INTERNATIONAL JOURNAL OF ADVANCED COMPUTING AND INFORMATICS
VOLUME 1, ISSUE 2, 2025, pp. 118 – 127
ISSN: 3089-7483, DOI: https://doi.org/10.71129/ijaci.v1i2.pp118-127

REVIEW ARTICLE

Single Shot Multibox Detector (SSD) in Object Detection: A Review

Juhartini [a]*, Dwinita Arwidiyarti [b] Desmiwati [c]

[a] Department of Informatic Engineering, Mataram University of Technology, NTB, Indonesia
[b] Faculty of Information and Communication Technology, Mataram University of Technology, NTB, Indonesia
[c] Faculty of Information Technology, Respati University of Indonesia, Jakarta, Indonesia

ABSTRACT - In computer vision, object detection is a core challenge, particularly in scenarios requiring real-time responses and limited computational resources. SSD is a single-pass object detection method that combines classification and localization, recognized for its rapid processing and efficiency. This systematic review analyzes the application and enhancement of SSD-based models across various domains between 2020 and 2025. The study focuses on three major challenges in object detection: real-time performance, small object detection, and deployment in limited-resource settings. It examines diverse strategies, including domain-specific data augmentation, architectural modifications such as deformable SSDs and DRMP modules, and the integration of lightweight backbones like MobileNet. The review highlights significant performance gains in accuracy and efficiency but also reveals persistent limitations, including difficulties in detecting small objects in complex scenes, vulnerability to occlusion, and challenges in model generalization. This study offers key insights into recent research trends and future directions for optimizing SSD-based frameworks to meet the demands of increasingly complex and dynamic detection scenarios.

Keywords: Computer Vision, Object Detection, SSD, Occlusion, Multibox Detector

1. Introduction

Identifying and finding objects in both images and videos is the essential but complex task of object identification in computer vision [1]. As a core function of CV, object detection encompasses various types such as general, salient, and category-specific detection, primarily using Convolutional Neural Networks (CNNs) and deep learning to extract and learn visual features from unprocessed data. These capabilities are applied in diverse domains, from automatic meter reading and traffic surveillance to urban tree detection using high-resolution imagery, often employing models like SSD for real-time object localization [2]. Despite significant advancements, object detection remains challenged by factors such as occlusion, poor lighting, and class variability [3]. Current research focuses on improving real-time performance, model efficiency for embedded systems, reducing reliance on labeled data, and enhancing robustness using techniques like GAN-based data augmentation and domain adaptation [4].

Deep learning, a key subfield of artificial intelligence, enables computers to automatically learn complex patterns from large datasets, reducing reliance on manual programming [5]. Convolutional Neural Networks (CNNs), particularly the Single Shot Multibox Detector (SSD), are widely applied in object detection tasks due to their speed and accuracy, with proven effectiveness in fields such as medical diagnostics and art analysis [6]. These models extract hierarchical features from basic edges to complex shapes through layered architectures and are trained using large datasets, data augmentation, and specialized loss functions. While advancements in GPUs and deep architectures like ResNet have propelled performance, challenges persist including high computational demands, long training times, reliance on balanced data, and limited interpretability [7]. Despite these limitations, advancements like ResNet and loss function optimization continue to enhance performance and scalability across diverse domains [8].

A single-stage object detection framework designed for real-time situations, the Single Shot Multibox Detector (SSD) combines object localization and classification into a single inference cycle. It uses a deep Convolutional Neural Network (CNN) to generate predictions from multiple feature maps, producing bounding boxes and class scores efficiently [7]. Unlike two-stage detectors, SSD omits the region proposal phase, which accelerates inference speed while preserving detection precision. Its architecture typically includes a backbone network like VGG16 or Feature Pyramid Networks (FPNs) are often employed to capture features across multiple spatial resolutions. While SSD is effective across various domains, including medical imaging, it struggles with detecting small objects due to weak feature fusion between low- and high-level layers [9]. To address this, researchers have proposed architectural enhancements and improved feature fusion strategies to strengthen its performance on small object detection tasks [10].

Although numerous studies have investigated object detection techniques and SSD-based models independently, thorough assessments of their practical deployment in real-world contexts remain scarce. This review aims to address this gap by systematically analyzing and comparing the implementation of SSD architectures in object detection research

Received 5 March 2025, Accepted 20 June 2025
Available online 01 July 2025

* **Corresponding Author**
E-mail: juhartini8815@gmail.com (Juhartini)
© The Author(s) 2025. This is an open-access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/)

conducted between 2020 and 2025. The focus of the analysis is on three main application domains where SSD models are commonly utilized to tackle key challenges: real-time detection, small object recognition, and constrained-resource environments. This study outlines methodological approaches and assesses reported performance metrics from a range of research works.

The following are the main contributions of this review:

1. The application of SSD-based models to object detection tasks from 2020 to 2025 is examined in this review.
2. This study examines how SSD-based architectures tackle three primary issues in object detection namely, performance under real-time constraints, recognition of small objects, and efficiency in environments with limited computational capacity.
3. This research offers a comparative examination of performance indicators for SSD-based object detection methodologies, showcasing the strengths and limitations of various modifications and enhancements across different application areas.
4. The paper highlights ongoing difficulties in implementing SSD models and proposes future research paths to improve their accuracy, efficiency, and flexibility in intricate detection environments.

2. Methods

Deep learning applications in object recognition have attracted considerable interest among researchers, leading to a growing number of review papers focused on various detection approaches. Although these surveys offer useful insights into object detection approaches, a concentrated analysis of the unique contributions and advancements of Single Shot Detector (SSD) designs in tackling critical detection difficulties is necessary. This section analyzes relevant surveys position this study within the existing body of research while highlighting its unique contributions, as summarized in Table 1.

