GUARDIAN: Guarding Against Uncertainty and Adversarial Risks in Robot-Assisted Surgeries

Ufaq Khan*✉, Umair Nawaz*, Tooba T. Sheikh,
Asif Hanif, and Mohammad Yaqub

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI),
Abu Dhabi, United Arab Emirates
{firstname.lastname}@mbzuai.ac.ae

Abstract. In the realm of robotic-assisted surgeries, like laparoscopic cholecystectomy, the integration of deep learning (DL) models marks a significant advancement in achieving surgical precision and minimal invasiveness, which in turn, elevates patient outcomes and reduces recovery times. However, the vulnerability of these DL models to adversarial attacks introduces a critical risk, emphasizing the need for enhanced model robustness. Our study addresses this challenge by proposing a comprehensive framework that not only fortifies surgical action recognition models against adversarial threats through adversarial training and pre-processing strategies but also incorporates uncertainty estimation to enhance prediction confidence and trustworthiness. Our framework demonstrates superior resilience against a wide spectrum of adversarial attacks and showcases improved reliability in surgical tool detection under adversarial conditions. It achieves an improvement from 8% to 23.58% in terms of triplet (instrument, verb, triplet) predictions. These contributions significantly enhance the security and reliability of deep learning applications in the critical domain of robotic surgery, offering an approach that safeguards advanced surgical technologies against malicious threats, thereby promising enhanced patient care and surgical precision. Code is available at https://github.com/umair1221/guardian.

Keywords: Robotic Surgery · Adversarial Attacks · Adversarial Training · Uncertainty Estimation · Trustworthy Robotic Surgery

1 Introduction

In modern surgical procedures, a wide array of specialized tools are utilized to ensure precision and efficacy, ranging from scalpels and forceps to advanced robotic instruments. With advancements in technology, there has been a noticeable shift towards integrating robotic-assisted procedures into surgical practices like laparoscopic cholecystectomy [16], offering enhanced dexterity and minimally invasive techniques for improved patient outcomes.

* Equal Contribution ✉ Corresponding Author

The integration of robotics into surgical procedures significantly enhances precision and patient outcomes. Combining deep learning algorithms with robotic platforms further augments their capabilities [13], enabling more effective task execution and real-time decision-making. Despite these advancements, robotic surgery faces unique security challenges due to its highly interactive operational environment within hospitals, making it more susceptible to adversarial attacks compared to less interactive specialties such as radiology. These attacks can exploit some vulnerabilities of deep learning models [11], potentially manipulating robotic actions and leading to severe consequences such as surgical inaccuracies, patient harm, and diminished trust in robotic-assisted surgeries.

These critical implications emphasize the imperative for effective defense methods to protect such complex medical systems. In a study by Han *et al.* [10], the authors highlight the vulnerabilities of deep learning models in ECG signal analysis. Additionally, Aguiar *et al.* [7] reveal similar susceptibilities in COVID-19 X-ray classification whereas Morshuis *et al.* [23] show the susceptibility of MRI reconstruction algorithms to adversarial perturbations, thereby, risking diagnostic accuracy. Furthermore, Rahman *et al.* [29] explore the vulnerabilities in COVID-19 diagnostics using medical IoT data while Puttagunta *et al.* [28] provide an overview of defensive measures in medical imaging. Almalik *et al.* [3] introduce SEViT, a technique enhancing the resilience of vision transformers in chest X-rays and fundoscopy. Moreover, similar attacks can also arise from cyber-attacks, camera feed manipulation, or sensor noise in robotic-assisted surgeries, similar to scenarios in autonomous vehicles [5]. Research on tele-operated robots like Raven-II [6] and [24,2] show vulnerabilities that can affect fully autonomous systems. Lastly, Cheng *et al.* [4] analyze the vulnerabilities of deep learning models in surgical action recognition under adversarial conditions, establishing a foundation for our study. However, Cheng *et al.* [4] did not provide a robust training framework to mitigate the impact of these adversarial attacks.

