ACT-Net: Anchor-Context Action Detection in Surgery Videos

Luoying Hao[1,2], Yan Hu[2(✉)], Wenjun Lin[2,3], Qun Wang[4], Heng Li[2], Huazhu Fu[5], Jinming Duan[1(✉)], and Jiang Liu[2(✉)]

[1] School of Computer Science, University of Birmingham, Birmingham, UK
j.duan@bham.ac.uk
[2] Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
{huy3,liuj}@sustech.edu.cn
[3] Department of Mechanical Engineering, National University of Singapore, Singapore, Singapore
[4] Third Medical Center of Chinese PLAGH, Beijing, China
[5] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

Abstract. Recognition and localization of surgical detailed actions is an essential component of developing a context-aware decision support system. However, most existing detection algorithms fail to provide high-accuracy action classes even having their locations, as they do not consider the surgery procedure's regularity in the whole video. This limitation hinders their application. Moreover, implementing the predictions in clinical applications seriously needs to convey model confidence to earn entrustment, which is unexplored in surgical action prediction. In this paper, to accurately detect fine-grained actions that happen at every moment, we propose an anchor-context action detection network (ACT-Net), including an anchor-context detection (ACD) module and a class conditional diffusion (CCD) module, to answer the following questions: 1) where the actions happen; 2) what actions are; 3) how confidence predictions are. Specifically, the proposed ACD module spatially and temporally highlights the regions interacting with the extracted anchor in surgery video, which outputs action location and its class distribution based on anchor-context interactions. Considering the full distribution of action classes in videos, the CCD module adopts a denoising diffusion-based generative model conditioned on our ACD estimator to further reconstruct accurately the action predictions. Moreover, we utilize the stochastic nature of the diffusion model outputs to access model confidence for each prediction. Our method reports the state-of-the-art performance, with improvements of 4.0% mAP against baseline on the surgical video dataset.

L. Hao and Y. Hu—Co-first authors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_19.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 196–206, 2023.
https://doi.org/10.1007/978-3-031-43996-4_19

Keywords: Action detection · Anchor-context · Conditional diffusion · Surgical video

1 Introduction

Surgery is often an effective therapy that can alleviate disabilities and reduce the risk of death from common conditions [17]. While surgical procedures are intended to save lives, errors within the surgery may bring great risks to the patient and even cause sequelae [18], which emphasizes the development of a computer-assisted system. A context-aware assistant system for surgery can not only decrease intraoperative adverse events, and enhance the quality of interventional healthcare [28], but also contribute to surgeon training, and assist procedure planning and retrospective analysis [9].

Designing intelligent assistance systems for operating rooms requires an understanding of surgical scenes and procedures [20]. Most current works pay attention to phase and step recognition [3,27], which is to get the major types of events that occurred during the surgery. They merely provided very coarse descriptions of scenes. As the granularity of action increases, the clinical utility becomes more valuable in providing an accurate depiction of detailed motion [13,19]. Recent studies focus on fine-grained action recognition by modelling action as a group of the instrument, its role, and its target anatomy and capturing their associations [7,26]. Recognizing targets in different methods is dependent on different surgical scenarios and it also significantly increases the complexity and time consumption for anatomy annotation [30]. In addition, although most existing methods can provide accurate action positions, the predicted action class is often inaccurate. Moreover, they do not provide any information about the reliability of their output, which is a key requirement for integrating into assistance systems of surgery [11]. Thus, we propose a reliable surgical action detection method in this paper, with high-accuracy action predictions and their confidence.

Mistrust is a major barrier to deep-learning-based predictions applied to clinical implementation [14]. Existing works measuring the model uncertainty [1,8] often need several-time re-evaluations, and store multiple sets of weights. It is hard for them to apply to surgery assistance applications to get confidence for each prediction directly [10], and they are limited to improving prediction performance. Conditional diffusion-based generative models have received significant attention due to their ability to accurately recover the full distribution of data guided by conditions from the perspective of diffusion probabilistic models [24]. However, they focus on generating high-resolution photo-realistic images. Instead, after observing our surgical video dataset, our conditional diffusion model aims to reconstruct accurately class distribution. We also access the estimation of confidence with the stochastic nature of the diffusion model.

