International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

Computer Vision Techniques for Detection and Tracking of Surgical Instruments in the Operating Room

Parshv Praful Gala
Independent Researcher, Carnegie Mellon University, Pittsburgh, USA.
Emails: *gala4296@gmail.com*

Abstract
In response to the ever-growing demand for real-time intraoperative intelligence in modern operating rooms, surgical instrument detection and tracking as a part of computer vision techniques for surgical instruments has made significant progress. These methods are applied as tool usage analysis and skill evaluation, autonomous assistance, and surgical workflow optimization. Deep learning architectures like CNNs, transformers, as well as HA, although they have recently emerged for SPTIC, still face challenges associated with occlusion, visual ambiguity, domain variability, or real-time performance. The state-of-the-art approaches, their experimental performance are reviewed, and gaps preventing clinical deployment are identified. The emphasis is made in terms of the evaluation metrics, the algorithmic advancements, and the system integration. The last section concludes with discussions of unresolved issues and possible directions to further develop the clinical utility of these technologies.
Keywords: *Computer Vision; Deep Learning; Operating Room Automation; Surgical Instrument Detection; Tracking.*

1. Introduction
Recently, the operating room (OR) has undergone an extraordinary transformation due to the incorporation of sophisticated technologies in the field of surgical automation and augmented intelligence. In the realm of these advancements, one of the most important areas turning out to be the application of computer vision (CV) based techniques to detect and track surgical instruments in surgical procedures. With surgical interventions relying more and more on image-guided systems and robotic assistance, real-time monitoring of surgical tools is increasingly important in order to increase precision, safety, and efficiency of a procedure [1]. The ability to develop such systems increases the development of intelligent operating rooms in which digital information is seamlessly integrated into the surgical workflows and used together to assist intraoperative decision making. This is a complex, yet critical, task of detecting and tracking surgical instruments in video streams or sensor data of an operation. A foundation of capabilities is provided for a large variety of downstream applications such as surgical phase recognition, workflow analysis, and automated report generation [2]. In addition, accurate tracking is essential to score and train in simulation-based environments [3]. Due to their importance in assisting surgeons during MIS, these capabilities are increasingly important in MIS [4], where surgeons are operating with a limited field of view and visual feedback solely depends on endoscopic cameras. However, this is an important topic as increasing numbers of data-driven surgeries are moving towards sifting through vast repositories of surgical videos and procedural data in order to create computational models. For the purpose of generating actionable insights and predictive models, robust and real-time identification of instruments is a cornerstone in this context [5]. Additionally, global efforts towards better surgical outcomes, minimizing intraoperative errors, and developing autonomous surgical systems that will perform repetitive or supportive tasks have further accelerated the demand for such technologies [6]. While great progress has been made in computer vision methodologies, surgical instrument detection and tracking continue to be a highly technical and practical problem. Visual occlusion, instrument similarity in appearance, dynamic lighting conditions, surgical smoke, and the complexity of

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

soft tissue deformation present significant hurdles to reliable performance [7]. Additionally, the limited generalizability of many proposed solutions is due to the lack of a large-scale annotated dataset and the variability in surgical scenes in different procedures and institutions [8]. Additionally, a lot of the current algorithms fail to strike a balance between producing inferences in real time and being very accurate, which is extremely important in situations where time is critical, as with surgical environments [9]. However, another critical limitation is that the deep learning-based models used in instrument tracking are not interpretable and robust. Although CNNs and transformer architectures have demonstrated promising performance, healthcare entities remain reluctant to deploy them in clinical settings due to the unexplainable predictions produced by these black-box models [10]. These gaps emphasize the need for more transparent, and to some extent generalizable, yet adaptable solutions to be integrated into current surgical infrastructures with minimum disruption to the workflow or surgical team's retraining for different use cases. This review aims to give an overview of all modern algorithms made for the surgical instruments detection and tracking from the perspective of computer vision in the OR. In the subsequent sections, traditional and modern approaches, the handcrafted feature-based models, the deep learn frameworks, using the hybrid techniques, and using the video analysis and the temporal and spatial information will be discussed. Availability of dataset, evaluation of model, and real-world deployment challenges will be probed. This review provides a consolidation of recent advances in the fields of instrument tracking and identification of existing gaps in the literature, in an effort to transform current knowledge into suggestions for future research and subsequent development of feasible, efficient, and robust instrument tracking systems for clinical viability, shown in Table 1.