We introduce a novel defense mechanism for deep learning-enabled robotic-assisted surgical systems against adversarial attacks. Our innovative framework, GUARDIAN, enhances decision-making accuracy through uncertainty estimation and augments the safety of robotic-assisted surgical procedures using adversarial training. Integrating adversarial training with refined object detection algorithms, our approach as illustrated in Fig. 1 increases the accuracy of surgical instrument identification under adversarial attacks. Through comprehensive evaluations across diverse datasets, we demonstrate the potential to elevate both the reliability and security of robotic-assisted surgical systems. This paper does not aim to provide a new technical contribution but rather perform an extensive analysis of how the existing methods can be further enhanced for robotic-assisted surgical applications. In summary, our main contributions are as follows:

– We propose a framework for enhancing adversarial resilience in surgical triplet recognition, incorporating in-depth adversarial training and reliable tool detection. Our approach also explores the cross-task transferability of adversarial attacks between recognition and detection tasks. To the best of

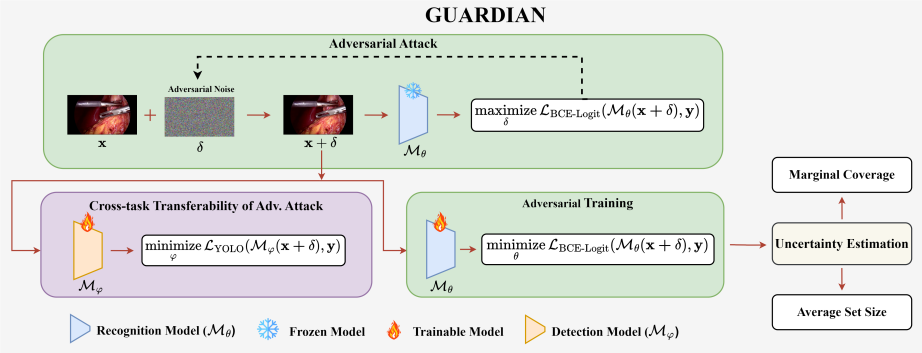


Fig. 1. GUARDIAN: A three-step approach to enhance model robustness and predictive accuracy. It begins with transforming clean samples into perturbed ones via adversarial attacks, posing it as a maximization problem for the model, M_θ. The process proceeds with two training phases: enhancing tool detection through cross-task transferability of adversarial examples and refining triplet recognition with live adversarial training. Here, $\mathcal{L}_{\text{YOLO}}$ denotes the combination of classification and bounding-box loss. The final step applies conformal prediction post-training, evaluating prediction reliability, with the dotted line indicating gradient updates.

our knowledge, this paper is the first to investigate this crucial task and devise a robust mechanism against adversarial attacks.
– We embed an inferential uncertainty estimation mechanism to support the model's predictive confidence fidelity, underpinning surgical decision-making with empirical insights.
– We conduct a comprehensive assessment, utilizing ablation studies to systematically delineate the robustness of our methodology against various adversarial challenges and hyperparameter settings.

2 Methodology

In this work, we consider two models: the surgical action triplet recognition model (denoted by $\mathcal{M}_\theta$) and the tool detection model (denoted by $\mathcal{M}_\varphi$) which are commonly performed tasks in such clinical practice. We use the recognition model $\mathcal{M}_\theta$ to generate adversarial samples and analyze their performance under various attacks. Tool detection model $\mathcal{M}_\varphi$ is used to study cross-task transferability (from action triplet recognition to tool detection) of the baseline attacks.

