundaries. By epoch 25, the precision value of the proposed method reaches 0.93, while the DeepLabV3+ model achieves 0.87. The improvement demonstrates that the ensemble learning framework produces more reliable segmentation predictions. Recall evaluates the ability of the segmentation system to identify all relevant pixels that belong to the target class.

Table.12. Recall comparison across segmentation models

Epoch	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.70	0.72	0.74	0.77
10	0.73	0.75	0.77	0.81
15	0.76	0.78	0.80	0.85
20	0.78	0.81	0.83	0.88
25	0.80	0.83	0.85	0.91

The recall performance results in Table.12 demonstrate that the proposed ADES-Net model identifies relevant segmentation regions more effectively than the baseline models. At epoch 5, the recall value for U-Net is 0.70, whereas Attention U-Net and DeepLabV3+ achieve values of 0.72 and 0.74 respectively. The proposed method achieves a recall value of 0.77 during the same stage. This improvement indicates that the ensemble architecture detects more relevant instrument and tissue pixels. At epoch 15, the recall values for U-Net, Attention U-Net, and DeepLabV3+ reach 0.76, 0.78, and 0.80 respectively. The ADES-Net method achieves a recall value of 0.85 at this stage, which demonstrates improved detection capability. By epoch 25, the recall value of the proposed method increases to 0.91, which exceeds the DeepLabV3+ performance by approximately 6%. These results indicate that the active learning mechanism improves selection that enhances the learning efficiency. The F1 score combines precision and recall into a balanced evaluation metric.

Table.13. F1 score comparison across segmentation models

Epoch	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.71	0.73	0.74	0.78
10	0.74	0.76	0.78	0.82
15	0.76	0.79	0.81	0.86
20	0.79	0.82	0.84	0.89
25	0.81	0.84	0.86	0.92

The F1 score results presented in Table.**13** demonstrate that the proposed ADES-Net framework provides a balanced improvement in both precision and recall performance. At epoch 5, the F1 score of the U-Net model is 0.71, while the Attention U-Net and DeepLabV3+ models achieve values of 0.73 and 0.74 respectively. The proposed model produces an F1 score of 0.78, which indicates superior segmentation accuracy. At epoch 15, the F1 score for DeepLabV3+ reaches 0.81, whereas the proposed ADES-Net model reaches 0.86. This improvement demonstrates that the ensemble architecture improves segmentation reliability by combining multiple model predictions. By epoch 25, the proposed model achieves an F1 score of 0.92, which exceeds the DeepLabV3+ performance of 0.86. The results confirm that the integration of ensemble learning and active frame selection enhances the overall segmentation performance across surgical video datasets.

5.1 RESULTS BASED ON ANNOTATION SIZE

The experimental evaluation analyzes the segmentation performance of the U-Net, Attention U-Net, DeepLabV3+, and the proposed ADES-Net framework across varying annotation sizes. The Dice Similarity Coefficient evaluates the spatial overlap between the predicted segmentation mask and the ground truth annotation. The metric measures segmentation accuracy that exists across varying numbers of annotated frames.

Table.14. Dice similarity coefficient across frame sizes

Annotated Frames	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.68	0.70	0.72	0.76
10	0.72	0.74	0.76	0.81
15	0.75	0.78	0.80	0.86
20	0.79	0.82	0.84	0.90
25	0.82	0.85	0.87	0.93

The Dice similarity coefficient results demonstrate that the segmentation accuracy increases as the number of annotated frames increases. As shown in Table.14, the U-Net method produces a Dice score of 0.68 when only five annotated frames exist within the training set. The Attention U-Net and DeepLabV3+ methods produce values of 0.70 and 0.72 respectively under the same condition. In contrast, the proposed ADES-Net method achieves a Dice value of 0.76, which indicates that the ensemble mechanism improves segmentation performance even when the available annotation data is limited. When the number of annotated frames increases to fifteen, the Dice score of U-Net increases to 0.75, while Attention U-Net and DeepLabV3+ reach 0.78 and 0.80 respectively. The proposed method achieves a Dice value of 0.86 at the same stage. This improvement of approximately 6% relative to DeepLabV3+ indicates that the ensemble architecture captures richer spatial features that enhance segmentation accuracy. When twenty-five annotated frames exist within the dataset, the proposed method achieves a Dice score of 0.93, which exceeds the DeepLabV3+ score of 0.87. The results therefore demonstrate that the active ensemble learning mechanism improves segmentation performance across limited annotation scenarios.

The Intersection over Union metric measures the ratio between the intersection region and the union region of the predicted segmentation mask and the reference annotation.