Table 1. Summary of relevant surveys

Ref	Year	Primary Focus	SSD Coverage	Main Contribution
[11]	2025	Open World Object Detection (OWOD), Involving Learning from Unknown Objects.	Minor	Cites surveys that include SSD but does not analyze SSD in OWOD. Focus is on open-world settings, OSR, ILOD, and the proposed OWOD taxonomy. First comprehensive OWOD survey
[12]	2024	Object Detection Algorithms: From Classical to CNN-based	Moderate	Offers a structured review of classical and CNN-based object detection methods, including SSD, and compares it with YOLO, Faster R-CNN, etc.
[13]	2023	Evolution of Object Detection	Moderate	Presents an extensive review of object detection techniques developed over the past 20 years, recognizing SSD as a pivotal point and examining notable progress in architectural design and computational performance
[14]	2022	Deep Learning for Maritime UAV Object Detection	Minor	Reviews UAV-based object detection methods for maritime applications; briefly mentions SSD for small object detection but not as a main focus
[15]	2022	Based Object Detectors	High	Delivers an in-depth evaluation of single-stage (e.g., SSD) versus two-stage detection frameworks, focusing on aspects such as operational performance, efficiency, and usage patterns in real-time environments

Numerous investigations have concentrated on general object detection methodologies and their development, providing various insights into the progression of detection techniques over time. For instance, Zou et al. [13] conducted a comprehensive 20-year retrospective on object detection methods, mapping the trajectory from handcrafted feature extractors (e.g., HOG, DPM) to learning-based methodologies like Faster R-CNN, SSD, and YOLO. In their analysis, SSD is highlighted as a critical advancement in one-stage detectors, offering a breakthrough in balancing speed and accuracy, particularly for real-time applications on constrained hardware. The authors position SSD as a bridge between the high accuracy of two-stage detectors and the low latency requirements of practical deployment environments. Zaidi et al. [15] delivered an in-depth survey of modern deep learning-based object detectors, classifying them into one- and two-stage frameworks. Their work provides a detailed comparison of the architectural design, loss functions, and performance trade-offs, where SSD is acknowledged for its simplicity and efficiency, making it a foundational method among one-stage detectors.

They also discussed the integration of SSD with lightweight backbones (e.g., MobileNet) to support deployment on embedded devices. Neha et al. [12] reviewed object detection algorithms from classical paradigms to contemporary CNN-based models. Their survey traces the transition from rule-based and template-matching techniques to neural architectures like SSD, YOLO, and RetinaNet. SSD is presented as a core one-stage framework that introduced multiscale detection using default anchor boxes, enabling faster detection without compromising significantly on accuracy. Although Zhao et al. [14] focused specifically on object detection in maritime UAV imagery, they included SSD in their comparative evaluation of detectors for small object detection under high-altitude and adverse weather conditions. While

not central to the paper, SSD is referenced for its baseline performance and its limitations in detecting small, densely packed objects, especially in cluttered oceanic backgrounds.

In relation to SSD-focused surveys, numerous significant studies have offered comprehensive analyses of object detection frameworks utilizing deep learning, with particular emphasis on the Single Shot MultiBox Detector (SSD). Zaidi et al. [15] present a thorough examination of key object detection models, encompassing both single-stage and two-stage approaches. SSD outperforms other advanced models like YOLO and Faster R-CNN in terms of speed and precision. While SSD initially faced challenges detecting small-scale objects, these issues were resolved by integrating enhanced backbone structures like ResNet and additional refinements. SSDLite is a computationally efficient variant of SSD. By employing MobileNetV2 as its feature extractor, the model reduces parameter count by a factor of eight compared to the standard SSD, yet retains strong accuracy. Edozie et al. [16] categorize current object detection methods into one-stage and two-stage frameworks, Assessing aspects such as design improvements, accuracy in detection, processing efficiency, and practical deployment capabilities. The review explicitly describes SSD as a single-stage (regression) model, characterized by its very high speed and medium to high accuracy, and notes its end-to-end nature.

However, many of these improvements compromise SSD's original speed advantage due to increased computational complexity, and challenges remain in ensuring robustness under conditions such as scale variation, occlusion, low contrast, or cluttered backgrounds. Consequently, a key research gap persists in designing SSD-based models that can effectively integrate semantic richness with high-resolution spatial detail for enhanced small object detection while preserving real-time performance, minimizing resource demands, and maintaining strong generalization across unconstrained, dynamic environments.

3. Results and Discussion

Significant progress has been made in the development of data augmentation techniques in SSD-based object detection with the goal of enhancing detection robustness, especially in difficult situations such small object sizes, occlusions, and complicated backdrops. Early SSD models primarily relied on basic geometric transformations like flipping, cropping, and scaling. However, more recent approaches have incorporated intelligent augmentation techniques tailored to specific domains. For instance, in safety helmet detection, K-means++ clustering was applied to anchor box optimization, aligning better with the distribution of helmet sizes in the dataset. Meanwhile, for occluded face recognition, synthetic occlusion and multiscale augmentation strategies were leveraged to enhance feature diversity and robustness against partial visibility. These domain-adaptive augmentations are often combined with architectural enhancements (e.g., multi-branch detection paths or context modules) to improve SSD's overall performance, especially in small or partially visible object detection scenarios. Improving object detection performance in complex visual environments often requires either enhancements in data preprocessing or architectural modifications to the detection network itself. The Single Shot MultiBox Detector (SSD), while effective and efficient, can struggle with noisy data, limited sample variability, or geometric distortions. To address these limitations, two notable strategies have emerged in the literature: data enhancement using PCA and CF algorithms and architectural modifications such as deformable convolution.

Wu et al. [17] conducted a series of experiments on prostate capsule detection that demonstrate the effectiveness of PCA+CF, a data preprocessing pipeline that combines image denoising using a collaborative filtering (CF) algorithm and Principal Component Analysis (PCA) for dimensionality reduction. When applied to a standard SSD, this enhancement improves the detection accuracy from 71.90% to 77.30%, indicating its value in noise-prone medical imagery. Similar improvements were observed when PCA+CF was applied to other SSD variants: the mAP of FSSD increased from 73.82% to 75.58%, and FFSSD improved to 83.58% with PCA+CF, slightly surpassing the performance of PCA+FFSSD alone (82.39%). The findings indicate that the PCA+CF combination is a viable preprocessing strategy applicable to various SSD-based configurations. Nonetheless, these preprocessing techniques may be redundant in the case of more sophisticated object detection models such as EfficientDet or Sparse R-CNN, which often incorporate their own internal feature optimization mechanisms. Furthermore, complex networks like FFSSD, which integrate feature pyramid networks (FPN) and multistage fusion, tend to require more computation and may experience reduced speed or generalization issues on smaller datasets.