Surgical Action Triplet Recognition - Primer: For surgical action triplet recognition in laparoscopic cholecystectomy surgery video frames, the model containing four classification heads is given a frame/image and it provides predictions of four entities: instruments (I), verb/actions (V), targets (T) and triplet combination of all these components (IVT). Let's denote the recognition model,

clean three-channel input image and categorical ground-truth labels with $\mathcal{M}_\theta$, $\mathbf{x} \in \mathbb{R}^{c,h,w}$ and $\mathbf{y} = \{\mathrm{y_I}, \mathrm{y_V}, \mathrm{y_T}, \mathrm{y_{IVT}}\}$ respectively. The number of classes in each component are denoted with $|\mathrm{I}|, |\mathrm{V}|, |\mathrm{T}|,$ and $|\mathrm{IVT}|$. When the input image is passed through the model, the prediction scores/logits $\hat{\mathbf{y}} = \mathcal{M}_\theta(\mathbf{x})$ are obtained where $\hat{\mathbf{y}} = \{\hat{\mathrm{y}}_\mathrm{I} \in \mathbb{R}^{|\mathrm{I}|}, \hat{\mathrm{y}}_\mathrm{V} \in \mathbb{R}^{|\mathrm{V}|}, \hat{\mathrm{y}}_\mathrm{T} \in \mathbb{R}^{|\mathrm{T}|}, \hat{\mathrm{y}}_\mathrm{IVT} \in \mathbb{R}^{|\mathrm{IVT}|}\}$.

For normal training with clean images, the following objective is optimized;

$$\underset{\theta}{\text{minimize}} \ \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{BCE-Logit}}(\mathcal{M}_\theta(\mathbf{x}_i), \mathbf{y}_i), \tag{1}$$

where $(\mathbf{x}_i, \mathbf{y}_i)$ is image-labels pair, $\mathcal{L}_{\text{BCE-Logit}}(\mathcal{M}_\theta(\mathbf{x}), \mathbf{y}) = \mathcal{L}(\hat{\mathrm{y}}_\mathrm{I}, \mathrm{y}_\mathrm{I}) + \mathcal{L}(\hat{\mathrm{y}}_\mathrm{V}, \mathrm{y}_\mathrm{V}) + \mathcal{L}(\hat{\mathrm{y}}_\mathrm{T}, \mathrm{y}_\mathrm{T}) + \mathcal{L}(\hat{\mathrm{y}}_\mathrm{IVT}, \mathrm{y}_\mathrm{IVT})$ is sum of component-wise binary cross-entropy loss and θ denotes model's weights.

Adversarial Attack on Recognition Model $\mathcal{M}_\theta$: An adversarial attack on a pre-trained deep learning model aims to find a human-imperceptible perturbation which when added to the image makes the model predict the wrong output. Let's denote the clean image with $\mathbf{x}$ and perturbed image with $\mathbf{x}' = \mathbf{x} + \delta$ where δ is the perturbation. In a vanilla adversarial attack, an adversarial image $\mathbf{x}'$ is obtained by optimizing the following objective:

$$\begin{aligned} \underset{\delta}{\text{maximize}} \ & \mathcal{L}_{\text{BCE-Logit}}(\mathcal{M}_\theta(\mathbf{x} + \delta), \mathbf{y}) \\ \text{s.t.} \ & -\epsilon \leq \|\delta\|_p \leq +\epsilon, \end{aligned} \tag{2}$$

where ϵ is the perturbation budget and "ℓ_p" denotes ℓ_p norm. In general, ℓ_2 and ℓ_∞ norms are used in adversarial attacks. While generating the adversarial sample $\mathbf{x}'$, the model $\mathcal{M}_\theta$ is kept frozen i.e. its weights are not updated.

