sensors

MDPI

Review

A Comprehensive Survey of Machine Learning Techniques and Models for Object Detection

Maria Trigka and Elias Dritsas *

Industrial Systems Institute, Athena Research and Innovation Center, 26504 Patras, Greece; trigka@isi.gr
* Correspondence: dritsas@isi.gr

Abstract: Object detection is a pivotal research domain within computer vision, with applications spanning from autonomous vehicles to medical diagnostics. This comprehensive survey presents an in-depth analysis of the evolution and significant advancements in object detection, emphasizing the critical role of machine learning (ML) and deep learning (DL) techniques. We explore a wide spectrum of methodologies, ranging from traditional approaches to the latest DL models, thoroughly evaluating their performance, strengths, and limitations. Additionally, the survey delves into various metrics for assessing model effectiveness, including precision, recall, and intersection over union (IoU), while addressing ongoing challenges in the field, such as managing occlusions, varying object scales, and improving real-time processing capabilities. Furthermore, we critically examine recent breakthroughs, including advanced architectures like Transformers, and discuss challenges and future research directions aimed at overcoming existing barriers. By synthesizing current advancements, this survey provides valuable insights for enhancing the robustness, accuracy, and efficiency of object detection systems across diverse and challenging applications.

Keywords: object detection; machine learning; deep learning; techniques; models

Academic Editors: Marco Leo and Zhe-Ming Lu

Received: 10 November 2024
Revised: 18 December 2024
Accepted: 30 December 2024
Published: 2 January 2025

Citation: Trigka, M.; Dritsas, E. A Comprehensive Survey of Machine Learning Techniques and Models for Object Detection. *Sensors* **2025**, *25*, 214. https://doi.org/10.3390/s25010214

1. Introduction

Object detection is a cornerstone of computer vision, focusing on the precise identification and localization of objects within images and videos. This sophisticated technology is fundamental to a wide array of applications across multiple industries. In autonomous vehicles, it is essential for detecting pedestrians, other vehicles, and potential obstacles, ensuring safe navigation. In medical imaging, object detection plays a crucial role in identifying tumours and other abnormalities, significantly aiding in early diagnosis and treatment. Surveillance systems rely on this technology to enhance the monitoring, analysis, and interpretation of activities, providing increased security and situational awareness. Additionally, in industrial settings, object detection is integral to quality control processes, ensuring product standards, and in robotics, where it enables machines to interact more intelligently and accurately with their environment [1].

In recent developments, exponential smoothing has emerged as a powerful image processing technique, particularly when combined with joint power and contrast shrinking in RGB images [2,3]. Moreover, exponential smoothing reduces noise and highlights critical features, leading to enhanced image clarity. This, in turn, allows object detection algorithms to operate more effectively in varied and challenging environments [4]. By refining the visual input, these techniques ensure more precise and reliable detection, which is crucial for maintaining high performance, especially in real-time applications where both accuracy and efficiency are paramount.

The evolution of object detection has been an intriguing journey, beginning with traditional computer vision techniques and advancing through the incorporation of ML and, more recently, DL. Initially, traditional methods were grounded in handcrafted features and simple classifiers, which, despite laying the groundwork for future advancements, struggled to handle the diversity of object appearances and the challenges posed by complex environments. The introduction of ML marked a significant leap forward, bringing with it feature learning methods and more advanced classifiers that enhanced detection capabilities. However, these approaches still encountered limitations when confronted with large-scale datasets and the demands of computational efficiency, necessitating further innovation [5].

DL has radically transformed the field of object detection, offering a powerful means to automatically learn and extract hierarchical features directly from raw data. Central to this revolution are convolutional neural networks (CNNs), which have emerged as the bedrock of contemporary object detection frameworks. CNNs excel in processing and interpreting intricate patterns and structures within images, enabling these systems to achieve unprecedented levels of accuracy and efficiency. This paradigm shift has not only enhanced the precision of object detection but has also dramatically improved processing speed, rendering these systems increasingly suitable for real-time applications across various domains [6].

Object detection encompasses several critical tasks that work in tandem to achieve a detailed and nuanced understanding of visual scenes. The first of these tasks is classification, where the primary objective is to accurately identify and assign a category or label to each detected object within the image. This step is crucial for distinguishing between different types of objects. Next is localization, which focuses on determining the precise location of each object by identifying its coordinates or bounding box within the image. This task ensures that the detected objects are not only recognized but also spatially situated. In some advanced scenarios, segmentation is also employed, which involves delineating the exact boundaries of each object at the pixel level, offering an even more refined understanding of the scene. By integrating these tasks—classification, localization, and segmentation—object detection becomes a comprehensive and sophisticated process, capable of interpreting complex visual environments with high accuracy. This makes it one of the most challenging and vital areas of computer vision [7,8].

The field has witnessed significant advancements through the establishment of influential benchmarks and competitions, including the PASCAL visual object classes (VOC) challenge, the Microsoft COCO (common objects in context) dataset, and the ImageNet large-scale visual recognition challenge (ILSVRC). These initiatives have been pivotal in propelling the progress of object detection, as they offer standardized datasets and rigorous evaluation metrics. By setting high standards and creating a competitive environment, these benchmarks have not only encouraged continuous improvement but also sparked innovation, pushing the boundaries of what is possible in the domain of computer vision [9,10].

Despite the significant advancements in object detection technology, several formidable challenges continue to persist. One of the primary obstacles is managing occlusions, where objects are partially hidden or obscured, making accurate detection difficult. Another challenge lies in addressing varying object scales, where objects appear in different sizes due to changes in perspective or distance, complicating the detection process. The demand for real-time processing adds another layer of complexity, particularly in critical applications such as autonomous driving and live surveillance, where speed and accuracy are paramount. Additionally, achieving consistent and reliable detection in diverse and unstructured environments—where lighting conditions, backgrounds, and object types can

vary widely—remains a substantial challenge that the field must overcome to realize its full potential [11,12].

To tackle these challenges head-on, the research community is persistently exploring cutting-edge techniques and models that can revolutionize object detection. Pioneering innovations, including attention mechanisms, Transformer architectures, and sophisticated ML methods, are poised to transcend current limitations, significantly enhancing the accuracy and efficiency of object detection systems. The impetus for conducting this survey is rooted in the fast-paced advancements and the dynamic evolution within the field of object detection. As the landscape continuously shifts with the emergence of novel models and techniques, there is a pressing need for a thorough and cohesive overview that not only synthesizes these developments but also delves into their practical applications and implications.

Several surveys have been conducted on object detection, as summarized in Table 1. These surveys cover a range of topics, such as DL-based object detection methods [13], benchmark datasets, evaluation metrics and lightweight architectures for edge devices [14], metrics for evaluating object detection algorithms [15], the categorization of methods into one-stage and two-stage detectors [16], challenges specific to small object detection, including the introduction of novel datasets [17], and generic object detection techniques focusing on frameworks, proposal generation, and training strategies [18]. However, many of these works focus narrowly on specific aspects of object detection, such as particular methodologies or application domains, without providing a unified perspective on the field's evolution. Moreover, emerging advancements like Transformer-based architectures and critical challenges, such as occlusion handling and ethical considerations, have not been comprehensively addressed.

Table 1. Overview of surveys covering related topics in object detection.

Reference	Description
[13]	Comprehensive survey of DL-based object detection methods, covering components, strategies, and applications. Highlights advancements in architectures, sampling techniques, and future research directions.
[14]	Discusses benchmark datasets, evaluation metrics, and lightweight networks for object detection, emphasizing performance on edge devices and comparing major architectures.
[15]	Explores metrics and their variants for evaluating object detection algorithms and proposes a standardized implementation for cross-dataset compatibility.
[16]	Reviews state-of-the-art object detection methods, categorizing them into one-stage and two-stage detectors, and discusses their applications in real-world scenarios.
[17]	Focuses on challenges in small object detection, introduces two novel datasets, and evaluates mainstream methods to facilitate progress in this sub-field.
[18]	Covers generic object detection techniques using DL, focusing on frameworks, proposal generation, context modelling, and training strategies. Identifies research trends and challenges.

To bridge the previous gaps, this survey presents a holistic overview of object detection, spanning traditional techniques, modern DL approaches, and future research directions, providing a comprehensive resource for both researchers and practitioners Figure 1 provides a structured roadmap of the object detection domain, showcasing the evolution from initial foundational methods to modern DL architectures. This survey, in particular, aspires to achieve the following objectives:

- Covering the evolution of object detection from traditional methods to modern DL approaches, the survey offers a complete picture of the field's progression and current state.

- Understanding the strengths and limitations of various techniques and models helps in identifying the most suitable approaches for different applications.
- A detailed examination of evaluation metrics ensures that researchers and practitioners can accurately assess the performance of object detection systems.
- By pinpointing the current challenges and proposing future research directions, the survey aims to inspire new solutions and advancements in the field.

The remainder of the paper is structured as follows. Section 2 provides background knowledge and traditional methods. Section 3 outlines the contribution of DL and its impact on object detection. Section 4 describes techniques and methods for the subject under consideration. In Section 5, the evaluation metrics are noted. Section 6 illustrates challenges and future directions. Section 7 provides a practical guide for implementing and experimenting with the object detection methodologies and techniques discussed in this survey. Finally, Section 8 concludes the present survey.

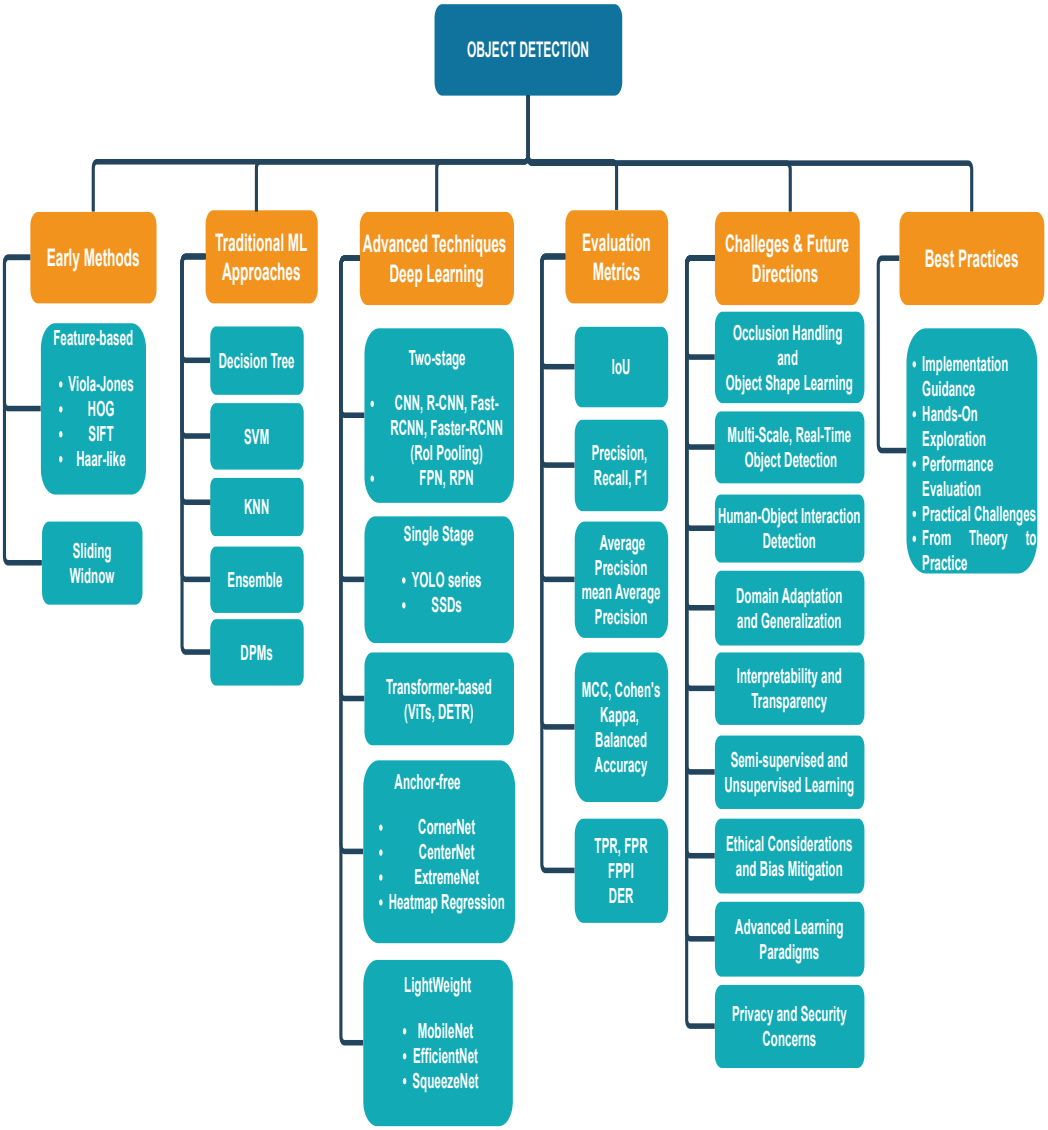

Figure 1. An overview of the object detection landscape.

