L-Spot: Slice-Aware Consensus for Hepatocellular Carcinoma Localisation in Contrast-Enhanced CT

Julia Rebecca Kampel[1], Tripti Singh[2], Farhan Akram[3], Anne Christina Melissa Jegiraj[1], Shriya Bommaraveni[4], Byron John Morris[1], Akshat Tiwari[1], Sheetal Santosh Deshpande[2], Khawaja Shahabuddin[5], Domenec Puig[2], and Vivek Kumar Singh[1]

[1] Barts Cancer Institute, Queen Mary University of London, London, UK
[2] Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, Spain
[3] Department of Pathology and Clinical Bioinformatics, Erasmus MC, Rotterdam, The Netherlands
[4] Queen's University Belfast, Northern Ireland, UK
[5] Barts Health NHS Trust, London, UK

Abstract. Automated localisation of hepatocellular carcinoma (HCC) on contrast-enhanced CT scans is difficult when lesions are small, weakly enhancing, or obscured by heterogeneous liver parenchyma. Although 2D detectors provide efficient tumour localisation, analysing each axial slice independently can produce inconsistent predictions across a CT volume. We address this limitation with L-Spot, a slice-aware consensus framework that refines detector outputs by combining complementary YOLO predictions with spatial agreement across neighbouring slices. The framework uses two trained detectors: YOLOv8m, evaluated under both standard and augmented inference, and an independently trained YOLOv10m model. Predicted lesions are retained only when multiple prediction sources agree at anatomically consistent locations. We evaluated L-Spot on arterial-phase CT scans from 94 HCC patients and found that it improved the precision–recall balance over the strongest single-detector baseline by suppressing isolated false positives while preserving true positive tumour detections. L-spot achieved an F1-score of 0.733 (precision: 0.820, recall: 0.664), representing an 11.8% improvement over the baseline YOLOv8m. These results demonstrate that detector consensus and inter-slice continuity can improve the robustness of HCC localisation without altering the underlying detection architectures.

Keywords: HCC · Contrast CT · Object detection · Deep learning

1 Introduction

Hepatocellular carcinoma (HCC) is a major contributor to cancer-related morbidity and mortality worldwide as the most common primary liver malignancy [5, 3]. Accurate detection and localisation of HCC are essential for diagnosis, staging, treatment selection, locoregional therapy planning, surgical assessment, and

longitudinal follow-up [15]. Contrast-enhanced computed tomography (CECT) plays a central role in evaluating HCC by capturing lesion morphology and vascular enhancement characteristics across different imaging phases [3],[15]. Despite its clinical importance, automated HCC detection in CECT remains challenging due to heterogeneous tumour appearance, variable enhancement patterns, irregular boundaries, small lesion size, and low contrast relative to the surrounding liver parenchyma [9, 2, 6]. These challenges are further amplified in patients with cirrhosis, fibrosis, or vascular abnormalities, where the visual distinction between tumor and non-tumor tissue can be subtle [4, 3].

Deep learning has substantially advanced automated lesion detection in medical imaging by enabling data-driven extraction of complex visual patterns from radiological images [9, 19]. Object detection networks are particularly relevant for tumour localisation because they provide both classification and spatial localisation, typically in the form of bounding boxes [14]. Among these methods, the You Only Look Once (YOLO) family [16] of detectors has gained increasing attention due to its efficient single-stage detection design, favourable inference speed, and strong localisation performance [13, 18]. These properties make YOLO-based models attractive for CT-based tumour detection, where computational efficiency and spatial accuracy are both important [7]. However, most YOLO detectors operate on two-dimensional (2D) images and therefore process CT volumes slice by slice. This formulation does not fully exploit the three-dimensional anatomical continuity of CT data and may lead to unstable predictions across neighbouring slices [8, 2]. Specifically, a detector may generate isolated false-positive boxes on individual slices, whereas true hepatic tumours usually persist over multiple consecutive slices at anatomically consistent locations [2, 6, 10]. Improving the robustness of slice-wise detection is therefore an important requirement for reliable CT-based tumour localisation, especially in complex areas like the liver [9, 3]. Several strategies can be used to increase prediction stability, including test-time augmentation (TTA), model ensembling, and spatial post-processing [11, 12]. TTA involves evaluating transformed versions of the same image, such as flipped or scaled inputs, which can reduce sensitivity to image orientation, scale variation, and local perturbations [12, 23]. Model ensembling or multi-detector inference can further improve robustness by combining predictions from independently trained or architecturally distinct models [8, 11]. Nevertheless, simply aggregating outputs from multiple detectors may also introduce additional false positives, particularly when uncertain or inconsistent predictions are retained [2, 8]. For volumetric medical imaging, a more clinically meaningful strategy should combine complementary detector evidence while also respecting the anatomical principle that real lesions maintain spatial continuity across adjacent slices [2, 6, 10]. This suggests the potential of developing post-processing techniques that enforce volumetric consistency to refine initial 2D detections. Volumetric approaches address the same problem at the architectural level. Self-configuring 3D frameworks such as nnDetection [1] apply 3D convolutions to CT sub-volumes, 2.5D operators such as A3D [22] fuse features across adjacent slices, and multitask detectors such as MULAN [21]