Here, to predict accurately micro-action (fine-grained action) categories happening every moment, we achieve it with two modules. Specifically, a novel anchor-context module for action detection is proposed to highlight the spatio-temporal regions that are interacted with the anchors (we extract instrument

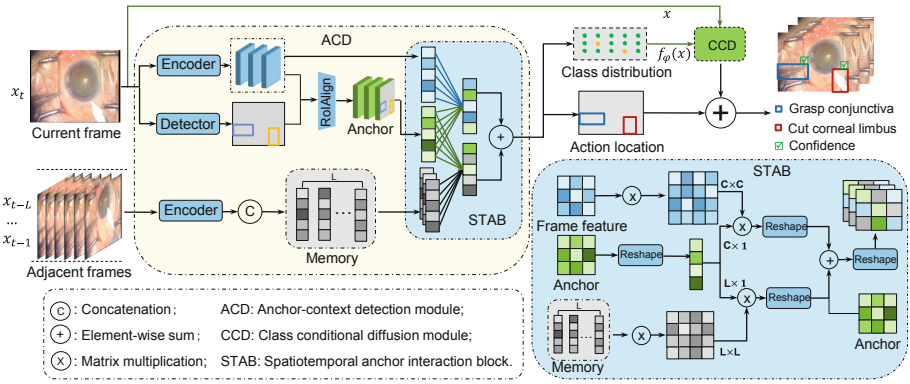

Fig. 1. The pipeline of our ACTNet includes ACD and CCD modules.

features as anchors), which includes surrounding tissues and movement information. Then, with the constraints of class distributions and the surgical videos, we propose a conditional diffusion model to cover the whole distribution of our data and to accurately reconstruct new predictions based on full learning. Furthermore, our class conditional diffusion model also accesses uncertainty for each prediction, through the stochasticity of outputs.

We summarize our main contributions as follows: 1) We develop an anchor-context action detection network (ACTNet), including an anchor-context detection (ACD) module and a class conditional diffusion (CCD) module, which combines three tasks: i) where actions locate; ii) what actions are; iii) how confident our model is about predictions. 2) For ACD module, we develop a spatio-temporal anchor interaction block (STAB) to spatially and temporally highlight the context related to the extracted anchor, which provides micro-action location and initial class. 3) By conditioning on the full distribution of action classes in the surgical videos, our proposed class conditional diffusion (CCD) model reconstructs better class prototypes in a stochastic fashion, to provide a more accurate estimations and push the assessment of the model confidence in its predictions. 4) We carry out comparison and ablation study experiments to demonstrate the effectiveness of our proposed algorithm based on cataract surgery.

2 Methodology

The overall framework of our proposed ACTNet for reliable action detection is illustrated in Fig. 1. Based on a video frame sequence, the ACD module extracts anchor features and aggregates the spatio-temporal interactions with anchor features by proposed STAB, which generates action locations and initial action class distributions. Then considering the full distribution of action classes in surgical videos, we use the CCD module to refine the action class predictions and access confidence estimations.

2.1 Our ACD Module

Anchor Extraction: Assuming a video X with T frames, denoted as $X = \{x_t\}_{t=1}^{T}$, where x_t is the t-th frame of the video. The task of this work is to estimate all potential locations and classes $P = \{box_n, c_n\}_{n=1}^{N}$ for action instances contained in video X, where box_n is the position of the n-th action happened, c_n is the action class of n-th action, and N is the number of action instances. For video representation, this work tries to encode the original videos into features based on the backbone ResNet50 [5] network to get each frame's feature $F = \{f_t\}_{t=1}^{T}$.

In surgical videos, the instruments, as action subjects, are significant to recognize the action. For instrument detection, it is very important but not very complicated. Existing excellent object detection method like Faster R-CNN [22] is enough to obtain results with high accuracy. After getting the detected instrument anchors, RoIAlign is applied to extract the instrument features from frame features. The instrument features are denoted as $I = \{i_t\}_{t=1}^{T}$. Since multiple instruments exist in surgeries, our action detection needs to solve the problem that related or disparate concurrent actions often lead to wrong predictions. Thus, in this paper, we propose to provide action location and class considering the spatio-temporal anchor interactions in the surgical videos, based on STAB.

Spatio-Temporal Action interaction Block (STAB): For several actions like pushing, pulling, and cutting, there is no difference just inferred from the local region in one frame. Thus we propose STAB to utilize spatial and temporal interactions with an anchor to improve the prediction accuracy of the action class, which finds actions with strong logical links to provide an accurate class. The structure of STAB is shown in Fig. 1. We introduce spatial and temporal interactions respectively in the following.