2. Background and Traditional Approaches

The field of object detection has experienced significant evolution, beginning with early techniques that laid the groundwork for future advancements. This section provides an overview of these foundational methods, emphasizing the initially designed approaches

before transitioning to more sophisticated ML models. These early techniques, while limited by the technological constraints of their time, were instrumental in addressing the core challenges of object detection, such as feature extraction and classification. The subsequent subsections will explore these pioneering methods, illustrating their development and their role in shaping the modern landscape of object detection technologies.

2.1. Early Object Detection Methods

The foundation of object detection in computer vision was built on traditional approaches that relied heavily on handcrafted features and basic classification techniques. In the earliest stages, the primary challenge was how to effectively represent objects in a way that was both computationally feasible and robust to variations in scale, rotation, and illumination.

One of the most influential early methods was the scale-invariant feature transform (SIFT). SIFT was designed to detect and describe local features in images that were invariant to scale and rotation. By identifying key points in an image and generating descriptors that capture local image gradients around those key points, SIFT became a powerful tool for matching objects across different images. Its robustness to changes in viewpoint and illumination made it a popular choice for early object recognition tasks. However, despite its accuracy, SIFT was computationally intensive, which limited its application in real-time systems [19–23].

Around the same time, another critical advancement was the introduction of the histogram of oriented gradients (HOG). HOG features were specifically designed for human detection and became widely used due to their simplicity and effectiveness. The HOG method works by dividing an image into small, connected regions called cells, computing a histogram of gradient directions or edge orientations within each cell, and then normalizing these histograms across blocks of cells. This approach provides a dense grid of feature vectors that capture the underlying structure of objects, making it particularly effective for detecting pedestrians in images. The HOG combination with linear support vector machines (SVMs) in a sliding window approach became the standard pipeline for pedestrian detection. Although HOG was faster than SIFT and more suitable for real-time applications, it still struggled with detecting objects in complex scenes with cluttered backgrounds or under varying lighting conditions [24–29].

Another significant development in early object detection was the introduction of Haar-like features. Haar-like features are simple rectangular features used to represent the difference in intensity between adjacent rectangular groups of pixels. These features could be computed rapidly using an integral image, allowing for real-time detection [30–34]. Viola and Jones combined these features with a cascaded classifier, where a series of increasingly complex classifiers were applied sequentially, allowing for efficient object detection, particularly in face detection. The Viola–Jones detector became one of the first successful real-time object detection systems, widely adopted in applications ranging from security cameras to consumer electronics. Despite its success, the Viola–Jones approach was limited by its reliance on rigid, manually designed features, which struggled with variations in pose, scale, and occlusion [35–39].

In parallel to the development of these feature-based methods, the sliding window technique was a common strategy used across various object detection tasks. The sliding window approach involves scanning the entire image at multiple scales and positions and applying a classifier to each window to determine whether an object is present. While conceptually simple, this brute-force method was computationally expensive, as it required evaluating thousands of windows per image, making it impractical for real-time applications. Additionally, the sliding window method was prone to generating a high

number of false positives, especially in cluttered environments, which further complicated its application [40–44].

These early methods, outlined in Table 2, laid the groundwork for modern object detection by providing essential insights into feature extraction and classification. However, they also highlighted the limitations of handcrafted features and the need for more sophisticated, automated approaches. The reliance on predefined features meant that these methods often lacked the flexibility to handle the diversity and complexity of real-world data. This recognition would eventually drive the shift toward ML and, later, DL approaches that could learn features directly from the data, leading to significant advancements in detection accuracy and efficiency.

Table 2. Overview of works with early methods in object detection.

Method	References	Description
SIFT	[19–23]	Discuss the development and applications of SIFT, a method for detecting and describing local features in images invariant to scale and rotation.
HOG	[24–29]	Focus on HOG, a feature descriptor used for object detection by computing gradients in localized portions of an image, particularly effective for human detection.
Haar-like Features	[30–34]	Explores the use of Haar-like features, which are simple rectangular features used in object detection by comparing pixel intensities, notably applied in face detection.
Viola–Jones detector	[35–39]	Details the Viola–Jones method, which combines Haar-like features with a cascaded classifier for real-time face detection and other object detection tasks.
Sliding Window Method	[40–44]	Describe the Sliding Window technique, a brute-force method for object detection by scanning an image at multiple scales and positions, typically using classifiers.

2.2. Traditional Machine Learning Approaches

As the limitations of purely feature-based methods became increasingly apparent, the object detection community began to explore more sophisticated techniques that could leverage the power of ML to improve detection accuracy and generalization. The integration of ML into object detection marked a significant turning point, as it allowed models to learn patterns and decision boundaries from data rather than relying solely on handcrafted features. The earliest forays into ML for object detection involved the use of simple classifiers, such as decision trees (DTs), K-nearest neighbors (KNN), and SVMs. These classifiers were often paired with features extracted using methods like SIFT or HOG.

DTs, for instance, provided a straightforward way to model decision boundaries based on feature values. However, while DTs were easy to interpret, they often lacked the robustness required for complex object detection tasks, as they were prone to overfitting, particularly when the training data were limited or noisy [45–48].

KNN offered another early ML approach, leveraging a non-parametric method that classified objects based on the closest labeled examples in the feature space. While simple and effective for certain tasks, KNN was computationally expensive during inference, as it required comparing each test instance to all training instances. This made KNN less suitable for large-scale or real-time applications, where speed was a critical factor [49–52].

SVMs emerged as a more powerful alternative, particularly when combined with high-dimensional feature spaces. SVMs aimed to find the optimal hyperplane that maximized the margin between different object classes in the feature space. This approach proved to be highly effective in many object detection tasks, especially when paired with robust feature descriptors like HOG. SVMs could handle non-linear decision boundaries through the use

of kernel functions, which allowed them to perform well on complex datasets. However, SVMs required careful tuning of hyperparameters and were computationally intensive, especially when dealing with large datasets or high-dimensional feature spaces [53–56].

As researchers sought to improve detection accuracy, they turned to ensemble methods, which combined the strengths of multiple classifiers to achieve better performance than any single model. One of the most influential ensemble methods was AdaBoost, a boosting technique that combined several weak classifiers to form a strong classifier. AdaBoost worked by iteratively adjusting the weights of training examples, focusing more on the examples that were misclassified by previous classifiers. This process resulted in a model that was more robust and capable of handling a variety of object detection tasks [57–60].

The combination of AdaBoost with Haar-like features became particularly famous through the work of Viola and Jones, who developed a highly efficient face detection system. The Viola–Jones detector uses a cascade of classifiers, each trained to detect increasingly complex features, allowing for rapid detection of faces in images. The success of the Viola–Jones detector demonstrated the potential of ML to significantly improve the speed and accuracy of object detection. However, despite its effectiveness for specific tasks like face detection, the Viola–Jones approach was still limited by its reliance on rigid, handcrafted features and struggled with variations in pose, scale, and occlusion [61–64].

Parallel to these developments, deformable part models (DPMs) introduced a more sophisticated approach by representing objects as collections of parts that could move relative to each other. This part-based model allowed for greater flexibility in detecting objects under different poses and occlusions, making it particularly effective for detecting complex objects like humans and animals. DPMs used latent SVMs to jointly learn the appearance of object parts and their spatial relationships, leading to improved detection accuracy compared to earlier methods. The success of DPMs on benchmarks like PASCAL VOC underscored the potential of combining ML with more structured representations of objects. However, DPMs were still constrained by their reliance on handcrafted features, which limited their ability to scale to more complex and diverse datasets [65–68]. In Table 3, we summarize recent works that were discussed above and utilize classical ML models for object detection.

The integration of ML into object detection represented a significant advance over previous methods, allowing for more automated and adaptive systems. However, these approaches were not without their limitations. The need for manually designed features and the computational complexity of training and inference posed significant challenges, particularly as datasets grew larger and more diverse. These challenges highlighted the need for further innovation, paving the way for the development of DL methods that could learn both features and classifiers directly from raw data, marking the next major evolution in object detection.

Table 3. Summary of works with traditional ML approaches.

Model	References	Purpose of Use
DT	[45–48]	DTs are used to model decision boundaries based on feature values. They are easy to interpret but can overfit, especially with limited or noisy data.
KNN	[49–52]	KNN is a non-parametric method that classifies objects based on the closest labeled examples in the feature space. It is effective but computationally expensive for large datasets.
SVM	[53–56]	SVMs are powerful for object detection, especially when combined with high-dimensional feature spaces and kernel functions. They require careful tuning and are computationally intensive.
Ensemble Methods	[57–64]	They discuss methods such as voting with NN detectors, ensembling based on class hierarchy, weighted ensemble block and non-maximum suppression ensembling which are used to combine the strengths of multiple classifiers for better performance in object detection. AdaBoost, which combines weak classifiers to form a strong classifier, is applied in various domains: face detection using the Viola–Jones detector, vehicle recognition, and iris localization based on Haar-like features and underwater object detection combined with DL.
DPMs	[65–68]	DPMs represent objects as collections of parts, allowing for flexibility in detecting objects under various poses and occlusions. They improve detection accuracy in complex scenarios.

3. Deep Learning and Its Impact on Object Detection

DL has revolutionized object detection by enabling models to learn features directly from raw data, vastly improving both accuracy and efficiency. This section explores the evolution of DL in object detection, beginning with the foundational CNNs and their groundbreaking applications, followed by the development of faster, more streamlined frameworks like single shot detectors (SSDs) and You Only Look Once (YOLO), and culminating in the latest advancements with anchor-free detection methods.

3.1. CNNs

The introduction of CNNs marked a transformative moment in the field of object detection, fundamentally altering the landscape of computer vision. Unlike traditional methods that depended on manually engineered features, CNNs offered the ability to automatically learn feature hierarchies directly from data. This capability not only improved detection accuracy but also enabled models to generalize better across varying datasets and conditions [69–73].

The breakthrough in using CNNs for object detection began with the development of the regions with convolutional neural networks (R-CNN) model. R-CNN represented a significant departure from earlier methods by leveraging CNNs for feature extraction. The model worked by first generating region proposals from an image, which were then passed through a CNN to extract features. These features were subsequently fed into a classifier to predict the presence and category of objects within each region. This approach demonstrated remarkable improvements in detection accuracy over traditional methods, as CNNs could capture more complex and abstract representations of objects, which were difficult to encode manually [74–78].

However, R-CNN's performance gains came at the cost of computational efficiency. The need to process each region proposal independently through a CNN led to significant redundancy, as many overlapping regions shared similar features. This inefficiency motivated the development of Fast R-CNN, which introduced several key innovations to streamline the detection pipeline [79–82]. By employing a technique known as region of interest (RoI) pooling, Fast R-CNN allowed the sharing of convolutional computations across all region proposals. Instead of feeding each proposal through the entire CNN,

the model first computed a single feature map for the entire image and then extracted features for each proposal from this shared map. This optimization not only reduced the computational burden but also enabled faster training and inference without sacrificing accuracy [83–86].

Building upon the successes of R-CNN and Fast R-CNN, the Faster R-CNN model introduced another critical advancement: the region proposal network (RPN). The RPN replaced the external region proposal generation step with a fully integrated, end-to-end trainable network. By generating region proposals directly from the CNN feature map, Faster R-CNN eliminated the need for external proposal algorithms like selective search, further speeding up the detection process. The RPN worked by sliding a small network over the feature map and predicting object bounds and scores at each position. This approach not only streamlined the pipeline but also improved the quality of the proposals, leading to even better detection performance [87–90]. Faster R-CNN quickly became the benchmark for high-accuracy object detection, setting new standards in various object detection challenges [91–95].

3.2. SSDs and YOLO

The pursuit of faster and more efficient object detection models led to the development of SSDs and the YOLO family of models, both of which redefined the landscape of real-time object detection. These models addressed the limitations of earlier region-based methods, such as the R-CNN family, which, despite their high accuracy, were computationally intensive and unsuitable for real-time applications due to their multi-stage pipelines.

SSDs introduced a groundbreaking approach by eliminating the need for separate region proposal stages. SSDs predict both the object classes and bounding boxes directly from feature maps in a single forward pass through the network. This innovation allows for the detection of objects at multiple scales by applying small convolutional filters to different layers of the feature pyramid, each layer responsible for detecting objects of varying sizes. The key advantage of SSDs is their ability to generate a large number of detections across the entire image simultaneously, significantly reducing the computational cost compared to previous methods. By leveraging a fixed set of default boxes with different aspect ratios and scales, SSDs effectively handle the detection of objects at various resolutions within a single network, enabling a more streamlined and efficient detection process [96–99].