combine 3D feature fusion with joint detection and segmentation. These methods exploit inter-slice context during training and require full volumetric supervision and retraining. We instead retain a 2D detector and recover volumetric consistency at inference time. This choice suits small clinical cohorts since added capacity is difficult to train reliably, and allows the method to be applied to deployed 2D pipelines without architectural retraining.

In this work, we propose L-Spot, a slice-aware consensus framework that incorporates volumetric anatomical consistency into slice-wise HCC localisation. Unlike existing approaches that seek performance gains through architectural modifications, increased model capacity, or complex training strategies, L-Spot improves prediction reliability by refining detector outputs entirely at inference time. Specifically, the proposed framework integrates complementary predictions from independently trained detectors with inter-slice spatial coherence, enabling the robust suppression of isolated false-positive detections while preserving anatomically consistent tumour localisations. A further key contribution is demonstrating that inference-time consensus provides a practical and effective alternative to increasing model complexity, achieving consistent improvements over strong YOLO baselines without modifying the underlying detector architectures or their training procedures. Ultimately, this work demonstrates that anatomical continuity can strengthen slice-wise detection in volumetric CT, providing a practical route to more reliable tumour localisation without increasing detector complexity.

2 Methods

2.1 Dataset and Pre-processing

We retrospectively included 94 patients diagnosed with primary HCC, obtained from a multi-phase 3D CECT collection [24]. Arterial-phase volumes were selected because this phase provides strong tumour-to-liver contrast and is clinically pivotal for HCC visualisation. Each 3D volume was sliced into an ordered sequence of 2D axial slices with a spatial resolution of 512×512 pixels. The original colour-coded segmentation masks represented the liver parenchyma, tumour, and background using green, red, and black labels, respectively. Tumour masks were converted into YOLO-format bounding-box annotations by detecting red-channel tumour pixels on each slice. For each tumour-containing slice, we generated the minimum enclosing axis-aligned bounding box and normalised the coordinates as $(x_{\text{centre}}, y_{\text{centre}}, w, h) \in [0, 1]$. Slices without tumour annotations were retained as negative examples to mitigate false-positive detections. The patient-level 70/10/20 split yielded 3,686 training slices (65 patients), 390 validation slices (10 patients), and 1,159 test slices (19 patients).

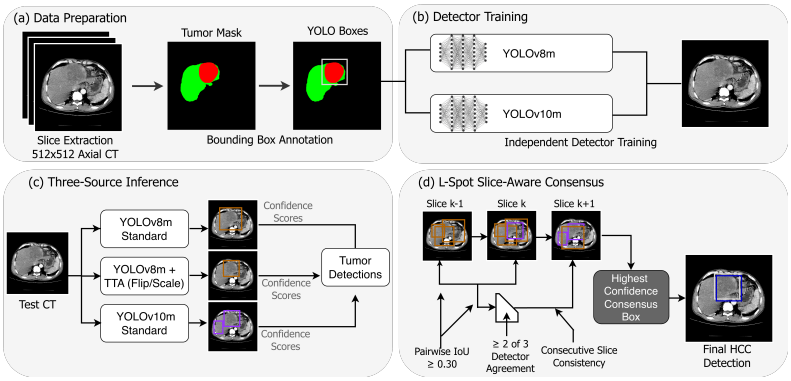

Fig. 1. Overview of the proposed L-Spot framework for HCC detection.