Defense: To counter the adversarial attack, we consider adversarial training [21] as a first measure of defense. In adversarial training, firstly the model is attacked in a frozen state to get adversarial samples (Eq. 2), and then its weights are updated against these samples (Eq. 3). Formally, adversarial training is a min-max objective that is optimized as follows:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \ & \left(\underset{\delta}{\text{maximize}} \ \mathcal{L}_{\text{BCE-Logit}}(\mathcal{M}_\theta(\mathbf{x} + \delta), \mathbf{y}) \right) \\ \text{s.t.} \ & -\epsilon \leq \|\delta\|_p \leq +\epsilon, \end{aligned} \tag{3}$$

Adversarial training makes the model robust to adversarial attacks. Since adversarial training is compute-intensive and requires many GPU hours, therefore, we also consider other defense mechanisms that perform preprocessing to neutralize the adversarial perturbations such as Spatial Smoothing (SS) [32], Pixel Defend (PD) [30], Feature Squeeze (FS) [32], and JPEG compression [8].

Cross-task Transferability of Adversarial Attack: We also study the transferability of the adversarial attack from recognition model $\mathcal{M}_\theta$ to tool detection model $\mathcal{M}_\varphi$. Cross-task transferability occurs when adversarial samples from one model (e.g., recognition) also fool another model (e.g., detection) trained on the

same input for different output tasks. For this purpose, we first generate adversarial samples by attacking $\mathcal{M}_\theta$ and then perform inference on pre-trained tool detection model $\mathcal{M}_\varphi$. Moreover, we adversarially train $\mathcal{M}_\varphi$ to analyze its robustness against the adversarial samples.

Uncertainty Estimation: Our methodology enhances decision-making under adversarial conditions by incorporating conformal prediction (CP), which provides statistically valid prediction intervals to gauge confidence. This method flags low-confidence predictions often linked to adversarial inputs, thus improving accuracy in surgical action recognition. We divide our dataset into training and calibration sets to train the model and then calibrate it for confidence scores. We define a nonconformity measure and set a threshold that ensures prediction sets accurately reflect the true label with a specified confidence level. For instance, using CP with a 10% miscoverage level (α=0.1), our model outputs probable actions like grasping (0.65), suturing (0.22), and cutting (0.08), capturing the true action 90% of the time and enhancing decision-making reliability.

3 Experimental Setup

Datasets: Our research utilizes two datasets: CholecT45 [25], and m2cai16-tool-locations [14]. CholecT45, a subset of CholecT80 [31], contains 45 videos annotated with 100 triplet classes (Instrument, Verb, Target) across 5 folds, using 4 for training/validation and 1 for testing. This dataset facilitates an in-depth analysis of surgical procedures, instruments, and actions. The m2cai16-tool-locations dataset provides 2,532 frames with bounding box annotations for seven surgical tools.

Baseline Models: Our baseline model, Rendezvous (RDV) [26], stands as a state-of-the-art (SOTA) solution, specializes in surgical action triplets (instrument, verb, target) recognition during laparoscopic cholecystectomy surgery. The RDV model inputs a laparoscopic image and predicts sets of triplets in the image. Furthermore, to address the cross-task transferability of adversarial attacks, we first fine-tuned the YOLOv8 model [15] for surgical tool detection on the m2cai16-tool-locations dataset. Then, we created adversarial images by attacking the RDV model and subsequently utilized them to test the detection capabilities of the YOLOv8 model [17,18].

Adversarial Attacks: We apply four adversarial attacks named Projected Gradient Descent (PGD) [22], Fast Gradient Sign Method (FGSM) [9], Basic Iterative Method (BIM) [20], and Gaussian Noise (GN) [9], adapted from the TorchAttack [19] library and customized to suit our model's specifications. We perform different experiments on these attacks across varying parameters (steps, step-size, perturbation budget, standard deviation): steps $\in \{5, 10, 15, 20\}; \alpha \in \{2, 4, 8, 16\}; \epsilon \in \{4, 8, 16, 32\}; \sigma \in \{0.05, 0.1, 0.5, 0.9\}$, to rigorously assess our model's resilience against adversarial inputs.