Zhang et al. [18] expanded on preprocessing techniques by introducing Deformable SSD, which incorporates architectural refinements to enhance spatial flexibility within the backbone network. By integrating deformable convolutions into the SSD's VGG16 structure, the model becomes better equipped to detect objects with irregular shapes or variable orientations. Evaluated on the PASCAL VOC2007 and a custom asphalt crack detection dataset, Deformable SSD achieved a 3.1% improvement in mAP over standard SSD on VOC2007. On the crack dataset, it reached 85.11% mAP, outperforming the original SSD and YOLOv4 by 0.55% and 10.4%, respectively. These enhancements in performance highlight the model's adaptability to real-world settings with erratic or unpredictable patterns, like the textures of roads.

In this study, Yamaguchi et al. [19] used a custom histopathological image dataset to detect breast lesions. The model employed random horizontal flipping as a form of data augmentation to enhance its generalization ability. It achieved 88.3% classification accuracy for a three-class setup and 90.5% accuracy for a binary (two-class) task. These results suggest that SSD can perform well in medical imaging when augmented appropriately. However, a significant drawback was noted in terms of precision for non-invasive carcinoma, which remained low at 68.0%, indicating that the

model struggles with certain subtypes of cancer. Moreover, the method showed limitations in detecting small lesions obscured by Hematoxylin and Eosin (H&E) staining, a common practice in histopathological preparation

For instance, Chen et al. [20] proposed a data augmentation strategy based on random image adjustment within an SSD+VGG-16 model to assist in gastrointestinal lesion detection from endoscopic images. Although this approach achieved a relatively high mean Average Precision (mAP) of 95.74%, its performance remains highly dependent on expert diagnosis and the availability of well-annotated images, which are often scarce due to the tedious and selective nature of endoscopy-based data acquisition. The limitation in dataset size, typically reduced to about 30 usable images per session, results in insufficient diversity, and the system's design for single-frame detection inhibits its adaptability to real-time endoscopic video analysis, an essential requirement for clinical applications.

Further exploring the optimization of SSD-based models in specific object detection domains, Ma et al. [21] developed the Lightweight Single Shot Multibox Detector (L-SSD) for garbage classification tasks. Utilizing a self-collected dataset comprising 9,000 images across categories such as cardboard, glass, paper, and plastic, the study used data augmentation methods like scaling, cropping, and random flipping. The L-SSD attained a mean Average Precision (mAP) of 83.48%, reflecting its potential effectiveness in identifying waste materials under practical environmental conditions. However, the method exhibited slight reductions in detection speed (40 FPS compared to 46 FPS for the conventional SSD) and faced challenges in low-light environments and with irregularly shaped waste materials, exposing issues with the model's generalization in various visual contexts.

In a separate effort, Cai et al. [22] introduced a refined SSD-based framework specifically adapted for fruit classification, utilizing a dataset comprising 2,000 self-gathered fruit images consisting of various fruits such as apples, pears, and persimmons. By applying rotation and translation-based data augmentation and incorporating Soft-NMS for enhanced anchor box selection, the improved SSD attained a mAP of 92.4%. Despite the high detection accuracy, the model's effectiveness diminished in complex natural environments where fruit appearances vary significantly due to lighting, occlusion, or deformation, highlighting the sensitivity of traditional detection frameworks to environmental variability.

In the industrial domain, Li et al. [23] combined SSD with a ResNet50 backbone and incorporated Context Fusion Structure (CFS) and a Feature Refinement Module (FRM) to mitigate semantic conflicts and enrich contextual cues. When tested on the NEU-DET dataset for detecting steel surface anomalies, the model achieved a mean Average Precision of 79.5% through the use of data augmentation techniques like random cropping and flipping. Nevertheless, certain defect types, especially *crazing* and *rolled-in scale,* remained difficult to detect, primarily due to their fine, scattered textures and low contrast. These results highlight the limitations of anchor-based SSD models when applied to irregular, small-scale targets.

Magalhães et al. [24] conducted a comparative study evaluating four SSD variants MobileNet v2, Inception v2, ResNet 50, and ResNet 101 on a custom greenhouse tomato dataset. Extensive augmentation techniques, including geometric and photometric transformations, were employed to simulate challenging agricultural conditions. Despite these efforts, the highest mAP achieved was 51.46% using SSD MobileNet v2, with deeper variants like SSD ResNet 101 experiencing performance degradation due to overfitting. This suggests that while augmentation improves robustness, architectural complexity must be balanced with data diversity to prevent over-specialization.

The Single Shot MultiBox Detector (SSD) is an object detection method known for its high-speed, real-time detection capabilities. It achieves this by combining predictions from multiple feature maps with various resolutions, allowing it to process objects of diversified sizes within a single network. Data augmentation is a common technique used during the training of SSD models to enhance performance and robustness. The table below highlights various approaches that build upon the SSD architecture, often incorporating architectural modifications, advanced feature fusion techniques, attention-based strategies or advanced data processing techniques beyond standard augmentation, aiming to enhance detection precision, especially in difficult contexts such as small or occluded objects, or specific domains such as medical imaging and remote sensing. While these enhancements often lead to significant gains in mean Average Precision (mAP), they can sometimes introduce drawbacks, such as increased computational cost, reduced inference speed, or limitations in generalizability to different object sizes or scene conditions.