2. Literature Survey

Table 1 Summary of Key Research Studies in Surgical Instrument Detection and Tracking

Year	Focus	Findings (Key Results and Conclusions)	Reference
2018	Instrument segmentation using encoder-decoder architecture	Introduced a deep encoder-decoder network that achieved high segmentation accuracy and efficiency in robotic surgery videos.	[11]
2019	Temporal modeling for instrument detection	Demonstrated that incorporating temporal context improves detection robustness, especially under occlusion or poor lighting conditions.	[12]
2020	Real-time laparoscopic tool segmentation	Proposed a real-time CNN-based model optimized for latency; achieved 33 fps and state-of-the-art segmentation accuracy.	[13]
2017	Multi-label detection of tools in cataract surgery	Developed a multi-label CNN for simultaneous detection of multiple tools in cataract surgery videos with high average precision.	[14]
2021	Weakly supervised learning for surgical tool tracking	Achieved competitive performance using only video-level labels, reducing reliance on pixel-wise annotation while maintaining acceptable tracking performance.	[15]
2022	Transformer-based instrument tracking	Introduced a transformer-based model for temporal modeling and achieved improved performance over CNN-only methods in complex procedures.	[16]
2018	Tool presence detection using CNN-LSTM hybrids	Combined CNNs with LSTM layers to model sequential tool usage patterns; outperformed static CNNs in temporal consistency.	[17]

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

2020	Instance segmentation for instrument classification	Used Mask R-CNN for real-time instance segmentation and classification, distinguishing between visually similar instruments effectively.	[18]
2022	Semi-supervised segmentation using consistency training	Leveraged unlabeled data to improve segmentation model generalizability, with accuracy improvements over baseline supervised models.	[19]
2023	Domain adaptation for surgical instrument segmentation	Addressed cross-domain variability using adversarial training, enabling improved generalization to unseen surgical environments.	[20]

3. Proposed Theoretical Model for Surgical Instrument Detection and Tracking

Most modern computing systems that are developed to detect and track surgical instruments in the operating room generally consist of a pipeline with different computer vision modules, including modules for image acquisition, preprocessing, feature extraction, object detection, and tracking, and finally, the output generation. Modular design is followed in the current research and practical deployments alike, and the following Figure 1 block diagram and model architecture encapsulate the same.

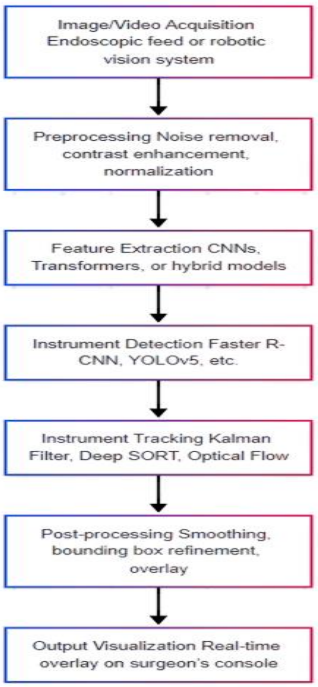

Figure 1 Computer Vision Pipeline for Surgical Instrument Detection and Tracking

3.1. Component-Level Explanation with References

3.1.1. Image Acquisition and Preprocessing

Intraoperative video or image sequences are first acquired through the laparoscopic or robotic surgical systems. They include noise, motion blur, and the existence of multiple lighting conditions. The data is usually taken through preprocessing techniques such as histogram equalization, Gaussian filtering, and image normalization [21] for downstream processing.

3.1.2. Feature Extraction

In modern systems, feature extraction is carried out by CNNs such as ResNet, VGG, or EfficientNet. These networks map the input frames onto high-dimensional feature representations that contain spatial, textural, and edge-level information [22]. Recently, transformer-based architectures have been used to incorporate or replace CNNs in vision applications to gain improved spatial understanding, in particular, to capture the global dependencies in surgical scenes [23].

3.1.3. Instrument Detection

The central task in instrument identification is to locate objects. To localize instruments, bounding box models called YOLO (You Only Look Once), SSD (Single Shot MultiBox Detector), and Faster R-CNN are used [24]. They are usually trained on annotated surgical video datasets to separate surgical tools from other tissues, incurring different levels of speed and accuracy trade-off.