Table.15. Intersection over Union performance comparison

Annotated Frames	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.58	0.61	0.63	0.68
10	0.62	0.65	0.68	0.73
15	0.66	0.69	0.72	0.78
20	0.69	0.73	0.76	0.82
25	0.72	0.76	0.79	0.86

The IoU results presented in Table.15 indicate that the proposed ADES-Net method consistently produces higher segmentation overlap across all annotation sizes. When five annotated frames exist in the dataset, the U-Net model produces an IoU value of 0.58, whereas the Attention U-Net and DeepLabV3+ models achieve values of 0.61 and 0.63 respectively. The proposed method achieves an IoU value of 0.68 under the same condition. As the number of annotated frames increases to fifteen, the IoU score of U-Net reaches 0.66, while Attention U-Net and DeepLabV3+ reach 0.69 and 0.72 respectively. The proposed ADES-Net achieves a value of 0.78 at the same stage. This improvement indicates that the ensemble prediction mechanism enhances boundary segmentation that improves spatial consistency. When the annotated frames increase to twenty-five, the IoU score of the proposed method reaches 0.86, whereas the DeepLabV3+ model achieves 0.79. The difference of approximately 7% demonstrates that the active frame selection mechanism improves the training efficiency and segmentation accuracy.

Precision measures the proportion of predicted positive pixels that are correctly classified by the segmentation model.

Table.16. Precision comparison across segmentation models

Annotated Frames	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.70	0.72	0.74	0.78
10	0.74	0.76	0.79	0.83
15	0.77	0.80	0.82	0.87
20	0.80	0.83	0.85	0.90
25	0.83	0.86	0.88	0.93

The precision results demonstrate that the proposed ADES-Net framework produces fewer false positive predictions than the baseline segmentation models. As shown in Table.16, the U-Net model achieves a precision value of 0.70 when the dataset contains five annotated frames. The Attention U-Net and DeepLabV3+ models produce values of 0.72 and 0.74 respectively under the same condition. The proposed ADES-Net achieves a precision score of 0.78 at this stage. As the number of annotated frames increases to fifteen, the precision score of U-Net increases to 0.77, while the Attention U-Net and DeepLabV3+ models achieve values of 0.80 and 0.82 respectively. The proposed method produces a precision score of 0.87 during the same stage. This improvement indicates that the ensemble learning strategy reduces segmentation errors that occur around surgical instrument boundaries. When the annotated frames increase to twenty-five, the precision score of the proposed method reaches 0.93, whereas DeepLabV3+ achieves 0.88. This difference indicates that the ensemble model improves prediction reliability that benefits surgical annotation tasks.

Recall evaluates the ability of the segmentation system to correctly identify all relevant pixels that belong to the target class.

Table.17. Recall comparison across segmentation models

Annotated Frames	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.67	0.70	0.72	0.76
10	0.71	0.74	0.76	0.81
15	0.75	0.78	0.80	0.85
20	0.78	0.81	0.83	0.88
25	0.81	0.84	0.86	0.91

The recall analysis presented in Table.17 indicates that the proposed ADES-Net model identifies relevant segmentation regions more effectively than the baseline methods. When the dataset contains five annotated frames, the U-Net model produces a recall value of 0.67, whereas Attention U-Net and DeepLabV3+ produce values of 0.70 and 0.72 respectively. The proposed ADES-Net achieves a recall value of 0.76 during the same stage. This improvement demonstrates that the ensemble architecture detects more relevant surgical instrument pixels. As the annotation size increases to fifteen frames, the recall values for U-Net, Attention U-Net, and DeepLabV3+ increase to 0.75, 0.78, and 0.80 respectively. The proposed method achieves a recall value of 0.85 under the same condition. When twenty-five annotated frames exist within the dataset, the recall value of the proposed model reaches 0.91, which exceeds the DeepLabV3+ performance by approximately 5%. These results demonstrate that the active learning mechanism selects informative frames that improve segmentation learning efficiency.

The F1 score combines precision and recall into a balanced metric that evaluates the overall segmentation performance.

Table.18. F1 score comparison across segmentation models

Annotated Frames	U-Net	Attention U-Net	DeepLabV3+	ADES-Net (Proposed)
5	0.69	0.71	0.73	0.77
10	0.73	0.75	0.78	0.82
15	0.76	0.79	0.81	0.86
20	0.79	0.82	0.84	0.89
25	0.82	0.85	0.87	0.92

The F1 score results demonstrate that the proposed ADES-Net framework achieves balanced improvements in both detection accuracy and segmentation completeness. As shown in Table.18, the U-Net model achieves an F1 score of 0.69 when the dataset contains five annotated frames. The Attention U-Net and DeepLabV3+ models achieve values of 0.71 and 0.73 respectively. The proposed ADES-Net method produces an F1 score of 0.77 during the same stage. As the annotated dataset size increases to fifteen frames, the DeepLabV3+ model achieves an F1 score of 0.81, whereas the proposed model achieves 0.86. This improvement indicates that the ensemble architecture integrates complementary segmentation predictions that improve accuracy. When the annotation size increases to twenty-five frames, the proposed method achieves an F1 score of 0.92, which exceeds the DeepLabV3+ performance of 0.87. These findings confirm that the integration of ensemble segmentation and active learning improves automated surgical video annotation performance.