These studies collectively highlight the substantial progress made in SSD-based object detection techniques, particularly in addressing challenges such as occlusion, small object detection, and complex background interference. By incorporating architectural innovations ranging from attention mechanisms and multiscale feature fusion to lightweight backbones and enhanced contextual reasoning, SSD-based models have demonstrated improved detection accuracy and real-time performance across diverse application domains. The effectiveness of various SSD-derived models in managing occlusion-related challenges, a synthesis of essential performance indicators from each referenced study is presented in Table 2.

Table 2. Performance summary of SSD-based data augmentation methods for object detection

Ref	Method	Dataset	Results (mAP %)	Drawbacks
[21]	L-SSD (Lightweight SSD) – Garbage Detection	9000 self-collected images (cardboard, glass, etc.)	83.48%	Detection speed is slightly reduced and performance drops under low lighting or with irregularly shaped waste materials.
[22]	Improved SSD – Fruit Image Recognition	2000 fruit images	92.4%	Ineffective in complex natural environments due to high variability in fruit appearance.
[19]	SSD with random horizontal flipping	Custom histopathological breast lesion dataset.	88.3% (3-class) 90.5% (2-class)	Low precision (68.0%) for non-invasive carcinoma; missed lesions under weak H&E staining.
[20]	SSD with VGG-16 backbone	Kvasir-SEG (GI endoscopic images)	95.74%	The pathology and experience of the physician are crucial to the detection process. Only ~30 usable images are available per case after screening from thousands, limiting dataset size. The method is designed for single-image detection and lacks real-time video processing capability. CNN-based feature extraction demands high-end hardware and introduces processing latency.
[23]	SSD with ResNet50	NEU-DET (steel surface defects)	79.5%	Detection performance remains low for certain defects (e.g., crazing at 47% accuracy) due to low contrast, resolution, and the fine, widely dispersed shapes of the defects—challenging for anchor-based SSD models.
[24]	Evaluated SSD MobileNet v2, Inception v2, ResNet50, ResNet101	Greenhouse Tomato Dataset	MobileNet v2: 51.46% Inception v2: 48.54% ResNet 50: 42.62% ResNet 101: 36.32%	All models showed reduced generalization, with test performance dropping significantly versus validation. ResNet101 was prone to overfitting due to its complexity. Dataset variability needs to be increased, and real-time performance should be evaluated in future studies.

Occlusion poses a persistent challenge in object detection tasks, especially in real-world scenarios where visual clutter, partial visibility, and environmental interference are common. Variants of the Single Shot Multibox Detector (SSD) have been widely adopted and enhanced to address this issue by improving real-time detection speed and robustness against occlusion, deformation, and background complexity.

Guo et al. [25] introduced FESSD, an SSD-based model optimized for detecting small and densely packed objects in remote sensing imagery. Operating at 35.6 frames per second (FPS), FESSD enhances discriminative target features while suppressing background noise. Despite its strong performance in real-time environments, the model still struggles to achieve high accuracy when detecting very small or overlapping targets due to low pixel resolution and occlusion among objects.

In the medical imaging domain, Souaidi et al. [26] proposed the MP-FSSD (Multiscale Pyramidal Fusion SSD) to detect small polyps in capsule endoscopy and colonoscopy datasets. By fusing features across multiple spatial scales, MP-FSSD achieves a processing speed of 62.5 FPS while significantly improving detection of polyps even in noisy and low-illumination scenarios. However, the model's performance is limited by the small size of medical datasets and variations in lighting and texture. Moreover, it may fail to detect polyps that are obscured by folds or resemble surrounding mucosa.

To tackle occlusion in facial recognition, particularly during the COVID-19 pandemic, Tsai et al. [7] developed a system that integrates SSD with an efficient pre-trained convolutional neural network. Applied to datasets with varying levels of facial occlusion (25–75%), this system achieves real-time detection at 8 FPS. The approach can identify masked faces in challenging conditions, although it remains susceptible to the general difficulties of occluding face recognition, such as the loss of essential facial features. Qin et al. [27] proposed FDA-SSD (Fast Depth-Assisted SSD), which incorporates depth information to improve detection of rigid indoor objects. Designed for the OTMP dataset, FDA-SSD runs at 28.7 FPS and leverages temporal depth cues to track and detect objects accurately. While effective under most conditions, it cannot extract depth information in the first frame and depends on prior frames for current predictions. Additionally, the model is unsuitable for deformable objects and requires normalization of depth-related parameters to handle inconsistencies.

In the context of industrial defect detection, Liu et al. [28] introduced LW-SSD, a lightweight version of SSD tailored for detecting small defects in fabric with complex textures. Achieving 48 FPS, this model significantly reduces computational cost while maintaining high detection accuracy. Nevertheless, its performance declines in low-light environments and is limited by the diversity and quantity of the dataset used for training. Kumar et al. [29] proposed an improved SSD variant trained on Pascal VOC and COCO datasets, achieving a remarkable 89 FPS. The model enhances recall and can detect multiple object categories in a single image. However, it continues to encounter challenges with occluded, corrupted, or deformable objects. Furthermore, architecture lacks generalizability in novel object classes and underperforms on small object detection due to inherent limitations of the base SSD structure. These SSD-based enhancements underscore the trade-offs between detection speed, complexity, and robustness under occluding or visually

degraded conditions. Despite progress, future research must address generalization to new object types, small object detection, and improved adaptability to occluded or partially visible targets.

Although SSD-based techniques have shown significant advancements in managing occlusion challenges in object detection, they are still accompanied by specific constraints. For instance, many approaches depend heavily on complex architectural modifications, such as attention mechanisms or multi-scale feature fusion, which can increase computational overhead and reduce inference speed. Moreover, these models may still struggle in scenarios involving severe occlusion, highly cluttered backgrounds, or diverse object scales. A comprehensive overview of SSD-oriented techniques designed to address occlusion-related issues is presented in Table 3.