For spatial interaction: The instrument feature i_t acts as the anchor. In order to improve the anchor features, the module has the ability to select value features that are highly active with the anchor features and merge them. The formulation is defined as: $a_t = \frac{1}{C(f_t)} \sum_{j \in S_j} h(f_{tj}, i_t) g(f_{tj})$, where j is the index that enumerates all possible positions of f_t. A pairwise function $h(\cdot)$ computes the relationship such as affinity between i_t and all f_{tj}. In this work, dot-product is employed to compute the similarity. The unary function $g(f_{tj})$ computes a representation of the input signal at the position j. The response is normalized by a factor $C(f_t) = \sum_{j \in S_j} h(f_{tj}, i_t)$. S_j represents the set of all positions j. Through the formulation, the output a_t obtains more information from the positions related to the instrument and catches interactions in space for the actions.

For temporal interaction: We build memory features consisting of features in consecutive frames: $m_t = [f_{t-L}, ..., f_{t-1}]$. To effectively model temporal interactions of the anchor, the network offers a powerful tool for capturing the complex and dynamic dependencies that exist between elements in sequential data and anchors. Same with the spatial interaction, we take i_t as an anchor and calculate the interactions between the memory features and the anchor. The formulation

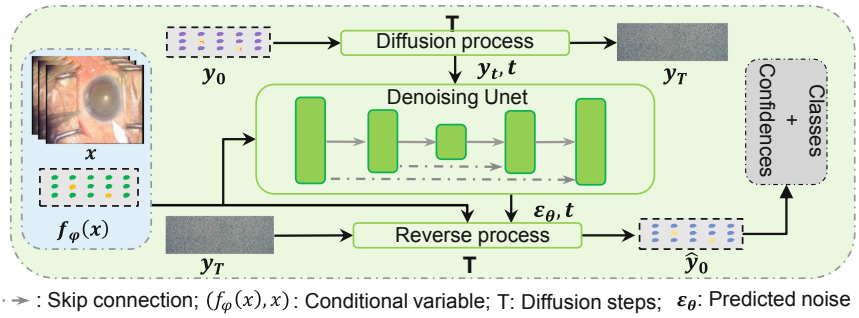

Fig. 2. Overview of the class conditional diffusion (CCD) model.

is defined as: $b_t = \frac{1}{C(m_t)} \sum_{j \in T_j} h(m_{tj}, i_t) g(m_{tj})$, where T_j refers to the set of all possible positions along the time series in the range of L. In this way, temporal interactions with anchors are obtained. Then a global average pooling is carried out on the spatial and temporal outputs. Action localizations and initial action class distributions are produced based on the fully-connected classifier layer.

2.2 CCD Module for Reliable Action Detection

Since the surgical procedures follow regularity, we propose a CCD module to reconstruct the action class predictions considering the full distribution of action classes in videos. The diffusion conditioned on the action classes and surgical videos is adopted in our paper. Let $y_0 \in \mathbb{R}^n$ be a sample from our data distribution. As shown in Fig. 2, a diffusion model specified in continuous time is a generative model with latent y_t, obeying a forward process $q_t(y_t|y_{t-1})$ starting at data y_0 [6]. y_0 indicates a one-hot encoded label vector. We treat each one-hot label as a class prototype, i.e., we assume a continuous data and state space, which enables us to keep the Gaussian diffusion model framework [2,6]. The forward process and reverse process of unconditional diffusion are provided in the supplementary material.

Here, for the diffusion model optimization can be better guided by meaningful information, we integrate the ACD and our surgical video data as priors or constraints in the diffusion training process. We design a conditional diffusion model $\hat{p}_\theta(y_{t-1}|y_t, x)$ that is conditioned on an additional latent variable x. Specifically, the model $\hat{p}_\theta(y_{t-1}|y_t, x)$ is built to approximate the corresponding tractable ground-truth denoising transition step $\hat{p}_t(y_{t-1}|y_t, y_0, x)$. We specify the reverse process with conditional distributions as [21]:

$$\hat{p}_t\left(y_{t-1}|y_t, y_0, x\right) = \hat{p}_t\left(y_{t-1}|y_t, y_0, f_\varphi(x)\right) = N\left(y_{t-1}; \hat{\mu}\left(y_t, y_0, f_\varphi(x)\right), \hat{\beta}_t I\right)$$

where $\hat{\mu}\left(y_t, y_0, f_\varphi(x)\right)$ and $\hat{\beta}_t$ are described in supplementary material. $f_\varphi(x)$ is the prior knowledge of the relation between x and y_0, i.e., the ACD module