Evaluation Metrics: We assess our model's performance and adversarial attack impact using metrics like mean Average Precision (mAP) for accuracy, along with LPIPS [33], PSNR [12], and SSIM [12] for evaluating perceptual effects of

perturbations. For predictive reliability, we apply conformal prediction metrics such as marginal coverage and average set size. The marginal coverage is the percentage of times the true label is included in the prediction set, reflecting reliability, whereas the average set size is the mean number of labels per set. These are essential in surgical contexts to assess the model's accuracy in generating prediction intervals and are crucial for enhancing the decision-making process.

Experimental Configuration: We utilize the PyTorch framework [27] with Python 3.10. The baseline experiments are conducted on workstations equipped with an Intel Xeon Silver 4215 processor, NVIDIA Quadro RTX 6000 GPU, and 128 GB RAM. The batch size of 32 is used for training and testing experiments. To optimize our framework, we fine-tune RDV and YOLOv8 hyperparameters to uncover the ideal settings. Fig. B in the Appendix showcases the performance of the RDV model against different learning rates. Using Optuna [1] library with a linear scheduler and SGD optimizer with weight decay, we identify optimal learning rates of 0.009108, 0.001168, and 0.001372 for the RDV model.

4 Results and Discussion

4.1 Adversarial Attack and Defense

We present a comprehensive evaluation of RDV model resilience against adversarial perturbations. In pristine conditions, the model demonstrates superior performance, as anticipated without adversarial interference. In Table 1, our comparison of various adversarial attacks on the CholecT45 dataset reveals that the GN attack induces the highest fooling rate. However, this results in significant image quality degradation, resulting in noticeable visual perturbations, an outcome generally to be avoided. We further provide an in-depth analysis of the performance of each attack against different metrics influenced by hyperparameter variations in the supplementary materials.

We train the RDV model against each identified attack, subsequently evaluating the model's performance on both clean images and those subjected to each attack. This process allows us to assess the model's robustness and its ability to transfer learning across various adversarial conditions. Typically, all models show a slight reduction in performance on clean images compared to models trained exclusively on unperturbed images. This phenomenon is commonly observed in adversarial training and thus emphasizes its effectiveness. The training is conducted with optimal epsilon values to fine-tune the process. Additionally, we apply pre-processing techniques to both clean and adversarially trained models. As outlined in Table 2, our comparative analysis between different methods indicates minor variations in performance across each attacked model, illustrating the adversarial training's capacity to enhance model resilience effectively.

4.2 Adversarial Object Detection

Quantitative Results: In our study, we also assess the YOLOv8 ($\mathcal{M}\varphi$) robustness to different adversarial attacks. Initially, the baseline model achieves around

Table 1. Comparative evaluation of image quality measures (PSNR, SSIM, LPIPS) and component accuracy (mAP for I, V, T, IVT) under clean and adversarial conditions, highlighting the model's resilience in surgical recognition tasks.

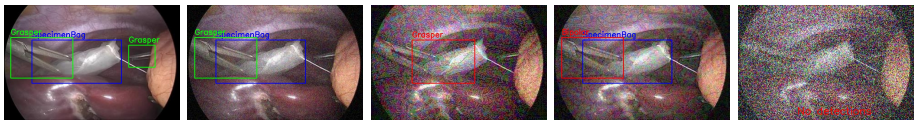
Evaluation Measures →	Image Quality Measure			Rendezvous Performance			
Attacks ↓	PSNR ↑	SSIM ↑	LPIPS ↑	I ↓	V ↓	T ↓	IVT ↓
Clean Images	-	-	-	81.29	55.21	31.16	25.12
PGD ($\epsilon = 8/16$)	**33.69**/**28.70**	**73.20**/**51.74**	**83.35**/**76.92**	66.06/51.21	42.16/31.31	23.37/16.45	17.05/12.26
FGSM ($\epsilon = 8/16$)	30.59/24.64	62.98/36.43	78.29/66.97	30.11/25.00	16.24/15.11	11.24/10.01	8.18/7.16
BIM ($\epsilon = 8/16$)	33.06/28.45	71.67/**52.91**	82.23/74.51	**17.28**/17.28	**10.98**/**10.12**	8.83/8.08	8.75/8.01
GN ($\sigma = 0.1/0.5$)	29.57/23.49	60.35/35.98	72.35/64.89	20.31/**15.17**	12.79/10.34	**7.36**/**5.18**	**6.45**/**4.27**

Table 2. Quantitative evaluation of mAP for triplet recognition (IVT) across adversarially trained models against PGD, FGSM, BIM, and GN attacks, along with preprocessing methods demonstrating robustness on both clean and adversarial images.

