IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 February 2025, accepted 6 March 2025, date of publication 20 March 2025, date of current version 31 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3553087

SURVEY

Deep Learning-Based Medical Object Detection: A Survey

MOHAMMADREZA SARAEI[1], MEHRSHAD LALINIA[2], AND EUNG-JOO LEE[1], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, The University of Arizona, Tucson, AZ 85721, USA
[2]Department of Biomedical Engineering, Amirkabir University of Technology, Tehran, Iran

Corresponding author: Eung-Joo Lee (eungjoolee@arizona.edu)

ABSTRACT Recent advancements in medical object detection (MOD) have been propelled by the rapid evolution of deep learning (DL) technologies, revolutionizing medical imaging and diagnostic workflows. This survey comprehensively reviews various studies across diverse imaging modalities, including X-Ray, CT, MRI, Ultrasound, and Histopathology. Notable improvements include integrating You Only Look Once (YOLO)-based architectures, Vision Transformers (ViT), and hybrid attention mechanisms, significantly enhancing detection accuracy and efficiency. Standout models, such as YOLOv8m, Hybrid YOLO-NAS, and YOLOv4+ViT, have demonstrated exceptional performance, achieving mean average precision (mAP) scores between 98.6% and 99.5%. These advancements leverage sophisticated features like Cross-Stage Partial (CSP) networks, Spatial Pyramid Pooling (SPP), and Bi-Directional Feature Pyramid Networks (BiFPN) to improve feature extraction and detection in medical images. Despite these successes, challenges remain in adapting these models to resource-limited settings and ensuring their outputs are interpretable for clinicians. This survey aims to bridge the gap between theoretical progress and practical implementation by aligning cutting-edge technological developments with clinical demands. It provides a certain roadmap for future innovation in MOD, with the overarching goal of improving patient care through enhanced diagnostic capabilities.

INDEX TERMS Medical object detection, deep learning, medical image analysis, medical imaging.

I. INTRODUCTION

Medical Object Detection (MOD) has emerged as a pivotal component of modern computer vision applications in healthcare, leveraging the transformative capabilities of Deep Learning (DL) within Artificial Intelligence (AI). By enabling the autonomous detection and precise localization of objects of interest, MOD addresses critical challenges in the diagnostic workflow, aiding clinical practitioners in interpreting medical images with enhanced accuracy and efficiency. From detecting tumors to identifying pathological masses and localizing anatomical structures, MOD contributes significantly to disease diagnosis and treatment planning [1], [2], [3].

The MOD model based on DL, illustrated in Fig. 1, typically consists of the backbone, the neck, and the head. The backbone extracts features from the input image, and the neck aggregates and refines features from different backbone levels to obtain more sophisticated representations. The head performs specific tasks such as object detection, including bounding box prediction and classification. These models are applied across various imaging modalities, including X-Ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US), Histopathology Imaging (HPI), Endoscopy, and Optical Coherence Tomography (OCT). Each modality presents unique challenges, such as variations in resolution, noise, and object scales, necessitating robust detection models tailored to specific clinical applications [1]. These challenges have driven extensive research into DL-driven approaches, which now form the foundation of advancements in MOD.

Over the past decade, DL has revolutionized the field of medical imaging, offering unparalleled accuracy and efficiency in object detection tasks. Early methods relied

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

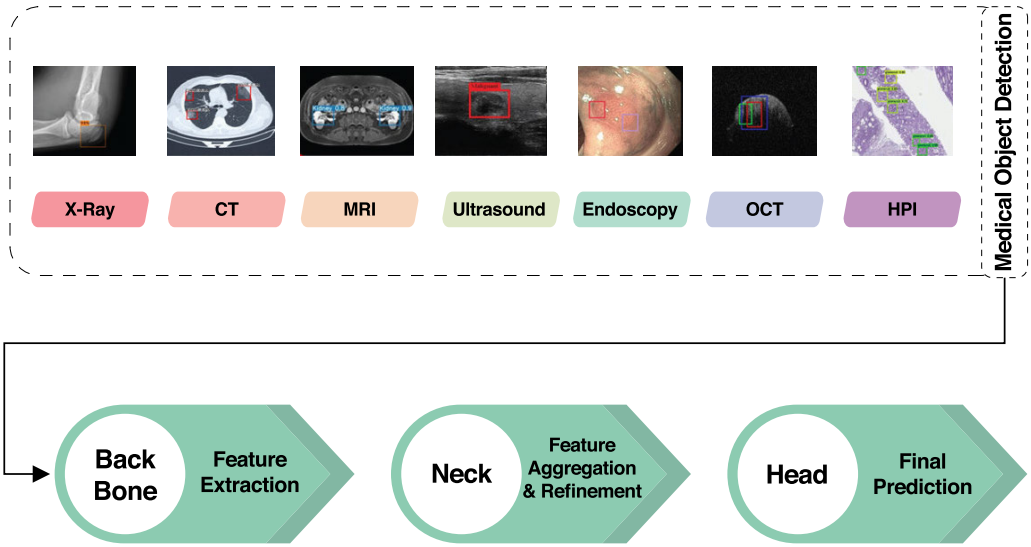

FIGURE 1. The MOD process begins with various imaging modalities, including X-Ray, CT, MRI, US, HPI, Endoscopy, and OCT. These images are processed through a deep neural network architecture comprising three primary components: the backbone, responsible for feature extraction; the neck, which performs feature aggregation and refinement; and the head, which handles final object detection and prediction.

on feature-engineered algorithms that required significant domain expertise and manual intervention. The advent of convolutional neural networks (CNN) marked a turning point, automating feature extraction and enabling more precise detection. Within MOD, CNN-based models have evolved into two primary architectural paradigms: single-stage and two-stage detectors [4].

Two-stage detectors, such as Faster R-CNN [5], and Cascade R-CNN [6], excel in detection accuracy by decoupling the region proposal and classification tasks. These models have demonstrated high efficacy in medical applications. However, their computational demands and slower inference times pose challenges for real-time and resource-constrained scenarios, such as edge device deployment. In contrast, single-stage detectors, including You Only Look Once (YOLO) [7] and its modified versions [8] along with RetinaNet [9], [10], [11], integrate detection tasks into a unified framework, significantly reducing inference times while maintaining competitive accuracy [12]. For instance, YOLOv4 strikes a balance between speed and precision, while YOLOX and YOLOR address resource-constrained applications with improved real-time performance [13], [14].

As the demand for more accurate and efficient MOD systems grows, the field has seen the adoption of novel architectural paradigms. Transformer-based models, such as the Vision Transformer (ViT) [15], Swin-Transformer [16], and their derivatives, have introduced global attention mechanisms, enabling models to capture long-range dependencies and achieve competitive performance in object detection tasks [17], [18], [19]. These models have been extended to medical imaging, with Detection Transformer (DETR) and its variants achieving state-of-the-art results by addressing challenges like data dependency and computational complexity [20], [21], [22], [23].

The EfficientDet family further exemplifies the trend toward scalable and lightweight architectures. EfficientDet models range from the compact and resource-efficient D0 variant to the more accurate but computationally intensive D6 variant. EfficientDet-Lite models are specifically designed for mobile and embedded systems, enabling real-time MOD in remote and resource-limited healthcare environments [24]. Similarly, advances in one-stage architectures, such as CenterNet and its derivatives (e.g., CenterNet++), focus on precise localization and robust performance, catering to the diverse needs of medical imaging applications [[25], [26], [27], [28].

This study examines recent advancements in DL models for MOD over the past two years. It considers peer-reviewed studies published in English and indexed in Google Scholar between January 2023 and December 2024. Research focused on segmentation, classification, or registration tasks, as well as studies outside this timeframe or those lacking sufficient assessment details—such as data type, data size, quantitative metrics, and technical components—have been excluded. As depicted in Fig. 2, the topics in this review are categorized based on imaging modalities and model architectures. Included studies focus on MOD using bounding boxes or employing hybrid approaches that combine masks and bounding boxes. The performance of these DL models is evaluated using the mean average precision (mAP) metric, calculated as detailed in Equation (1). Moreover, the technical features and components influencing model performance

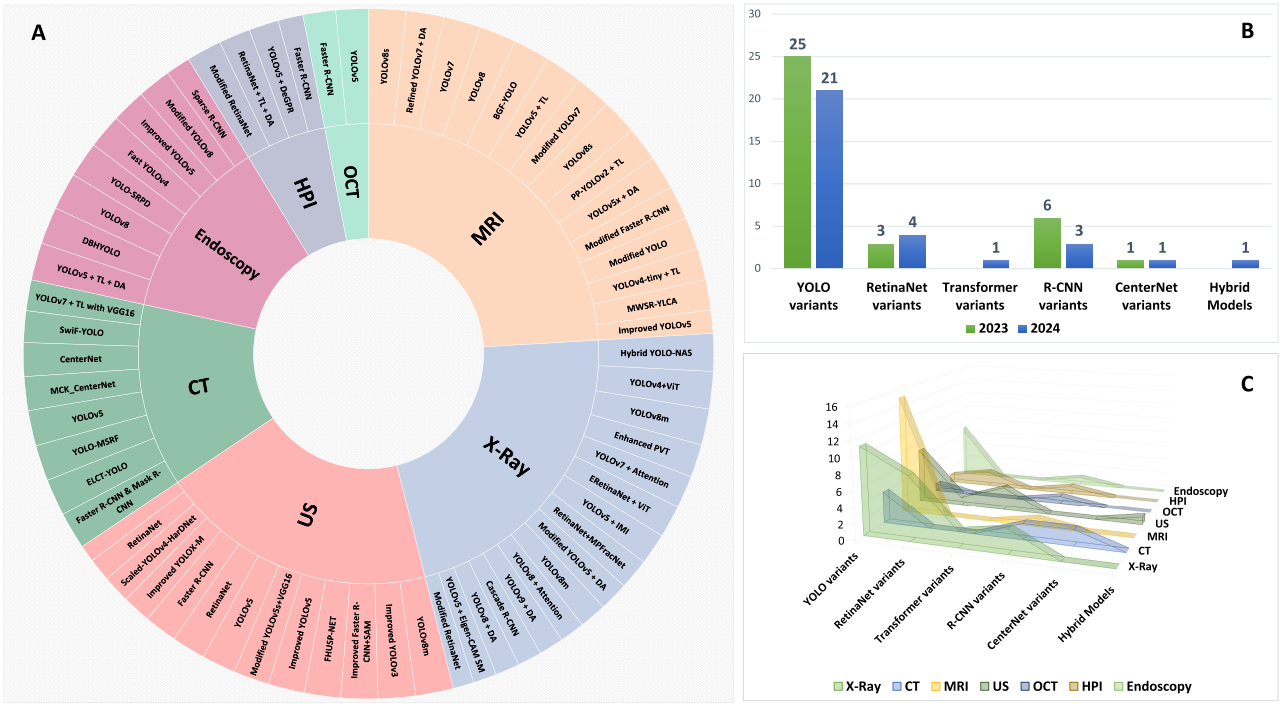

FIGURE 2. Categorization of our review, highlighting: (A) the distribution of DL-based models across different medical imaging modalities, (B) the temporal distribution of DL-based models within the study's timeframe, and (C) the breakdown of DL-based models for each medical imaging modality.

are analyzed. By aligning model architectures with imaging characteristics and clinical demands, this research bridges the gap between theoretical progress and practical implementation. It provides valuable insights into optimizing MOD for enhanced diagnostic performance and clinical applicability, ultimately guiding future research efforts.

$$mAP = \frac{1}{C}\sum_{i}^{n} AP_c \tag{1}$$

where C is the total number of object classes in your dataset, $\sum$ indicates calculating the sum of an arithmetic progression for the first n terms, and AP_c is the Average Precision (AP) for a specific class c.

The AP for a class represents the area under the Precision-Recall Curve (PRC) for that class. The PRC plots the trade-off between Precision (the percentage of detected objects that are truly the class of interest) and Recall (the percentage of all objects detected).

II. LITERATURE REVIEW

A. X-RAY IMAGING

As shown in Fig. 3, recent DL advances have significantly enhanced X-Ray imaging diagnostic capabilities. These developments, driven by innovations in model architectures, optimization strategies, and multimodal approaches, underscore the collaborative potential of diverse methodologies.

YOLO-based architectures have emerged as dominant tools in MOD, showcasing remarkable versatility and adaptability across diverse diagnostic tasks. Their efficacy is exemplified in various studies, highlighting their capability to achieve high accuracy when appropriately tailored to specific datasets. For instance, Aldughayfiq et al. modified YOLOv5 for pressure ulcer detection, achieving a mAP of 76.9% through data augmentation to address limited dataset size [29]. Similarly, Ahmed et al. evaluated YOLOv8 variants for wrist fracture detection, with YOLOv8m outperforming YOLOv8x, achieving a mAP of 95.0% versus 77.0% [30]. These examples underscore the critical need for dataset-specific tuning to unlock YOLO's full potential in MOD applications.

Advancing the YOLO paradigm, Chien et al. introduced YOLOv9, which integrates programmable gradient information and efficient layer aggregation, resulting in a mAP of 65.4%. They further enhanced performance with YOLOv8-AM by incorporating attention mechanisms such as the Convolutional Block Attention Module (CBAM), improving accuracy to 65.8% [31], [32]. These modifications highlight the growing importance of attention modules in refining feature representation and improving diagnostic precision.