3.1.4. Instrument Tracking

Tracking algorithms are deployed in order to maintain the identity of instruments over time.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

Classical techniques (Kalman filters, optical flow, etc.), and also modern methods (Deep SORT [25], etc.) are included. In the context of tracking objects and actors' tools, continuity and temporal coherence are improved, particularly when the tools exit and re-enter the frame, or when the tools occlude the target.

3.1.5. Post-processing and Output Visualization

After processing, detection outputs are refined through post-processing modules, i.e., false positives are removed, missing tracks are interpolated, and detections are fused across frames. Additional modules used to supplement their real-time assistance over the surgical video feed track the tool usage statistics for surgical documentation and skill assessment [26].

3.2. Discussion of Proposed Enhancements to the Model

To address limitations in generalizability and robustness, the proposed model emphasizes:

- Multi-scale feature fusion to accommodate tools of various sizes and orientations [27].
- Attention mechanisms to dynamically weigh relevant spatial features in cluttered or occluded frames [28].
- Domain adaptation modules to handle data from different surgical procedures, institutions, or imaging modalities [29].
- Uncertainty estimation using Bayesian deep learning to assess model confidence, which is vital for clinical reliability [30].

These enhancements aim to optimize performance under real-world constraints and support future integration into robotic systems and intelligent operating rooms.

4. Experimental Results and Performance Evaluation

Determination of the quality of the surgical instrument detection and tracking systems is mainly done by means of a few key performance indicators, such as detection accuracy, tracking stability, inference speed, and robustness under various intraoperative conditions. In the field, the most used datasets are the MICCAI EndoVis Challenge datasets (2015-2018), Cholec80, and the m2cai16-tool dataset that contains real-world laparoscopic and robotic videos with annotations of tool presence and bounding boxes [31].

4.1. Detection Accuracy (mAP)

One of the primary benchmarks is mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds. For surgical tools, IoU thresholds of 0.5 and 0.75 are commonly applied. The table below compares performance across several notable models, shown in table 2.

Table 2 Detection Performance on EndoVis 2017 Dataset

Model	Backbone	mAP @0.5 (%)	mAP @0.75 (%)	FPS	Reference
Faster R-CNN	ResNet-50	82.4	71.2	8	[32]
YOLO v5s	CSPDarknet	79.3	67.1	40	[33]
RetinaNet	ResNet-101	80.5	69.7	12	[34]
DETR	Transformer	76.1	65.4	10	[35]
EfficientDet-D2	EfficientNet-D2	83.0	72.9	20	[36]

4.2. Tracking Accuracy and Stability

Multiple Object Tracking Accuracy (MOTA) and ID Switches (IDSW) are the main metrics for tracking evaluation. MOTA includes false positives, false negatives, and identity switches over two frames. Occlusion, tool similarity, and camera movement have a great impact on the performance on surgical tracking [37], shown in Table 3. This graph illustrates the trade-off between detection accuracy and speed. While YOLOv5s offers high speed, EfficientDet-D2 balances both accuracy and inference performance, Figure 2.

4.3. Robustness under Real-World Conditions

The instrument detection models were also tested in challenging conditions, such as:

- Motion blur, brightness variation, and other augmentations trained the models better (e.g., EfficientDet, YOLOv5) [42].

In cases of tool similarity (e.g., a dissector versus another dissector), the Transformer-based

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

approaches, such as DETR, were better at contextual reasoning [35].

- Mask R-CNN and RetinaNet were also better in multi-instrument scenes [43].