Table 3. Performance summary of SSD-based methods for handling occlusions in object detection

Ref	Method	Dataset	Result	Drawbacks
[25]	FESSD	SD-RSI, DIOR (remote sensing with dense/small targets)	35.6 FPS	For very small/dense targets, AP still unsatisfactory due to low pixel size and overlap.
[26]	MP-FSSD	WCE (PillCam COLON 2), CVC-ClinicDB, ETIS-Larib	62.5 FPS	Small datasets. Performance may vary with image conditions (illumination, texture). May miss hidden or lighting-similar regions. Background noise (debris) unavoidable.
[7]	SSD + Efficient Pre-trained CNN (Occluded Face Recognition)	CelebA (with 25–75% occlusion); self-collected masked face dataset.	8 FPS	General challenges in occluded face detection and feature loss.
[27]	FDA-SSD	OTMP (depth-based, indoor rigid objects)	28.7 FPS	Depth missing in first frame; uses prior depth info. Not suitable for violently deformable objects. Needs parameter normalization.
[28]	LW-SSD	Fabric defect dataset (custom). Small defects, deformities, complex texture.	48 FPS	Dataset limited in defect types and size. mAP drops in low light.
[29]	Improved SSD	Pascal VOC, COCO	89 FPS	Issues with occlusion, deformation, corrupt/interlaced objects. Cannot generalize to new classes. SSD weak on small objects.

Small object detection using SSD-based models continues to pose considerable challenges due to the limited pixel information, background clutter, and scale variance present in typical detection scenarios. These difficulties are especially noticeable in applications where the objects of interest frequently take up a small percentage of the image, like surveillance, autonomous driving, and aerial imagery. To address these issues, recent research has proposed various enhancements to the original SSD framework, such as feature pyramid architecture, context-aware modules, and resolution-preserving feature extraction. These modifications aim to better preserve spatial detail and improve feature representation of small-scale objects, leading to improved detection accuracy and robustness. A consolidated overview of SSD-based approaches focused on detecting small-scale objects is presented in Table 4.

Table 4. Performance summary of SSD-based methods for small object detection

Ref	Method	Performance	Drawbacks
[30]	Enhanced SSD with Inception, IFPN, and DFE	- 90.2% accuracy overall- 3.4% mAP gain on DOTA- High precision on small/medium objects (e.g., oil tanks 90.7%)	Slightly lower FPS and prone to missing small objects in dense scenes due to sampling loss.
[31]	ISSD (Improved SSD with DSDCM, Inception & Recalibration Modules)	- 0.912 F-measure- 77 FPS- ~13% accuracy gain over SSD	Struggles with fine crack details and has high computational cost.
[32]	SSDMNV2 (SSD + MobileNetV2) for Mask Detection	- 92.64% accuracy- F1 score: 0.93- Works across various mask sizes	Dataset noise and variability in masks reduce detection precision.
[33]	Puppis (Hardware-accelerated SSD post-processing)	- 33.34× speed-up- <0.3% mAP drop after quantization- Supports multi-scale detection	Prioritizes speed and lacks detailed evaluation for small object accuracy.
[34]	SSD + DRMP (Deconvolutional Region Magnification Procedure)	42.4% mAP for small objects; +27.1% improvement over SSD; improves 'person' class TPR by 10%.	Accuracy gain comes with reduced speed (24 FPS), increasing computational demand.
[35]	Optimized IFSSD / Modified SSD (MSSD)	94.8% accuracy on smoke detection; outperforms Faster R-CNN and YOLOv2; 97.5% overall average accuracy.	Struggles with very small features (<10×10 px) and similar-looking object confusion.

Wang et al. [30] proposed an enhanced SSD architecture integrating an improved Inception network, a modified Feature Pyramid Network (IFPN), and a Deep Feature Enhancement (DFE) module to explicitly boost small object feature extraction capabilities. Designed for complex scenarios like remote sensing with densely packed small and medium-sized objects, their method significantly improved detection performance. The model attained 90.7% accuracy on oil tank images and 85.8% for playground targets within the DOTA dataset, indicating significant enhancement in detecting small

objects. The IFPN module contributed to a 3.4% increase in mean Average Precision (mAP), raising it from 85.9% to 89.3%. The overall detection accuracy reached 90.2%. However, the method exhibited a lower frame rate (49.7 FPS) compared to traditional SSD and YOLOv3 (60.8–67.5 FPS), owing to higher model complexity. Additionally, small objects remain susceptible to missed detections due to pixel sparsity and feature loss during downsampling, especially in high-density scenes like DOTA.

Lu et al. [31] introduced ISSD, an improved SSD variant tailored for detecting narrow and elongated bridge cracks. Their design incorporates three key components: a Depth Separable Deformation Convolution Module (DSDCM) for handling irregular crack structures, an Inception Module (IM) for expanding the network width while reducing computational load, and a Feature Recalibration Module (FRM) to enhance critical feature emphasis. The ISSD model delivered a notable F-measure of 0.912 at 77 FPS and outperformed the baseline SSD by approximately 13% in detection accuracy. It also demonstrated the ability to preserve and accurately represent complex crack shapes. However, challenges persisted in capturing very fine crack details and resisting interference under noisy backgrounds. The model's structural complexity also posed limitations for deployment on lightweight or portable detection devices.

Nagrath et al. [32] developed SSDMNV2, which integrates SSD with MobileNetV2 for real-time face mask detection across varying object sizes. This model effectively handled both small and large facial regions and mask types, achieving an overall accuracy of 92.64% and an F1 score of 0.93. Despite its strong performance, the model faced limitations due to the lack of robust datasets for masked faces. Noise arising from mask variability and obstructions further complicated detection, necessitating manual data cleaning to maintain dataset quality and precision. Vrbaski et al. [33] addressed performance bottlenecks in SSD-based systems by introducing *Puppis*, a hardware accelerator focused on optimizing post-processing operations such as softmax, bounding box extraction, non-maximum suppression, and top-K sorting. While the study does not directly target small object detection, the underlying SSD architecture inherently supports multi-scale object recognition, including small objects. Puppis achieved a 97% reduction in execution time, translating to a 33.34× speedup, with only a minor drop in mAP (<0.3%) following quantization and hardware migration. This makes it highly suitable for edge-based applications requiring real-time performance.