The introduction of YOLO marked another significant advancement in object detection, fundamentally altering how detection tasks were approached. YOLO reframed object detection as a single regression problem, predicting bounding boxes and class probabilities directly from the entire image in one unified process. This end-to-end approach allows YOLO to process images in a single pass, dramatically increasing the speed of detection. The model divides the input image into a grid, with each grid cell predicting a fixed number of bounding boxes and confidence scores for the presence of objects. The simplicity of this approach allows YOLO to operate at an unprecedented speed, achieving real-time performance even on relatively modest hardware. One of the major contributions of YOLO is its unified architecture, which treats object detection as a global problem rather than focusing on individual regions of interest. This holistic view enables YOLO to capture contextual information from the entire image, leading to fewer false positives compared to methods that rely heavily on localized region proposals. However, this design choice also introduced challenges. Early versions of YOLO struggled with small object detection and accurately localizing objects close to each other, primarily due to the coarse grid used for predictions. Additionally, the rigid structure of the grid limited the model's ability to adapt to objects that did not fit neatly within its predefined cells [100–104].

Despite these challenges, YOLO has seen significant improvements through subsequent iterations. YOLOv2 introduced batch normalization and anchor boxes, which enhanced the model's ability to detect objects at different scales and improved localization accuracy [105–109]. YOLOv3 [110–114] further refined the architecture by incorporating multi-scale predictions and a deeper, more powerful feature extractor, Darknet-53, allowing the model to balance speed and accuracy more effectively [115–118].

Moreover, YOLOv4 marked a significant evolution in the YOLO family of object detection models, bringing together a suite of innovations designed to enhance both speed and accuracy [119,120]. The architecture of YOLOv4 integrated cross-stage partial (CSP) connections, which improved the learning process by enabling gradient flow through the network while reducing computation [121]. The backbone of YOLOv4, based on CSPDarknet53, was designed to balance the need for high accuracy and real-time performance [122]. Additionally, YOLOv4 incorporated multiple improvements like the Mish activation function, spatial pyramid pooling (SPP), and path aggregation network (PAN) for better feature fusion, making it one of the most powerful real-time object detectors at its time of release [123–125].

YOLOv5, YOLOv6, YOLOv7, and YOLOv8 each built upon the foundation laid by YOLOv4, iterating on architectural advancements to push the boundaries of object detection further. YOLOv5 introduced a more streamlined and user-friendly implementation, focusing on efficiency with a leaner model architecture that maintained competitive performance [126,127]. YOLOv6 continued this trend, emphasizing modularity and scalability, which allowed for tailored deployment in various industrial applications [128,129]. YOLOv7 made strides in optimizing model complexity and inference speed, introducing new techniques like the extended efficient layer aggregation network (E-ELAN) to maximize the use of parameters and computational resources [130,131]. Finally, YOLOv8 integrated these advances into a more sophisticated architecture, refining the network layers and training methods to achieve superior accuracy and faster inference times, positioning it as one of the most advanced models in the YOLO series [132–134]. Each iteration of YOLO has pushed the boundaries of what is possible in real-time object detection, making it a popular choice for applications where speed is critical, such as in autonomous driving, robotics, and real-time video analysis.

Both SSD and YOLO represent a shift toward more practical and deployable object detection systems. Their ability to operate at high frame rates without significant sacrifices in accuracy has made these systems highly desirable in environments where computational resources are limited and latency is a critical factor. This has opened up new possibilities for real-time applications, from drone navigation to augmented reality, where traditional, slower object detection methods would be impractical.

3.3. Advances in Anchor-Free Detectors

The evolution of object detection has seen a significant shift toward eliminating the reliance on predefined anchor boxes, a core component in many earlier detection frameworks. Anchor boxes, though effective in traditional detectors like Faster R-CNN and SSD, introduced considerable complexity and required meticulous tuning of hyperparameters such as scale, aspect ratio, and the number of anchor boxes per location. Moreover, anchor-based methods often struggled with the inefficiency of matching these predefined boxes to the ground truth, especially when dealing with objects of varying sizes, shapes, and densities. This inherent limitation prompted researchers to explore anchor-free detection methods, which aimed to streamline the detection process and enhance the model's ability to generalize across different object types and environments [135–139].

The emergence of anchor-free detectors was driven by the need to simplify the detection pipeline while maintaining, or even improving, detection accuracy. These methods

bypass the anchor box mechanism entirely by predicting key points, centroids, or other critical features directly from the image, thus reducing the dependence on predefined spatial configurations. One of the pioneering approaches in this category was CornerNet, which re-imagined object detection by framing it as a key point detection problem. Instead of using anchors, CornerNet predicted the top-left and bottom-right corners of bounding boxes and then linked these pairs to form the final detection. This approach offered a more flexible and direct way of localizing objects, particularly excelling in scenarios where objects appeared in non-standard shapes or orientations. By focusing on corners, CornerNet reduced the computational overhead associated with anchor box generation and matching, while also enhancing localization precision [140–143].

Building upon the concepts introduced by CornerNet, CenterNet further advanced the anchor-free paradigm by simplifying the process to focus on predicting the center point of objects along with their dimensions. CenterNet's approach was to identify the center of the bounding box and then regress the object's height and width, effectively capturing the object's spatial extent without the need for anchors. This method not only simplified the detection process but also improved the model's efficiency, particularly in detecting objects of various scales within the same image. The elimination of anchors allowed CenterNet to avoid the pitfalls of anchor box design, such as overlapping boxes and the challenge of balancing positive and negative samples during training. This led to more robust and reliable detections across a range of object categories [144–148].

Another notable advance in the anchor-free domain is the concept of dense keypoint estimation, where the model predicts a dense grid of key points across the image, each potentially corresponding to an object's significant feature, such as its center, edges, or extremities. This approach, exemplified by models like ExtremeNet, extends the idea of corner detection to include additional key points along the boundaries of objects, providing a richer set of cues for object localization. By focusing on key points, these models can effectively handle objects with irregular shapes and complex boundaries, which are often challenging for anchor-based methods [149–152].

The shift to anchor-free detectors also opened up new possibilities for incorporating advanced techniques like heatmap regression and feature pyramid networks (FPNs), which further enhanced the models' ability to detect objects at multiple scales and with high precision. Heatmap regression, for instance, allows the model to predict the probability distribution of object key points across the image, effectively capturing the spatial uncertainty and improving localization accuracy. Combined with multi-scale feature extraction techniques, anchor-free models can achieve state-of-the-art performance in scenarios involving small objects, dense object clusters, and varying object sizes [153–161].

A synopsis of the above-mentioned classification of DL models for object detection is presented and described in Table 4.

Table 4. Summary of works related to DL and object detection.

Model	References	Description
CNN	[69–73]	Discuss the foundational role of CNNs in object detection, focusing on models, methodologies, and various applications.
R-CNN	[74–78]	Cover the evolution of R-CNN models including original R-CNN, Fast R-CNN, and related advancements in region-based detection.
Fast R-CNN	[79–82]	Examine Fast R-CNN, focusing on improvements in efficiency and accuracy by sharing computations across regions.
Faster R-CNN	[91–95]	Discuss Faster R-CNN, which uses RPN for faster and more accurate object detection.
RoI Pooling	[83–86]	Detail the use of RoI pooling in enhancing object detection efficiency and performance in various models.
RPN	[87–90]	Focuses on advancements in RPN technology, contributing to faster and more integrated object detection pipelines.
SSD	[96–99]	Discuss SSDs, which eliminate region proposal stages for faster object detection across multiple scales.
YOLO Family	[100–114,119–134]	Cover the YOLO series, from the original YOLO to YOLOv8, detailing the innovations that allow for real-time object detection.
Darknet-53	[115–118]	Describe the Darknet-53 architecture, highlighting its balance of speed and accuracy.
Anchor-Free Detectors	[135–139]	Discuss the shift toward anchor-free object detection methods, focusing on reducing complexity and improving accuracy.
CornerNet	[140–143]	Focuses on CornerNet, an anchor-free detector that uses keypoint detection to identify object corners.
CenterNet	[144–148]	Examines CenterNet, which simplifies object detection by focusing on predicting the center of objects.
ExtremeNet	[149–152]	Describes ExtremeNet, a model that extends keypoint detection to include object extremities for better localization.
Heatmap regression	[153–156]	Discusses heatmap regression techniques used for predicting key points in anchor-free models.
FPN	[157–161]	Covers advancements in feature pyramid networks, which enhance multi-scale object detection, particularly for small objects.

4. Advanced Techniques and Methods

As object detection has matured, researchers have sought to push the boundaries of what is achievable by exploring innovative architectures and methodologies that go beyond traditional DL models. These advanced techniques and methods have not only refined detection accuracy and efficiency but also opened new avenues for handling increasingly complex scenarios and challenges in the real world.

One of the most transformative advancements in recent years has been the application of Transformer architectures to object detection [162,163]. Originally developed for natural language processing [164], Transformers have revolutionized tasks such as machine translation and text generation by leveraging self-attention mechanisms to capture long-range dependencies within data [165,166]. Recognizing the potential of these mechanisms, the computer vision community adapted Transformers for image-based tasks, leading to the development of Vision Transformers (ViTs) [167,168] and, more specifically, Detection Transformers (DETR) [169,170]. DETR fundamentally rethinks the object detection pipeline by treating it as a direct set prediction problem, where the model predicts a fixed number

of objects without the need for traditional components like anchor boxes or non-maximum suppression. The self-attention mechanism within DETR enables the model to capture complex relationships between objects across the entire image, making it particularly effective in scenarios involving cluttered scenes or overlapping objects. This approach has shown promise in simplifying the object detection pipeline while achieving competitive performance, especially in tasks that require a holistic understanding of the scene [171–174].

Another significant advancement is the integration of multi-scale and context-aware detection techniques, which address the perennial challenge of detecting objects that vary dramatically in size and appearance. Multi-scale detection, a concept that has evolved over time, has become increasingly sophisticated with the introduction of FPNs. FPNs leverage the inherent hierarchical structure of convolutional neural networks to construct feature maps at multiple levels of abstraction, enabling the detection of small, medium, and large objects with equal proficiency. This approach is particularly advantageous in scenarios such as autonomous driving, where objects of interest can range from distant pedestrians to nearby vehicles. By effectively capturing features across different scales, FPNs enhance the model's robustness and accuracy, making them a critical component in modern object detection frameworks [175–178].

The demand for real-time object detection on resource-constrained devices, such as mobile phones and embedded systems, has spurred the development of lightweight models and optimization techniques that maintain high accuracy while minimizing computational requirements [179,180]. Lightweight architectures like MobileNet [181,182], SqueezeNet [183,184], and EfficientNet [185,186] have been specifically designed to operate efficiently on devices with limited processing power and memory. These models achieve remarkable performance by using techniques such as depthwise separable convolutions, which reduce the number of parameters and computations without sacrificing representational power. Additionally, recent advancements in neural architecture search (NAS) have automated the process of discovering optimal network architectures tailored to specific hardware constraints, further enhancing the deployment of object detection models on edge devices [187–189].

Optimization techniques such as model pruning, quantization, and knowledge distillation have also become indispensable tools for adapting high-performance models to run on constrained hardware. Model pruning involves systematically removing less critical weights from a neural network, thereby reducing its size and computational complexity [190–192]. Quantization reduces the precision of the model's weights and activations, enabling faster inference with minimal loss of accuracy [193,194]. Knowledge distillation, on the other hand, transfers knowledge from a large, complex model (the teacher) to a smaller, more efficient model (the student), allowing the student model to achieve comparable performance with a fraction of the resources. These techniques collectively enable the deployment of powerful object detection models in real-world applications where computational resources are limited, such as in autonomous drones, IoT devices, and wearable technology [195–198].

Finally, the exploration of novel modalities and the fusion of multimodal data have opened new frontiers in object detection. While traditional object detection relies primarily on RGB images, integrating additional data types, such as depth maps, infrared imagery, and LiDAR point clouds, can significantly enhance detection performance, particularly in challenging environments. For instance, in low-light conditions or foggy weather, infrared or thermal images can provide critical information that is not visible in the RGB spectrum. Similarly, LiDAR data offer precise distance measurements that can improve the detection and localization of objects in 3D space, which is crucial for applications like autonomous driving and robotics. Multimodal fusion techniques combine these diverse data sources to

create a more comprehensive representation of the environment, enabling more robust and accurate object detection across a wide range of scenarios [199–203].