In addition to standalone YOLO enhancements, hybrid approaches have demonstrated superior results by leveraging complementary strengths from multiple architectures. Medaramatla et al. achieved a mAP of 99.1% for hand fracture detection by combining YOLO-NAS, EfficientDet, DETR, and attention mechanisms, illustrating the power of architectural fusion [33]. Similarly, Guan et al. adopted a hybrid ViT-CNN model for thighbone fracture detection,

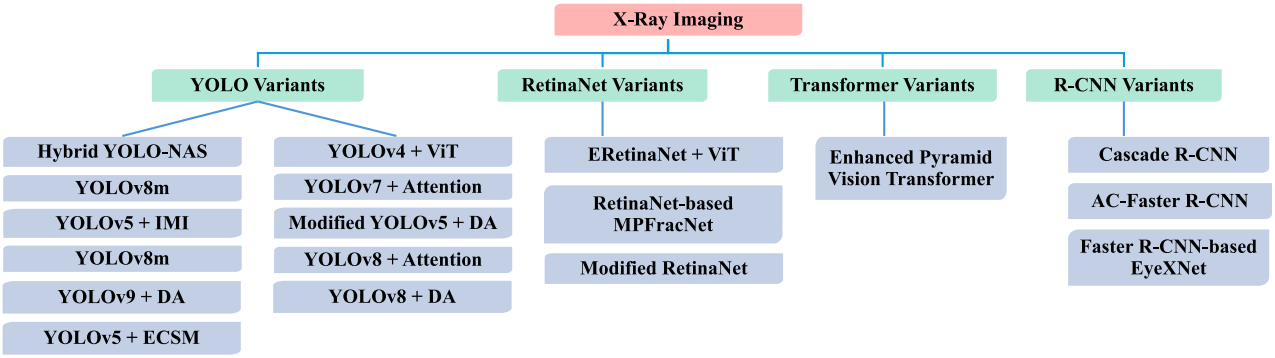

FIGURE 3. Schematic representation of recent advances in DL-based object detection techniques aimed at enhancing diagnostic capabilities in X-Ray imaging. IMI denotes Integrated Multimodal Information, DA denotes Data Augmentation, and ECSM denotes Eigen-CAM Saliency Maps.

achieving a mAP of 87.0%, showcasing the advantages of integrating global and local feature extraction mechanisms [34]. These hybrid frameworks reflect a transformative shift toward combining diverse methodologies to enhance accuracy.

Other studies have explored hybrid architectures for specific MODs. Hassan et al. achieved a mAP of 98.69% by integrating YOLOv4 with ViT for breast mass detection [35], while Chen et al. developed ERetinaNet with channel attention modules, reaching a mAP of 85.01% [36]. These results emphasize the potential of hybrid models in improving both diagnostic precision and feature extraction capabilities.

While YOLO-based models have gained prominence, other frameworks like Cascade R-CNN and RetinaNet have also demonstrated significant potential. Rathinakumar et al.'s CovidXDetector, based on the Cascade R-CNN framework, achieved a mAP of 64.2%, outperforming YOLOv5 and Single-Shot Detector (SSD) in COVID-19 diagnostics [34]. Similarly, Qin et al.'s MPFracNet, which incorporated deformable bottlenecks and specialized loss functions into a RetinaNet foundation, achieved a mAP of 80.4% for metacarpophalangeal fracture detection. These approaches illustrate the benefits of cascaded architectures and tailored loss functions in addressing specific challenges such as precise localization and class imbalance [37].

Explainability is another critical aspect of MOD applications. Prinzi et al. demonstrated this by integrating Eigen-CAM with YOLOv5 for mammogram-based breast cancer detection, achieving a mAP of 62.1% while enhancing model interpretability [38]. Chen et al. further highlighted the value of multimodal integration by incorporating patient metadata, such as age and gender, into a YOLOv5 model for diagnosing Developmental Dysplasia of the Hip (DDH), achieving a mAP of 83.1% [39]. These efforts underscore the importance of making AI models more transparent and relevant for clinical decision-making.

Preprocessing techniques, particularly data augmentation, have played a pivotal role in addressing dataset limitations. Ju et al. utilized advanced data augmentation methods to improve YOLOv8's performance for pediatric wrist trauma detection, achieving a mAP of 63.8% [40]. Similarly, Razaghi et al. leveraged YOLOv8 for dental radiograph analysis, achieving a mAP of 71.6%, underscoring the importance of validating models across heterogeneous imaging conditions [41].

Across these varied applications and architectural innovations, several trends emerge. Hybrid models consistently deliver state-of-the-art performance but face challenges related to computational complexity and scalability. Attention mechanisms, as seen in YOLOv8-AM and ViT-CNN integrations, have proven instrumental in enhancing accuracy by focusing on critical image regions. Furthermore, specialized loss functions and dataset-specific optimizations, such as those employed in MPFracNet and DDH diagnostics, highlight the importance of tailoring methods to specific tasks. Despite these advances, challenges remain, including variability in dataset quality, limited generalizability across diverse imaging conditions, and the ongoing need for improved model explainability. Future research should prioritize the development of lightweight, scalable models that integrate multimodal data while maintaining high performance and interpretability.

B. COMPUTERIZED TOMOGRAPHY (CT)

Recent betterments in DL have remarkably enriched the diagnostic capabilities of CT imaging, as illustrated in Fig. 4. By integrating diverse architectural innovations and leveraging domain-specific adaptations, these models have achieved great performance across various diagnostic tasks. The following analysis explores key contributions in the field, focusing on their strengths, limitations, and collective insights.

Hybrid approaches have proven highly effective in MOD by combining the strengths of different models, thereby achieving exceptional accuracy across various diagnostic tasks. For example, Sahin et al. employed a fusion of Faster R-CNN and Mask R-CNN for COVID-19 diagnosis, achieving a remarkable mAP of 97.72%. This hybrid model substantially outperformed baseline approaches, showcasing the potential of combining complementary features

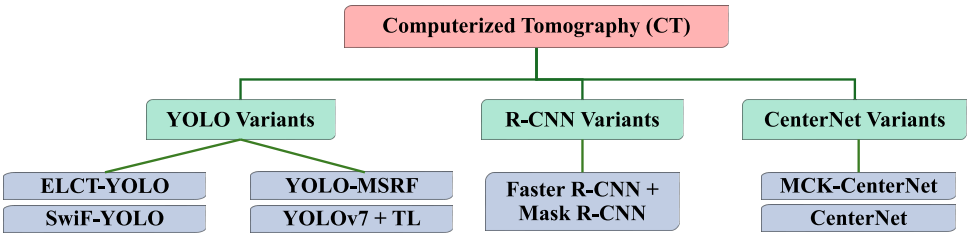

FIGURE 4. Schematic representation of recent betterment in DL-based object detection strategies aimed at improving diagnostic capabilities in CT imaging. TL denotes Transfer Learning.

for automating complex diagnostic processes. However, while the results are promising, scalability and real-world deployment remain significant challenges that require further exploration [42].

Similarly, Ji et al. introduced the ELCT-YOLO model for lung tumor detection, which integrates cascaded refinement schemes and receptive field enhancement modules to improve multiscale feature representation, achieving a mAP of 97.1%. This work highlights the precision that can be attained with hybrid designs, though it also underscores the difficulty in ensuring robustness across diverse clinical datasets, pointing to the need for broader validation [43].

YOLO-based models have seen significant advancements in the detection of specific medical conditions, such as lung nodules. Wu et al. enhanced YOLOv7 by adding a small object detection layer and a multi-scale receptive field module, which allowed the model to achieve a mAP of 95.26%. This innovation demonstrates YOLO's adaptability, particularly in detecting small and irregular patterns. However, consistency across heterogeneous datasets remains a persistent challenge [44]. In a complementary approach, Mammeri et al. combined YOLOv7 with VGG16-based transfer learning for lung nodule classification, achieving a mAP of 81.02%. While the results are promising, the computational demands of this method may hinder its deployment in resource-limited settings, highlighting the need for lightweight but effective alternatives [45].

Anatomical localization tasks have also benefitted from YOLO-based advancements. Yaseen et al. demonstrated the efficacy of YOLOv5 in detecting cervical spine bones in CT images, achieving a mAP of 93.0%. This success illustrates YOLO's robustness in anatomical localization tasks, although further optimizations are necessary for real-time clinical applications [46]. In a related development, Ren et al. introduced the SwiF-YOLO framework, which combines a Swin-Transformer, Adaptively Spatial Feature Fusion (ASFF), and Generalized Intersection over Union (GIoU) loss. Achieving a mAP of 83.1%, the model strikes a balance between computational efficiency and accuracy. However, the use of transformer-based architectures may face scalability limitations in low-resource environments [47].

Fracture detection has also seen significant improvements through architectural innovations. Su et al. proposed a CenterNet-based model with a pyramidal heatmap structure for rib fracture detection, achieving a mAP of over 89.0%. While effective, the trade-off between model complexity and real-time usability remains a key concern [48]. Similarly, Zhou et al. introduced MCK-CenterNet, which incorporates multiscale feature capture, channel-spatial fusion modules, and knowledge distillation for mandibular fracture detection. This model achieved a mAP of 89.3%, demonstrating the importance of multiscale and knowledge-driven enhancements. However, broader validation across diverse clinical scenarios is needed to fully assess its generalizability [49].

The recent advancements in DL for CT imaging reveal several recurring themes and challenges that are pivotal to the field's progress. Hybrids and attention-enhanced architectures, such as the fusion of Faster R-CNN and Mask R-CNN or the incorporation of advanced attention mechanisms in YOLO-based frameworks, consistently yield high accuracy. However, these models often face scalability issues due to their computational complexity. Additionally, efforts to improve small object and multiscale detection underscore the importance of multiscale receptive fields in enhancing sensitivity.

A critical challenge remains balancing accuracy with efficiency, as lightweight and resource-efficient designs are essential for real-world adoption. Furthermore, domain-specific customization, which includes dataset-specific tuning, is crucial to optimizing performance for particular tasks. Moving forward, addressing the key limitations of scalability, generalizability, explainability, and cost-effectiveness will be vital. Interdisciplinary collaboration and innovative design approaches will be necessary to bridge the gap between experimental breakthroughs and clinical integration, paving the way for scalable, interpretable, and clinically applicable solutions.

C. MAGNETIC RESONANCE IMAGING (MRI)

As indicated in Fig. 5, the transformative potential of DL for enhancing diagnostic accuracy in MOD has been widely demonstrated, with numerous studies showcasing the adaptability and performance of YOLO-based architectures and related models across various modalities. These approaches emphasize the balance between computational efficiency, detection accuracy, and scalability. A key example is the enhanced YOLOv7 model proposed by Abdusalomov et al., achieving a remarkable mAP of 98.9% for brain tumor

detection in MRI scans. This achievement is attributed to innovations such as CBAM, Spatial Pyramid Pooling Fast Plus (SPPF+), decoupled heads, and Bi-Directional Feature Pyramid Network (BiFPN) (Fig. 6). While these modifications illustrate the efficacy of multiscale feature fusion, the associated computational cost raises scalability concerns for real-time clinical use [50]. Complementing this, Kang et al. introduced BGF-YOLO, a YOLOv8 variant, incorporating bi-level routing attention and generalized feature pyramid networks, achieving 97.4% mAP for similar applications. Both studies underline the significance of attention mechanisms and advanced feature extraction in boosting diagnostic accuracy while grappling with computational overheads [51].

Exploring lightweight alternatives, Rahimi et al. applied transfer learning within a YOLOv4-tiny framework, achieving mAP values of 80.74% for raw and 83.2% for preprocessed brain tumor imaging datasets. This approach highlights the trade-off between architectural simplicity and diagnostic performance, especially when compared to newer YOLO variants with superior multiscale detection capabilities [52]. Similarly, N et al. demonstrated the utility of YOLO-tiny integrated with the Nvidia Jetson Nano for prostate cancer detection, achieving an impressive mAP of 99.0%. These studies emphasize the potential of lightweight models for real-time diagnostics, while also highlighting the need for broader validation [53].

Beyond brain imaging, the adaptability of YOLO models has been demonstrated in other modalities. Kulavuz et al. applied YOLOv7 to detect malignant breast lesions on dynamic contrast-enhanced MRI, achieving 98.54% mAP, showcasing the versatility of DL frameworks. Anari et al. also used YOLOv7 for kidney detection with a mAP of 95.0%, underscoring the model's cross-domain utility. However, these applications reveal a recurring theme: the importance of external validation and dataset diversity to ensure clinical robustness [54], [55].

The challenge of detecting subtle and small-scale lesions was highlighted by Zhou et al., who modified YOLOv5 with CBAM and Swin-Transformer blocks for brain metastasis detection. Despite a modest mAP of 61.2%, their focus on small metastases and the use of the F1-score underscores the need for alternative metrics to evaluate clinical relevance [56]. Cengil et al. similarly reported a mAP of 85.6% for glioma and meningioma detection, emphasizing the potential of DL-based MOD for tumor localization while identifying room for improvement in feature representation [57].

In spinal disease diagnostics, Xuan et al. demonstrated the effectiveness of PP-YOLOv2, achieving 90.08% mAP with a drastically reduced diagnosis time of 14.5 seconds compared to manual interpretation. Likewise, Shao et al. employed Faster R-CNN with VGG16 and ResNet-50 backbones for spinal cord injury detection, achieving 88.6% mAP. Both studies highlight the potential of efficient DL models to augment radiological workflows, though reliability in high-stakes scenarios remains a concern [58], [59].

In the domain of Nasopharyngeal Carcinoma (NPC) detection, Wu et al. developed MWSR-YLCA, a hybrid model integrating Multi-Window Settings Resampling (MWSR) and Coordinate Attention Mechanisms (YLCA), achieving a mAP of 80.1%. This modular approach underscores the importance of customization for domain-specific tasks while revealing the limitations of adapting generic frameworks to highly specialized applications [60].

Finally, the influence of preprocessing and dataset characteristics on model performance is exemplified by Priscilla et al., who compared YOLO variants for lumbar disc herniation detection, with YOLOv5x achieving a mAP of 89.30% on a relatively small dataset. The study underscores the critical role of dataset size and augmentation in optimizing model accuracy while raising concerns about scalability to larger, more diverse datasets [61].