Wang et al. [34] addressed the limitations of the original Single Shot Multibox Detector (SSD) in detecting small objects by introducing a Deconvolutional Region Magnification Procedure (DRMP). Their approach enhances the feature representation of low-resolution feature maps through deconvolution, enabling better localization of small and medium objects. Experimental results demonstrated a notable improvement, with a 42.4% mean Average Precision (mAP) for small object detection—an increase of 27.1% over the baseline SSD model. Additionally, the method improved the True Positive Rate (TPR) for the 'person' class by 10%, showcasing its efficacy in detecting finer targets. However, this accuracy improvement came at the cost of detection speed, with the frame rate reduced to 24 FPS compared to the original SSD's 55 FPS, thus increasing the model's computational burden during inference.

Similarly, Han et al. [35] proposed an optimized version of SSD, referred to as IFSSD or Modified SSD (MSSD), which enhances multi-scale detection by integrating fused feature maps. Their method achieved impressive results in urban fire smoke detection, attaining 94.8% accuracy and outperforming other mainstream detectors like Faster R-CNN (87.0%) and YOLOv2-Darknet19 (85.4%). The model also reported an overall average accuracy of 97.5% across detection scenarios. Despite its effectiveness, the approach showed limitations in detecting very small features, particularly those with a spatial resolution under 10×10 pixels—and encountered misclassification issues when dealing with objects that exhibit similar visual characteristics, such as motorcycles and people.

These advancements demonstrate the ongoing progress in small object detection using SSD-based architecture. Nevertheless, trade-offs between accuracy, speed, and resource efficiency continue to shape the development and deployment of these systems in practical applications. Shi et al. [36] introduced SsdNet, an enhanced SSD architecture designed for 3D machining feature recognition and localization from 2D STL models. The system was trained on 24,000 single-feature models and tested on 1,000 multi-feature STL models that included highly intersecting and visually degraded shapes. SSDNet achieved an F-score of 95.20% for recognition and 90.62% for localization, processing each model in approximately 243.85 ms. Notably, the model was specifically built to handle geometric complexity, including overlapping or intersecting machining features, and no significant performance drawbacks were reported. Its robustness in the presence of shape degradation affirms its suitability for manufacturing and CAD-based automation.

In the medical domain, Wu et al. [37] employed a standard SSD model for esophageal neoplasm detection using endoscopic images, comprising 936 training and 264 testing samples collected from 46 patients. The dataset included both white-light imaging (WLI) and narrow-band imaging (NBI) modalities. The model achieved 90.9% overall accuracy, 96.2% sensitivity, and 92.7% positive predictive value (PPV), with 100% sensitivity for squamous cell carcinoma (SCC) detection in NBI images. This performance is particularly notable because it was achieved without the need for magnified endoscopy or expert interpretation. However, the dataset was derived from a single medical center, raising concerns about data imbalance, selection bias, and the lack of inclusion of other types of esophageal lesions, such as Barrett's esophagus or adenocarcinoma. Wu et al. [2] also developed an optimized SSD model to detect individual trees in high-resolution urban plantation imagery, including lychee and palm plantations. The model achieved 96.32% accuracy and effectively detected densely packed tree canopies, outperforming conventional methods such as watershed segmentation and local maxima detection. This lightweight model is ideal for remote sensing applications, particularly in urban forestry. Importantly, no specific limitations were reported regarding its performance under occlusion or complex canopy arrangements.

Lastly, Ciampi et al. [38] deployed an SSD-based system in a multi-camera urban traffic surveillance setup, incorporating footage from two different cameras (TLC1 and TLC2) monitoring an intersection. The dual-camera system improved precision (68%) and recall (84%) over single-camera setups, yielding 40% more true positive detections by leveraging spatiotemporal fusion. This method highlights the benefit of combining visual and temporal redundancy to reduce false negatives in dynamic environments. However, the study only used decision-level fusion, lacked testing under diverse lighting/weather conditions, and did not address security or adversarial threats, which are essential considerations for real-world deployment.

Challenges in SSD-based small object detection primarily revolve around the inherent trade-off between detection accuracy and processing speed, especially for small, cluttered, or occluded objects in complex environments. The loss of original information during continuous down-sampling and up-sampling processes and the insufficient reuse of shallow feature layers, which are rich in object information but often lack semantic depth, continue to compromise detection performance for small objects. Furthermore, these models frequently demand high computational power and memory resources, limiting their deployment on portable or less-configured hardware. Complex backgrounds, varying illumination, and occlusions (e.g., in remote sensing images, pavement cracks, or face detection) introduce significant interference, leading to missed detections or false positives. Issues like misalignment between classification and regression features and performance degradation on larger objects when models are optimized for smaller ones also pose considerable challenges. The limited availability of large, publicly annotated medical datasets can lead to subjective performance evaluations and hinder model generalizability. Additionally, the anti-interference ability of networks in noisy environments needs substantial improvement.

Future directions aim to mitigate these limitations by focusing on several key areas. Researchers are working to expand and optimize optical remote sensing image databases and detection algorithms to handle diverse and complex scenes more effectively. Developing more powerful and lightweight backbone networks, such as DenseNet or ResNet-152, is crucial to improve feature extraction capabilities while ensuring models are suitable for deployment on portable or resource-constrained devices, addressing the speed-accuracy dilemma. Efforts include designing more efficient attention modules and adaptive anchor boxes to better capture contextual information and refine object localization. Exploring advanced feature fusion techniques beyond simple concatenation or element-wise summation is vital to create more balanced and semantically rich feature maps across multiple layers, thereby enhancing small object detection and reducing aliasing effects.

Future work also involves investigating strategies to regularize edge-pooling layers and filter out background interference to improve robustness. For applications like hand and gesture recognition, developing systems capable of detecting and tracking multiple objects and handling distant or very small instances is a priority. Integrating object detection with elevator scheduling algorithms or real-time traffic monitoring systems by leveraging temporal continuity in video streams and minimizing power consumption also represents promising avenues for real-world application. Finally, addressing security vulnerabilities against adversarial attacks and improving noise reduction in single-shot images for applications like astronomical imaging are also critical areas for research.