The previously discussed works that analyzed advanced techniques beyond classical DL models (focusing on innovative architectures and methodologies) are presented in Table 5.

Table 5. A list of advanced techniques and methods beyond classical DL models.

Technique	References	Description
Transformer Architectures	[162–174]	Focus on the application of Transformer models to object detection, such as ViTs and DETR, revolutionizing detection by leveraging self-attention mechanisms.
Multi-Scale and Context-Aware Detection	[175–178]	Discuss the integration of multi-scale detection techniques like FPNs to handle objects of varying sizes and appearances efficiently.
Lightweight Models and Optimization	[179–189]	Cover the development of lightweight architectures like MobileNet, SqueezeNet, EfficientNet, and optimization techniques such as pruning, quantization, and NAS for real-time object detection on resource-constrained devices.
Model Pruning and Quantization	[190–194]	Detail methods for simplifying models through pruning and quantization to enable deployment on devices with limited computational resources.
Knowledge Distillation	[195–198]	Explore techniques for transferring knowledge from larger models to smaller ones to maintain performance while reducing computational demands.
Multimodal Fusion	[199–203]	Discuss the fusion of data from different modalities, such as LiDAR, infrared, and RGB, to improve detection in challenging environments.

5. Evaluation Metrics

Evaluation metrics play a pivotal role in the development and assessment of object detection models, providing quantitative measures to compare performance across different methodologies. These metrics provide insight into how well a model detects, localizes, and classifies objects within images, balancing factors such as prediction confidence and error rates. The following illustration delves into the key metrics used in object detection, offering mathematical formulations and theoretical underpinnings to facilitate a deeper understanding of model evaluation [15,204–207].

Intersection over Union (*IoU*) is a fundamental metric in object detection that measures the overlap between the predicted bounding box and the ground truth bounding box. It is mathematically defined as the ratio of the area of intersection to the area of union between the predicted bounding box (B_p) and the ground truth bounding box (B_{gt}). The formula for IoU is given by the following:

$$IoU = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|},$$

where $\cap$ denotes the intersection (common area) and $\cup$ represents the union (total area covered by both boxes). A higher *IoU* indicates a better match between the predicted and actual objects, with values typically ranging from 0 to 1, where 1 represents a perfect match.

Precision quantifies the proportion of true positive predictions among all positive predictions made by the model. It is a measure of the accuracy of positive predictions. Mathematically, precision is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP), given by the following:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Precision ranges from 0 to 1, where a precision of 1 indicates that every positive prediction made by the model is correct. Precision is crucial when the cost of false positives is high.

Recall, or sensitivity, measures the proportion of actual positives that are correctly identified by the model. It is a measure of the model's ability to capture all relevant instances. Recall is mathematically defined as the ratio of true positives (TP) to the sum of true positives and false negatives (FN), expressed as follows:

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Recall ranges from 0 to 1, with a recall of 1 indicating that the model correctly identifies all positive instances. High recall is critical when the cost of false negatives is high, ensuring that the model captures as many true positives as possible.

Average precision (AP) is a summary metric that represents the precision across different recall levels, integrating the precision–recall curve. AP is calculated as the area under the precision–recall curve, which is computed as the weighted mean of precisions achieved at each threshold, with the change in recall serving as the weight. The formula for AP is given by the following:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n,$$

where R_n is the recall at the nth threshold and P_n is the precision at the nth threshold. AP is a key metric in object detection as it accounts for the trade-off between precision and recall across all thresholds.

Mean average precision (mAP) extends the concept of AP to multiple object classes, providing a single metric that summarizes the performance across all classes. Mathematically, mAP is the mean of the AP values calculated for each class, given by the following:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i,$$

where N is the number of classes and AP_i is the average precision for the ith class. mAP is widely used in object detection challenges as it provides a comprehensive measure of the model's ability to detect and localize objects across different categories.

The $F1$ score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful when precision and recall are both critical and need to be optimized simultaneously. The $F1$ score is mathematically defined as follows:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

which can also be expressed as follows:

$$F1 = \frac{2TP}{2TP + FP + FN}.$$

The $F1$ score ranges from 0 to 1, where a value of 1 indicates perfect precision and recall. It is an important metric when there is an uneven class distribution, as it provides a balanced measure.

The Matthews correlation coefficient (MCC) is a metric that provides a more balanced measure of classification quality, taking into account all four quadrants of the confusion matrix (true positives, true negatives, false positives, and false negatives). It is particularly useful for evaluating binary classification tasks with imbalanced classes. MCC is mathematically defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

MCC returns a value between -1 and +1, where +1 indicates perfect prediction, 0 signifies no better prediction than random, and -1 indicates complete disagreement between prediction and observation.

Cohen's Kappa (κ) is a statistical measure of inter-rater agreement that accounts for the possibility of agreement occurring by chance. It is commonly used to assess the performance of models in multi-class classification tasks. The formula for Cohen's Kappa is given by the following:

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the observed agreement (the accuracy of the model) and P_e is the expected agreement by chance, calculated as follows:

$$P_e = \frac{1}{N^2} \sum_{i=1}^{k} (A_i \cdot B_i).$$

Here, A_i and B_i are the marginal totals of the contingency table, N is the total number of instances, and k is the number of categories. Kappa values range from -1 to 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and negative values indicate agreement worse than chance.

Balanced accuracy is a metric designed to address the limitations of traditional accuracy when dealing with imbalanced datasets. It is the average of recall obtained in each class, ensuring that each class contributes equally to the final measure. Balanced accuracy is mathematically defined as follows:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right).$$

This metric ranges from 0 to 1, where 1 indicates perfect classification across all classes. Balanced accuracy is particularly valuable in scenarios where some classes are underrepresented, providing a more reliable evaluation of model performance.

Localization error measures the discrepancy between the predicted and ground truth bounding boxes. Although no single formula universally defines localization error, it typically involves calculating the Euclidean distance between the centers of the predicted and ground truth boxes, the difference in aspect ratios, and the size differences. This error metric is crucial in object detection tasks where not only the detection but also the precise localization of objects is important.

The confusion matrix is a table that provides a comprehensive summary of the prediction results of a classification problem. It shows the number of correct and incorrect predictions, organized into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The confusion matrix is particularly useful for visualizing the performance of a model, especially in multi-class classification problems, by showing the exact count of predicted versus actual instances for each class.

The ROC curve is a graphical representation of the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different thresholds. The area under the curve (AUC) quantifies the overall ability of the model to discriminate between the positive and negative classes. The true positive rate (TPR) is calculated as follows:

$$TPR = \frac{TP}{TP + FN},$$

and the false positive rate (*FPR*) is calculated as follows:

$$FPR = \frac{FP}{FP + TN}.$$

The AUC is typically computed using numerical integration methods such as the trapezoidal rule, and it ranges from 0 to 1, where 1 indicates a perfect model and 0.5 indicates a model with no discriminative ability.

False positive per image (*FPPI*) measures the average number of false positives detected per image, providing insight into the model's tendency to make incorrect positive predictions. The formula for FPPI is given by the following:

$$FPPI = \frac{FP}{\text{Number of Images}}.$$

A lower *FPPI* indicates that the model generates fewer false positives, which is desirable in object detection tasks, particularly when the cost of false positives is high.

The objectness score is a metric that represents the likelihood that a region contains an object as opposed to just background. This score is typically output by object detection models like YOLO, which evaluate whether a bounding box contains an object based on the learned features. Although there is no specific formula for the objectness score, it is usually a probability value ranging from 0 to 1, with higher scores indicating a higher likelihood of an object being present.

The detection error rate (*DER*) quantifies the percentage of images in which the model fails to detect all objects or incorrectly detects objects. It is calculated as the sum of false positives and false negatives divided by the total number of detections, expressed as follows:

$$DER = \frac{FP + FN}{\text{Total number of detections}}.$$

DER provides a quick summary of the model's performance, with lower values indicating fewer errors in object detection.

These evaluation metrics provide a comprehensive framework for assessing the performance of object detection models, each offering unique insights into various aspects of model accuracy and reliability. Understanding and utilizing these metrics is crucial for the development of robust and effective object detection systems. This detailed examination of evaluation metrics aims to equip researchers and practitioners with the necessary tools to rigorously assess object detection models, fostering the advancement of the field through precise and meaningful evaluations.

Table 6 organizes the discussed metrics into categories such as classification, localization, detection, and advanced metrics, providing a structured framework to analyze their roles in evaluating object detection models while illustrating the inherent trade-offs between accuracy, computational speed, and error minimization. In particular, Faster R-CNN achieves high mAP and IoU, excelling in classification and localization precision, but its slower inference limits real-time applications. Also, YOLO balances speed and accuracy, achieving competitive precision and recall with lower FPPI, making it suitable for real-time tasks. Moreover, SSD optimizes multi-scale detection, offering strong computational efficiency and accuracy. Error metrics like FPPI and DER provide insights into model reliability, with modern models like Faster R-CNN and DETR showing lower detection errors. Localization metrics, particularly IoU, emphasize the spatial precision achieved by models like Faster R-CNN and DETR. By correlating these metrics with model performance, a comprehensive framework is derived helping to understand the strengths and trade-offs of different detection methods.

Table 6. Grouped evaluation metrics in object detection.

Category	Metrics
Classification	Precision Recall (Sensitivity) F1-Score
Localization	IoU
Detection	mAP, DER, FPPI
Agreement	κ, MCC
Advanced	Balanced Accuracy ROC-AUC

6. Challenges and Future Directions

Object detection has seen significant advancements over the years, yet it still faces numerous challenges that must be addressed to enable more effective and reliable applications in real-world scenarios.

One of the most persistent challenges in object detection is handling occlusions, where objects are partially obscured by other objects or the environment. Effective occlusion handling requires models to have a robust understanding of context and the ability to infer missing parts of objects. Recent research efforts have focused on learning occluded shapes to enhance 3D object detection, which remains a crucial area for improvement [208–211].

Another key challenge is detecting objects at multiple scales. In real-world applications, objects often appear in various sizes and perspectives within the same image, making it difficult for models to detect small objects within large scenes and vice versa. Techniques such as multi-scale interactive networks and deep feature learning are being developed to address this issue, offering more sophisticated methods for multi-scale object detection [212–214].

The need for real-time processing is particularly critical in applications like autonomous driving, real-time surveillance, and robotics. Balancing accuracy and speed remains a persistent issue, as more complex models tend to slow down processing. Innovations in lightweight networks and optimized inference algorithms are essential to achieve high-performance real-time object detection without sacrificing accuracy [215–217].

In the realm of human–object interaction detection, another significant challenge is understanding and predicting interactions between humans and objects. Advances in functional generalization techniques are helping models detect these interactions more accurately, which is essential for applications such as surveillance and human-computer interaction [218–221].

Object detection in adverse conditions, such as low-light environments or harsh weather, also presents challenges. Thermal object detection using models like YOLO has shown promise in these challenging conditions, enabling better detection performance when traditional methods might fail [222–226]. Domain adaptation and generalization of models across different environments and conditions remain significant hurdles. Models trained on specific datasets often fail to perform well in diverse settings, such as varying lighting conditions, weather, and camera angles. Progressive domain adaptation techniques are being explored to enhance model robustness and adaptability in different environments [227–229].

Another pressing issue is the interpretability and transparency of DL models in object detection. As these models are often treated as black boxes, understanding their decision-making processes is challenging. The emerging field of Explainable AI (XAI) introduces methods that enhance the interpretability of object detection models by visualizing and explaining learned features, thereby improving their transparency. This increased transparency aids in diagnosing errors, fostering trust in model predictions, and enhancing overall robustness [230–233].

The process of data annotation for training high-quality object detection models is both labor-intensive and time-consuming, which poses a bottleneck in model development. To mitigate this, semi-supervised and unsupervised learning methods are being explored. These approaches leverage large amounts of unlabeled data, reducing dependency on extensive labeled datasets while maintaining high performance [234–237].

Ethical considerations and bias in object detection systems are critical issues that need to be addressed. Bias in training data can lead to models that perform poorly on certain demographic groups or object types, raising concerns about fairness and ethical implications. Ensuring unbiased performance in object detection models requires careful dataset curation and the development of algorithms to detect and mitigate biases [238–240].

Looking forward, there are exciting opportunities for integrating advanced learning paradigms such as reinforcement learning, meta-learning, and multi-task learning into object detection. These techniques can help models learn more effectively from their interactions and adapt to new tasks with minimal additional training, pushing the boundaries of what is possible in object detection [241–243].

As object detection technology continues to evolve, privacy and security concerns are becoming increasingly important, especially in surveillance and autonomous systems. Addressing these concerns involves developing models that respect privacy while maintaining security, ensuring that the technology is used responsibly [244–247]. The list of discussed challenges in key topics related to object detection is summarized in Table 7.