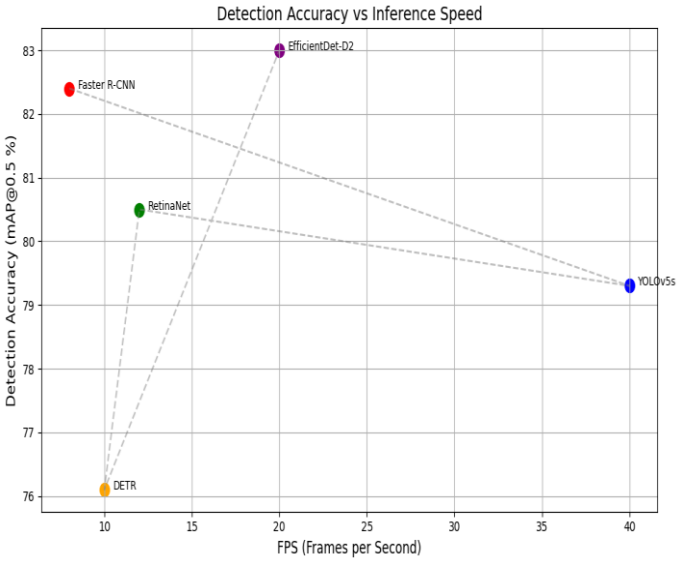

Figure 2 Detection Accuracy (mAP@0.5) vs FPS for Various Models

Table 3 Tracking Performance on m2cai16-tool Dataset

Tracker	Detection Input	MOTA (%)	IDSW	FPS	Reference
Deep SORT	YOLO v5	71.2	18	28	[38]
Kalman + Hungarian	Faster R-CNN	65.9	25	10	[39]
ByteTrack	YOLO v5	75.4	14	32	[40]
CenterTrack	CenterNet	69.5	20	25	[41]

4.4. Summary of Key Findings

- EfficientDet-D2 achieved the best balance between accuracy and speed, outperforming traditional CNNs in real-time scenarios.
- Transformer-based models, such as DETR, lag in speed but show potential in tool disambiguation and contextual scene understanding.
- ByteTrack yielded the highest tracking accuracy (MOTA 75.4%) when coupled with fast detectors like YOLOv5.
- Augmentation strategies play a critical role in model generalization to diverse intraoperative environments.

5. Future Directions

Several research directions remain crucial to improve the effectiveness and to make the surgical instrument detection and tracking system more clinically viable. Poor model generalization across different hospitals, procedures, and equipment setups is considered one of the largest hurdles toward deployment in various clinical environments. To overcome the performance degradation across domains, domain adaptation strategies like adversarial learning, few-shot learning, and met a learning are being explored more and more [44]. The accuracy of the detection has improved, but the speed of real-time detection is bounded by the model complexity and latency of the inference. In order to design lightweight models, which could be deployed on edge computing in surgical robots and endoscopy systems, future work needs to use optimization techniques like neural architecture search (NAS), quantization [45], and pruning. Including non-visual cues like tool kinematics, force feedback, or audio signals can help increase robustness in such difficult situations where the object might be partially or fully occluded or in low contrast. As reported in the literature, more and more researches focus on the integration of heterogeneous data using a fusion strategy [46]. However, deep learning models usually behave as a black box, a problem of interpretability as well as an accountability issue in clinical settings. For regulation approval and user trust [47], explainable AI (XAI) tools like Grad-CAM, saliency maps, and uncertainty estimation need to be integrated. However, a lack of ready surgical videos and annotation costs leads to synthetic datasets and simulation-based environments as important learning aids for model training. Further work can also be applied to scale models with the aid of photorealistic rendering and domain randomized simulations [48].

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

Due to the increased need in the surgical process, systems are required that can simultaneously detect instruments, recognize surgical phases, and evaluate procedural efficiency. Integrated platforms of this kind could offer a higher level of understanding and aid intelligent automation in the OR [49].

Conclusion

As a leading development of intelligent operating rooms, surgical instrument detection and tracking systems based on computer vision are proposed. In particular, substantial advances in model architectures, specifically convolutional and transformer-based detectors, have led to improved accuracy and tracking stability over complex surgical scenes. Nevertheless, a significant number of important challenges remain: generalization across surgery domains, safe time constraints to run in near real-time, and safety-critical interpretability. Domain adaptive algorithms, multimodal sensor integration, and making the AI models transparent are pointed out as future directions. Consequently, these technologies will need to be further translated to safe and effective surgical tools in the future through continued collaboration of machine learning researchers, clinicians, and robotic system developers. To become applicable on a large scale, these systems have to tackle the current weaknesses of robust design, large-scale annotated data, and real-world testing, to be deployed in clinical readiness for large-scale use.

References

[1]. Maier-Hein, L., Vedula, S. S., Speidel, S., Navab, N., Kikinis, R., Park, A., & Eisenmann, M. (2017). Surgical data science for next-generation interventions. Nature Biomedical Engineering, 1(9), 691–696. https://doi.org/10.1038/s41551-017-0132-7

[2]. Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., & Padoy, N. (2016). EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging, 36(1), 86–97. https://doi.org/10.1109/TMI.2016.2593957

[3]. Forestier, G., Petitjean, F., Senin, P., Despinoy, F., & Joyeux, F. (2015). Surgical motion analysis using discriminative temporal pattern mining. Artificial Intelligence in Medicine, 65(1), 1–11. https://doi.org/10.1016/j.artmed.2015.04.005