2.2 Detector Training

Figure 1 shows the proposed *L-Spot* framework for slice-wise HCC localisation in CECT. HCC localisation was formulated as a single-class 2D object detection task on axial CT slices. We independently trained two medium-scale object detectors, YOLOv8m [17] and YOLOv10m [20] using annotated tumour bounding boxes. Before training, each CT slice was resized to the network input resolution and intensity-normalised to ensure consistent image representation. YOLOv8m was selected for its efficient multi-scale, anchor-free tumour detection and classification. YOLOv10m complemented the primary detector by increasing prediction diversity and sensitivity to subtle tumours; both models were trained using standard localisation and classification losses.

2.3 Test-Time Inference and Consensus Filtering

After training the individual detectors, we applied a test-time consensus strategy to improve prediction stability across different models and adjacent CT slices. The proposed *L-Spot* framework generated three prediction sources: standard YOLOv8m inference, YOLOv8m with TTA, and an independently trained YOLOv10m inference model. Standard inference was performed on the original test slices, whereas TTA inference used transformed image variants, with predicted boxes mapped back to the original coordinate system before post-processing. We evaluated the predicted tumour detections using both inter-source agreement and inter-slice persistence. Spatial agreement between bounding boxes was measured using the IoU metric, and detections were considered concordant when the IoU ≥ 0.30. A predicted detection was retained only when at least two of the three prediction sources produced spatially concordant bounding boxes at the same anatomical location, and the detection persisted across consecutive CT slices. When multiple boxes satisfied the consensus criterion for the same tumour region, the highest-confidence box was selected as the representative detection.

2.4 Implementation Details

We trained the YOLOv8m and YOLOv10m detectors separately before applying the *L-Spot* consensus procedure. Both detectors used an input resolution of 512×512 pixels. Standard inference was performed on the original test slices, while TTA inference was performed on flipped and scaled variants, with predicted bounding boxes transformed back to the original coordinate system before voting. For slice-consistency and ensemble-based post-processing, bounding-box matching was performed using an IoU threshold of 0.30. Detections were grouped when they occurred at spatially corresponding locations across prediction sources and neighbouring slices. When multiple boxes satisfied the agreement criterion for the same lesion candidate, the final retained box was defined by the highest-confidence prediction among the agreeing detections. The same post-processing procedure was applied consistently to all test volumes. Models were trained for up to 50 epochs with a batch size of 16, early stopping, and AdamW (lr = 0.0002). Experiments used Python 3.10, PyTorch 2.1.0, and an NVIDIA A100 GPU.

3 Experimental Results

3.1 Baseline Detector Performance

Table 1 shows the test set performance of the evaluated YOLO architectures for 2D slice-wise HCC detection in CECT. YOLOv8m achieved the best overall performance, demonstrating the highest precision (0.612), F1-score (0.540), mAP_{50} (0.491), and $mAP_{50–95}$ (0.182). This indicates the most favourable balance between localisation accuracy and false-positive suppression. Meanwhile, YOLOv10m achieved the highest recall (0.502), suggesting greater sensitivity in tumour detection, while maintaining a comparable F1-score and mAP. YOLOv11m showed competitive precision and $mAP_{50–95}$ but lower recall and mAP_{50} comparable to the other variants. Based on these results, YOLOv8m was selected as the primary detector, complemented by the recall-oriented YOLOv10m.

Table 1. Test set performance of the evaluated YOLO architectures.

Model	Precision	Recall	F1-score	mAP_{50}	$mAP_{50–95}$
YOLOv5m	0.516	0.427	0.464	0.388	0.148
YOLOv8m	**0.612**	0.483	**0.540**	**0.491**	**0.182**
YOLOv9m	0.490	0.425	0.455	0.393	0.167
YOLOv10m	0.568	**0.502**	0.533	0.477	0.178
YOLOv11m	0.584	0.446	0.506	0.435	0.179

3.2 Ablation Study

Table 2 presents the ablation study of different YOLOv8m configurations on the test set. The baseline model with a 512×512 input resolution achieved the best performance across all metrics, with a precision, recall, F1-score, mAP_{50},

and $mAP_{50–95}$ of 0.612, 0.483, 0.540, 0.491, and 0.182, respectively. All evaluated modifications reduced the mAP_{50} and F1-score relative to the baseline, including using a higher input resolution, the large model variant, extended training, additional augmentation, weighted loss, the Convolutional Block Attention Module (CBAM), and the combined augmentation–weighted loss configuration. Increased complexity did not improve generalisation; therefore, baseline YOLOv8m at 512×512 resolution was selected as the primary *L-Spot* detector.