Table 7. A taxonomy of challenges, key topics, and reference works.

Challenge	References	Key Topics/Descriptions
Occlusion Handling and Object Shape Learning	[208–211]	Learning occluded shapes for 3D object detection.
Multi-scale object detection	[212–214]	Multi-scale interactive networks and deep feature learning for object detection.
Real-time object detection	[215–217]	Lightweight networks and optimized inference algorithms for real-time object detection.
Human–object interaction detection	[218–221]	Detecting human–object interactions via functional generalization.
Thermal Object Detection	[222–226]	Using YOLO for object detection in challenging weather conditions.
Domain adaptation and Generalization	[227–229]	Progressive domain adaptation for object detection in various environments.
Interpretability and transparency	[230–233]	Interpretable DL models and methods for visualizing and interpreting features.
Semi-supervised and Unsupervised Learning	[234–237]	Leveraging unlabeled data to reduce dependency on labeled datasets for object detection.
Ethical considerations and Bias mitigation	[238–240]	Addressing bias in object detection systems and ensuring fairness across different groups.
Advanced learning paradigms	[241–243]	Integration of reinforcement learning, meta-learning, and multi-task learning in object detection.
Privacy and security concerns	[244–247]	Addressing privacy and security concerns in object detection applications.

Finally, the future of object detection will likely see a convergence of multiple disciplines and technologies. Research will increasingly focus on creating holistic vision systems that integrate object detection with other computer vision tasks, such as instance segmen-

tation and object tracking. This integration could lead to more versatile and powerful systems capable of operating in complex and dynamic environments. Moreover, fostering collaboration between academia, industry, and interdisciplinary fields will be crucial in driving innovation and establishing standardized benchmarks for the future of object detection [248–252].

7. Best Practices for Object Detection

This section provides a practical guide for implementing and experimenting with the object detection methodologies and techniques discussed in this paper. While Section 6 outlines the challenges and future directions in the field, the following best practices aim to help researchers and practitioners navigate these obstacles effectively. By applying these recommendations, readers can better understand existing approaches, evaluate their performance, and address common challenges, such as occlusions, multi-scale detection, real-time constraints, and data annotation.

The proposed practices are derived from the methods presented earlier, ranging from traditional approaches to advanced DL models. By systematically engaging with these techniques, researchers can gain hands-on experience, identify the most suitable methods for specific tasks, and contribute to the advancement of robust and efficient object detection systems. These practices provide a structured approach for navigating the complexities of object detection and fostering a deeper understanding of the methodologies surveyed in this paper.

A strong foundation begins with understanding the evolution of object detection, from traditional methods such as Viola–Jones, HOG, and SIFT to modern ML approaches. Revisiting these earlier techniques helps contextualize the strengths and limitations that led to the adoption of DL methods. Researchers are encouraged to begin experimentation with pre-trained models like Faster R-CNN, SSD, and the YOLO series, which represent significant milestones in object detection. These models can be evaluated on benchmark datasets, such as PASCAL VOC and COCO, to gain insights into their performance across different scenarios. Performance evaluation is critical to understanding detection accuracy and localization quality. Researchers should rely on established metrics, such as IoU, AP, and mAP, which have been thoroughly presented in the paper. Using these metrics ensures consistency when comparing models and evaluating improvements.

For addressing the challenges of multi-scale detection, methods like FPN, which have been discussed extensively, provide solutions for detecting objects of varying sizes. Data augmentation strategies—such as scaling, cropping, and flipping—further enhance the robustness of models when dealing with images containing objects of different scales. Handling occlusions, as highlighted among the key challenges, can be approached by exploring advanced techniques such as anchor-free methods like CenterNet, which predict object key points and offer greater flexibility in challenging scenes. Combining these techniques with training strategies that simulate occlusions in the dataset enhances the ability of models to detect partially visible objects effectively.

For real-time applications, lightweight architectures such as MobileNet and EfficientNet offer a balance between speed and accuracy. These methods, coupled with optimization techniques like model pruning and quantization, enable the deployment of high-performing models in resource-constrained environments—essential for applications such as mobile devices or edge computing. Moreover, to overcome the labor-intensive nature of data annotation, leveraging semi-supervised learning methods can reduce dependency on large labeled datasets, making object detection more feasible for researchers with limited resources. Finally, staying current with recent advancements, including the emergence of Transformer-based methods like DETR, aligns with the future directions discussed pre-

viously. These methods demonstrate promising potential for addressing challenges such as complex object relationships and cluttered scenes. To ensure transparency and reproducibility, researchers should document their findings and share code using platforms like GitHub or similar repositories.

By following these best practices, researchers can systematically engage with object detection methods, evaluate their performance, and address practical challenges. This structured approach provides a pathway for advancing knowledge and contributing to the ongoing evolution of object detection systems.

8. Conclusions

The field of object detection has experienced significant advancements due to the integration of ML and DL techniques. These advancements have led to substantial improvements in accuracy, speed, and robustness, transforming applications ranging from autonomous vehicles to medical imaging. Traditional approaches, which depended on handcrafted features, have been outpaced by techniques that leverage the power of learning algorithms to extract meaningful features and detect objects in complex environments.

DL, particularly through neural network architectures, has driven much of the recent progress. These models have shown remarkable capability in learning hierarchical features directly from data, making them highly effective for diverse and challenging object detection tasks. The continuous innovation in this area includes exploring new architectures and learning paradigms that push the boundaries of performance and efficiency.

Despite the progress, several challenges persist, such as dealing with occlusions, varying object scales, and the demand for real-time processing. Addressing these challenges requires ongoing research into more sophisticated feature extraction methods, improved model interpretability, and the adoption of unsupervised and semi-supervised learning techniques.

In conclusion, the advancements in object detection through ML and DL have been transformative, yet there is still room for improvement. Future research will likely focus on overcoming existing challenges and further refining these technologies to enhance their applicability and effectiveness across various domains. This survey underscores the rapid evolution of the field and provides insights into potential future directions for continued innovation and improvement in object detection systems.

Author Contributions: E.D. and M.T. conceived the idea, designed and performed the experiments, analyzed the results, drafted the initial manuscript, and revised the final manuscript. All authors have read and agreed to the submitted version of the manuscript.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Amit, Y.; Felzenszwalb, P.; Girshick, R. Object detection. In *Computer Vision: A Reference Guide*; Springer: Cham, Switzerland, 2021; pp. 875–883.
2. Trigka, M.; Dritsas, E.; Moustakas, K. Joint Power and Contrast Shrinking in RGB Images with Exponential Smoothing. In Proceedings of the 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Nafplio, Greece, 26–29 June 2022; IEEE: New York, NY, USA, 2022; pp. 1–5.
3. Dritsas, E.; Trigka, M. A methodology for extracting power-efficient and contrast enhanced rgb images. *Sensors* **2022**, *22*, 1461. [CrossRef] [PubMed]
4. An, L.; Chen, L.; Hao, X. Indoor Fire Detection Algorithm Based on Second-Order Exponential Smoothing and Information Fusion. *Information* **2023**, *14*, 258. [CrossRef]
5. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
6. Xiao, Y.; Tian, Z.; Yu, J.; Zhang, Y.; Liu, S.; Du, S.; Lan, X. A review of object detection based on deep learning. *Multimed. Tools Appl.* **2020**, *79*, 23729–23791. [CrossRef]

7. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10186–10195.
8. Cao, J.; Cholakkal, H.; Anwer, R.M.; Khan, F.S.; Pang, Y.; Shao, L. D2det: Towards high quality object detection and instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11485–11494.
9. Tong, K.; Wu, Y. Rethinking PASCAL-VOC and MS-COCO dataset for small object detection. *J. Vis. Commun. Image Represent.* **2023**, *93*, 103830. [CrossRef]
10. Salari, A.; Djavadifar, A.; Liu, X.; Najjaran, H. Object recognition datasets and challenges: A review. *Neurocomputing* **2022**, *495*, 129–152. [CrossRef]
11. Saleh, K.; Szénási, S.; Vámossy, Z. Occlusion handling in generic object detection: A review. In Proceedings of the 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 21–23 January 2021; IEEE: New York, NY, USA, 2021; pp. 000477–000484.
12. Khosravian, A.; Amirkhani, A.; Kashiani, H.; Masih-Tehrani, M. Generalizing state-of-the-art object detectors for autonomous vehicles in unseen environments. *Expert Syst. Appl.* **2021**, *183*, 115417. [CrossRef]
13. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [CrossRef]
14. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
15. Padilla, R.; Netto, S.L.; Da Silva, E.A. A survey on performance metrics for object-detection algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói, RJ, Brazil, 1–3 July 2020; IEEE: New York, NY, USA, 2020; pp. 237–242.
16. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
17. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards large-scale small object detection: Survey and benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]
18. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
19. Burger, W.; Burge, M.J. Scale-invariant feature transform (SIFT). In *Digital Image Processing: An Algorithmic Introduction*; Springer: Cham, Switzerland, 2022; pp. 709–763.
20. Hossein-Nejad, Z.; Agahi, H.; Mahmoodzadeh, A. Detailed review of the scale invariant feature transform (sift) algorithm; concepts, indices and applications. *J. Mach. Vis. Image Process.* **2020**, *7*, 165–190.
21. Pinthong, T.; Yimyam, W.; Chumuang, N.; Ketcham, M.; Pramkeaw, P.; Utakrit, N. Image Classification of Forage Plants in Fabaceae Family Using Scale Invariant Feature Transform Method. In Proceedings of the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 18–20 November 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
22. ALAMRI, J.; HARRABI, R.; CHAABANE, S.B. Face recognition based on convolution neural network and scale invariant feature transform. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 644–654. [CrossRef]
23. Chumuang, N.; Hiranchan, S.; Ketcham, M.; Yimyam, W.; Pramkeaw, P.; Jensuttiwetchakult, T. Face detection system for public transport service based on scale-invariant feature transform. In Proceedings of the 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 18–20 November 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
24. Ghaffari, S.; Soleimani, P.; Li, K.F.; Capson, D.W. Analysis and comparison of FPGA-based histogram of oriented gradients implementations. *IEEE Access* **2020**, *8*, 79920–79934. [CrossRef]
25. Girsang, N.D. Classification Of Batik Images Using Multilayer Perceptron With Histogram Of Oriented Gradient Feature Extraction. In Proceedings of the International Conference on Science and Engineering, Banda Aceh, Indonesia, 29–30 September 2021; Volume 4, pp. 197–204.
26. Zhou, W.; Gao, S.; Zhang, L.; Lou, X. Histogram of oriented gradients feature extraction from raw Bayer pattern images. *IEEE Trans. Circuits Syst. II Express Briefs* **2020**, *67*, 946–950. [CrossRef]
27. Patel, C.I.; Labana, D.; Pandya, S.; Modi, K.; Ghayvat, H.; Awais, M. Histogram of oriented gradient-based fusion of features for human action recognition in action video sequences. *Sensors* **2020**, *20*, 7299. [CrossRef]
28. Hosseini-Fard, E.; Roshandel-Kahoo, A.; Soleimani-Monfared, M.; Khayer, K.; Ahmadi-Fard, A.R. Automatic seismic image segmentation by introducing a novel strategy in histogram of oriented gradients. *J. Pet. Sci. Eng.* **2022**, *209*, 109971. [CrossRef]
29. Ahmed, N.; Rabbi, S.; Rahman, T.; Mia, R.; Rahman, M. Traffic sign detection and recognition model using support vector machine and histogram of oriented gradient. *Int. J. Inf. Technol. Comput. Sci.* **2021**, *13*, 61–73. [CrossRef]