4. Conclusions

The development and improvement of SSD-based object identification techniques between 2020 and 2025 have been thoroughly reviewed in this study, shedding light on major advancements in feature representation, multi-scale processing, and real-time performance. Through a systematic analysis of diverse strategies such as super-resolution modules, feature fusion techniques, and deconvolutional layers, the review has highlighted how SSD has been continually adapted to tackle persistent challenges in small object detection, aerial imagery, and low-resolution environments. Despite notable gains in detection accuracy and computational efficiency, issues remain, including limitations in scale variance, difficulty in dense scene interpretation, and tradeoffs between speed and precision. The integration of SSD with novel architectural improvements and auxiliary modules continues to evolve, pointing toward promising avenues such as transformer-based feature enhancement, context-aware detection, and hybrid fusion frameworks. As SSD-based methods mature, future research is expected to deliver more robust and adaptable solutions tailored for real-world deployments, especially under resource-constrained scenarios, domain adaptation requirements, and dynamically changing visual contexts.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] A. Rohan, M. S. Rafaq, Md. J. Hasan, F. Asghar, A. K. Bashir, and T. Dottorini, “Application of deep learning for livestock behaviour recognition: A systematic literature review,” Computers and Electronics in Agriculture, vol. 224, p. 109115, Sep. 2024, doi: 10.1016/j.compag.2024.109115.

[2] Y. Zheng and G. Wu, “Single Shot MultiBox Detector for Urban Plantation Single Tree Detection and Location With High-Resolution Remote Sensing Imagery,” vol. 9, no. November, pp. 1–14, 2021, doi:

10.3389/fenvs.2021.755587.
[3] A. Iqbal, A. Basit, I. Ali, J. Babar, and I. Ullah, "Automated meter reading detection using inception with single shot multi-box detector," Intell. Autom. Soft Comput., vol. 27, no. 2, pp. 299–309, 2021, doi:10.32604/iasc.2021.014250.
[4] M. Ahmed, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Survey and Performance Analysis of Deep Learning Based Object Detection in Challenging Environments," Sensors, vol. 21, no. 15, p. 5116, Jul. 2021, doi: 10.3390/s21155116.
[5] Y. Mori, T. Oichi, M. Enomoto-iwamoto, and T. Saito, "Automatic Detection of Medial and Lateral Compartments from Histological Sections of Mouse Knee Joints Using the Single-Shot Multibox Detector Algorithm," 2022, doi: 10.1177/19476035221074009.
[6] Z. Zhao, "Enhancing artistic analysis through deep learning : a graphic art element recognition model based on SSD and FPT," 2024, doi: 10.7717/peerj-cs.1761.
[7] A. Tsai, S. Member, Y. Ou, and W. Wu, "Integrated Single Shot Multi-Box Detector and Efficient Pre-Trained Deep Convolutional Neural Network for Partially Occluded Face Recognition System," IEEE Access, vol. 9, pp. 164148–164158, 2021, doi: 10.1109/ACCESS.2021.3133446.
[8] Indrabayu, Taslinda, R. Irianty, and S. W. Sidehabi, "Various Obstacles Detection Systems using Single Shot Multi-Box Detector (SSD) for Autonomous-Driving Vehicles," Int. J. Eng. Trends Technol., vol. 71, no. 5, pp. 1–8, 2023, doi: 10.14445/22315381/IJETT-V71I5P201.
[9] Q. Li et al., "Detection of Tailings Dams Using High-Resolution Satellite Imagery and a Single Shot Multibox Detector in the Jing–Jin–Ji Region, China," Remote Sensing, vol. 12, no. 16, p. 2626, Aug. 2020, doi: 10.3390/rs12162626.
[10] M. Souaidi, S. Lafraxo, Z. Kerkaou, M. El Ansari, and L. Koutti, "A Multiscale Polyp Detection Approach for GI Tract Images Based on Improved DenseNet and Single-Shot Multibox Detector," Diagnostics, vol. 13, no. 4, p. 733, Feb. 2023, doi: 10.3390/diagnostics13040733.
[11] Y. Li, Y. Wang, W. Wang, D. Lin, B. Li, and K.-H. Yap, "Open World Object Detection: A Survey," IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 2, pp. 988–1008, Feb. 2025, doi: 10.1109/tcsvt.2024.3480691.
[12] F. Neha, D. Bhati, D. K. Shukla, and M. Amiruzzaman, "From classical techniques to convolution-based models: A review of object detection algorithms," 2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS), pp. 1–6, Jan. 2025, doi: 10.1109/ipas63548.2025.10924494.
[13] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," Proceedings of the IEEE, vol. 111, no. 3, pp. 257–276, Mar. 2023, doi: 10.1109/jproc.2023.3238524.
[14] C. Zhao, R. W. Liu, J. Qu, and R. Gao, "Deep learning-based object detection in maritime unmanned aerial vehicle imagery: Review and experimental comparisons," Engineering Applications of Artificial Intelligence, vol. 128, p. 107513, Feb. 2024, doi: 10.1016/j.engappai.2023.107513.
[15] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," Digital Signal Processing, vol. 126, p. 103514, Jun. 2022, doi: 10.1016/j.dsp.2022.103514.
[16] E. Edozie, A. N. Shuaibu, U. K. John, and B. O. Sadiq, "Comprehensive review of recent developments in visual object detection based on deep learning," Artificial Intelligence Review, vol. 58, no. 9, Jun. 2025, doi: 10.1007/s10462-025-11284-w.
[17] S. Wu, X. Wang, and C. Guo, "Application of Feature Pyramid Network and Feature Fusion Single Shot Multibox Detector for Real-Time Prostate Capsule Detection," Electronics, vol. 12, no. 4, p. 1060, Feb. 2023, doi: 10.3390/electronics12041060.
[18] Z. Zhang, "Automated Asphalt Highway Pavement Crack Detection Based on Deformable Single Shot Multi-Box Detector Under a Complex Environment," IEEE Access, vol. 9, pp. 150925–150938, 2021, doi: 10.1109/ACCESS.2021.3125703.
[19] M. Yamaguchi, T. Sasaki, K. Uemura, Y. Tajima, S. Kato, and K. Takagi, "Automatic breast carcinoma detection in histopathological micrographs based on Single Shot Multibox Detector," J. Pathol. Inform., vol. 13, no. July, p. 100147, 2022, doi: 10.1016/j.jpi.2022.100147.
[20] X. Chen, K. Zhang, S. Lin, K. F. Dai, and Y. Yun, "Single Shot Multibox Detector Automatic Polyp Detection Network Based on Gastrointestinal Endoscopic Images," vol. 2021, 2021, doi: 10.1155/2021/2144472.
[21] W. Ma, X. Wang, and J. Yu, "A lightweight feature fusion single shot multibox detector for garbage detection," IEEE Access, vol. 8, pp. 188577–188586, 2020, doi: 10.1109/ACCESS.2020.3031990.
[22] J. Cai, J. Tao, Y. Ma, X. Fan, and L. Cheng, "Fruit image recognition and classification method based on improved single shot multi-box detector," J. Phys. Conf. Ser., vol. 1629, no. 1, 2020, doi: 10.1088/1742-6596/1629/1/012010.
[23] Y. Li, L. He, M. Zhang, Z. Cheng, W. Liu, and Z. Wu, "Improving the Performance of the Single Shot Multibox Detector for Steel Surface Defects with Context Fusion and Feature Refinement," Electronics, vol. 12, no. 11, p. 2440, May 2023, doi: 10.3390/electronics12112440
[24] S. A. Magalhães et al., "Evaluating the Single-Shot MultiBox Detector and YOLO Deep Learning Models for the Detection of Tomatoes in a Greenhouse," Sensors, vol. 21, no. 10, p. 3569, May 2021, doi: 10.3390/s21103569.