Table 2. YOLOv8m ablation study on the test set.

Configuration	Precision	Recall	F1-score	mAP_{50}	$mAP_{50–95}$
Baseline (512×512)	**0.612**	**0.483**	**0.540**	**0.491**	**0.182**
Baseline (640×640)	0.411	0.352	0.379	0.306	0.124
Large variant	0.442	0.392	0.415	0.329	0.139
Extended training	0.555	0.343	0.424	0.374	0.171
Augmented	0.598	0.383	0.468	0.413	0.171
Weighted loss	0.584	0.429	0.495	0.430	0.174
CBAM attention	0.410	0.380	0.394	0.345	0.139
Combined augmentation–weighted loss	0.513	0.290	0.370	0.276	0.131

3.3 Slice-Consistency Post-processing

Tables 3 and 4 were produced with a custom evaluation protocol that differs from the Ultralytics evaluator used in Tables 1 and 2. Detections were retained at a confidence threshold of 0.10, matched greedily one-to-one against ground-truth boxes at IoU$\geq$0.30, and scored at that fixed operating point. The Ultralytics evaluator instead matches at IoU 0.50 and reports precision and recall at the F1-optimal point of the precision–recall curve. The same YOLOv8m predictions therefore yield different values in the two sets of tables, and figures should be compared only within a protocol. Table 3 summarises the effects of applying slice-consistency post-processing to YOLOv8m detections at a confidence threshold of 0.10. As CT scans comprise volumetric image sequences, true hepatic tumours are expected to persist across adjacent slices, whereas isolated false-positive detections are less likely to exhibit such continuity [2]. The evaluated rules therefore provide a training-free approach for suppressing spurious detections that lack inter-slice persistence. All rules improved precision, F1-score, and AP0.30 relative to the unfiltered YOLOv8m baseline, which achieved an F1-score of 0.615 and AP0.30 of 0.437, with 599 false positives. In contrast, Rule B, which required detections to persist across at least three consecutive slices, achieved the best overall performance, with the highest precision (0.632), F1-score (0.636), and AP0.30 (0.441), as well as the fewest false positives (431). This improvement was accompanied by a moderate reduction in recall, reflecting the expected trade-off between false-positive suppression and sensitivity. Although Rule A achieved slightly higher recall (0.657), Rule B offered the best overall precision–recall balance. Rule C performed similarly to Rule A but did not surpass Rule B in either F1-score or AP0.30. Therefore, Rule B was selected for subsequent *L-Spot* ensemble experiments.

Table 3. Test-set results for slice-consistency post-processing of YOLOv8m detections at a confidence threshold of 0.10, evaluated at IoU≥ 0.30 ($AP_{0.30}$)

Configuration	Precision	Recall	F1-score	$AP_{0.30}$	TP	FP	FN
Unfiltered YOLOv8m baseline	0.566	0.674	0.615	0.437	781	599	378
Rule A (≥2 consecutive slices)	0.610	0.657	0.632	0.437	761	487	398
Rule B (≥3 consecutive slices)	**0.632**	0.639	**0.636**	**0.441**	741	431	418
Rule C (conf>0.5 OR ≥2 slices)	0.608	0.660	0.633	0.435	765	493	394