30. Oualla, M.; Ounachad, K.; Sadiq, A. The fast integration of a rotated rectangle applied to the rotated Haar-like features for rotated objects detection. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 4055–4062. [CrossRef]
31. Indra, E.; Yasir, M.; Andrian, A.; Sitanggang, D.; Sihombing, O.; Tamba, S.P.; Sagala, E. Design and implementation of student attendance system based on face recognition by Haar-like features methods. In Proceedings of the 2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT), Medan, Indonesia, 25–27 June 2020; IEEE: New York, NY, USA, 2020; pp. 336–342.
32. Wu, H.; Cao, Y.; Wei, H.; Tian, Z. Face recognition based on Haar like and Euclidean distance. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 1813, p. 012036.
33. Arreola, L.; Gudiño, G.; Flores, G. Object recognition and tracking using Haar-like Features Cascade Classifiers: Application to a quad-rotor UAV. In Proceedings of the 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), Istanbul, Turkey, 17–20 May 2022; IEEE: New York, NY, USA, 2022; Volume 1, pp. 45–50.
34. Arfi, A.M.; Bal, D.; Hasan, M.A.; Islam, N.; Arafat, Y. Real time human face detection and recognition based on Haar features. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 5–7 June 2020; IEEE: New York, NY, USA, 2020; pp. 517–521.
35. Prasanna, G.S.; Pavani, K.; Singh, M.K. Spliced images detection by using Viola-Jones algorithms method. *Mater. Today Proc.* **2022**, *51*, 924–927. [CrossRef]
36. Jauhari, A.; Anamisa, D.; Negara, Y. Detection system of facial patterns with masks in new normal based on the Viola Jones method. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 1836, p. 012035.
37. Obaida, T.H.; Jamil, A.S.; Hassan, N.F. Real-time face detection in digital video-based on Viola-Jones supported by convolutional neural networks. *Int. J. Electr. Comput. Eng. (IJECE)* **2022**, *12*, 3083. [CrossRef]
38. Fatima, B.; Shahid, A.R.; Ziauddin, S.; Safi, A.A.; Ramzan, H. Driver fatigue detection using viola jones and principal component analysis. *Appl. Artif. Intell.* **2020**, *34*, 456–483. [CrossRef]
39. Al-Tuwaijari, J.M.; Shaker, S.A. Face detection system based viola-jones algorithm. In Proceedings of the 2020 6th International Engineering Conference "Sustainable Technology and Development"(IEC), Erbil, Iraq, 26–27 February 2020; IEEE: New York, NY, USA, 2020; pp. 211–215.
40. Hou, C.; Liu, G.; Tian, Q.; Zhou, Z.; Hua, L.; Lin, Y. Multisignal modulation classification using sliding window detection and complex convolutional network in frequency domain. *IEEE Internet Things J.* **2022**, *9*, 19438–19449. [CrossRef]
41. Lian, R.; Huang, L. DeepWindow: Sliding window based on deep learning for road extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1905–1916. [CrossRef]
42. Bao, S.; Lee, H.H.; Yang, Q.; Remedios, L.W.; Deng, R.; Cui, C.; Cai, L.Y.; Xu, K.; Yu, X.; Chiron, S.; et al. Alleviating tiling effect by random walk sliding window in high-resolution histological whole slide image synthesis. *Proc. Mach. Learn. Res.* **2024**, *227*, 1406. [PubMed]
43. Qin, Y.; Yan, Y.; Ji, H.; Wang, Y. Recursive correlative statistical analysis method with sliding windows for incipient fault detection. *IEEE Trans. Ind. Electron.* **2021**, *69*, 4185–4194. [CrossRef]
44. Savva, A.D.; Kassinopoulos, M.; Smyrnis, N.; Matsopoulos, G.K.; Mitsis, G.D. Effects of motion related outliers in dynamic functional connectivity using the sliding window method. *J. Neurosci. Methods* **2020**, *330*, 108519. [CrossRef]
45. Aboah, A. A vision-based system for traffic anomaly detection using deep learning and decision trees. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 4207–4212.
46. Pal, K.; Goswami, S.; Benia, T.K.; Chowdhury, S.; Mukherjee, M.; Sharma, M. Decision Tree Based Method for Detecting Islanding using Pattern Recognition with HOG Features. In Proceedings of the 2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE), Bangalore, India, 16–17 February 2024; IEEE: New York, NY, USA, 2024; pp. 507–512.
47. Johora, F.T.; Mahbub-Or-Rashid, M.; Yousuf, M.A.; Saha, T.R.; Ahmed, B. Diabetic retinopathy detection using PCA-SIFT and weighted decision tree. In Proceedings of the International Joint Conference on Computational Intelligence: IJCCI 2018, Seville, Spain, 18–20 September 2018; Springer: Berlin/Heidelberg, Germany, 2020; pp. 25–37.
48. Alam, F.; Mehmood, R.; Katib, I. Comparison of decision trees and deep learning for object classification in autonomous driving. In *Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 135–158.
49. Rajput, S.S.; Manimaran, S.; Satapathy, S.K.; Mishra, S.; Mohanty, S.N. Feature Extraction and Recognition of Fingerprint Using KNN-SIFT Algorithm. In Proceedings of the 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 11–12 March 2024; IEEE: New York, NY, USA, 2024; pp. 247–252.
50. Putra, F.A.I.A.; Utaminingrum, F.; Mahmudy, W.F. HOG feature extraction and KNN classification for detecting vehicle in the highway. *IJCCS (Indones. J. Comput. Cybern. Syst.)* **2020**, *14*, 231–242. [CrossRef]
51. Demir, K.; Yaman, O. A HOG Feature Extractor and KNN-Based Method for Underwater Image Classification. *Firat Univ. J. Exp. Comput. Eng.* **2024**, *3*, 1–10. [CrossRef]

52. Fan, Z.; Xie, J.k.; Wang, Z.y.; Liu, P.C.; Qu, S.j.; Huo, L. Image classification method based on improved KNN algorithm. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 1930, p. 012009.
53. Sharma, S.; Raja, L.; Bhatnagar, V.; Sharma, D.; Bhagirath, S.N.; Poonia, R.C. Hybrid HOG-SVM encrypted face detection and recognition model. *J. Discret. Math. Sci. Cryptogr.* **2022**, *25*, 205–218. [CrossRef]
54. Liu, Y.; Zhong, W. A novel SVM network using HOG feature for prohibition traffic sign recognition. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 6942940. [CrossRef]
55. Tomikj, N.; Kulakov, A. Vehicle detection with HOG and linear SVM. *J. Emerg. Comput. Technol.* **2021**, *1*, 6–9.
56. Yang, J.; Chen, Z.; Zhang, J.; Zhang, C.; Zhou, Q.; Yang, J. HOG and SVM algorithm based on vehicle model recognition. In Proceedings of the MIPPR 2019: Pattern Recognition and Computer Vision, Wuhan, China, 2–3 November 2019; SPIE: Bellingham, WA, USA, 2020; Volume 11430, pp. 162–168.
57. Casado-García, Á.; Heras, J. Ensemble methods for object detection. In *ECAI 2020*; IOS Press: Amsterdam, The Netherlands, 2020; pp. 2688–2695.
58. Xu, J.; Wang, W.; Wang, H.; Guo, J. Multi-model ensemble with rich spatial information for object detection. *Pattern Recognit.* **2020**, *99*, 107098. [CrossRef]
59. Mujkic, E.; Christiansen, M.P.; Ravn, O. Object Detection for Agricultural Vehicles: Ensemble Method Based on Hierarchy of Classes. *Sensors* **2023**, *23*, 7285. [CrossRef] [PubMed]
60. Körez, A.; Barışçı, N.; Çetin, A.; Ergün, U. Weighted ensemble object detection with optimized coefficients for remote sensing images. *Isprs Int. J. Geo-Inf.* **2020**, *9*, 370. [CrossRef]
61. Zhang, L.; Wang, J.; An, Z. Vehicle recognition algorithm based on Haar-like features and improved Adaboost classifier. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 807–815. [CrossRef]
62. Lin, Y.N.; Hsieh, T.Y.; Huang, J.J.; Yang, C.Y.; Shen, V.R.; Bui, H.H. Fast Iris localization using Haar-like features and AdaBoost algorithm. *Multimed. Tools Appl.* **2020**, *79*, 34339–34362. [CrossRef]
63. Rani, K.H.; Chakkaravarthy, M. Improving Accuracy in Facial Detection Using Viola-Jones Algorithm AdaBoost Training Method. In *Intelligent Systems and Sustainable Computing: Proceedings of the ICISSC 2021, Fairfield, OH, USA, 12–18 October 2021*; Springer: Singapore, 2022; pp. 127–137.
64. Chen, L.; Liu, Z.; Tong, L.; Jiang, Z.; Wang, S.; Dong, J.; Zhou, H. Underwater object detection using Invert Multi-Class Adaboost with deep learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: New York, NY, USA, 2020; pp. 1–8.
65. Zhang, J.; Li, X.; Sun, L.; Bai, C. DPM-Det: Diffusion Model Object Detection Based on DPM-Solver++ Guided Sampling. In *International Conference on Multimedia Modeling*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 379–393.
66. Jiang, J.; Zhong, X.; Chang, Z.; Gao, X. Object Detection of Transmission Tower Based on DPM. In Proceedings of the 4th International Conference on Information Technologies and Electrical Engineering, Changde, China, 29–31 October 2021; pp. 1–5.
67. Bae, S.H. Deformable part region learning and feature aggregation tree representation for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10817–10834. [CrossRef]
68. Zhang, L.; Li, H.; Zhang, X.; Lou, X. RAW Images-Based Motion-Assisted Object Detection Accelerator Using Deformable Parts Models Features on 1080p Videos. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2024**, *71*, 5054–5066. [CrossRef]
69. Dhillon, A.; Verma, G.K. Convolutional neural network: A review of models, methodologies and applications to object detection. *Prog. Artif. Intell.* **2020**, *9*, 85–112. [CrossRef]
70. Sultana, F.; Sufian, A.; Dutta, P. A review of object detection models based on convolutional neural network. In *Intelligent Computing: Image Processing Based Applications*; Springer: Singapore, 2020; pp. 1–16.
71. Yang, R.; Yu, Y. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front. Oncol.* **2021**, *11*, 638182. [CrossRef] [PubMed]
72. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
73. Reddy, S.; Pillay, N.; Singh, N. Comparative Evaluation of Convolutional Neural Network Object Detection Algorithms for Vehicle Detection. *J. Imaging* **2024**, *10*, 162. [CrossRef] [PubMed]
74. Guo, H.; Yang, X.; Wang, N.; Song, B.; Gao, X. A rotational libra R-CNN method for ship detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5772–5781. [CrossRef]
75. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3520–3529.
76. Beery, S.; Wu, G.; Rathod, V.; Votel, R.; Huang, J. Context r-cnn: Long term temporal context for per-camera object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13075–13085.