pre-trained with our surgical video dataset. The x indicates the input surgical video frames. Since ground-truth step $\hat{p}_t(y_{t-1}|y_t, y_0, x)$ cannot get directly, the model $\hat{p}_\theta(y_{t-1}|y_t, x)$ are trained by following loss function for estimating ϵ_θ to approximate the ground truth:

$$\hat{L}(\theta) = E_{t,y_0,\epsilon,x}\left[\|\epsilon - \epsilon_\theta(x, \sqrt{\bar{\alpha}_t}y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon + (1-\sqrt{\bar{\alpha}_t})f_\varphi(x), f_\varphi(x), t)\|^2\right]$$

where $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$, $\epsilon \sim N(0,1)$ and $\epsilon_\theta(\cdot)$ estimates ϵ using a time-conditional network parameterized by θ. β_t is a constant hyperparameter.

To produce model confidence for each action instance, we mainly calculate the prediction interval width (IW). Specifically, we first sample N class prototype reconstruction with the trained diffusion model. Then calculate the IW between the 2.5^{th} and 97.5^{th} percentiles of the N reconstructed values for all test classes. Compared with traditional classifiers to get deterministic outputs, the denoising diffusion model is a preferable modelling choice due to its ability to produce stochastic outputs, which enables confidence generation.

3 Experimental Results

Cataract Surgical Video Dataset: To perform reliable action detection, we build a cataract surgical video dataset. Cataract surgery is a procedure to remove the lens of the eyes and, in most cases, replace it with an artificial lens. The dataset consists of 20 videos with a frame rate of 1 fps (a total of 17511 frames and 28426 action instances). Under the direction of ophthalmologists, each video is labelled frame by frame with the categories and locations of the actions. 49 types of action bounding boxes as well as class labels are included in our dataset. The surgical video dataset is randomly split into a training set with 15 videos (13583 frames) and a testing set with 5 videos (3928 frames).

Implementation Details: The proposed architecture is implemented using the publicly available Pytorch Library. A model with ResNet50 backbone from Faster R-CNN-benchmark [23] is adopted for our instrument anchor detection. In STAB, we use ten adjacent frames. During inference, detected anchor boxes with a confidence score larger than 0.8 are used. More implementation details are listed in the supplementary material. The performances are evaluated with official metric frame level mean average precision (mAP) at IoU = 0.1, 0.3, and 0.5, respectively, obtaining figures in the following named mAP_{10}, mAP_{30} and mAP_{50} with their mean mAP_{mean}.

Method Comparison: In order to demonstrate the superiority of the proposed method for surgical action detection, we carry out a comprehensive comparison between the proposed method and the following state-of-the-art methods: 1) single-stage algorithms, including the Single Shot Detector (SSD) [16], SSDLite [25] and RetinaNet [12]. 2) two-stage algorithms, including Faster R-CNN [23], Mask R-CNN [4], Dynamic R-CNN [29] and OA-MIL [15]. The data presented in Table 1 clearly demonstrate that our method outperforms other approaches,

Table 1. Methods comparison and ablation study on cataract video dataset.

Methods	mAP_{10}	mAP_{30}	mAP_{50}	mAP_{mean}
Faster R-CNN [23]	0.388	0.384	0.371	0.381
SSD [16]	0.360	0.358	0.350	0.356
RetinaNet [12]	0.358	0.356	0.347	0.354
Mask R-CNN [4]	0.375	0.373	0.363	0.370
SSDlite [25]	0.305	0.304	0.298	0.302
Dynamic R-CNN [29]	0.315	0.310	0.296	0.307
OA-MIL [15]	0.395	0.394	0.378	0.389
backbone	0.373	0.365	0.360	0.366
+temporal	0.385	0.378	0.372	0.378
+spatial	0.394	0.385	0.377	0.385
+STAB	0.400	0.393	0.385	0.393
+CCD (ACTNet)	**0.415**	**0.406**	**0.397**	**0.406**

irrespective of the IoU threshold being set to 0.1, 0.3, 0.5, or the average values. Notably, the results obtained after incorporating diffusion even surpass Faster R-CNN by 2.5% and baseline by 4.0% in terms of average mAP. This finding provides compelling evidence for the efficacy of our method in integrating spatio-temporal interactive information under the guidance of anchors and leveraging diffusion to optimize the category distribution. The quantitative results further corroborate the effectiveness of our approach in Fig. 3, which shows that our model does not only improve the performance of the baseline models but also localizes accurately the regions of interest of the actions. More results are listed in the material.