3.4 Consensus Ensemble Post-processing

Building on the slice-consistency analysis, we evaluated whether TTA and multi-model agreement could further improve 2D slice-wise HCC detection. Table 4 presents the step-wise performance from the unfiltered YOLOv8m baseline to the final three-source *L-Spot* consensus ensemble. AP0.30 summarises performance across the full precision–recall curve, whereas the F1-score reflects a single operating point; therefore, the two metrics may not change consistently. Precision rises from 0.566 to 0.820 while recall falls only from 0.674 to 0.664. A confidence threshold alone cannot achieve this, so the suppressed detections were confident predictions that no second source reproduced. Combining TTA (Rule B) with slice-consistency filtering increased the F1-score from 0.636 to 0.653. Enforcing agreement between baseline YOLOv8m and TTA predictions further improved the F1-score to 0.675 and precision to 0.729, demonstrating the greater reliability of detections reproduced under transformed inputs. The final *L-Spot* configuration retained detections supported by at least two of the three sources at IoU≥0.30. The final *L-Spot* configuration achieved the best overall performance, with an F1-score of 0.733, precision of 0.820, and recall of 0.664, representing an absolute F1-score improvement of 11.8 percentage points over the unfiltered YOLOv8m baseline. In contrast, the stricter "all-three" consensus achieved the highest precision (0.897) but reduced recall to 0.525 and $AP_{0.30}$ to 0.375, indicating a highly conservative operating point. Crucially, these gains were achieved solely through test-time consensus filtering, without architectural modifications or model retraining. As shown in Figure 2, TTA and YOLOv10m recovered lesions missed by baseline YOLOv8m, while *L-Spot* retained the correct detections. In panel (d), *L-Spot* also suppressed a TTA false positive that was not consistently reproduced across sources.

Table 4. Step-wise test-set results from baseline detection to the proposed *L-Spot* three-source consensus ensemble, evaluated at IoU≥ 0.30 ($AP_{0.30}$).

Approach	Sources	Precision	Recall	F1-score	$AP_{0.30}$
Unfiltered YOLOv8m baseline	YOLOv8m	0.566	0.674	0.615	0.437
YOLOv8m + Rule B	YOLOv8m	0.632	0.639	0.636	0.441
TTA (conf≥0.30) + Rule B	YOLOv8m + TTA	0.655	0.651	0.653	0.458
YOLOv8m ∩ TTA agreement	YOLOv8m + TTA	0.729	0.628	0.675	0.445
***L-Spot* (≥2 of 3)**	YOLOv8m + TTA + YOLOv10m	**0.820**	**0.664**	**0.733**	0.439
All-three agreement	YOLOv8m + TTA + YOLOv10m	0.897	0.525	0.663	0.375

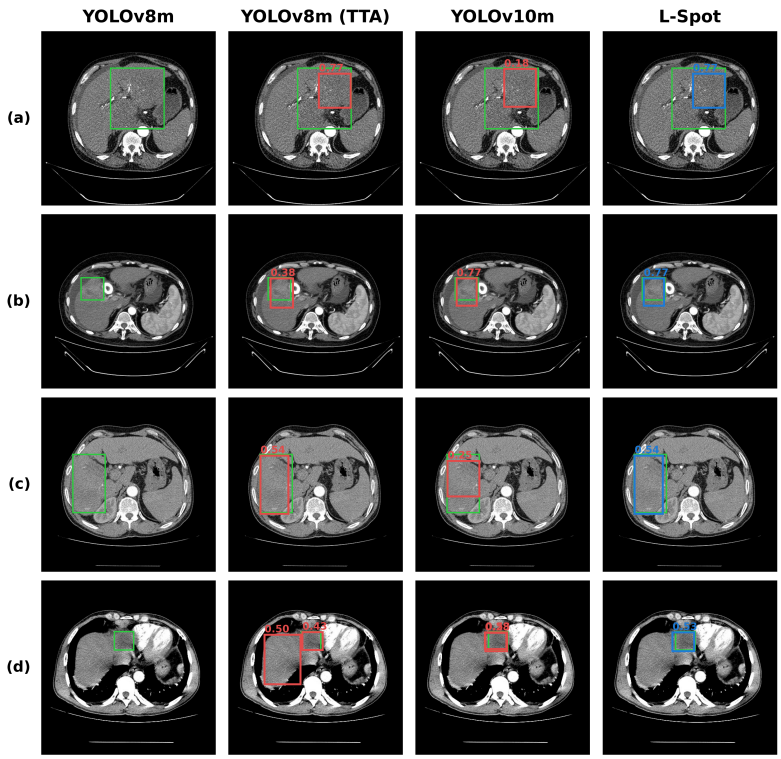

Fig. 2. Qualitative comparison of HCC detections from YOLOv8m, YOLOv8m-TTA, YOLOv10m, and the proposed L-Spot method.