77. Hmidani, O.; Alaoui, E.I. A comprehensive survey of the R-CNN family for object detection. In Proceedings of the 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet), Marrakech, Morocco, 12–14 December 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.
78. Mao, J.; Niu, M.; Bai, H.; Liang, X.; Xu, H.; Xu, C. Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2723–2732.
79. Rani, S.; Ghai, D.; Kumar, S. Object detection and recognition using contour based edge detection and fast R-CNN. *Multimed. Tools Appl.* **2022**, *81*, 42183–42207. [CrossRef]
80. Arora, N.; Kumar, Y.; Karkra, R.; Kumar, M. Automatic vehicle detection system in different environment conditions using fast R-CNN. *Multimed. Tools Appl.* **2022**, *81*, 18715–18735. [CrossRef]
81. Mijwil, M.M.; Aggarwal, K.; Doshi, R.; Hiran, K.K.; Gök, M. The Distinction between R-CNN and Fast RCNN in Image Analysis: A Performance Comparison. *Asian J. Appl. Sci.* **2022**, *10*, 429–437.
82. Jiang, L.; Chen, J.; Todo, H.; Tang, Z.; Liu, S.; Li, Y. Application of a fast RCNN based on upper and lower layers in face recognition. *Comput. Intell. Neurosci.* **2021**, *2021*, 9945934. [CrossRef]
83. Raja, R.; Kumar, S.; Mahmood, M.R. Color object detection based image retrieval using ROI segmentation with multi-feature method. *Wirel. Pers. Commun.* **2020**, *112*, 169–192. [CrossRef]
84. Rossi, L.; Karimi, A.; Prati, A. A novel region of interest extraction layer for instance segmentation. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: New York, NY, USA, 2021; pp. 2203–2209.
85. Guo, H.; Yao, S.; Yang, Z.; Zhou, Q.; Nahrstedt, K. CrossRoI: Cross-camera region of interest optimization for efficient real time video analytics at scale. In Proceedings of the 12th ACM Multimedia Systems Conference, Istanbul, Turkey, 28 September–1 October 2021; pp. 186–199.
86. Tang, G.; Liu, S.; Fujino, I.; Claramunt, C.; Wang, Y.; Men, S. H-YOLO: A single-shot ship detection approach based on region of interest preselected network. *Remote Sens.* **2020**, *12*, 4192. [CrossRef]
87. Tao, X.; Gong, Y.; Shi, W.; Cheng, D. Object detection with class aware region proposal network and focused attention objective. *Pattern Recognit. Lett.* **2020**, *130*, 353–361. [CrossRef]
88. Wang, R.; Jiao, L.; Xie, C.; Chen, P.; Du, J.; Li, R. S-RPN: Sampling-balanced region proposal network for small crop pest detection. *Comput. Electron. Agric.* **2021**, *187*, 106290. [CrossRef]
89. Steno, P.; Alsadoon, A.; Prasad, P.; Al-Dala'in, T.; Alsadoon, O.H. A novel enhanced region proposal network and modified loss function: Threat object detection in secure screening using deep learning. *J. Supercomput.* **2021**, *77*, 3840–3869. [CrossRef]
90. Qing, C.; Xiao, T.; Zhang, S.; Li, P. Region Proposal Networks (RPN) Enhanced Slicing for Improved Multi-Scale Object Detection. In Proceedings of the 2024 7th International Conference on Communication Engineering and Technology (ICCET), Tokyo, Japan, 22–24 February 2024; IEEE: New York, NY, USA, 2024; pp. 66–70.
91. Li, W. Analysis of object detection performance based on Faster R-CNN. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 1827, p. 012085.
92. Avola, D.; Cinque, L.; Diko, A.; Fagioli, A.; Foresti, G.L.; Mecca, A.; Pannone, D.; Piciarelli, C. MS-Faster R-CNN: Multi-stream backbone for improved Faster R-CNN object detection and aerial tracking from UAV images. *Remote Sens.* **2021**, *13*, 1670. [CrossRef]
93. Xiao, Y.; Wang, X.; Zhang, P.; Meng, F.; Shao, F. Object detection based on faster R-CNN algorithm with skip pooling and fusion of contextual information. *Sensors* **2020**, *20*, 5490. [CrossRef] [PubMed]
94. Li, C.j.; Qu, Z.; Wang, S.y.; Liu, L. A method of cross-layer fusion multi-object detection and recognition based on improved faster R-CNN model in complex traffic environment. *Pattern Recognit. Lett.* **2021**, *145*, 127–134. [CrossRef]
95. Wang, Y.; Wang, K.; Zhu, Z.; Wang, F.Y. Adversarial attacks on Faster R-CNN object detector. *Neurocomputing* **2020**, *382*, 87–95. [CrossRef]
96. Kumar, A.; Zhang, Z.J.; Lyu, H. Object detection in real time based on improved single shot multi-box detector algorithm. *Eurasip J. Wirel. Commun. Netw.* **2020**, *2020*, 204. [CrossRef]
97. Juneja, A.; Juneja, S.; Soneja, A.; Jain, S. Real time object detection using CNN based single shot detector model. *J. Inf. Technol. Manag.* **2021**, *13*, 62–80.
98. Jin, Y.; Fu, Y.; Wang, W.; Guo, J.; Ren, C.; Xiang, X. Multi-feature fusion and enhancement single shot detector for traffic sign recognition. *IEEE Access* **2020**, *8*, 38931–38940. [CrossRef]
99. Zhou, Z.; Sanders, J.W.; Johnson, J.M.; Gule-Monroe, M.K.; Chen, M.M.; Briere, T.M.; Wang, Y.; Son, J.B.; Pagel, M.D.; Li, J.; et al. Computer-aided detection of brain metastases in T1-weighted MRI for stereotactic radiosurgery using deep learning single-shot detectors. *Radiology* **2020**, *295*, 407–415. [CrossRef]
100. Diwan, T.; Anirudh, G.; Tembhurne, J.V. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimed. Tools Appl.* **2023**, *82*, 9243–9275. [CrossRef]

101. Ahmad, T.; Ma, Y.; Yahya, M.; Ahmad, B.; Nazir, S.; Haq, A.u. Object detection through modified YOLO neural network. *Sci. Program.* **2020**, *2020*, 8403262. [CrossRef]
102. Sirisha, U.; Praveen, S.P.; Srinivasu, P.N.; Barsocchi, P.; Bhoi, A.K. Statistical analysis of design aspects of various YOLO-based deep learning models for object detection. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 126. [CrossRef]
103. Tianjiao, L.; Hong, B. A optimized YOLO method for object detection. In Proceedings of the 2020 16th International Conference on Computational Intelligence and Security (CIS), Guangxi, China, 27–30 November 2020; IEEE: New York, NY, USA, 2020; pp. 30–34.
104. Lee, J.; Hwang, K.i. YOLO with adaptive frame control for real-time object detection applications. *Multimed. Tools Appl.* **2022**, *81*, 36375–36396. [CrossRef]
105. Gupta, S.; Devi, D.T.U. YOLOv2 based real time object detection. *Int. J. Comput. Sci. Trends Technol. IJCST* **2020**, *8*, 26–30.
106. Han, X.; Chang, J.; Wang, K. Real-time object detection based on YOLO-v2 for tiny vehicle object. *Procedia Comput. Sci.* **2021**, *183*, 61–72. [CrossRef]
107. Li, X.; Shi, B.; Nie, T.; Zhang, K.; Wang, W. Multi-object recognition method based on improved yolov2 model. *Inf. Technol. Control* **2021**, *50*, 13–27.
108. Boudjit, K.; Ramzan, N. Human detection based on deep learning YOLO-v2 for real-time UAV applications. *J. Exp. Theor. Artif. Intell.* **2022**, *34*, 527–544. [CrossRef]
109. Safaldin, M.; Zaghden, N.; Mejdoub, M. Moving object detection based on enhanced Yolo-V2 model. In Proceedings of the 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Istanbul, Turkey, 8–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–8.
110. Zhao, L.; Li, S. Object detection algorithm based on improved YOLOv3. *Electronics* **2020**, *9*, 537. [CrossRef]
111. Ezat, W.A.; Dessouky, M.M.; Ismail, N.A. Evaluation of deep learning yolov3 algorithm for object detection and classification. *Menoufia J. Electron. Eng. Res.* **2021**, *30*, 52–57. [CrossRef]
112. Shen, L.; Tao, H.; Ni, Y.; Wang, Y.; Stojanovic, V. Improved YOLOv3 model with feature map cropping for multi-scale road object detection. *Meas. Sci. Technol.* **2023**, *34*, 045406. [CrossRef]
113. Viraktamath, S.; Yavagal, M.; Byahatti, R. Object detection and classification using YOLOv3. *Int. J. Eng. Res. Technol. (IJERT)* **2021**, *10*, 2278-0181.
114. Wang, K.; Liu, M.; Ye, Z. An advanced YOLOv3 method for small-scale road object detection. *Appl. Soft Comput.* **2021**, *112*, 107846. [CrossRef]
115. Pathak, D.; Raju, U. Content-based image retrieval using feature-fusion of GroupNormalized-Inception-Darknet-53 features and handcraft features. *Optik* **2021**, *246*, 167754. [CrossRef]
116. Pathak, D.; Raju, U. Shuffled-Xception-DarkNet-53: A content-based image retrieval model based on deep learning algorithm. *Comput. Electr. Eng.* **2023**, *107*, 108647. [CrossRef]
117. Rajkumar, R.; Gopalakrishnan, S.; Praveena, K.; Venkatesan, M.; Ramamoorthy, K.; Hephzipah, J.J. DARKNET-53 Convolutional Neural Network-Based Image Processing for Breast Cancer Detection. *Mesopotamian J. Artif. Intell. Healthc.* **2024**, *2024*, 59–68. [CrossRef]
118. Nisar, D.-E.M.; Mahum, R.; Azim, T.; Shah, N.-U.H. Proteins Classification Using An Improve Darknet-53 Deep Learning Model. In Proceedings of the 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 27–28 October 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.
119. Ning, M.; Lu, Y.; Hou, W.; Matskin, M. Yolov4-object: An efficient model and method for Object Discovery. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021; IEEE: New York, NY, USA, 2021; pp. 31–36.
120. Gao, C.; Cai, Q.; Ming, S. YOLOv4 object detection algorithm with efficient channel attention mechanism. In Proceedings of the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 25–27 December 2020; IEEE: New York, NY, USA, 2020; pp. 1764–1770.
121. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13029–13038.
122. Mahasin, M.; Dewi, I.A. Comparison of CSPDarkNet53, CSPResNeXt-50, and EfficientNet-B0 backbones on YOLO v4 as object detector. *Int. J. Eng. Sci. Inf. Technol.* **2022**, *2*, 64–72. [CrossRef]
123. Yang, Y.; Xie, G.; Qu, Y. Real-time detection of aircraft objects in remote sensing images based on improved YOLOv4. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021; IEEE: New York, NY, USA, 2021; pp. 1156–1164.
124. Humayun, M.; Ashfaq, F.; Jhanjhi, N.Z.; Alsadun, M.K. Traffic management: Multi-scale vehicle detection in varying weather conditions using yolov4 and spatial pyramid pooling network. *Electronics* **2022**, *11*, 2748. [CrossRef]
125. Yu, H.; Li, X.; Feng, Y.; Han, S. Multiple attentional path aggregation network for marine object detection. *Appl. Intell.* **2023**, *53*, 2434–2451. [CrossRef]

126. Kim, J.H.; Kim, N.; Park, Y.W.; Won, C.S. Object detection and classification based on YOLO-V5 with improved maritime dataset. *J. Mar. Sci. Eng.* **2022**, *10*, 377. [CrossRef]
127. Zhang, Y.; Guo, Z.; Wu, J.; Tian, Y.; Tang, H.; Guo, X. Real-time vehicle detection based on improved yolo v5. *Sustainability* **2022**, *14*, 12274. [CrossRef]
128. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
129. Gupta, C.; Gill, N.S.; Gulia, P.; Chatterjee, J.M. A novel finetuned YOLOv6 transfer learning model for real-time object detection. *J. Real-Time Image Process.* **2023**, *20*, 42. [CrossRef]
130. Yang, H.; Liu, Y.; Wang, S.; Qu, H.; Li, N.; Wu, J.; Yan, Y.; Zhang, H.; Wang, J.; Qiu, J. Improved apple fruit target recognition method based on YOLOv7 model. *Agriculture* **2023**, *13*, 1278. [CrossRef]
131. Feng, G.; Yang, Q.; Tang, C.; Liu, Y.; Wu, X.; Wu, W. Mask-Wearing Detection in Complex Environments Based on Improved YOLOv7. *Appl. Sci.* **2024**, *14*, 3606. [CrossRef]
132. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef] [PubMed]
133. Li, Y.; Fan, Q.; Huang, H.; Han, Z.; Gu, Q. A modified YOLOv8 detection network for UAV aerial image recognition. *Drones* **2023**, *7*, 304. [CrossRef]
134. Sohan, M.; Sai Ram, T.; Reddy, R.; Venkata, C. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 529–545.
135. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1922–1933. [CrossRef]
136. Zand, M.; Etemad, A.; Greenspan, M. Objectbox: From centers to boxes for anchor-free object detection. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 390–406.
137. Xin, Y.; Wang, G.; Mao, M.; Feng, Y.; Dang, Q.; Ma, Y.; Ding, E.; Han, S. PAFNet: An efficient anchor-free object detector guidance. *arXiv* **2021**, arXiv:2104.13534.
138. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
139. Sun, Z.; Dai, M.; Leng, X.; Lei, Y.; Xiong, B.; Ji, K.; Kuang, G. An anchor-free detection method for ship targets in high-resolution SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7799–7816. [CrossRef]
140. Zhao, G.; Dong, T.; Jiang, Y. Corner-based object detection method for reactivating box constraints. *IET Image Process.* **2022**, *16*, 3446–3457. [CrossRef]
141. Duan, K.; Xie, L.; Qi, H.; Bai, S.; Huang, Q.; Tian, Q. Corner proposal network for anchor-free, two-stage object detection. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 399–416.
142. Ma, T.; Tian, W.; Kuang, P.; Xie, Y. An anchor-free object detector with novel corner matching method. *Knowl.-Based Syst.* **2021**, *224*, 107083. [CrossRef]
143. Wu, X.; Xue, Q. An improved CornerNet-lite method for pedestrian detection of unmanned aerial vehicle images. In Proceedings of the 2021 China Automation Congress (CAC), Beijing, China, 22–24 October 2021; IEEE: New York, NY, USA, 2021; pp. 2322–2327.
144. Xu, Z.; Hrustic, E.; Vivet, D. Centernet heatmap propagation for real-time video object detection. In *Computer Vision–ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 220–234.
145. He, H.; Na, Q.; Su, D.; Zhao, K.; Lou, J.; Yang, Y. Improved CenterNet for Accurate and Fast Fitting Object Detection. *Discret. Dyn. Nat. Soc.* **2022**, *2022*, 8417295. [CrossRef]
146. Zou, J.; Ge, B.; Zhang, B. An Improved Object Detection Algorithm Based on CenterNet. In Proceedings of the Artificial Intelligence and Security: 7th International Conference, ICAIS 2021, Dublin, Ireland, 19–23 July 2021; Proceedings, Part I 7; Springer: Berlin/Heidelberg, Germany, 2021; pp. 455–467.
147. Wang, H.; Xu, Y.; Wang, Z.; Cai, Y.; Chen, L.; Li, Y. Centernet-auto: A multi-object visual detection algorithm for autonomous driving scenes based on improved centernet. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 742–752. [CrossRef]
148. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet++ for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 3509–3521. [CrossRef]
149. Pang, S.; Liu, X.; Mao, S.; Jia, H.; Liu, B. Advanced-ExtremeNet: Combined with Depthwise Separable Convolution for the Detection of Steel Bars. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Information Systems, Chongqing, China, 28–30 May 2021; pp. 1–6.
150. Liu, R.; Ao, B.; Wen, Q.; Wu, X.; Yin, J.; Li, K. Combining ExtremeNet with Shape Constraints and Re-Discrimination to Detect Cells from CD56 Images. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; IEEE: New York, NY, USA, 2022; pp. 4587–4593.