Despite the remarkable advancements demonstrated in these studies, several challenges hinder their seamless adoption in clinical practice. A recurring issue is the computational complexity of advanced YOLO variants, such as YOLOv7 and YOLOv8, which integrate sophisticated attention mechanisms and feature fusion techniques but impose significant processing overhead, limiting their scalability for real-time diagnostics. The generalizability of models remains another critical concern, as many studies rely on relatively small or homogeneous datasets, which may not capture the variability of real-world clinical scenarios. Additionally, the trade-offs between model simplicity and accuracy are evident, particularly in lightweight architectures like YOLOv4-tiny, which, while computationally efficient, often struggle with detecting small or subtle lesions. Furthermore, studies focusing on specialized tasks, such as detecting small metastases or subtle abnormalities, highlight the need for improved feature extraction and tailored metrics beyond traditional mAP values to assess clinical relevance.

D. ULTRASOUND IMAGING (US)

As illustrated in Fig. 7, recent studies in MOD demonstrate the transformative potential of DL architectures in enhancing the diagnostic performance of US imaging. These advancements are characterized by variations in architectural choices, dataset applicability, and clinical feasibility, as summarized below.

Tian et al. significantly improved the Faster R-CNN architecture by incorporating a ResNet-50 backbone and deformable convolutions, achieving a mAP of 97.4%, a 10% enhancement over the baseline. This improvement was driven by the integration of FPN and ROI Align, which boosted precision for small thyroid nodules. However, the model's computational complexity and reliance on advanced optimizers present challenges for real-time clinical applications. In a subsequent study, the same authors explored a Swin-Transformer backbone with Faster R-CNN, achieving a lower mAP of 44.8%. Despite outperforming traditional CNN-based baselines, this model's reduced accuracy

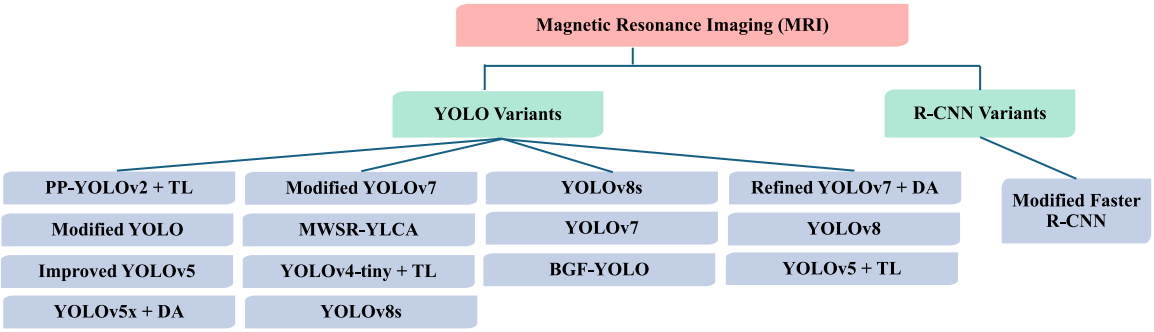

FIGURE 5. Schematic expression of the recent improvement in DL-based object detection approaches aspired to enhance diagnostic abilities in MR imaging. DA denotes Data Augmentation and ECSM denotes Eigen-CAM Saliency Maps.

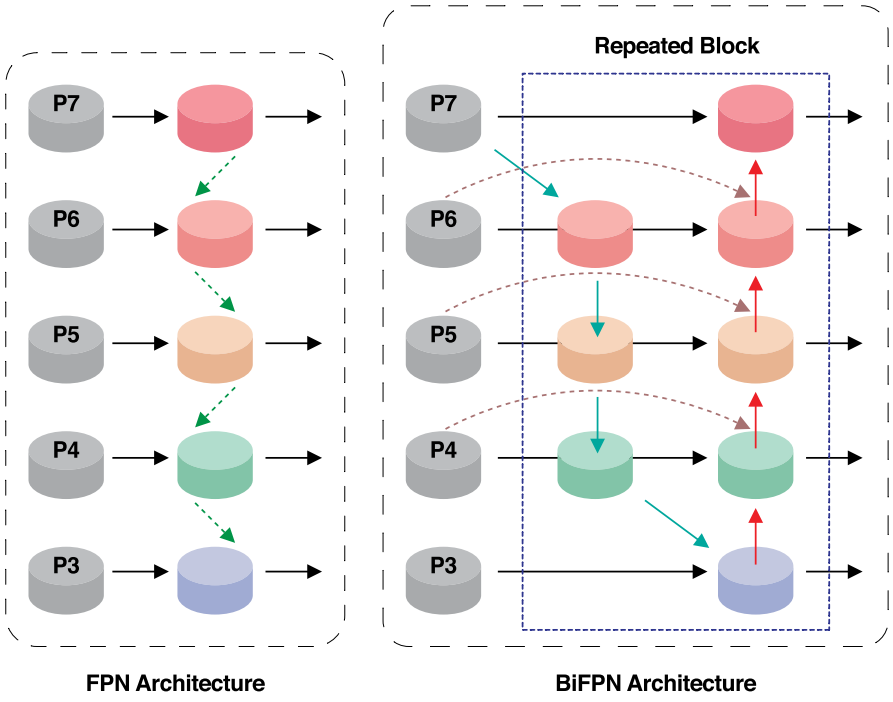

FIGURE 6. Schematic of bi-directional feature pyramid network. FPN incorporates a top-down pathway to combine multi-scale features from different levels (from level 3 to level 7, denoted as P3 to P7). Additionally, BiFPN is introduced, which offers improved accuracy-efficiency trade-offs compared to the original FPN approach.

highlighted the difficulty of adapting transformers to highly variable US datasets [62], [63].

Yang et al. achieved notable success in optimizing YOLOv3 for uterine fibroid detection by integrating Darknet-53 and EfficientNet. The resulting model reached an exceptional mAP of 98.38%, demonstrating the effectiveness of backbone refinement in stabilizing gradients and enhancing feature extraction. However, questions about the model's scalability to diverse datasets remain unresolved [64]. Similarly, Inui et al. leveraged YOLOv8 variants for detecting elbow Osteochondritis Dissecans (OCD), achieving mAPs of 99.4% and 99.5% with YOLOv8n and YOLOv8m, respectively. These results underscore YOLOv8's robustness for high-resolution detection tasks, although its generalizability to complex modalities remains to be explored [65].

Other studies explored YOLO-based architectures for specialized tasks. Ghabri et al. employed YOLOv5 for real-time thyroid nodule localization, achieving a mAP of 89.2%. While this model demonstrated practicality for deployment, its accuracy lagged behind more advanced YOLO versions, indicating room for improvement [66]. Kuo et al. utilized a scaled YOLOv4+HarDNet model for cardiac US imaging, achieving a mAP of 72.63%. Despite excelling at small feature detection, its performance was constrained by the complexity of noisy cardiac data [67]. Mahajan et al. combined YOLOv5s with a VGG16 backbone for OCD classification in the humeral capitellum, achieving over 95% mAP for detection and 89% accuracy for classification. This study illustrated the versatility of pairing YOLO models with traditional backbones for specific applications [68].

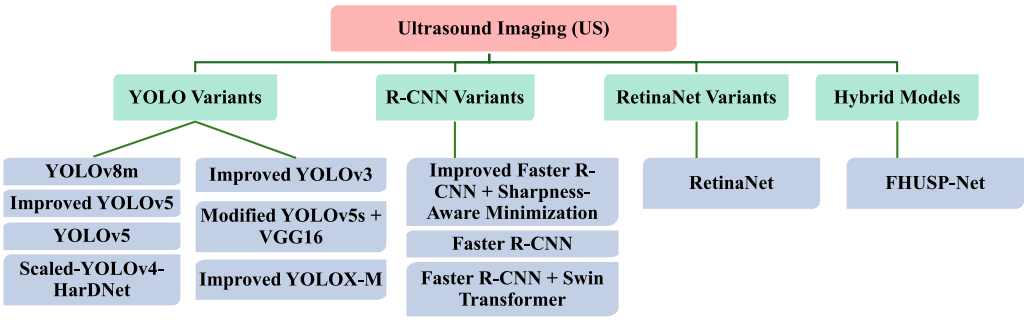

FIGURE 7. Schematic articulation of the recent progress in DL-based object detection methods desired to enrich diagnostic capacities in US imaging.

Beyond YOLO architectures, other approaches have also demonstrated potential. Daoud et al. fine-tuned RetinaNet for breast US analysis, achieving a mAP of 89.0% after preprocessing with contrast enhancement. While the model excelled in detection, its dependence on preprocessing highlighted the need for architectures that inherently adapt to diverse image quality [69]. Bassiouny et al. designed a system for detecting seven lung US features, where Faster R-CNN outperformed RetinaNet with a mAP of 86.57%. This comparative result emphasized the limitations of general-purpose detection models like RetinaNet when handling complex US data [70].

Task-specific architectures have also shown promise. Li et al. introduced FHUSP-NET for detecting fetal heart US standard planes, achieving a mAP of 95.5% by incorporating Spatial Pyramid Pooling (SPP) and SE networks. Despite its high accuracy, real-time efficiency and generalization across patient populations remain challenges [71]. Similarly, Yang et al. enhanced YOLOX-M for thyroid nodule detection by integrating Involution layers and deformable convolutions, achieving a mAP of 75.7%. Although effective, the model faced difficulties with complex US backgrounds. In another effort, Yang et al. improved YOLOv5 by adding Coordinate Attention (CA) and Label Smoothing Regularization (LSR), achieving a mAP of 95.3%. This approach balanced performance gains with minimal increases in inference time, offering practicality for clinical adoption [72].

These studies reveal key challenges in applying DL architectures for MOD in US imaging. Computational complexity limits real-time clinical use of models like the improved Faster R-CNN and Swin-Transformer. While transformer-based models show promise, they struggle with the variability of US data. YOLO variants like YOLOv8 perform well in high-resolution tasks but may not generalize effectively to complex US modalities. Task-specific models like FHUSP-NET excel in performance but face issues with real-time efficiency and patient population diversity. Preprocessing-dependent models like RetinaNet highlight the need for more adaptable architectures, while models using CA and LSR offer a balance of accuracy and efficiency but need further refinement for broader scalability.

E. OTHER MEDICAL IMAGING MODALITIES

As shown in Fig. 8, recent studies on the diagnostic capabilities in MOD, specifically in OCT, HPI, and Endoscopic techniques, have led to advancements in automated object detection. Jasitha et al. evaluated the effectiveness of Faster R-CNN and CenterNet for glomeruli detection in whole-slide images (WSI) and single-glomeruli patches for renal disease diagnosis. Their study demonstrated that Faster R-CNN achieved a mAP of 65.7% on WSI patches, offering promising results while also revealing the complexities of scaling object detection for large-scale histopathological images [73]. In comparison, Tyagi et al. introduced Guided Posterior Regularization (DeGPR), a model-agnostic technique aimed at enhancing object detection by focusing on discriminative cellular features. Their application of DeGPR to YOLOv5, Faster R-CNN, and EfficientDet resulted in mAP improvements of up to 9.0%, with YOLOv5 reaching the highest mAP of 78.7%. This illustrates the potential of feature-guided regularization in improving detection accuracy, although the increased computational demands present a significant challenge [74].

The application of YOLO models in MOD has also seen significant success. He et al. developed a YOLO-based model for detecting NPC during endoscopy, achieving an impressive mAP of 97.7% and a fast inference speed of 52.9 FPS. These results highlight the effectiveness of YOLO models in real-time medical applications. However, further validation is needed to confirm their robustness across diverse patient populations and Endoscopic imaging conditions [75]. In a similar vein, Aliyi et al. implemented a YOLOv5-based system for colorectal cancer detection in colonoscopy images, achieving a mAP of 98.8%. This model benefited from transfer learning and data augmentation strategies, underscoring the importance of optimal training approaches to mitigate challenges posed by small datasets. However, generalizing the model to diverse clinical settings remains an ongoing challenge [76].

Further advancements have been made through the optimization of YOLO models for specific medical tasks. Carrinho et al. enhanced YOLOv4 for polyp detection by integrating regularization techniques, custom anchor

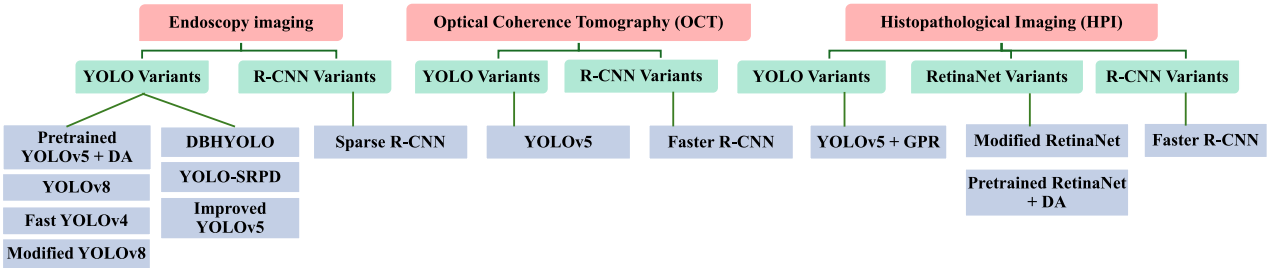

FIGURE 8. Schematic representation of the recent headway in DL-based object detection techniques aimed to increase diagnostic capabilities in HPI, Endoscopy, and OCT imaging. DA denotes Data Augmentation and GPR denotes Guided Posterior Regularization.

boxes, and advanced loss functions, achieving mAP values of 82.93% and 90.96% across two datasets. While these results demonstrate the model's effectiveness, the variability between datasets indicates a need for more robust approaches to handle imaging heterogeneity [77]. Similarly, Fan et al. evaluated YOLOv5, Faster R-CNN, and RetinaNet for detecting dental caries in OCT images. Both YOLOv5 and Faster R-CNN achieved an mAP of 86.0%, while RetinaNet performed slightly worse with a mAP of 75.0%. This highlights the unique challenges of adapting standard object detection models to OCT data, which often contains subtle and irregular features [78].