6. CONCLUSION

The proposed ADES-Net demonstrates a significant improvement in automated surgical video segmentation annotation. The experimental evaluation shows that integrating deep ensemble learning with an active frame selection mechanism enhances segmentation accuracy while reducing manual annotation effort. Across multiple performance metrics, the proposed framework consistently outperforms baseline methods such as U-Net, Attention U-Net, and DeepLabV3+. Specifically, the Dice similarity coefficient reaches 0.93, IoU achieves 0.86, precision and recall reach 0.93 and 0.91 respectively, and the F1 score attains 0.92 when trained with twenty-five annotated frames. These improvements indicate that the ensemble models have captured diverse spatial and contextual features that improve the detection of surgical instruments and anatomical structures. The active learning component prioritizes frames with high uncertainty, which reduces redundant labeling while it maintains a robust model performance. The iterative training strategy has proved continuous improvement as newly annotated frames are incorporated into the training set. The evaluation also confirms that the proposed framework maintains consistent performance even with limited annotation samples, highlighting its efficiency for clinical settings where expert annotation is costly.

REFERENCES

[1] F. Wu, P. Marquez-Neila, M. Zheng, H. Rafii-Tari and R. Sznitman, "Correlation-Aware Active Learning for Surgery Video Segmentation", *Proceedings of International Conference on Applications of Computer Vision*, pp. 2010-2020, 2024.

[2] H. Peng, S. Lin, D. King, Y.H. Su, W.M. Abuzeid, R.A. Bly and B. Hannaford, "Reducing Annotating Load: Active Learning with Synthetic Images in Surgical Instrument Segmentation", *Medical Image Analysis*, Vol. 97, pp. 1-8, 2024.

[3] L. Yang, Y. Zhang, J. Chen, S. Zhang and D.Z. Chen, "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation", *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 399-407, 2017.

[4] S. Zhao, Y. Zhou and J. Chen, "Active Learning Pipeline for Biomedical Image Instance Segmentation with Minimal Human Intervention", *BVM Workshop*, pp. 217-222, 2025.

[5] J.M. Brandenburg, A.C. Jenke, A. Stern, M.T. Daum, A. Schulze, R. Younis and M. Wagner, "Active Learning for Extracting Surgomic Features in Robot-Assisted Minimally Invasive Esophagectomy: A Prospective Annotation Study", *Surgical Endoscopy*, Vol. 37, No. 11, pp. 8577-8593, 2023.

[6] J. Aklilu and S. Yeung, "ALGES: Active Learning with Gradient Embeddings for Semantic Segmentation of Laparoscopic Surgical Images", *Proceedings of International Conference on Machine Learning for Healthcare*, pp. 892-911, 2022.

[7] S. Ayache and G. Quenot, "Video Corpus Annotation using Active learning", *Proceedings of International Conference on Information Retrieval*, pp. 187-198, 2008.

[8] A. Murali, A. Garg, S. Krishnan, F.T. Pokorny, P. Abbeel, T. Darrell and K. Goldberg, "Tsc-dl: Unsupervised Trajectory Segmentation of Multi-Modal Surgical Demonstrations with Deep Learning", *Proceedings of International Conference on Robotics and Automation*, pp. 4150-4157, 2016.

[9] X. Xiao, J. Zhang, Y. Shao, J. Liu, K. Shi, C. He and D. Kong, "Deep Learning-based Medical Ultrasound Image and Video Segmentation Methods: Overview, Frontiers and Challenges", *Sensors*, Vol. 25, No. 8, pp. 1-13, 2025.

[10] M. Gorriz, A. Carlier, E. Faure and X. Giro-I-Nieto, "Cost-Effective Active Learning for Melanoma Segmentation", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-6, 2017.

[11] W. Zhang, L. Zhu, J. Hallinan, S. Zhang, A. Makmur, Q. Cai and B.C. Ooi, "Boostmis: Boosting Medical Image Semi-Supervised Learning with Adaptive Pseudo Labeling and Informative Active Annotation", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 20666-20676, 2022.

[12] M.A. Shoaib, R. Ali, S.U. Bazai and T. Mir, "Deep Learning Techniques for Image Segmentation and Data Annotation", *Modern Intelligent Techniques for Image Processing*, pp. 63-94, 2025.

[13] I. Kansizoglou, L. Bampis and A. Gasteratos, "An Active Learning Paradigm for Online Audio-Visual Emotion Recognition", *IEEE Transactions on Affective Computing*, Vol. 13, No. 2, pp. 756-768, 2019.

[14] O.R. Meireles, G. Rosman, M.S. Altieri, L. Carin, G. Hager and A. Madani, "SAGES Consensus Recommendations on an Annotation Framework for Surgical Video", *Surgical Endoscopy*, Vol. 35, No. 9, pp. 4918-4929, 2021.

[15] K.C. Santosh and S. Nakarmi, "*Active Learning to Minimize the Possible Risk of Future Epidemics*", 2023.