Models →	GUARDIAN (ours)					Defense Methods			
Attacks ↓	$\mathcal{M}^{\text{Clean}}$	$\mathcal{M}_{\text{🛡}}^{\text{PGD}}$	$\mathcal{M}_{\text{🛡}}^{\text{FGSM}}$	$\mathcal{M}_{\text{🛡}}^{\text{BIM}}$	$\mathcal{M}_{\text{🛡}}^{\text{GN}}$	SS	PD	FS	JPEG
Clean Images	25.21	23.84	23.47	23.65	23.12	-	-	-	-
PGD	**17.31**	21.32	22.76	**23.58**	**21.89**	18.98	14.75	19.87	21.04
FGSM	8.64	**21.48**	23.02	22.47	21.67	17.23	12.86	18.29	20.58
BIM	8.29	20.95	**23.17**	22.84	21.53	15.58	13.64	17.45	19.75
GN	6.84	21.25	22.65	22.95	20.98	10.04	8.63	16.37	18.22

Table 3. Comparative analysis of mAP for the YOLOv8 model in the tool detection task on the m2cai16-tool-locations dataset [14], contrasting conventional training against adversarial training with GUARDIAN. Here, the bold values represent the best mAP performance against the adversarial attacks.

Models →	Baseline Model $\mathcal{M}_{\varphi}$			GUARDIAN (ours) $\mathcal{M}_{\varphi}^{\text{🛡}}$		
Attacks ↓	mAP@[.5, .95]	mAP@.5	mAP@.7	mAP@[.5, .95]	mAP@.5	mAP@.7
Clean Images	61.62	95.48	70.24	52.87	95.52	68.62
PGD ($\epsilon = 32$)	26.25	53.35	23.55	**57.85**	**93.24**	**63.28**
FGSM ($\epsilon = 32$)	3.56	6.89	3.64	54.94	90.55	58.98
BIM ($\epsilon = 32$)	8.97	19.45	7.79	56.79	92.14	61.33
GN ($\sigma = 0.5$)	0.00	0.00	0.00	43.63	80.54	43.75



Clean **PGD** **FGSM** **BIM** **GN**

Fig. 2. Comparing YOLOv8's object detection mAP@.5 on clean versus perturbed images, with correct predictions highlighted in green/blue and incorrect ones in red.

95% mAP score at an IoU threshold of 0.5 on clean images. However, its performance significantly degrade under the influence of different attacks. Remarkably, after undergoing adversarial training with GUARDIAN, $\mathcal{M}^{\text{🛡}}\varphi$ demonstrates significant recovery in mAP scores against these attacks particularly against the GN attack, as detailed in Table 3. Here, 🛡 represents our model after the adversarial training.

Table 4. Results for conformal prediction. The marginal coverage and the corresponding average set size are reported for different confidence intervals.