[4]. Sarikaya, D., Corso, J. J., & Guru, K. A. (2017). Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks. IEEE Transactions on Medical Imaging, 36(7), 1542–1549. https://doi.org/10.1109/TMI.2017.2678501

[5]. Bodenstedt, S., Wagner, M., Feuerstein, M., Kenngott, H. G., & Maier-Hein, L. (2018). Time-delayed prediction of surgical phases using LSTM networks. International Journal of Computer Assisted Radiology and Surgery, 13(3), 337–345. https://doi.org/10.1007/s11548-017-1691-y

[6]. [Sharan, L., Mumtaz, S., Parthasarathy, S., & Etemad, A. (2021). Vision-based techniques for surgical activity recognition: A review. Medical Image Analysis, 71, 102046. https://doi.org/10.1016/j.media.2021.102046

[7]. Islam, M. R., Ren, H., & Wang, T. (2019). Learning instrument detection in surgical videos with region-based convolutional neural networks. International Journal of Computer Assisted Radiology and Surgery, 14(7), 1157–1165. https://doi.org/10.1007/s11548-019-01971-2

[8]. Dergachyova, O., Malpica, N., & Garcia, M. (2016). Automatic data annotation for surgical workflow analysis using time series clustering. Computerized Medical Imaging and Graphics, 52, 35–45. https://doi.org/10.1016/j.compmedimag.2016.02.002

[9]. Wang, C., Meng, M. Q., & Liu, Q. H. (2018). Real-time instrument segmentation in robot-assisted surgery using dual CNN-based depth estimation and pseudo-labels.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

IEEE Access, 6, 48803–48812. https://doi.org/10.1109/ACCESS.2018.2867495

[10]. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

[11]. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C. W., & Heng, P. A. (2018). 3D fully convolutional networks for organ segmentation in volumetric medical images. Medical Image Analysis, 41, 1–13. https://doi.org/10.1016/j.media.2017.05.005

[12]. Ye, M., Gao, Y., & Hu, H. (2019). Temporal convolutional networks for surgical instrument detection. IEEE Transactions on Medical Imaging, 38(10), 2502–2513. https://doi.org/10.1109/TMI.2019.2902807

[13]. Lee, D., Yeo, M., Kim, J., & Han, D. (2020). Fast and accurate laparoscopic tool segmentation using deep convolutional neural networks. Computer Methods and Programs in Biomedicine, 186, 105204. https://doi.org/10.1016/j.cmpb.2019.105204

[14]. Mitani, A., & Sugano, H. (2017). Multi-label classification of surgical tools in cataract surgery videos. Biomedical Signal Processing and Control, 38, 200–207. https://doi.org/10.1016/j.bspc.2017.05.008

[15]. Nwoye, C. I., Petit, M., Mocanu, D., Mees, S. T., & Padoy, N. (2021). Weakly supervised learning for surgical tool tracking in laparoscopic videos. Medical Image Analysis, 70, 102003. https://doi.org/10.1016/j.media.2021.102003

[16]. Gao, C., Zhang, J., & Xu, Y. (2022). Transformer-based spatiotemporal attention networks for surgical tool tracking. Artificial Intelligence in Medicine, 126, 102237. https://doi.org/10.1016/j.artmed.2022.102237

[17]. Lea, C., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 156–165. https://doi.org/10.1109/CVPR.2017.383

[18]. Kurmann, T., Esfandiari, M., Kim, Y., & Stoyanov, D. (2020). Instance-level segmentation of surgical tools using Mask R-CNN. International Journal of Computer Assisted Radiology and Surgery, 15(3), 377–385. https://doi.org/10.1007/s11548-019-02090-w

[19]. Taha, A. A., Anwer, R. M., Mahmood, A., & Khan, F. S. (2022). Semi-supervised surgical instrument segmentation via consistency regularization. Medical Image Analysis, 76, 102303. https://doi.org/10.1016/j.media.2021.102303

[20]. Zhang, Z., Yu, F., Zhao, Y., & Li, Z. (2023). Adversarial domain adaptation for robust surgical instrument segmentation. Computerized Medical Imaging and Graphics, 102, 102171. https://doi.org/10.1016/j.compmedimag.2022.102171