[25] J. Guo, Z. Wang, and S. Zhang, “FESSD: Feature Enhancement Single Shot MultiBox Detector Algorithm for Remote Sensing Image Target Detection,” Electronics, vol. 12, no. 4, p. 946, Feb. 2023, doi: 10.3390/electronics12040946.

[26] M. Souaidi and M. El Ansari, “A New Automated Polyp Detection Network MP-FSSD in WCE and Colonoscopy Images Based Fusion Single Shot Multibox Detector and Transfer Learning,” IEEE Access, vol. 10, pp. 47124–47140, 2022, doi: 10.1109/ACCESS.2022.3171238.

[27] Z. Wang, S. Yang, M. Shi, and K. Qin, “FDA-SSD: Fast Depth-Assisted Single-Shot MultiBox Detector for 3D Tracking Based on Monocular Vision,” Applied Sciences, vol. 12, no. 3, p. 1164, Jan. 2022, doi: 10.3390/app12031164

[28] S. Liu, L. Huang, Y. Zhao, and X. Wu, “Lightweight Single Shot Multi-Box Detector : A fabric defect detection algorithm incorporating parallel dilated convolution and dual channel attention,” 2024, doi: 10.1177/00405175231202817.

[29] A. Kumar, Z. J. Zhang, and H. Lyu, “Object detection in real time based on improved single shot multi-box detector algorithm,” EURASIP Journal on Wireless Communications and Networking, vol. 2020, no. 1, Oct. 2020, doi: 10.1186/s13638-020-01826-x.

[30] L. Wang et al., “DATA A RT I C L E A novel deep learning- - based single shot multibox detector model for object detection in optical remote sensing images,” no. March 2022, pp. 237–251, 2024, doi: 10.1002/gdj3.162.

[31] G. Lu, X. H. Id, Q. Wang, F. Shao, J. Wang, and Q. Jiang, “Bridge crack detection based on improved single shot multi-box detector,” pp. 1–21, 2022, doi: 10.1371/journal.pone.0275538.

[32] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth, “SSDMNV2 : A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2,” Sustain. Cities Soc., vol. 66, no. December 2020, p. 102692, 2021, doi: 10.1016/j.scs.2020.102692.

[33] V. Vrbaski, S. Josic, V. Vranjkovic, P. Teodorovic, and R. Struharik, “Puppis: Hardware Accelerator of Single-Shot Multibox Detectors for Edge-Based Applications,” Aug. 2023, doi: 10.20944/preprints202308.0832.v1.

[34] Y. Wang, P. Niu, X. Guo, G. Yang, and J. Chen, “Single Shot Multibox Detector with Deconvolutional Region Magnification Procedure,” IEEE Access, vol. 9, pp. 47767–47776, 2021, doi: 10.1109/ACCESS.2021.3068486.

[35] H. Han, “A Novel Single Shot-multibox Detector Based on Multiple Gaussian Mixture Model for Urban Fire Smoke Detection,” Comput. Sci. Inf. Syst., vol. 20, no. 4, pp. 1819–1843, 2023, doi: 10.2298/CSIS221218032H.

[36] P. Shi, Q. Qi, Y. Qin, P. J. Scott, and X. Jiang, “Intersecting Machining Feature Localization and Recognition via Single Shot Multibox Detector,” vol. 17, no. 5, pp. 3292–3302, 2021, doi: 10.1109/TII.2020.3030620.

[37] Y.-K. Wang et al., “Endoscopic Images by a Single-Shot Multibox Detector for the Identification of Early Cancerous Lesions in the Esophagus: A Pilot Study,” Cancers, vol. 13, no. 2, p. 321, Jan. 2021, doi: 10.3390/cancers13020321.

[38] M. Lenatti, S. Narteni, A. Paglialonga, V. Rampa, and M. Mongelli, “Dual-View Single-Shot Multibox Detector at Urban Intersections: Settings and Performance Evaluation,” Sensors, vol. 23, no. 6, p. 3195, Mar. 2023, doi: 10.3390/s23063195.