Confidence Level (%) →	Marginal Coverage (%)				Average Set Size			
Models ↓	97 ↑	95 ↑	90 ↑	80 ↑	97 ↓	95 ↓	90 ↓	80 ↓
$\mathcal{M}_{\text{🛡}}^{\text{PGD}}$	**78.57**	71.57	60.47	52.83	**30.42**	52.33	64.65	72.74
$\mathcal{M}_{\text{🛡}}^{\text{FGSM}}$	76.42	70.85	58.97	51.29	29.88	51.17	63.45	71.59
$\mathcal{M}_{\text{🛡}}^{\text{BIM}}$	77.95	69.23	61.34	53.17	31.07	53.42	65.89	73.98
$\mathcal{M}_{\text{🛡}}^{\text{GN}}$	75.68	68.74	57.39	50.21	28.56	50.79	62.34	70.21

Qualitative Results: Fig. 2 illustrates the YOLOv8's object detection results at an IoU threshold of 0.5, comparing clean images with those perturbed by PGD, FGSM, BIM, and GN attacks, highlighting its detection performance in adversarial scenarios. Fig. A in the supplementary material provides further qualitative insights into the image quality degradation caused by attacks.

4.3 Uncertainty Estimation

Our adoption of conformal prediction (CP) highly elevates the robustness assessment of the RDV model under adversarial conditions. As shown in Table 4, CP maintains high marginal coverage at a 97% confidence level, affirming model reliability. However, with lower confidence levels, the performance predictably declines, demonstrating CP's precision-reliability balance. Additionally, the reduced average set size at higher confidence levels highlights CP's effectiveness in improving predictive accuracy.

Table 5. Comparison (mAP %) of our method with others for surgical component recognition. AT refers to adversarial training, and GUARDIAN combines all methodologies. BIM was chosen for its ability to reduce performance while maintaining perceptual integrity.

Methods ↓	I ↑	V ↑	T ↑	IVT ↑
RDV (No Attack)	81.29	55.21	31.16	25.12
RDV + BIM	17.15	10.05	8.32	8.01
RDV + BIM + AT	74.07	46.31	28.37	22.23
Ours (GUARDIAN)	**76.68**	**48.41**	**29.49**	**23.04**

4.4 Ablation Study

We also conduct several ablations to assess the impact of learning rate variations on adversarial training effectiveness (Fig. B in Appendix), as well as the influence of adversarial attack hyperparameters on image quality (Table A,B,C of Appendix) and the recognition of surgical components as shown in Fig. C of the

Appendix. The results of these studies are detailed in the supplementary section. Moreover, Table 5 also provides the comparison of different methods against our robust framework for surgical component recognition. Here, in Guardian, CP is used post-adversarial training to quantify uncertainty in predictions. CP generates a set of possible labels for each prediction, but for mAP calculation, we consider the highest confidence label from each set.

5 Conclusion

In conclusion, our work develops a framework that improves the resilience of deep learning models against adversarial attacks in robotic-assisted surgery by incorporating adversarial training along with conformal prediction for precise uncertainty estimation. This method significantly enhances the accuracy and reliability of surgical predictions crucial for decision-making. Our evaluations across various datasets confirm the effectiveness of our framework in enhancing predictive accuracy and ensuring surgical safety. Currently, we focus on established approaches for surgical action recognition. Future work will explore developing specialized attacks targeting key areas such as instruments, verbs, and targets to refine the precision of adversarial challenges.

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

References

1. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
2. Homa Alemzadeh, Daniel Chen, Xiao Li, Thenkurussi Kesavadas, Zbigniew T Kalbarczyk, and Ravishankar K Iyer. Targeted attacks on teleoperated surgical robots: Dynamic model-based detection and mitigation. In *2016 46th annual IEEE/IFIP international conference on dependable systems and networks (DSN)*, pages 395–406. IEEE, 2016.
3. Faris Almalik, Mohammad Yaqub, and Karthik Nandakumar. Self-ensembling vision transformer (sevit) for robust medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 376–386. Springer, 2022.
4. Yanqi Cheng, Lihao Liu, Shujun Wang, Yueming Jin, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Why deep surgical models fail?: Revisiting surgical action triplet recognition through the lens of robustness. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 177–189. Springer, 2023.
5. Yushi Cheng, Xiaoyu Ji, Wenjun Zhu, Shibo Zhang, Kevin Fu, and Wenyuan Xu. Adversarial computer vision via acoustic manipulation of camera sensors. *IEEE Transactions on Dependable and Secure Computing*, 21(4):3734–3750, 2024.