In addition to YOLO-based systems, other innovative methods have also contributed to the advancement of MOD. Chi et al. introduced a Sparse R-CNN-based method, enhanced by a Multiscale Detail Enhancement Pyramid Network (MDEPN), for detecting esophageal diseases on Lugol chromoendoscopy images, achieving a mAP of 65.0%. Although this represents a modest improvement, it reflects the inherent difficulty of detecting lesions in chromoendoscopy images and underscores the need for more specialized detection architectures [79]. Likewise, Prabhu et al. enhanced RetinaNet with an attention mechanism for detecting keratin pearls in squamous cell carcinoma, achieving a mAP of 92.99%. This study highlights the value of attention mechanisms in improving the detection of small, irregular objects, often difficult to identify in HPI [80].

Integrating multiple technologies has proven to be an effective strategy for improving detection performance. Wang et al. introduced YOLO-SRPD, which combines Super-Resolution Generative Adversarial Networks (SRGAN) with YOLOv5 for polyp detection in low-resolution colonoscopy images. By incorporating Attention-ConvMix (ACmix), CBAM modules, and $C3_{Res}2Net$, their model achieved a mAP of 94.2%. This demonstrates the potential of super-resolution techniques in improving performance in low-quality medical imaging scenarios [81]. Pan et al. developed DBH-YOLO, an end-to-end system for surgical instrument detection, achieving mAP values of 96.8%, 95.6%, and 98.4% across three datasets. The integration of a Dual Branch Head (DBH) and Overall-IoU loss contributed to enhanced detection precision, although further exploration is required to evaluate its robustness in diverse surgical contexts [82].

Additionally, the incorporation of attention mechanisms into object detection models has proven to enhance model performance. Yan et al. improved YOLOv8 for bronchi detection in bronchoscopy images by integrating a CBAM module and Feature Pyramid-Path Aggregation Network (PAFPN) (Fig. 9), which resulted in an increase in mAP from 87.09% to 88.27%. This highlights the potential of attention mechanisms in refining predictions, especially for complex regions in MOD [83]. Similarly, Liang et al. integrated advanced attention mechanisms and ConvNeXt blocks into a YOLOv5-based system for detecting ureteral orifices, achieving a mAP of 89.6%. This approach benefited from BiFormer and Wise-IoU loss, enabling the model to capture long-range dependencies and improve box localization. However, the added complexity raises questions regarding its feasibility in resource-constrained settings [84].

While improvements in DL-based object detection for OCT, HPI, and Endoscopic imaging methods have made substantial progress, challenges remain in scaling models for large, complex datasets, improving generalization across diverse clinical settings, and managing computational overhead. Techniques such as feature-guided regularization, attention mechanisms, and the integration of super-resolution technologies have demonstrated great promise in enhancing model performance across a variety of medical applications. However, further research is needed to address the variability in results and ensure robustness across different patient populations, imaging conditions, and medical contexts.

III. RESULTS

A. X-RAY IMAGING

As summarized in Table 1, recent progress in DL has markedly boosted the diagnostic accuracy of X-Ray imaging, though challenges related to variability, scalability, and practical deployment persist. In the context of lesion detection, The CovidXDetector, which employs a Cascade R-CNN, achieves moderate accuracy for COVID-19 detection in chest radiographs with a mAP of 0.642, yet its generalizability to other pathologies remains insufficient.

In the domain of bone and fracture analysis, Hybrid YOLO-NAS exhibits exceptional accuracy (mAP: 0.991) in

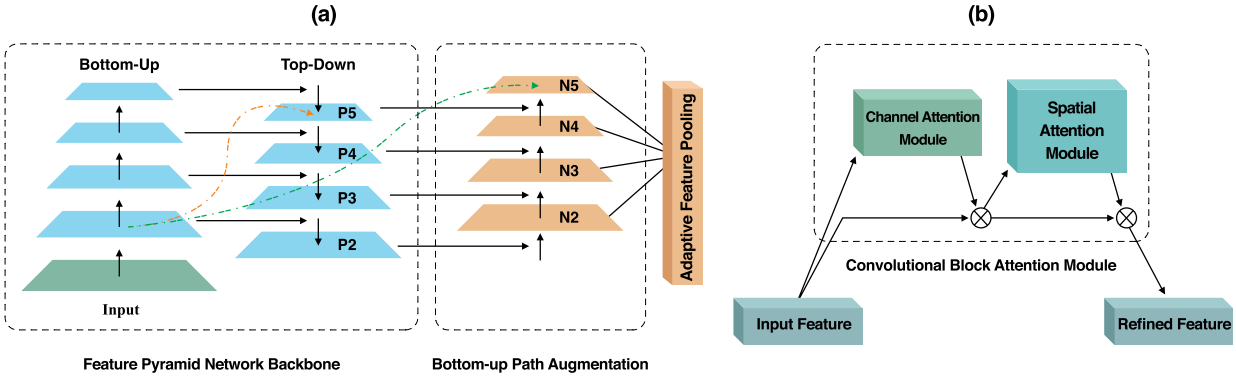

FIGURE 9. Schematic of feature pyramid-path aggregation network and convolutional block attention module: (a) PAFPN is a feature pyramid module used in path aggregation networks that combines FPNs with bottom-up path augmentation, which shortens the information path between lower layers and topmost features. FPN backbone that generates feature maps P_i and Bottom-up path augmentation module that generates feature maps N_i from the corresponding P_i feature maps. (b) CBAM has two sequential sub-modules: channel and spatial. The intermediate feature map is adaptively refined through this module at every convolutional block of deep networks.

detecting hand bone fractures by integrating YOLO-NAS, EfficientDet, and transformer-based attention mechanisms. This model achieves a balance between computational efficiency and performance through advanced feature fusion techniques. For wrist abnormalities, as shown in Fig. 10, YOLOv8m delivers high accuracy (mAP: 0.950) by leveraging Cross-Stage Partial (CSP), SPP, and a self-attention-based anchor-free design. However, in detecting pediatric wrist fractures, YOLOv8-AM (mAP: 0.658) and YOLOv9 (mAP: 0.654) achieve only moderate performance, highlighting the need for improved architectures to address the complexity of such cases.

For other fracture detection tasks, MPFracNet, a RetinaNet-based model, performs effectively in detecting metacarpophalangeal fractures, achieving a mAP of 0.804 through the integration of deformable bottleneck blocks and advanced feature fusion modules, albeit at the cost of high computational demands. In detecting thighbone fractures, the Enhanced Pyramid Vision Transformer (Enhanced-PVT) delivers reliable performance (mAP: 0.870) by employing scale-aware and spatial-aware attention mechanisms alongside overlapping patch embeddings.

In other applications, YOLOv5 demonstrates promise in diagnosing DDH with a performance score of 0.831 by incorporating multimodal data integration, SPP, and Path Aggregation Networks (PAN). For mammography breast mass detection, ERetinaNet surpasses YOLO-based methods with an accuracy of 0.850, utilizing Faster RepVGG and Feature Pyramid Networks (FPN) to improve detection accuracy, though at the expense of increased computational complexity. In dental X-Ray analysis, YOLOv8m achieves moderate performance (mAP: 0.716) but faces challenges in addressing intricate dental tasks. Similarly, Modified YOLOv5 performs reasonably well (mAP: 0.769) for pressure ulcer detection, supported by data augmentation techniques.

Emerging architectures, including transformer-integrated detectors and hybrid models, are establishing new benchmarks for accuracy and efficiency. Hybrid YOLO-NAS, which combines YOLO-NAS with DETR, exemplifies precision and adaptability. The YOLOv4+ViT model, specifically tailored for breast cancer detection and classification, achieves remarkable accuracy (0.986) by integrating CSP-Darknet53 with ViT. Advanced attention mechanisms, such as SPP and PAN, further enhance multi-scale feature representation, making this model a robust solution for complex mammography analyses.

B. COMPUTERIZED TOMOGRAPHY (CT)

Based on Table 2, in the field of MOD, advanced DL architectures have shown significant promise across various tasks involving CT images. For example, the integration of Faster R-CNN with Mask R-CNN has resulted in impressive diagnostic performance for detecting COVID-19 in CT scans, achieving a high mAP of 0.977. These models utilize robust region proposals and precise segmentation pipelines, with residual networks serving as effective backbones. While these results highlight the models' success in diagnosing COVID-19, further investigation is necessary to evaluate their applicability to other disease-specific scenarios.

In lung-specific applications, YOLO-based architectures have shown exceptional performance. For example, ELCT-YOLO, designed for lung tumor detection, achieves a remarkable mAP of 0.971 by combining the YOLOv7-tiny backbone with the Extended Efficient Layer Aggregation Network (E-ELAN), Spatial Pyramid Pooling Fast (SPPF), a decoupled neck, and a cascaded refinement scheme. Similarly, for lung nodule detection, YOLO-MSRF, an enhanced YOLOv8 variant, incorporates Multi-Scale Receptive Fields (MSRF) and Efficient Omni-dimensional Convolution (EODConv) to improve small object detection, achieving a mAP of 0.946. Another noteworthy approach is SwiF-YOLO, which builds on the YOLOx architecture by integrating a Swin-Transformer backbone with ASFF, resulting in a mAP of 0.831. Despite their high efficacy in

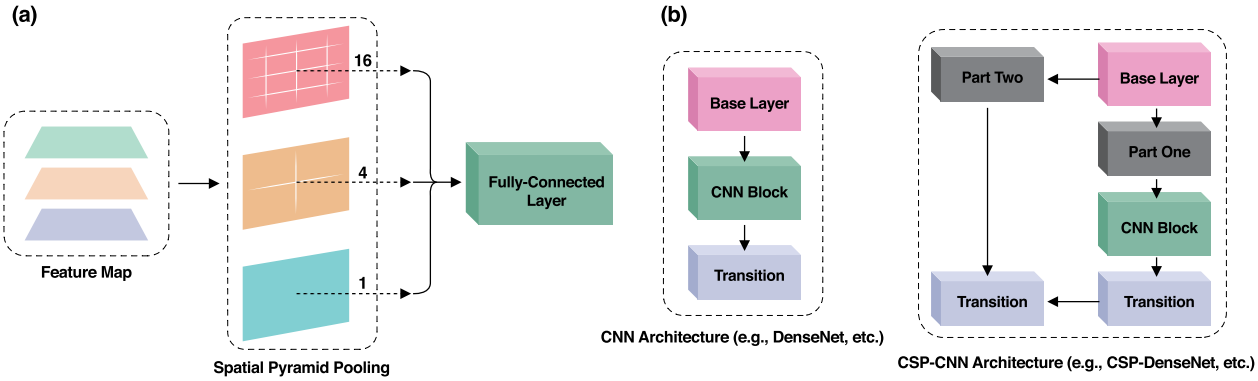

FIGURE 10. Schematic of spatial pyramid pooling and cross-stage partial network: (a) SPPNet captures spatial information by partitioning the input feature maps into grids of varying sizes (e.g., 16(4 × 4), 4(2 × 2), 2(1 × 1)) and applying average pooling within each region. This multi-scale spatial pooling preserves the spatial relationships in the input while providing a fixed-length feature representation. (b) CSPNet separates the feature map of the base layer into two parts, one part will go through a CNN block and a transition layer; the other part is then combined with the transmitted feature map to the next stage.

lung-specific applications, these YOLO-based models often struggle to generalize to tasks involving other anatomical regions.

Broadening the scope to other anatomical contexts, YOLOv5 has shown adaptability in detecting cervical spine bones in CT images, achieving a mAP of 0.930. This model employs a pretrained EfficientNet backbone alongside SPP for enhanced feature representation. In fracture detection, CenterNet-based networks have achieved notable success in handling complex tasks such as rib and mandibular fracture detection. For example, CenterNet, combined with a hierarchical fusion hourglass network and multi-branch residual blocks, achieved mAP scores of 0.890 for rib fractures and 0.893 for mandibular fractures. These models are further enhanced by advanced components such as non-local dual spatial attention mechanisms and corner point prediction modules, which enable them to effectively address challenges posed by occlusions and anatomical deformations.

C. MAGNETIC RESONANCE IMAGING (MRI)

As summarized in Table 3, one-stage object detection models, including YOLOv5 through YOLOv8, have demonstrated remarkable performance across a variety of MOD tasks using MRI modalities, achieving mAP values ranging from 0.801 to 0.990. These models have been successfully employed to detect lesions such as spinal cord injuries, brain tumors, kidney abnormalities, prostate cancer, and lumbar disc herniation. For instance, in prostate cancer detection, a YOLOv8s variant enhanced with CSPDarknet53, SPP, Efficient Channel Split (ECS), and PAFPN, combined with self-attention mechanisms, achieved an impressive mAP of 0.990, enabling real-time detection capabilities. Similarly, brain tumor detection has benefited from refined YOLOv7 models, which utilize data augmentation, CBAM, and advanced feature fusion techniques such as SPPF+ and BiFPN, achieving mAP values as high as 0.989. Further advancements, such as the BGF-YOLO model, which integrates multiscale attentional feature fusion and an additional detection head, have achieved mAP values of 0.974.

In spinal lesion analysis, Faster R-CNN, equipped with VGG16 and ResNet-50 backbones, has demonstrated efficacy in localizing spinal injuries with a mAP of 0.886. Fine-tuned YOLOv5x models with transfer learning and modified anchor systems have performed slightly better, achieving a mAP of 0.893 for lumbar disc herniation detection. Kidney lesion detection tasks have also seen significant improvements with modified YOLOv7 architectures incorporating extended E-ELAN and dynamic label assignment strategies, reaching a mAP of 0.950.

Modifications to the YOLO family, including attention mechanisms, transformer blocks, and specialized prediction heads, have proven effective in detection capabilities. For example, an improved version of YOLOv5, which incorporates additional prediction heads and Swin-Transformer blocks, achieved a mAP of 0.612 in detecting brain metastases.

While YOLO models provide an optimal balance between efficiency and accuracy, more complex architectures, such as transformer-based detectors, have demonstrated superior performance in tasks requiring advanced feature extraction and attention mechanisms. For example, transformer-based models excel in rib fracture identification by leveraging sophisticated attention mechanisms and multiscale feature representation. However, these performance gains often come at the cost of higher computational requirements, underscoring a trade-off between precision and resource efficiency in MOD tasks.