Ablation Study: To validate the effectiveness of our ACTNet, we have done some ablation studies. We train and test the model with spatial interaction, temporal interaction, spatio-temporal interaction (STAB), and finally together with our CCD model. The testing results are shown in Fig. 3 and Table 1. For our backbone, it is achieved by concatenating the anchor features through RoIAlign and the corresponding frame features to get the detected action classes.

The results reveal that the spatial and temporal interactions for instruments can provide useful information to detect the actions. What's more, spatial interaction has slightly better performance than temporal interaction. It may be led by the number of spatially related action categories being slightly more than that of temporally related action categories. It is worth noting that spatial interaction and temporal interaction can be enhanced by each other and achieve optimal performance. After being enhanced by the diffusion model conditioned on our obtained class distributions and video frames, we get optimal performance.

Confidence Analysis: To analyze the model confidence, we take the best prediction for each instance to calculate the instance accuracy. We can observe

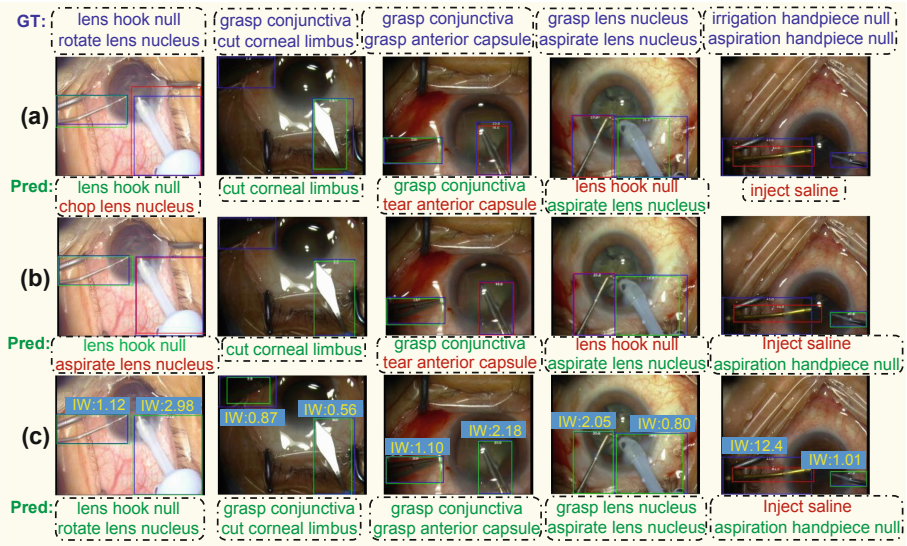

Fig. 3. Visualization on cataract dataset. We choose different actions to show the results of (a) Faster R-CNN, (b) OA_MIL, and (c) our ACTNet. For each example, we show the ground-truth (Blue), right predictions (Green) and wrong predictions (Red). The actions are labelled from left to right. IW values (multiplied by 100) mean prediction interval width to show the level of confidence. (Color figure online)

from Table 2 across the test set, the mean_IW of the class label among correctly classified instances by ACTNet is significantly narrower compared to that of incorrectly classified instances. This observation indicates that the model is more confident in its accurate predictions and is more likely to make errors when its predictions are vague. Furthermore, upon comparing the mean_IW at the true class level, we find that a more precise class tends to exhibit a larger disparity between the correct and incorrect predictions. Figure 3 also shows the

Table 2. The mean_IW (multiplied by 100) results from our CCD module.

Class	Instance	Accuracy	Mean_IW (Correct)	Mean_IW (Incorrect)
grasp conjunctiva	487	0.702	0.91 (342)	9.78 (145)
aspirate lens cortex	168	0.613	1.37 (103)	15.82 (65)
chop lens nucleus	652	0.607	0.54 (396)	9.73 (256)
polish intraocular lens	222	0.572	0.90 (127)	8.48 (95)
aspirate lens nucleus	621	0.554	0.76 (344)	10.30 (277)
inject viscoelastic	112	0.536	2.17 (60)	9.14 (52)
Remove lens cortex	174	0.471	0.42 (82)	5.84 (92)
forceps null	280	0.464	2.67 (130)	8.38 (150)

confidence estimations for some samples. We can see the correct prediction gets smaller IW values compared with the incorrect one (The rightmost figure in column (c)), which means it has more uncertainty for the incorrect prediction.