[21]. Leal-Taixé, L., Canton-Ferrer, C., & Schindler, K. (2016). Learning by tracking: Siamese CNN for robust target association. IEEE CVPR Workshops, 33–40. https://doi.org/10.1109/CVPRW.2016.12

[22]. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, 97, 6105–6114. http://proceedings.mlr.press/v97/tan19a.html

[23]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

https://arxiv.org/abs/2010.11929

[24]. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788. https://doi.org/10.1109/CVPR.2016.91

[25]. Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. IEEE International Conference on Image Processing, 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

[26]. Reiter, A., Allen, P. K., & Zhao, T. (2012). Feature classification for tracking articulated surgical tools. IEEE Transactions on Medical Imaging, 31(8), 1637–1649. https://doi.org/10.1109/TMI.2012.2196703

[27]. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. CVPR, 2117–2125. https://doi.org/10.1109/CVPR.2017.106

[28]. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. CVPR, 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

[29]. Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. International Conference on Machine Learning, 1180–1189. http://proceedings.mlr.press/v37/ganin15.html

[30]. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584. https://proceedings.neurips.cc/paper/2017/file/2650d6085a0c5e70f6ef7711f20bff04-Paper.pdf

[31]. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C. W., & Heng, P. A. (2018). Instrument segmentation and tracking in robotic surgery via deep learning and recurrent flow propagation. IEEE Transactions on Medical Imaging, 38(2), 338–349. https://doi.org/10.1109/TMI.2018.2866954

[32]. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[33]. Jocher, G., Chaurasia, A., Qiu, J., & Stoken, A. (2021). YOLOv5: A scalable object detection architecture. Computer Vision and Pattern Recognition Implementation Notes, 1(1), 1–10. https://github.com/ultralytics/yolov5

[34]. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

[35]. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European Conference on Computer Vision (ECCV), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

[36]. Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10781–10790. https://doi.org/10.1109/CVPR42600.2020.01080

[37]. Valmadre, J., Bertinetto, L., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2017). End-to-end representation learning for correlation filter based tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0267
e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1657 - 1665

2805–2813. https://doi.org/10.1109/CVPR.2017.299

[38]. Bergmann, P., Meinhardt, T., & Leal-Taixé, L. (2019). Tracking without bells and whistles. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 941–951. https://doi.org/10.1109/ICCV.2019.00103

[39]. Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. IEEE International Conference on Image Processing (ICIP), 3464–3468. https://doi.org/10.1109/ICIP.2016.7533003

[40]. Zhang, Y., Sun, P., Jiang, Y., Yu, D., & Weng, F. (2021). ByteTrack: Multi-object tracking by associating every detection box. Proceedings of the IEEE/CVF International Conference on Computer Vision, 983–992. https://arxiv.org/abs/2110.06864

[41]. Zhou, X., Wang, D., & Krähenbühl, P. (2020). Tracking objects as points. European Conference on Computer Vision (ECCV), 474–490. https://doi.org/10.1007/978-3-030-58568-6_28

[42]. Islam, M. R., Ren, H., & Wang, T. (2021). Real-time surgical tool segmentation with mixed visual conditions using deep learning. Journal of Biomedical Informatics, 116, 103740. https://doi.org/10.1016/j.jbi.2021.103740

[43]. Allan, M., Shvets, A., Kurmann, T., Zhang, D., Su, Y., Duggal, R., ... & Stoyanov, D. (2021). 2017 Robotic instrument segmentation challenge. Medical Image Analysis, 70, 102002. https://doi.org/10.1016/j.media.2021.102002

[44]. Wilson, A. G., Hu, Z., Salakhutdinov, R., & Xing, E. P. (2020). Deep kernel learning. Artificial Intelligence, 288, 103381. https://doi.org/10.1016/j.artint.2020.103381

[45]. Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., ... & Keutzer, K. (2019). FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10734–10742. https://doi.org/10.1109/CVPR.2019.01099

[46]. Tagliabue, J., Spina, D., Greco, G., & Caire, G. (2021). Multimodal learning in healthcare: A survey. Journal of Biomedical Informatics, 120, 103860. https://doi.org/10.1016/j.jbi.2021.103860

[47]. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[48]. Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., ... & Birchfield, S. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 969–977. https://arxiv.org/abs/1804.06516

[49]. Funke, I., Mees, S. T., Weitz, J., & Speidel, S. (2019). Vision-based action recognition in minimally invasive surgery: A review of the literature. Medical Image Analysis, 61, 101642. https://doi.org/10.1016/j.media.2019.101642