D. ULTRASOUND IMAGING (US)

As shown in Table 4, in thyroid nodule detection, advancements in object detection models have demonstrated significant potential. An improved Faster R-CNN model, leveraging a ResNet-50 backbone with deformable convolution, an FPN,

TABLE 1. Summary of the latest advances in DL-based MOD methodologies on X-Ray images.

Algorithm	Modality	Lesion	Dataset (Size)	Backbone	Neck	Head	mAP
Modified YOLOv5 + Data Augmentation (2023)	X-Ray	Pressure Ulcers	Custom (200)	Modified EfficientNet	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Spatial Pyramid Pooling (SPP) + Modified Anchor System	0.769
Modified RetinaNet (2023)	X-Ray	Chest Abnormality	Custom (15000)	ResNet101	Feature Pyramid Network (FPN) with Top-Down Pathway and Lateral Connections	RetinaNet Head with Focal Loss	0.550
Cascade R-CNN (2024)	X-Ray	COVID-19	COVID-19 CXR (6334)	Pre-trained Convolutional Neural Network (CNN)	Feature Pyramid Network (FPN) / Spatial Pyramid Pooling (SPP)	Region Proposal Network (RPN) + Region of Interest (ROI) Pooling	0.642
Hybrid YOLO-NAS	X-Ray	Hand Bone Fractures	Custom (4736)	YOLO NAS + EfficientDet	Feature Fusion	YOLO NAS + DETR Transformer + Attention	0.991
YOLOv8m (2024)	X-Ray	Wrist Abnormality	GRAZPEDWRI-DX (20327)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP) and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.950
YOLOv5 + Eigen-CAM Saliency Maps (2024)	X-Ray	Breast Cancer	Custom (278)	Pre-trained Convolutional Neural Network (CNN)	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	YOLOv3 Head	0.621
YOLOv9 + Data Augmentation (2024)	X-Ray	Wrist Fracture	GRAZPEDWRI-DX (20327)	Pre-trained Convolutional Neural Network (CNN)	Feature Pyramid Network (FPN) / Spatial Pyramid Pooling (SPP)	Convolutional Layers (Anchor-Free) with Self-Attention	0.654
YOLOv8 + Data Augmentation (2024)	X-Ray	Wrist Fracture	GRAZPEDWRI-DX (20327)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.638
YOLOv8m (2024)	X-Ray	Teeth Abnormality	Custom (316)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP) and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.716
YOLOv5 + Integrated Multimodal Information (2024)	X-Ray	Developmental Dysplasia of the Hip	Custom (7750)	Pre-trained Convolutional Neural Network (CNN)	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	YOLO Head (Anchor-Free)	0.831
YOLOv8 + Attention (2024)	X-Ray	Wrist Fracture	GRAZPEDWRI-DX (20327)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.658
YOLOv4 + ViT (2024)	X-Ray	Breast Cancer	INbreast (410)	CSPDarknet53	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	YOLOv3 Head	0.986
YOLOv7 + Attention (2024)	X-Ray	Bone Fracture	Science Research 2022 (1001)	Pre-trained Convolutional Neural Network (CNN)	Extended Efficient Layer Aggregation Network (E-ELAN)	Lead Head + Auxiliary Heads + Dynamic Label Assignment	0.862
ERetinaNet + ViT (2024)	X-Ray	Breast Mass	DDSM+MIAS (1930)	Faster RepVGG (FRepVGG)	Feature Pyramid Network (FPN)	Simplified RetinaNet Head	0.850
RetinaNet-based MPFracNet (2024)	X-Ray	Metacarpophalangeal Fractures	Custom (6268)	ResNet with Deformable Bottleneck Block (DBB)	Integrated Feature Fusion Module (IFFM)	RetinaNet Head with Focal Loss	0.804
Enhanced Pyramid Vision Transformer (PVT) (2024)	X-Ray	Thighbone Fracture	Custom (4000)	Pre-Trained Pyramid Vision Transformer (PVT) + Overlapping Patch Embedding (OPE)	Scale-Aware Attention + Spatial-Aware Attention	Simple Neural Network Layer	0.870

and an RPN, achieved a high mAP of 0.974, showcasing its effectiveness in detecting irregularly shaped thyroid nodules. Similarly, enhanced YOLOv5 models integrated with EfficientNet and coordinate attention mechanisms yielded a mAP of 0.953, offering a user-friendly solution for automated thyroid screening. In contrast, while promising for multiscale feature extraction, a Swin-Transformer-based Faster R-CNN achieved a comparatively lower mAP of 0.448. Further refinement in detection performance was observed with an improved YOLOX-M model incorporating involution modules and deformable convolution, which attained a mAP of 0.757.

In orthopedic applications, particularly for elbow osteochondritis dissecans (OCD) detection, YOLOv8m excelled with an exceptional mAP of 0.995, utilizing CSPDarknet53, a PAFPN, and self-attention mechanisms for enhanced feature extraction. A modified YOLOv5s model, augmented with VGG16, also performed well, achieving a mAP of 0.950.

For uterine fibroid detection, an improved YOLOv3 model integrated Darknet53 with EfficientNet to optimize feature extraction, resulting in a high mAP of 0.983. In breast tumor detection, RetinaNet fine-tuned with pretrained CNNs and enhanced through contrast and sharpening techniques achieved a mAP of 0.890, effectively identifying ROI.

In the detection of coronary lesions associated with Kawasaki disease, the scaled YOLOv4-HarDNet model, featuring a CSPHarDNet backbone along with FPN and PAN modules, demonstrated strong performance in identifying small objects, achieving a mAP of 0.726. In neonatal lung ultrasound imaging, Faster R-CNN outperformed RetinaNet, achieving a mAP of 0.865 compared to RetinaNet's 0.611. Both models were focused on detecting lung sliding and extracting relevant features.

TABLE 2. Summary of the recent improvements in DL-based MOD methods on CT images.

Algorithm	Modality	Lesion	Dataset (Size)	Backbone	Neck	Head	mAP
Hybrid R-CNN (Faster + Mask) (2023)	CT	COVID-19	YBUFM (42342)	ResNet-50	Region Proposal Network (RPN)	Faster R-CNN Head + Mask Branch	0.977
ELCT-YOLO (2023)	CT	Lung Tumor	Lung-PET-CT-Dx (2324)	YOLOv7-tiny + Extended Efficient Layer Aggregation Network (E-ELAN) + Spatial Pyramid Pooling-Fast (SPPF)	Decoupled Neck (DENeck) with Cascaded Refinement Scheme (CRS)	YOLO Head (Anchor-Free)	0.971
CenterNet (2023)	CT	Chest Rib Fracture	Custom (13230)	Pre-trained Convolutional Neural Network (CNN)	Hierarchical Fusion Hourglass Network + Multi-Branch Residual Blocks	Heatmap Pyramid + Non-Local Dual Spatial Attention + Corner Point Prediction	0.890
YOLOv7 + Transfer Learning with VGG16 (2024)	CT	Lung Nodules	LIDC-IDRI (1148)	Pre-trained Convolutional Neural Network (CNN)	Extended Efficient Layer Aggregation Network (E-ELAN)	YOLO Head with Coarse-to-Fine Label Assignment	0.812
YOLO-MSRF (2024)	CT	Lung Nodules	Luna16 (1186)	Improved YOLOv8	Multiscale Receptive Field (MSRF)	Small Object Detection Layer (SODL) + Efficient Omnidimensional Convolution (EODConv)	0.946
YOLOv5 (2024)	CT	Cervical Spine Bone	CSFDC (9170)	Pre-trained EfficientNet	Spatial Pyramid Pooling (SPP)	YOLO Head (Anchor-Free)	0.930
SwiF-YOLO (2024)	CT	Lung Nodules	Luna16 (888)	Swin-Transformer	Adaptively Spatial Feature Fusion (ASFF)	YOLOx-m Head with Generalized Intersection Over Union (GIoU) loss	0.831
MCK-CenterNet (2024)	CT	Mandibular Fracture	Custom (5861)	Pre-trained Convolutional Neural Network (CNN)	Multi-scale Deformable Recalibration Module	Parallel Attention Enhancement Residual Module + Center Prediction and Classification Head	0.893

In fetal heart ultrasound imaging, the FHUSP-NET model displayed robust performance, achieving a mAP of 0.955. This model employed a pretrained CNN, SPP, and multi-task learning to accurately recognize standard planes and detect key anatomical structures, contributing significantly to precise fetal heart evaluation.

E. OTHER MEDICAL IMAGING MODALITIES

As shown in Table 5, in the domains of HPI, OCT, and endoscopy, YOLOv5 consistently displays outstanding performance and adaptability. For example, it achieved an impressive mAP of 0.988 when detecting gastrointestinal disorders from colonoscopy images by employing transfer learning and data augmentation techniques. Additionally, YOLOv5 proved effective in detecting dental caries from OCT images, attaining a mAP of 0.860. In the detection of celiac disease, combining YOLOv5 with DeGPR resulted in a mAP of 0.787, showcasing its versatility across various medical applications.

Advancements with other YOLO-based models have also yielded significant results. For example, YOLOv8, the successor to YOLOv5, achieved a mAP of 0.977 for real-time detection of NPC during nasopharyngeal endoscopy, leveraging advanced architectural features such as CSP and ECS. Additionally, YOLO-SRPD, which integrates super-resolution reconstruction techniques, reported a mAP of 0.942 for colorectal polyp detection, demonstrating the potential of hybrid approaches in improving detection accuracy.

When compared to other models, YOLOv5 consistently outperforms competitors in terms of speed and versatility. For example, YOLOv5 surpassed Faster R-CNN in several tasks, including glomeruli detection, where Faster R-CNN achieved a mAP of 0.657. Although Faster R-CNN matched YOLOv5's performance in dental caries detection (mAP: 0.860), YOLOv5's efficiency makes it more suitable for real-time applications. RetinaNet, another competitive model, showed promise in specific histopathology imaging tasks such as ovarian follicle detection, achieving a mAP of 0.830, but it was outperformed by YOLOv5 in tasks like celiac disease detection.

Sparse R-CNN, evaluated for esophageal lesion detection, demonstrated a lower baseline performance with a mAP of 0.650. However, the application of innovative mechanisms such as MDEPN for this task suggests potential pathways for enhancing its capabilities. DeGPR, a generalized post-processing refinement technique, has also emerged as a significant enhancement across multiple models, including YOLOv5, Faster R-CNN, and EfficientDet, contributing to improved performance in tasks such as celiac disease detection.

IV. DISCUSSION

Since 2023, the field of MOD has seen significant advancements, driven particularly by DL architectures that have revolutionized medical imaging. The advent of hybrid models has further accelerated progress, combining diverse architectural strengths to achieve remarkable outcomes. For example, the hybrid YOLO-NAS model achieved an impressive mAP of 0.991 in detecting hand bone fractures in X-Ray images, leveraging the capabilities of the YOLO framework alongside neural architecture search. However, the computationally intensive nature of such techniques restricts their accessibility in resource-limited settings. Future research should prioritize the development of lightweight and energy-efficient architectures to make such models feasible

TABLE 3. Summary of the latest advancements in DL-based MOD models on MRI images.

Algorithm	Modality	Lesion	Dataset (Size)	Backbone	Neck	Head	mAP
Refined YOLOv7 + Data Augmentation (2023)	MRI	Brain Tumor	BT Kaggle (10288)	Pre-trained Convolutional Neural Network (CNN) + Convolutional Block Attention Module (CBAM)	Spatial Pyramid Pooling Fast Plus (SPPF+) + Bi-directional Feature Pyramid Network (BiFPN)	Decoupled Heads	0.989
YOLOv4-tiny + Transfer Learning (2023)	MRI	Brain Tumor	Figshare (3064)	Pre-trained Convolutional Neural Network (CNN)	-	YOLOv4 Head	0.832
Modified YOLOv7 (2023)	MRI	Kidney Lesion	Custom (5657)	Pre-trained Convolutional Neural Network (CNN)	Extended Efficient Layer Aggregation Network (E-ELAN)	Lead Head + Auxiliary Heads + Dynamic Label Assignment	0.950
PP-YOLOv2 + Transfer Learning (2023)	MRI	Spinal Lesion	Custom (604)	Pre-trained Convolutional Neural Network (CNN)	Path Aggregation Network (PAN)	Region Proposal Network (RPN) + Mish Activation Function (MAF)	0.900
Improved YOLOv5 (2023)	MRI	Brain Metastasis	Custom (2269)	Pre-trained EfficientNet	Convolutional Block Attention Model (CBAM)	Additional Prediction Head + Swin-Transformer Block	0.612
YOLOv8s (2023)	MRI	Brain Cancer	YOLO-CVP (300)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.941
YOLOv8s (2023)	MRI	Prostate Cancer	Custom (3585)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.990
YOLOv8 (2023)	MRI	Brain Tumor	BR35h (801)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.976
YOLOv5x + Data Augmentation (2023)	MRI	Lumbar Disc Herniation	Lumbar Spine MRI (550)	Pre-trained EfficientNet	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Spatial Pyramid Pooling (SPP) + Modified Anchor System	0.893
Modified YOLO (2023)	MRI	Brain Tumor	Custom (1420)	Pre-trained EfficientNet	Path Aggregation Network (PAN)	Simplified YOLO Head	0.856
BGF-YOLO (2023)	MRI	Brain Tumor	Br35H (801)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Generalized Feature Pyramid Networks (GFPN)	Bi-level Routing Attention (BRA) + Fourth Detecting Head	0.974
YOLOv5 + Transfer Learning (2023)	MRI	Brain Tumor	BraTS21 (1251)	Pre-trained EfficientNet	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Spatial Pyramid Pooling (SPP) + Modified Anchor System	0.972
YOLOv7 (2023)	MRI	Breast Lesions	DCE-MRI (2400)	Pre-trained Convolutional Neural Network (CNN)	Extended Efficient Layer Aggregation Network (E-ELAN)	Lead Head + Auxiliary Heads + Dynamic Label Assignment	0.985
MWSR-YLCA (2023)	MRI	Nasopharyngeal Carcinoma	Custom (26000)	Pre-trained Convolutional Neural Network (CNN) + Multi-Window Settings Resampling (MWSR)	Extended Efficient Layer Aggregation Network (E-ELAN)	YOLOv7 Head + Coordinate Attention Mechanism (YLCA)	0.801
Modified Faster R-CNN (2024)	MRI	Spinal Lesion	Custom (1500)	VGG16/ResNet-50	Region Proposal Network (RPN)	Faster R-CNN Head	0.886

for broader deployment [15]. Similarly, the YOLOv4+ViT model demonstrated a high mAP of 0.986 for breast cancer detection in mammography images, highlighting the potential of integrating YOLOv4 with transformers. Nevertheless, transformer-based models often encounter challenges such as high computational overhead and memory demands, hindering their feasibility for real-time deployment. Addressing these challenges through efficient transformer designs and memory optimization techniques could pave the way for their wider adoption in clinical workflows.