6. Keywhan Chung, Xiao Li, Peicheng Tang, Zeran Zhu, Zbigniew T Kalbarczyk, Ravishankar K Iyer, and Thenkurussi Kesavadas. Smart malware that uses leaked control data of robotic applications: The case of {Raven-II} surgical robots. In *22nd International symposium on research in attacks, intrusions and defenses (RAID 2019)*, pages 337–351, 2019.
7. Erikson J. de Aguiar, Karem D. Marcomini, Felipe A. Quirino, Marco A. Gutierrez, Caetano Traina Jr, and Agma JM Traina. Evaluation of the impact of physical adversarial attacks on deep learning models for classifying covid cases. In *Medical Imaging 2022: Computer-Aided Diagnosis*, volume 12033, pages 722–728. SPIE, 2022.
8. Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of jpg compression on adversarial images, 2016.
9. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2014.
10. Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature medicine*, 26(3):360–363, 2020.
11. Asif Hanif, Muzammal Naseer, Salman Khan, Mubarak Shah, and Fahad Shahbaz Khan. Frequency domain adversarial training for robust volumetric medical segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–467. Springer, 2023.
12. Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.
13. Sardar Mehboob Hussain, Antonio Brunetti, Giuseppe Lucarelli, Riccardo Memeo, Vitoantonio Bevilacqua, and Domenico Buongiorno. Deep learning based image processing for robot assisted surgery: A systematic literature survey. *IEEE Access*, 10:122627–122657, 2022.
14. Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 691–699. IEEE, 2018.
15. Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.
16. S. Kalata, J. R. Thumma, E. C. Norton, J. B. Dimick, and K. H. Sheetz. Comparative safety of robotic-assisted vs laparoscopic cholecystectomy. *JAMA Surgery*, 158(12):1303–1310, 2023.
17. Ufaq Khan, Mustaqeem Khan, Abdulmotaleb Elsaddik, and Wail Gueaieb. Ddnet: Diabetic retinopathy detection system using skip connection-based upgraded feature block. In *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE, 2023.
18. Ufaq Khan, Umair Nawaz, Mustaqeem Khan, Abdulmotaleb El Saddik, and Wail Gueaieb. Fetr: A weakly self-supervised approach for fetal ultrasound anatomical detection. In *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6, 2024.
19. Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.
20. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR)*, 2016.

21. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
22. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.
23. Jan Nikolas Morshuis, Sergios Gatidis, Matthias Hein, and Christian F Baumgartner. Adversarial robustness of mr image reconstruction under realistic perturbations. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 24–33. Springer, 2022.
24. Subash Neupane, Shaswata Mitra, Ivan A Fernandez, Swayamjit Saha, Sudip Mittal, Jingdao Chen, Nisha Pillai, and Shahram Rahimi. Security considerations in ai-robotics: A survey of current methods, challenges, and opportunities. *IEEE Access*, 2024.
25. Chinedu Innocent Nwoye and Nicolas Padoy. Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235*, 2022.
26. Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.
27. A Paszke, S Gross, F Massa, A Lerer, Jea PyTorch Bradbury, G Chanan, T Killeen, Z Lin, N Gimelshein, L Antiga, et al. An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst*, 32:8026.
28. Murali Krishna Puttagunta, S Ravi, and C Nelson Kennedy Babu. Adversarial examples: attacks and defences on medical deep learning systems. *Multimedia Tools and Applications*, pages 1–37, 2023.
29. Abdur Rahman, M Shamim Hossain, Nabil A Alrajeh, and Fawaz Alsolami. Adversarial examples—security threats to covid-19 deep learning systems in medical iot devices. *IEEE Internet of Things Journal*, 8(12):9603–9610, 2020.
30. Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples, 2018.
31. Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos, 2016.
32. Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018.
33. Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.