4 Conclusion

This paper introduces *L-Spot*, a slice-aware consensus framework for improving 2D HCC localisation in CECT. By combining predictions from complementary YOLO models with spatial consistency across adjacent CT slices, *L-Spot* suppresses isolated false positives while preserving true tumour detections, achieving a better precision–recall balance than the strongest single-detector baseline. The framework requires no additional training, is readily integrated into existing 2D detection pipelines and applies to different object detectors. Future work will assess its generalisability on larger multi-centre datasets, extend it to multi-phase CT, and investigate its integration with volumetric detection frameworks.

Acknowledgment

This work was supported by the Barts Charity (G-003271 and MGU0504).

References

1. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nndetection: A self-configuring method for medical object detection. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 12905, pp. 530–539. Springer (2021)
2. Bilic, P., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
3. Calderaro, J., et al.: Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma. Journal of Hepatology **80**(6), 1005–1019 (2024)
4. Chen, M., Zhang, B., Topatana, W., Cao, J., Zhu, H., Juengpanich, S., Cai, X.: Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning. NPJ Precision Oncology **4**(1), 14 (2020)
5. Chen, Z., et al.: Global epidemiology of hepatocellular carcinoma: trends, risk factors and prevention. Nature Reviews Gastroenterology and Hepatology **17**(12), 888–902 (2020)
6. Christ, P., et al.: Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks. arXiv:1702.05970 (2017)
7. Diwan, T., Anirudh, G., Tembhurne, J.: Object detection using yolo: challenges, architectural successors, datasets and applications. Multimedia Tools and Applications **82**, 9243–9275 (2023)
8. Ghouse, H., Behzad, M.: Mosaic: A multi-view 2.5d organ slice selector with cross-attentional reasoning for anatomically-aware ct localization in medical organ segmentation. Computer Vision and Image Understanding (2025). https://doi.org/10.1016/j.cviu.2025.104302
9. Lakshimioriya, S., et al.: A systematic review of deep learning techniques for liver tumour diagnosis. Diagnostics **13**(4), 653 (2023)
10. Li, X., et al.: Two-stage liver and tumour segmentation algorithm based on convolutional neural network. Diagnostics **11**(10), 1806 (2021)
11. Piffer, S., et al.: Tackling the small data problem in medical image classification with artificial intelligence: a systematic review. Progress in Biomedical Engineering **6**(3), 032001 (2024). https://doi.org/10.1088/2516-1091/ad525b
12. Ramos-Soto, O., et al.: Miafex: An attention-based feature extraction method for medical image classification. arXiv:2501.08562 (2025)
13. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
14. 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 (NeurIPS). pp. 91–99 (2015)
15. Singal, A., et al.: Advancing surveillance strategies for hepatocellular carcinoma: a new era of efficacy and precision. Journal of Clinical and Experimental Hepatology (2024), https://pmc.ncbi.nlm.nih.gov/articles/PMC11214318/
16. Terven, J., Córdova-Esparza, D.M., Romero-González, J.A.: A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. Machine learning and knowledge extraction **5**(4), 1680–1716 (2023)
17. Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International conference on advances in data engineering and intelligent computing systems (ADICS). pp. 1–6. IEEE (2024)

18. Vijayakumar, V., Vairavasundaram, S.: Yolo-based object detection models: a review and its applications. Multimedia Tools and Applications **83**, 83187–83257 (2024)
19. Vorontsov, E., et al.: Deep learning for automated segmentation of liver lesions at ct in patients with colorectal cancer liver metastases. Radiology: Artificial Intelligence **1**(2), e180014 (2019)
20. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: Yolov10: Real-time end-to-end object detection. Advances in neural information processing systems **37**, 107984–108011 (2024)
21. Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M.: Mulan: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 11769. Springer (2019)
22. Yang, J., He, Y., Kuang, K., Lin, Z., Pfister, H., Ni, B.: Asymmetric 3d context fusion for universal lesion detection. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 12905. Springer (2021)
23. Zhao, Y., Agyemang, D., Liu, Y., Mahoney, M., Li, S.: Quantifying interpretation reproducibility in vision transformer models with tavac. Science Advances **10**(51), eabg0264 (2024). https://doi.org/10.1126/sciadv.abg0264
24. Zhu, H., et al.: Comprehensive multi-phase 3d contrast-enhanced ct imaging for primary liver cancer. Scientific Data (2025). https://doi.org/10.1038/s41597-025-05125-2