The YOLOv8m model has shown notable performance, achieving a mAP of 0.950 for wrist abnormality detection on X-Ray images. As shown in Fig. 11, this success is attributed to its innovative architecture, including distribution focal loss, completed-IoU loss, and binary cross-entropy for classification, which optimizes the detection of smaller objects. However, the model's reliance on balanced datasets highlights its limitations in handling imbalanced or occluded objects. Future work should explore advanced data augmentation strategies and adaptive loss functions to mitigate these limitations. Similarly, the Enhanced-PVT model demonstrated its versatility with a mAP of 0.870 for thighbone fracture detection in X-Ray images, emphasizing the promise of transformer-based architectures. Yet, like other transformers, its susceptibility to overfitting on small datasets remains a challenge. Research into transfer learning and pretraining on larger, diverse datasets may help overcome this issue [85].

The YOLO family continues to dominate MOD tasks with consistent results across applications. YOLOv7 and its attention-enhanced variant achieved a mAP of 0.862 for bone fracture detection, while YOLOv5 with integrated multimodal information reached a mAP of 0.831 for DDH. While resource-efficient, lightweight models such as YOLOv7-tiny may compromise accuracy for complex lesions, as seen in ELCT-YOLO's performance (mAP: 0.971) for CT lung tumor detection. In MRI, advanced models have pushed performance boundaries. YOLOv8s, equipped with CSPDarknet-53 and tailored enhancements, achieved a mAP of 0.990 for prostate cancer detection. Similarly, refined YOLOv7 reached a mAP of 0.989 for brain tumor detection, solidifying its competitive edge. However, these

TABLE 4. Summary of the recent improvements in DL-based MOD approaches on US images.

Algorithm	Modality	Lesion	Dataset (Size)	Backbone	Neck	Head	mAP
Improved Faster R-CNN + Sharpness-Aware Minimization (SAM) (2023)	Ultrasound	Thyroid Nodule	Custom (6261)	ResNet50 + Deformable Convolution	Feature Pyramid Network (FPN)	Region Proposal Network (RPN)	0.974
Faster R-CNN + Swin-Transformer (2023)	Ultrasound	Thyroid Nodule	Custom (3853)	Swin-Transformer	Feature Pyramid Network (FPN)	Detection Head	0.448
Improved YOLOv3	Ultrasound	Uterine Fibroids	Custom (871)	Darknet53	-	EfficientNet	0.983
YOLOv8m (2023)	Ultrasound	Elbow Osteochondritis Dissecans	Custom (2430)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Feature pyramid pool aggregate network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.995
YOLOv5 (2023)	Ultrasound	Thyroid	DDTI (347)	EfficientNet	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Spatial Pyramid Pooling (SPP) + Modified Anchor System	0.892
Scaled-YOLOv4-HarDNet (2023)	Ultrasound	Coronary Lesions	Custom (1283)	CSPHarDNet	Feature Pyramid Network (FPN) + Path Aggregation Network (PAN)	YOLOv3 Head	0.726
RetinaNet (2023)	Ultrasound	Breast Tumor	Custom (380)	Pre-trained Convolutional Neural Network (CNN)	Feature Pyramid Network (FPN) + Top-Down Pathway and Lateral Connection	Simplified RetinaNet Head	0.890
Faster R-CNN (2023)	Ultrasound	Neonatal Lung	Custom (26632)	Pre-trained Convolutional Neural Network (CNN)	-	Region Proposal Network (RPN)	0.865
RetinaNet (2023)	Ultrasound	Neonatal Lung	Custom (26632)	Pre-trained Convolutional Neural Network (CNN)	Feature Pyramid Network (FPN) + Top-Down Pathway and Lateral Connection	Simplified RetinaNet Head	0.611
Improved YOLOX-M (2024)	Ultrasound	Thyroid Nodule	Custom (3042)	Pre-trained Convolutional Neural Network (CNN)	Involution Module	Deformable Convolution + Attention	0.757
FHUSP-NET (2024)	Ultrasound	Fetal Heart	Custom (3360)	Pre-trained Convolutional Neural Network (CNN)	Spatial Pyramid Pooling (SPP) + Fully-Connected Spatial Pyramid Convolution (FSPC)	FHUSP Recognition + Key Anatomical Structure Detection	0.955
Improved YOLOv5 (2024)	Ultrasound	Thyroid Nodule	Custom (191)	Pre-trained Convolutional Neural Network (CNN)	Coordinate Attention (CA	Label Smoothing Regularization (LSR)	0.953
Modified YOLOv5s + VGG16 (2024)	Ultrasound	Elbow Osteochondritis Dissecans	Custom (5328)	CSPDarknet53	Path Aggregation Network (PAN)	YOLOv3 Head + VGG16	0.950

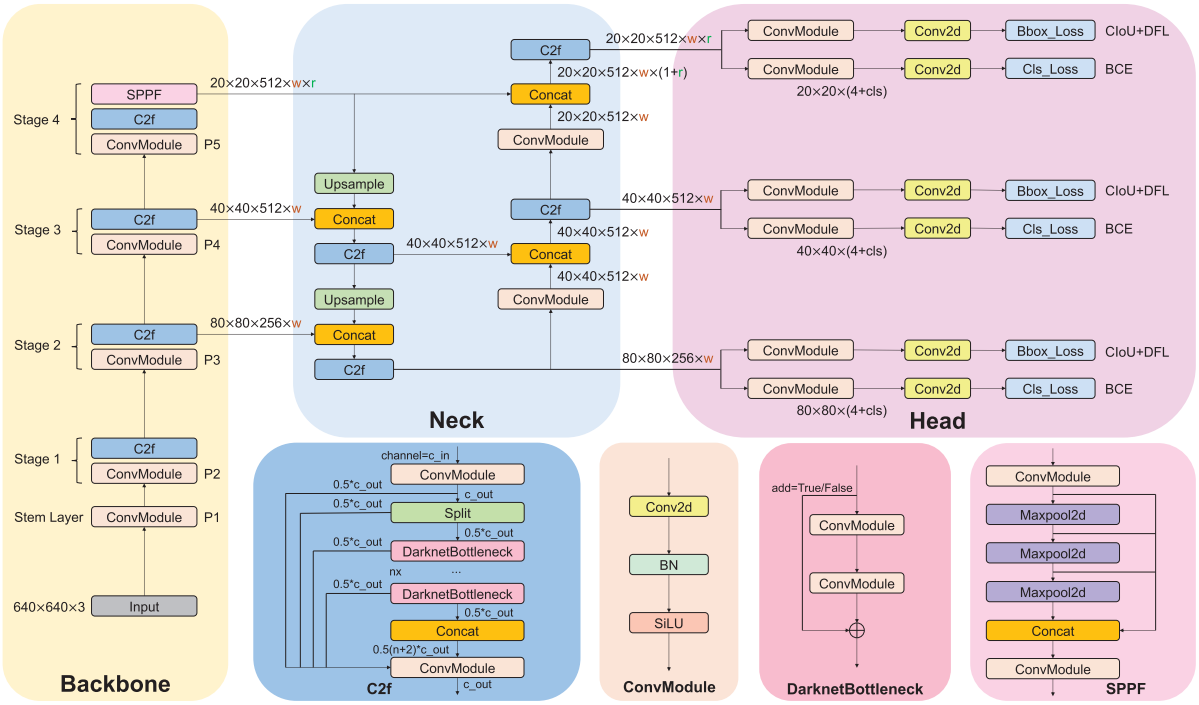

FIGURE 11. The architecture of the YOLOv8 model. This structure for efficient object detection includes several key components: ConvModule for feature extraction with convolution operations, Batch Normalization, and Sigmoid Linear Unit, activation; Cross-Stage Partial with Fusion (C2F) for improved gradient flow by processing and concatenating parts of the feature map through a DarknetBottleneck block; Spatial Pyramid Pooling Fast (SPPF) for detecting objects at various sizes by aggregating features at multiple scales with max pooling; and DarknetBottleneck for deeper feature extraction with residual connections that address the vanishing gradient problem.

models often struggle with variations in image quality and resolution, affecting robustness in clinical scenarios. Future research could address this through adaptive preprocessing pipelines and model training on diverse, real-world datasets.

TABLE 5. Summary of the latest progress in DL-based MOD techniques on OCT, HPI, and Endoscopy modalities.

Algorithm	Modality	Lesion	Dataset (Size)	Backbone	Neck	Head	mAP
YOLOv5 + Transfer Learning + Data Augmentation (2023)	Endoscopy	Lower Gastrointestinal Tract (LGIT)	Custom (6322)	Pre-trained Convolutional Neural Network (CNN)	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Spatial Pyramid Pooling (SPP) + Modified Anchor System	0.988
DBHYOLO (2024)	Endoscopy	Surgical Instrument	m2cai16-tool+Onyeogulu (6274)	CSPDarknet53	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Dual-Branched Head (DBH)	0.984
YOLOv8 (2023)	Endoscopy	Nasopharyngeal Carcinoma	Custom (2429)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP), and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention	0.977
YOLO-SRPD (2024)	Endoscopy	Colorectal Polyps	NSBEID (28000)	$C3-Res2Net$ + SRGAN with ACmix	Convolutional Block Attention Module (CBAM)	YOLOv5 Head	0.942
Modified RetinaNet (2024)	Histopathology	Squamous Cell Carcinoma (SCC)	Custom (101)	ResNet-50 + Attention	Feature Pyramid Network (FPN)	RetinaNet head	0.929
Fast YOLOv4 (2023)	Endoscopy	Colorectal Polyps	CVC-ClinicDB (612)	Pre-trained Convolutional Neural Network (CNN)	-	Custom Anchor Boxes + Optimizers + Cross-Iteration Batch Normalization (CBN) + Distance-Intersection Over Union Non-Maximum Suppression (DIOU-NMS)	0.909
Improved YOLOv5 (2024)	Endoscopy	Ureteral Orifice	Custom (1043)	ConvNeXt Blocks	Global Repositioning Network (GRN) + Squeeze and Excitation Network (SE Block)	BiFormer Attention	0.896
Modified-YOLOv8 (2024)	Endoscopy	Bronchi	Custom (2419)	CSPDarknet53 with Cross-Stage Partial (CSP), Spatial Pyramid Pooling (SPP) and Efficient Channel Split (ECS)	Feature Pyramid-Path Aggregation Network (PAFPN)	Convolutional Layers (Anchor-Free) with Self-Attention + Convolutional Block Attention Module (CBAM)	0.882
YOLOv5 (2023)	OCT	Dental Caries	Custom (500)	Pre-trained Convolutional Neural Network (CNN)	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Spatial Pyramid Pooling (SPP) + Modified Anchor System	0.860
Faster R-CNN (2023)	OCT	Dental Caries	Custom (500)	Pre-trained Convolutional Neural Network (CNN)	-	Region Proposal Network (RPN)	0.860
RetinaNet + Transfer Learning + Data Augmentation (2024)	Histopathology	Follicles and Corpus-Luteum	Custom (1209)	MobileNetV3	Feature Pyramid Network (FPN)	RetinaNet head	0.830
YOLOv5 + Guided Posterior Regularization (DeGPR) (2023)	Histopathology	Celiac Disease	MuCeD (55)	Pre-trained Convolutional Neural Network (CNN)	Spatial Pyramid Pooling (SPP) + Path Aggregation Network (PAN)	Spatial Pyramid Pooling (SPP) + Modified Anchor System	0.787
Faster R-CNN (2023)	Histopathology	Glomeruli (Kidney)	Custom (165)	Pre-trained Convolutional Neural Network (CNN)	-	Region Proposal Network (RPN)	0.657
Sparse R-CNN (2024)	Endoscopy	Multiple Esophageal Diseases	Custom (N/A)	Pre-trained Convolutional Neural Network (CNN)	Multi-scale Detail Enhancement Pyramid Network (MDEPN) with Gabor-Modulated Convolution and Directional Channel Pooling Modules	Sparse R-CNN Head	0.650

US imaging has seen a broad application of models, including YOLO-based architectures, Faster R-CNN, RetinaNet, and custom designs like FHUSP-NET. These models have addressed diverse tasks such as detecting thyroid nodules, uterine fibroids, elbow osteochondritis dissecans, and coronary lesions, with mAP values ranging from 0.448 to 0.995. High-performing models like YOLOv8m (mAP: 0.995) and improved YOLOv3 (mAP: 0.983) demonstrate exceptional detection accuracy. However, their reliance on task-specific annotations and high-quality data highlights a major limitation. Incorporating self-supervised or unsupervised learning methods could improve model performance in under-annotated scenarios, enhancing their generalizability.