4 Conclusions

In this paper, we propose a conditional diffusion-based anchor-context spatio-temporal action detection network (ACTNet) to achieve recognition and localization of every occurring action in the surgical scenes. ACTNet improves the accuracy of the predicted action class from two considerations, including spatio-temporal interactions with anchors by the proposed STAB and full distribution of action classes by class conditional diffusion (CCD) module, which also provides uncertainty in surgical scenes. Experiments based on cataract surgery demonstrate the effectiveness of our method. Overall, the proposed ACTNet presents a promising avenue for improving the accuracy and reliability of action detection in surgical scenes.

Acknowledgement. This work was supported in part by General Program of National Natural Science Foundation of China (82272086 and 82102189), Guangdong Basic and Applied Basic Research Foundation (2021A1515012195), Shenzhen Stable Support Plan Program (20220815111736001 and 20200925174052004), and Agency for Science, Technology and Research (A*STAR) Advanced Manufacturing and Engineering (AME) Programmatic Fund (A20H4b0141) and Central Research Fund (CRF).

References

1. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059. PMLR (2016)
2. Han, X., Zheng, H., Zhou, M.: CARD: classification and regression diffusion models. arXiv preprint arXiv:2206.07275 (2022)
3. Hashimoto, D.A., et al.: Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. Ann. Surg. **270**(3), 414 (2019)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
7. Islam, M., Seenivasan, L., Ming, L.C., Ren, H.: Learning and reasoning with the graph structure representation in robotic surgery. In: Martel, A.L., et al. (eds.) MICCAI 2020, Part III. LNCS, vol. 12263, pp. 627–636. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_60
8. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

9. Lalys, F., Jannin, P.: Surgical process modelling: a review. Int. J. Comput. Assist. Radiol. Surg. **9**, 495–511 (2014). https://doi.org/10.1007/s11548-013-0940-5
10. Lee, Y., et al.: Localization uncertainty estimation for anchor-free object detection. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022, Part VIII. LNCS, vol. 13808, pp. 27–42. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25085-9_2
11. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. **7**(1), 1–14 (2017)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
13. Lin, W., et al.: Instrument-tissue interaction quintuple detection in surgery videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 399–409. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_38
14. Linegang, M.P., et al.: Human-automation collaboration in dynamic mission planning: a challenge requiring an ecological approach. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 2482–2486. SAGE Publications Sage, Los Angeles (2006)
15. Liu, C., Wang, K., Lu, H., Cao, Z., Zhang, Z.: Robust object detection with inaccurate bounding boxes. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022, Part X. LNCS, vol. 13670, pp. 53–69. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20080-9_4
16. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part I. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
17. Mersh, A.T., Melesse, D.Y., Chekol, W.B.: A clinical perspective study on the compliance of surgical safety checklist in all surgical procedures done in operation theatres, in a teaching hospital, Ethiopia, 2021: a clinical perspective study. Ann. Med. Surg. **69**, 102702 (2021)
18. Nepogodiev, D., et al.: Global burden of postoperative death. The Lancet **393**(10170), 401 (2019)
19. Nwoye, C.I., et al.: Rendezvous: attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med. Image Anal. **78**, 102433 (2022)
20. Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minim. Invasive Ther. Allied Technol. **28**(2), 82–90 (2019)
21. Pandey, K., Mukherjee, A., Rai, P., Kumar, A.: DiffuseVAE: efficient, controllable and high-fidelity generation from low-dimensional latents. arXiv preprint arXiv:2201.00308 (2022)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: 2022 IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685 (2022)

25. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
26. Seenivasan, L., Mitheran, S., Islam, M., Ren, H.: Global-reasoned multi-task learning model for surgical scene understanding. IEEE Robot. Autom. Lett. **7**(2), 3858–3865 (2022). https://doi.org/10.1109/LRA.2022.3146544
27. 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 Trans. Med. Imaging **36**(1), 86–97 (2016)
28. Vercauteren, T., Unberath, M., Padoy, N., Navab, N.: CAI4CAI: the rise of contextual artificial intelligence in computer-assisted interventions. Proc. IEEE **108**(1), 198–214 (2019)
29. Zhang, H., Chang, H., Ma, B., Wang, N., Chen, X.: Dynamic R-CNN: towards high quality object detection via dynamic training. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part XV. LNCS, vol. 12360, pp. 260–275. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_16
30. Zhang, J., et al.: Automatic keyframe detection for critical actions from the experience of expert surgeons. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 8049–8056. IEEE (2022)