Other imaging modalities, such as HPI, Endoscopy, and OCT, have also benefited from DL advancements. Models like YOLOv5 and YOLOv8 have demonstrated strong performance, achieving mAP values of 0.860 for OCT dental caries detection and 0.988 for endoscopic lower gastrointestinal tract analysis. These results highlight YOLO's versatility, but the computational demands and reliance on specialized annotations remain critical bottlenecks. Exploring federated learning approaches and edge-optimized model designs could help address these constraints, enabling broader deployment in real-world clinical environments.

While the technical performance of these models is impressive, interpretability remains a critical barrier to clinical adoption. In MOD, clinicians must understand why a model made a particular decision to trust its predictions. Current DL models often operate as "black boxes," limiting their usability in clinical workflows. Integrating explainable AI techniques, such as attention heatmaps or gradient-based saliency maps, can help elucidate the reasoning behind model outputs. For example, visualizing ROI identified by a model for abnormalities could increase confidence in its utility for diagnostic or therapeutic decisions [86]. Furthermore, interpretability can assist in identifying model biases and inaccuracies, which are especially crucial in healthcare, where errors could have severe consequences. Emphasizing explainability is not merely an academic concern but a prerequisite for safe and effective clinical integration [87].

Another limitation is generalizability. MOD models often struggle to perform consistently across diverse clinical datasets, limiting their applicability to varied populations and imaging conditions. This lack of robustness hinders their deployment in real-world clinical settings, where

TABLE 6. Summary of the most effective DL-based object detection models based on their mAP performance in MOD tasks.

Modality	Models	Applications	Challenges	Future Direction
X-Ray	Hybrid YOLO-NAS	Fracture detection	High computational demand, limiting use in resource-constrained settings	Develop lightweight and energy-efficient architectures
	YOLOv8m	Wrist abnormality detection	Dependency on balanced datasets	Develop advanced data augmentation and adaptive loss mechanisms
	YOLOv4+ViT	Breast cancer detection	High computational cost	Develop lightweight and energy-efficient architectures
CT	Hybrid R-CNN	COVID-19 diagnosis	High model complexity and computational demand	Simplify architectures while maintaining accuracy by Knowledge Distillation
	ELCT-YOLO	Lung tumor detection	Balancing accuracy and resource efficiency	Develop resource-efficient designs with minimal accuracy trade-off
	YOLO-MSRF	Nodule identification	Dependence on annotated datasets	Develop self-supervised learning for under-annotated data
MRI	YOLOv8s	Prostate cancer detection	Sensitivity to variable image quality	Develop accurate preprocessing and diverse real-world training data
	Refined YOLOv7	Brain tumor detection	Challenges with low-resolution images	Improve robustness through advanced training methods
US	YOLOv8m	Elbow osteochondritis dissecans detection	Dependence on annotated datasets	Develop self-supervised learning for under-annotated data
	Improved YOLOv3	Thyroid nodule detection	Limited flexibility for multiple tasks	Develop modular designs for multi-task adaptability
	Improved Faster R-CNN	Uterine fibroid detection	Computational demands in real-time use cases	Develop optimization methods for edge deployment
HPI	Modified-RetinaNet	Squamous cell carcinoma detection	Overfitting due to small datasets	Develop pretraining models using larger, diverse datasets
	Modified-RetinaNet	Follicles and corpus luteum detection	Dependence on annotated datasets	Develop self-supervised learning for under-annotated data
OCT	YOLOv5	Dental caries detection	High computational demand for real-time detection	Employ federated learning for distributed training
	Faster R-CNN	Dental caries detection	Dependence on annotated datasets	Develop self-supervised learning for under-annotated data
Endoscopy	Enhanced YOLOv5	Lower gastrointestinal tract analysis	Sensitivity to variable image quality	Develop accurate preprocessing and diverse real-world training data
	DBHYOLO	Surgical instrument detection	Requires domain-specific fine-tuning	Develop domain adaptation for improved generalization
	YOLOv8	Nasopharyngeal carcinoma detection	Detecting subtle abnormalities is challenging	Multi-scale feature enhancement for detecting small objects

data variability is inevitable. Federated learning presents a transformative solution to enhance model robustness by enabling collaborative training across decentralized datasets. This approach preserves patient privacy while leveraging knowledge from diverse populations, resulting in more generalizable models better equipped to handle the variability inherent in real-world medical imaging scenarios [88].

Moreover, the high computational demands of advanced architectures, particularly transformer-based models, pose significant challenges for deployment in resource-constrained environments. These limitations restrict the use of MOD technologies on portable or low-cost devices, which are often essential in remote or underserved areas. Efficient model design is crucial for improving accessibility. Techniques such as model quantization, pruning, and knowledge distillation can significantly reduce computational and storage requirements, enabling the deployment of powerful DL models on edge devices, such as smartphones or portable imaging systems. These advancements facilitate real-time analysis, which is critical for timely medical interventions. Addressing these multifaceted challenges is essential for successfully translating MOD research into impactful clinical solutions [89].

V. CONCLUSION

The MOD domain has undergone remarkable advancements since 2023, spurred by innovations in DL architectures. The advanced models such as YOLOv8m, Hybrid YOLO-NAS, YOLOv8s, Refined YOLOv7, Enhanced YOLOv5, and YOLOv4+ViT have achieved mAP scores between 0.986 and 0.995 across diverse MOD tasks and datasets. The YOLO family, in particular, stands out for its exceptional

performance, leveraging advanced architectures like CSP networks, SPP, ECS, PAFPN, and BiFPN to enhance feature extraction and representation. Table 6 outlines the top DL models for each modality, based on their mAP performance in MOD tasks. Despite these strides, several challenges persist, including the need for scalable models in resource-constrained environments and interpretability within clinical workflows. Addressing these issues will require targeted strategies to enhance MOD systems.

Future research should prioritize the development of lightweight models optimized for edge devices, enabling real-time deployment in point-of-care settings. Models such as YOLO-Lite and EfficientDet-Lite illustrate the potential of resource-efficient architectures. However, further exploration is needed to balance computational efficiency and diagnostic accuracy, particularly for complex imaging modalities.

Integrating explainable AI techniques is another critical avenue. Methods such as Grad-CAM can provide clinicians with interpretable visualizations of model decisions, fostering trust and improving clinical adoption. Developing standardized benchmarks to evaluate interpretability across different MOD tasks will also help align model outputs with clinical expectations. Moreover, adopting domain-specific pretraining approaches, including self-supervised learning (SSL) tailored to medical imaging, can enhance model robustness in data-scarce scenarios. SSL techniques, leveraging unlabeled datasets, can minimize the reliance on extensively annotated datasets while maintaining high diagnostic performance.

Collaboration across disciplines, combining expertise in DL, radiology, and healthcare systems, will be vital in translating these advancements into clinical practice. Establishing a framework for regulatory compliance and ethical AI deployment will ensure that MOD technologies are effective but also equitable and safe. In summary, the transformative potential of DL in MOD lies in addressing its current limitations while exploring innovative directions. By focusing on lightweight architectures, interpretability, and domain-specific techniques, the field is poised to redefine diagnostic capabilities, ultimately improving patient outcomes and advancing global healthcare.

DECLARATIONS

- Funding: No Funding
- Competing interests: The authors declare no conflict of interest.
- Ethics approval: The paper complies with ethical requirements
- Availability of data and materials: Not Applicable
- Authors' contributions: M.S. and E.L.: Conceptualization, Methodology, Visualization, Formal Analysis, Writing, Review, and Editing. M.L.: Visualization, Writing, Review, and Editing. As well as All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] R. Yang and Y. Yu, "Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis," *Frontiers Oncol.*, vol. 11, Mar. 2021, Art. no. 638182.

[2] M. Saraei and S. Liu, "Attention-based deep learning approaches in brain tumor image analysis: A mini review," *Frontiers Health Informat.*, vol. 12, p. 164, Oct. 2023.

[3] F. Sheikhi, L. Fakher, and D. Chekani, "Depression detection on e-risk 2017 using long short-term memory models," in *Proc. 20th CSI Int. Symp. Artif. Intell. Signal Process. (AISP)*, Feb. 2024, pp. 1–6.

[4] P. Soviany and R. T. Ionescu, "Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction," in *Proc. 20th Int. Symp. Symbolic Numeric Algorithms Scientific Comput. (SYNASC)*, Sep. 2018, pp. 209–214.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Dec. 2015, pp. 91–99.

[6] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[7] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022.

[8] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[11] H. Zhang, H. Chang, B. Ma, S. Shan, and X. Chen, "Cascade RetinaNet: Maintaining consistency for single-stage object detection," 2019, *arXiv:1907.06881*.

[12] M. G. Ragab, S. J. Abdulkadir, A. Muneer, A. Alqushaibi, E. H. Sumiea, R. Qureshi, S. M. Al-Selwi, and H. Alhussian, "A comprehensive systematic review of YOLO for medical object detection (2018 to 2023)," *IEEE Access*, vol. 12, pp. 57815–57836, 2024.

[13] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023.

[14] M. Hussain, "YOLOv1 to v8: Unveiling each variant—A comprehensive review of YOLO," *IEEE Access*, vol. 12, pp. 42816–42833, 2024.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[17] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 10347–10357.

[18] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[19] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 569–578.

[20] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[21] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse DETR: Efficient end-to-end object detection with learnable sparsity," 2021, *arXiv:2111.14330*.

[22] X. Chen, F. Wei, G. Zeng, and J. Wang, "Conditional DETR v2: Efficient detection transformer with box queries," 2022, *arXiv:2207.08914*.

[23] Q. Chen, X. Chen, J. Wang, S. Zhang, K. Yao, H. Feng, J. Han, E. Ding, G. Zeng, and J. Wang, "Group DETR: Fast DETR training with group-wise one-to-many assignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6610–6619.

[24] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[25] Q. Wang, J. Chen, J. Deng, and X. Zhang, "3D-CenterNet: 3D object detection network for point clouds with center estimation priority," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107884.

[26] H. Guo, X. Yang, N. Wang, and X. Gao, "A CenterNet++ model for ship detection in SAR images," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107787.

[27] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[28] J. Zou, B. Ge, and B. Zhang, "An improved object detection algorithm based on CenterNet," in *Proc. 7th Int. Conf. Artif. Intell. Secur.*, Dublin, Ireland. Cham, Switzerland: Springer, Jan. 2021, pp. 455–467.

[29] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "YOLO-based deep learning model for pressure ulcer detection and classification," *Healthcare*, vol. 11, no. 9, p. 1222, Apr. 2023.

[30] A. Ahmed, A. S. Imran, A. Manaf, Z. Kastrati, and S. M. Daudpota, "Enhancing wrist abnormality detection with YOLO: Analysis of state-of-the-art single-stage detection models," *Biomed. Signal Process. Control*, vol. 93, Jul. 2024, Art. no. 106144.

[31] C.-T. Chien, R.-Y. Ju, K.-Y. Chou, and J.-S. Chiang, "YOLOv9 for fracture detection in pediatric wrist trauma X-ray images," *Electron. Lett.*, vol. 60, no. 11, 2024, Art. no. e13248.

[32] C.-T. Chien, R.-Y. Ju, K.-Y. Chou, E. Xieerke, and J.-S. Chiang, "YOLOv8-AM: YOLOv8 based on effective attention mechanisms for pediatric wrist fracture detection," 2024, *arXiv:2402.09329*.

[33] S. C. Medaramatla, C. V. Samhitha, S. D. Pande, and S. R. Vinta, "Detection of hand bone fractures in X-ray images using hybrid YOLO NAS," *IEEE Access*, vol. 12, pp. 57661–57673, 2024.

[34] A. P. Rathinakumar, N. Yuvaraj, P. Duraisamy, A. G. Joseph, A. F. Khan, and K. R. S. Preethaa, "CovidXDetector: Deep learning based chest abnormality detection for COVID radiography diagnosis," *AIP Conf. Proc.*, vol. 3097, Jan. 2024, Art. no. 020260.

[35] N. M. Hassan, S. Hamad, and K. Mahar, "YOLO-based CAD framework with ViT transformer for breast mass detection and classification in CESM and FFDM images," *Neural Comput. Appl.*, vol. 36, no. 12, pp. 6467–6496, Apr. 2024.

[36] L. Chen, Y. Zhou, and S. Xu, "ERetinaNet: An efficient neural network based on RetinaNet for mammographic breast mass detection," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 5, pp. 2866–2878, May 2024.

[37] G. Qin, P. Luo, K. Li, Y. Sun, S. Wang, X. Li, S. Liu, and L. Xue, "MPFracNet: A deep learning algorithm for metacarpophalangeal fracture detection with varied difficulties," *Comput., Mater. Continua*, vol. 75, no. 1, pp. 999–1015, 2023.

[38] F. Prinzi, M. Insalaco, A. Orlando, S. Gaglio, and S. Vitabile, "A YOLO-based model for breast cancer detection in mammograms," *Cognit. Comput.*, vol. 16, no. 1, pp. 107–120, Jan. 2024.

[39] J. Chen, X. Fan, Z. Chen, Y. Peng, L. Liang, C. Su, Y. Chen, and J. Yao, "Enhancing YOLOv5 for the assessment of irregular pelvic radiographs with multimodal information," *J. Imag. Informat. Med.*, vol. 37, no. 2, pp. 744–755, 2024.

[40] R.-Y. Ju and W. Cai, "Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm," 2023, *arXiv:2304.05071*.

[41] M. Razaghi, H. E. Komleh, F. Dehghani, and Z. Shahidi, "Innovative diagnosis of dental diseases using YOLO V8 deep learning model," in *Proc. 13th Iranian/3rd Int. Mach. Vis. Image Process. Conf. (MVIP)*, Mar. 2024, pp. 1–5.

[42] M. E. Sahin, H. Ulutas, E. Yuce, and M. F. Erkoc, "Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images," *Neural Comput. Appl.*, vol. 35, no. 18, pp. 13597–13611, Jun. 2023.

[43] Z. Ji, J. Zhao, J. Liu, X. Zeng, H. Zhang, X. Zhang, and I. Ganchev, "ELCT-YOLO: An efficient one-stage model for automatic lung tumor detection based on CT images," *Mathematics*, vol. 11, no. 10, p. 2344, May 2023.

[44] X. Wu, H. Zhang, J. Sun, S. Wang, and Y. Zhang, "YOLO-MSRF for lung nodule detection," *Biomed. Signal Process. Control*, vol. 94, Aug. 2024, Art. no. 106318.

[45] S. Mammeri, M. Amroune, M.-Y. Haouam, I. Bendib, and A. C. Silva, "Early detection and diagnosis of lung cancer using YOLO v7, and transfer learning," *Multimedia Tools Appl.*, vol. 83, no. 10, pp. 30965–30980, Sep. 2023.

[46] M. Yaseen, M. Ali, S. Ali, A. Hussain, A. Athar, and H.-C. Kim, "Deep learning based cervical spine bones detection: A case study using YOLO," in *Proc. 26th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2024, pp. 01–05.

[47] C. Ren, S. Hou, J. Hou, and Y. Pang, "SwiF-YOLO: A deep learning method for lung nodule detection," *Int. J. Biol. Life Sci.*, vol. 5, no. 2, pp. 20–27, Mar. 2024.

[48] Y. Su, X. Zhang, H. Shangguan, and R. Li, "Rib fracture detection in chest CT image based on a centernet network with heatmap pyramid structure," *Signal, Image Video Process.*, vol. 17, no. 5, pp. 2343–2350, Jul. 2023.

[49] T. Zhou, Y. Du, J. Mao, C. Peng, H. Wang, and Z. Zhou, "Parallel attention multi-scale mandibular fracture detection network based on CenterNet," *Biomed. Signal Process. Control*, vol. 95, Sep. 2024, Art. no. 106338.

[50] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain tumor detection based on deep learning approaches and magnetic resonance imaging," *Cancers*, vol. 15, no. 16, p. 4172, Aug. 2023.

[51] M. Kang, C.-M. Ting, F. Fung Ting, and R. C.-W. Phan, "BGF-YOLO: Enhanced YOLOv8 with multiscale attentional feature fusion for brain tumor detection," 2023, *arXiv:2309.12585*.

[52] M. Rahimi, M. Mostafavi, and A. Arabameri, "Automatic detection of brain tumor on MRI images using a YOLO-based algorithm," in *Proc. 13th Iranian/3rd Int. Mach. Vis. Image Process. Conf. (MVIP)*, Mar. 2024, pp. 1–5.

[53] N. Aishwarya and G. S. Y. Kannaa, "Real-time prostate cancer detection via YOLO-tiny variants," in *Proc. 7th Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud) (I-SMAC)*, Oct. 2023, pp. 657–662.

[54] B. Kulavuz, M. Çavusoglu, B. Bayram, T. Bakirman, S. Sahin, N. Araz, G. Orhan, H. E. Surmeli, and T. Çakar, "Breast lesion detection from DCE-MRI using YOLOv7," *AIP Conf. Proc.*, vol. 3030, Jan. 2024, Art. no. 030006.

[55] P. Y. Anari, F. Obiezu, N. Lay, F. D. Firouzabadi, A. Chaurasia, M. Golagha, S. Singh, F. Homayounieh, A. Zahergivar, S. Harmon, E. Turkbey, R. Gautam, K. Ma, M. Merino, E. C. Jones, M. W. Ball, W. M. Linehan, B. Turkbey, and A. A. Malayeri, "Using YOLO v7 to detect kidney in magnetic resonance imaging," 2024, *arXiv:2402.05817*.

[56] Z. Zhou, Q. Qiu, H. Liu, X. Ge, T. Li, L. Xing, R. Yang, and Y. Yin, "Automatic detection of brain metastases in T1-weighted construct-enhanced MRI using deep learning model," *Cancers*, vol. 15, no. 18, p. 4443, Sep. 2023.

[57] E. Cengil, Y. Eroğlu, A. Çinar, and M. Yildirim, "Detection and localization of glioma and meningioma tumors in brain MR images using deep learning," *Sakarya Univ. J. Sci.*, vol. 27, no. 3, pp. 550–563, Jun. 2023.

[58] J. Xuan, B. Ke, W. Ma, Y. Liang, and W. Hu, "Spinal disease diagnosis assistant based on MRI images using deep transfer learning methods," *Frontiers Public Health*, vol. 11, Feb. 2023, Art. no. 1044525.

[59] T. Shao, J. Xu, and Y. Dai, "Spinal cord injury identification and localization detection based on MRI imaging and deep learning technology," *Traitement du Signal*, vol. 41, no. 2, pp. 693–703, Apr. 2024.

[60] H. Wu, X. Zhao, G. Han, H. Li, Y. Kong, and J. Li, "MWSR-YLCA: Improved YOLOv7 embedded with attention mechanism for nasopharyngeal carcinoma detection from MR images," *Electronics*, vol. 12, no. 6, p. 1352, Mar. 2023.

[61] A. A. Prisilla, Y. L. Guo, Y.-K. Jan, C.-Y. Lin, F.-Y. Lin, B.-Y. Liau, J.-Y. Tsai, P. Ardhianto, Y. Pusparani, and C.-W. Lung, "An approach to the diagnosis of lumbar disc herniation using deep learning models," *Frontiers Bioeng. Biotechnol.*, vol. 11, Sep. 2023, Art. no. 1247112.

[62] T. Zheng, N. Yang, S. Geng, X.-Y. Zhao, Y. Wang, D. Cheng, and L. Zhao, "[An improved object detection algorithm for thyroid nodule ultrasound image based on faster R-CNN]," *J. Sichuan Univ. Med. Sci. Ed.*, vol. 54, no. 5, pp. 915–922, Sep. 2023.

[63] Y. Tian, J. Zhu, L. Zhang, L. Mou, X. Zhu, Y. Shi, B. Ma, and W. Zhao, "A Swin transformer-based model for thyroid nodule detection in ultrasound images," *J. Visualized Exp.*, vol. 10, no. 194, Apr. 2023, Art. no. e64480.

[64] T. Yang, L. Yuan, P. Li, and P. Liu, "Real-time automatic assisted detection of uterine fibroid in ultrasound images using a deep learning detector," *Ultrasound Med. Biol.*, vol. 49, no. 7, pp. 1616–1626, Jul. 2023.

[65] A. Inui, Y. Mifune, H. Nishimoto, S. Mukohara, S. Fukuda, T. Kato, T. Furukawa, S. Tanaka, M. Kusunose, S. Takigami, Y. Ehara, and R. Kuroda, "Detection of elbow OCD in the ultrasound image by artificial intelligence using YOLOv8," *Appl. Sci.*, vol. 13, no. 13, p. 7623, Jun. 2023.

[66] H. Ghabri, W. Fathallah, M. Hamroun, S. B. Othman, H. Bellali, H. Sakli, and M. N. Abdelkrim, "AI-enhanced thyroid detection using YOLO to empower healthcare professionals," in *Proc. IEEE Int. Workshop Mech. Syst. Supervision*, Nov. 2023, pp. 1–6.

[67] H.-C. Kuo, S.-H. Chen, Y.-H. Chen, Y.-C. Lin, C.-Y. Chang, Y.-C. Wu, T.-D. Wang, L.-S. Chang, I.-H. Tai, and K.-S. Hsieh, "Detection of coronary lesions in Kawasaki disease by scaled-YOLOv4 with HarDNet backbone," *Frontiers Cardiovascular Med.*, vol. 9, Jan. 2023, Art. no. 1000374.

[68] K. Sasaki, D. Fujita, K. Takatsuji, Y. Kotoura, M. Minami, Y. Kobayashi, T. Sukenari, Y. Kida, K. Takahashi, and S. Kobashi, "Deep learning-based osteochondritis dissecans detection in ultrasound images with humeral capitellum localization," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 19, no. 11, pp. 2143–2152, Jan. 2024.

[69] M. I. Daoud, A. Al-Ali, M. Z. Ali, I. Hababeh, and R. Alazrai, "Detecting the regions-of-interest that enclose the tumors in breast ultrasound images using the RetinaNet model," in *Proc. 10th Int. Conf. Electr. Electron. Eng. (ICEEE)*, May 2023, pp. 36–40.

[70] R. Bassiouny, A. Mohamed, K. Umapathy, and N. Khan, "An interpretable neonatal lung ultrasound feature extraction and lung sliding detection system using object detectors," *IEEE J. Translational Eng. Health Med.*, vol. 12, pp. 119–128, 2024.

[71] F. Li, P. Li, X. Wu, P. Zeng, G. Lyu, Y. Fan, P. Liu, H. Song, and Z. Liu, "FHUSP-NET: A multi-task model for fetal heart ultrasound standard plane recognition and key anatomical structures detection," *Comput. Biol. Med.*, vol. 168, Jan. 2024, Art. no. 107741.

[72] D. Yang, J. Xia, R. Li, W. Li, J. Liu, R. Wang, D. Qu, and J. You, "Automatic thyroid nodule detection in ultrasound imaging with improved YOLOv5 neural network," *IEEE Access*, vol. 12, pp. 22662–22670, 2024.

[73] P. Jasitha and P. Pournami, "Glomeruli detection using faster R-CNN and CenterNet," in *Proc. 3rd Asian Conf. Innov. Technol. (ASIANCON)*, Aug. 2023, pp. 1–6.

[74] A. K. Tyagi, C. Mohapatra, P. Das, G. Makharia, L. Mehra, and Mausam, "DeGPR: Deep guided posterior regularization for multi-class cell detection and counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23913–23923.

[75] Z. He, K. Zhang, N. Zhao, Y. Wang, W. Hou, Q. Meng, C. Li, J. Chen, and J. Li, "Deep learning for real-time detection of nasopharyngeal carcinoma during nasopharyngeal endoscopy," *iScience*, vol. 26, no. 10, Oct. 2023, Art. no. 107463.

[76] S. Aliyi, K. Dese, and H. Raj, "Detection of gastrointestinal tract disorders using deep learning methods from colonoscopy images and videos," *Sci. Afr.*, vol. 20, Jul. 2023, Art. no. e01628.

[77] P. Carrinho and G. Falcao, "Highly accurate and fast YOLOv4-based polyp detection," *Expert Syst. Appl.*, vol. 232, Dec. 2023, Art. no. 120834.

[78] S. Fan, H. Yu, Z. Guan, F. Lv, Z. Zhou, and C. Dai, "Diagnosis of dental caries in OCT images based on deep learning," in *Proc. Asia Commun. Photon. Conference/ Int. Photon. Optoelectronics Meetings (ACP/POEM)*, Nov. 2023, pp. 1–5.

[79] L. Chi, Z. Yingyue, Y. Hanmin, L. Xiaoxia, Q. Jiamin, Z. Ming, and W. Liming, "Multi-scale detail enhanced pyramid network for esophageal lesion detection," *J. Comput. Eng. Appl.*, vol. 60, no. 4, p. 229, 2024.

[80] S. Prabhu, K. Prasad, X. Lu, A. Robels-Kelly, and T. Hoang, "Single-stage object detector with attention mechanism for squamous cell carcinoma feature detection using histopathological images," *Multimedia Tools Appl.*, vol. 83, no. 9, pp. 27193–27215, Aug. 2023.

[81] S. Wang, J. Xie, Y. Cui, and Z. Chen, "Colorectal polyp detection model by using super-resolution reconstruction and YOLO," *Electronics*, vol. 13, no. 12, p. 2298, Jun. 2024.

[82] X. Pan, M. Bi, H. Wang, C. Ma, and X. He, "DBH-YOLO: A surgical instrument detection method based on feature separation in laparoscopic surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 19, no. 11, pp. 2215–2225, Apr. 2024.

[83] J. Yan, Y. Zeng, J. Lin, Z. Pei, J. Fan, C. Fang, and Y. Cai, "Enhanced object detection in pediatric bronchoscopy images using YOLO-based algorithms with CBAM attention mechanism," *Heliyon*, vol. 10, no. 12, Jun. 2024, Art. no. e32678.

[84] L. Liang and W. Yuanjun, "UO-YOLO: Ureteral orifice detection network based on YOLO and biformer attention mechanism," *Appl. Sci.*, vol. 14, no. 12, p. 5124, Jun. 2024.

[85] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[86] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimedia Syst.*, vol. 28, no. 6, pp. 2335–2355, Dec. 2022.

[87] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, "Explainable and interpretable artificial intelligence in medicine: A systematic bibliometric review," *Discover Artif. Intell.*, vol. 4, no. 1, p. 15, Feb. 2024.

[88] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognit.*, vol. 151, Jul. 2024, Art. no. 110424.

[89] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," 2022, *arXiv:2205.04766*.

MOHAMMADREZA SARAEI received the B.Sc. degree in biomedical engineering from Azad University, Tabriz, Iran, in 2014, and the M.Sc. degree in biomedical engineering from Seraj University, Tabriz, in 2022. Currently, he is pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, The University of Arizona, Tucson, AZ, USA. From 2014 to 2024, he gained professional experience as a Supervisor Clinical Engineer with Tabriz University of Medical Sciences.

MEHRSHAD LALINIA received the B.S. degree in electrical engineering (telecommunication) from Shahid Beheshti University, Tehran, Iran, in 2016, and the M.S. degree in biomedical engineering (bioelectric) from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, in 2019, where he is currently pursuing the Ph.D. degree in biomedical engineering (bioelectric). From 2023 to 2024, he was a Guest Researcher with the SDU Digital and High-Frequency Electronics Section, Southern Denmark University (SDU), Odense, Denmark. His research focuses on medical image processing and analysis, AI and deep learning, and biomedical devices and applications.

EUNG-JOO LEE (Member, IEEE) received the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2021.

He conducts research on the computer-aided design and implementation of digital signal processing systems, with a specific focus on real-time computer vision and medical imaging applications under challenging constraints, utilizing deep learning techniques. Following his doctoral studies, he was a Postdoctoral Research Fellow with Massachusetts General Hospital/Harvard Medical School. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, The University of Arizona, Tucson, AZ, USA.