151. Zhang, X.; Wang, L.; Ren, Y.; Mao, Z. Research and Implementation of Remote Sensing Image Object Detection Algorithm Based on Improved ExtremeNet. In *International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 859–868.
152. Jiang, H.; Li, S.; Liu, W.; Zheng, H.; Liu, J.; Zhang, Y. Geometry-aware cell detection with deep learning. *Msystems* **2020**, *5*, 10–1128. [CrossRef] [PubMed]
153. Wang, Z.; Zhou, Y.; Wang, F.; Wang, S.; Xu, Z. SDGH-Net: Ship detection in optical remote sensing images based on Gaussian heatmap regression. *Remote Sens.* **2021**, *13*, 499. [CrossRef]
154. Liu, Y.B.; Zeng, M.; Meng, Q.H. Unstructured road vanishing point detection using convolutional neural networks and heatmap regression. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–8. [CrossRef]
155. Li, J.; Wang, M. Multi-person pose estimation with accurate heatmap regression and greedy association. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5521–5535. [CrossRef]
156. Luo, Z.; Wang, Z.; Huang, Y.; Wang, L.; Tan, T.; Zhou, E. Rethinking the heatmap regression for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 11–17 October 2021; pp. 13264–13273.
157. Zhao, G.; Ge, W.; Yu, Y. GraphFPN: Graph feature pyramid network for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2763–2772.
158. Zhu, L.; Lee, F.; Cai, J.; Yu, H.; Chen, Q. An improved feature pyramid network for object detection. *Neurocomputing* **2022**, *483*, 127–139. [CrossRef]
159. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimed.* **2021**, *24*, 1968–1979. [CrossRef]
160. Wang, C.; Zhong, C. Adaptive feature pyramid networks for object detection. *IEEE Access* **2021**, *9*, 107024–107032. [CrossRef]
161. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic feature pyramid network for object detection. In Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, Oahu, HI, USA, 1–4 October 2023; IEEE: New York, NY, USA, 2023; pp. 2184–2189.
162. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
163. Fang, Y.; Liao, B.; Wang, X.; Fang, J.; Qi, J.; Wu, R.; Niu, J.; Liu, W. You only look at one sequence: Rethinking transformer in vision through object detection. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26183–26197.
164. Tunstall, L.; Von Werra, L.; Wolf, T. *Natural Language Processing with Transformers*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2022.
165. Li, B.; Lv, C.; Zhou, Z.; Zhou, T.; Xiao, T.; Ma, A.; Zhu, J. On vision features in multimodal machine translation. *arXiv* **2022**, arXiv:2203.09173.
166. Raisi, Z.; Naiel, M.A.; Younes, G.; Wardell, S.; Zelek, J.S. Transformer-based text detection in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3162–3171.
167. Beal, J.; Kim, E.; Tzeng, E.; Park, D.H.; Zhai, A.; Kislyuk, D. Toward transformer-based object detection. *arXiv* **2020**, arXiv:2012.09958.
168. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.
169. Choi, H.K.; Paik, C.K.; Ko, H.W.; Park, M.C.; Kim, H.J. Recurrent DETR: Transformer-Based Object Detection for Crowded Scenes. *IEEE Access* **2023**, *11*, 78623–78643. [CrossRef]
170. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A survey of visual transformers. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 7478–7498. [CrossRef] [PubMed]
171. Wang, Y.; Zhang, X.; Yang, T.; Sun, J. Anchor DETR: Query design for transformer-based object detection. *arXiv* **2021**, arXiv:2109.07107.
172. Chen, X.; Wei, F.; Zeng, G.; Wang, J. Conditional detr v2: Efficient detection transformer with box queries. *arXiv* **2022**, arXiv:2207.08914.
173. Jiang, H.; Zhang, X.; Xiang, S. Non-Maximum Suppression Guided Label Assignment for Object Detection in Crowd Scenes. *IEEE Trans. Multimed.* **2023**, *26*, 2207–2218. [CrossRef]
174. Tang, X.s.; Xie, X.; Hao, K.; Li, D.; Zhao, M. A line-segment-based non-maximum suppression method for accurate object detection. *Knowl.-Based Syst.* **2022**, *251*, 108885. [CrossRef]
175. Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; Sun, Q. Feature pyramid transformer. In *Computer Vision–ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXVIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 323–339.
176. Min, K.; Lee, G.H.; Lee, S.W. Attentional feature pyramid network for small object detection. *Neural Netw.* **2022**, *155*, 439–450. [CrossRef] [PubMed]

177. Wang, X.; Wang, S.; Ning, C.; Zhou, H. Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7918–7932. [CrossRef]
178. Huang, X.w.; Chen, X.; Zhang, S.f. A deep learning model based on improved feature pyramid networks for small object detection. *Comput. Eng. Sci.* **2023**, *45*, 734.
179. Walambe, R.; Marathe, A.; Kotecha, K.; Ghinea, G. Lightweight object detection ensemble framework for autonomous vehicles in challenging weather conditions. *Comput. Intell. Neurosci.* **2021**, *2021*, 5278820. [CrossRef] [PubMed]
180. Amudhan, A.; Sudheer, A. Lightweight and computationally faster Hypermetropic Convolutional Neural Network for small size object detection. *Image Vis. Comput.* **2022**, *119*, 104396.
181. Chiu, Y.C.; Tsai, C.Y.; Ruan, M.D.; Shen, G.Y.; Lee, T.T. Mobilenet-SSDv2: An improved object detection model for embedded systems. In Proceedings of the 2020 International conference on system science and engineering (ICSSE), Kagawa, Japan, 31 August–3 September 2020; IEEE: New York, NY, USA, 2020; pp. 1–5.
182. Wang, W.; Li, Y.; Zou, T.; Wang, X.; You, J.; Luo, Y. A novel image classification approach via dense-MobileNet models. *Mob. Inf. Syst.* **2020**, *2020*, 7602384. [CrossRef]
183. Ucar, F.; Korkmaz, D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Med. Hypotheses* **2020**, *140*, 109761. [CrossRef]
184. Hassanpour, M.; Malek, H. Learning document image features with SqueezeNet convolutional neural network. *Int. J. Eng.* **2020**, *33*, 1201–1207.
185. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
186. Bloch, L.; Boketta, A.; Keibel, C.; Mense, E.; Michailutschenko, A.; Pelka, O.; Rückert, J.; Willemeit, L.; Friedrich, C.M. Combination of Image and Location Information for Snake Species Identification using Object Detection and EfficientNets. In Proceedings of the CLEF (Working Notes), Thessaloniki, Greece, 22–25 September 2020.
187. Wang, N.; Gao, Y.; Chen, H.; Wang, P.; Tian, Z.; Shen, C.; Zhang, Y. NAS-FCOS: Fast neural architecture search for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11943–11951.
188. Zhang, H.; Wang, L.; Sun, J.; Sun, L.; Kobashi, H.; Imamura, N. NAS-EOD: An end-to-end Neural Architecture Search method for Efficient Object Detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: New York, NY, USA, 2021; pp. 1446–1451.
189. Jiang, C.; Xu, H.; Zhang, W.; Liang, X.; Li, Z. SP-NAS: Serial-to-parallel backbone search for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11863–11872.
190. Pasandi, M.M.; Hajabdollahi, M.; Karimi, N.; Samavi, S. Modeling of pruning techniques for deep neural networks simplification. *arXiv* **2020**, arXiv:2001.04062.
191. Zhang, Y.; Yao, Y.; Ram, P.; Zhao, P.; Chen, T.; Hong, M.; Wang, Y.; Liu, S. Advancing model pruning via bi-level optimization. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 18309–18326.
192. Roy, S.; Panda, P.; Srinivasan, G.; Raghunathan, A. Pruning filters while training for efficiently optimizing deep learning networks. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: New York, NY, USA, 2020; pp. 1–7.
193. Siddegowda, S.; Fournarakis, M.; Nagel, M.; Blankevoort, T.; Patel, C.; Khobare, A. Neural network quantization with ai model efficiency toolkit (aimet). *arXiv* **2022**, arXiv:2201.08442.
194. Chen, W.; Wang, P.; Cheng, J. Towards mixed-precision quantization of neural networks via constrained optimization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5350–5359.
195. Tang, J.; Shivanna, R.; Zhao, Z.; Lin, D.; Singh, A.; Chi, E.H.; Jain, S. Understanding and improving knowledge distillation. *arXiv* **2020**, arXiv:2002.03532.
196. Abbasi, S.; Hajabdollahi, M.; Karimi, N.; Samavi, S. Modeling teacher-student techniques in deep neural networks for knowledge distillation. In Proceedings of the 2020 International Conference on Machine Vision and Image Processing (MVIP), Qom, Iran, 18–20 February 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
197. Stanton, S.; Izmailov, P.; Kirichenko, P.; Alemi, A.A.; Wilson, A.G. Does knowledge distillation really work? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 6906–6919.
198. Li, Z.; Xu, P.; Chang, X.; Yang, L.; Zhang, Y.; Yao, L.; Chen, X. When object detection meets knowledge distillation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10555–10579. [CrossRef]
199. Wu, Y.; Wang, Y.; Zhang, S.; Ogai, H. Deep 3D object detection networks using LiDAR data: A review. *IEEE Sens. J.* **2020**, *21*, 1152–1171. [CrossRef]

200. Xu, S.; Zhou, D.; Fang, J.; Yin, J.; Bin, Z.; Zhang, L. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; IEEE: New York, NY, USA, 2021; pp. 3047–3054.
201. Wang, L.; Zhang, X.; Song, Z.; Bi, J.; Zhang, G.; Wei, H.; Tang, L.; Yang, L.; Li, J.; Jia, C.; et al. Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. *IEEE Trans. Intell. Veh.* **2023**, *8*, 3781–3798. [CrossRef]
202. Zamanakos, G.; Tsochatzidis, L.; Amanatiadis, A.; Pratikakis, I. A comprehensive survey of LIDAR-based 3D object detection methods with deep learning for autonomous driving. *Comput. Graph.* **2021**, *99*, 153–181. [CrossRef]
203. Sudharshan, V.; Seidel, P.; Ghamisi, P.; Lorenz, S.; Fuchs, M.; Fareedh, J.S.; Neubert, P.; Schubert, S.; Gloaguen, R. Object detection routine for material streams combining RGB and hyperspectral reflectance data based on guided object localization. *IEEE Sens. J.* **2020**, *20*, 11490–11498. [CrossRef]
204. Zhu, H.; Wei, H.; Li, B.; Yuan, X.; Kehtarnavaz, N. A review of video object detection: Datasets, metrics and methods. *Appl. Sci.* **2020**, *10*, 7834. [CrossRef]
205. Padilla, R.; Passos, W.L.; Dias, T.L.; Netto, S.L.; Da Silva, E.A. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics* **2021**, *10*, 279. [CrossRef]
206. Sharma, V.K.; Mir, R.N. A comprehensive and systematic look up into deep learning based object detection techniques: A review. *Comput. Sci. Rev.* **2020**, *38*, 100301. [CrossRef]
207. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
208. Xu, Q.; Zhong, Y.; Neumann, U. Behind the curtain: Learning occluded shapes for 3d object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 2893–2901.
209. Yang, C.; Chong, P.; Lam, P. An occlusion handling evaluation criterion for deep learning object segmentation. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 1880, p. 012008.
210. Zhang, T.; Huang, B.; Wang, Y. Object-occluded human shape and pose estimation from a single color image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 7376–7385.
211. Mehmood, K.; Jalil, A.; Ali, A.; Khan, B.; Murad, M.; Khan, W.U.; He, Y. Context-aware and occlusion handling mechanism for online visual object tracking. *Electronics* **2020**, *10*, 43. [CrossRef]
212. Pang, Y.; Zhao, X.; Zhang, L.; Lu, H. Multi-scale interactive network for salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9413–9422.
213. Ma, W.; Wu, Y.; Cen, F.; Wang, G. Mdfn: Multi-scale deep feature learning network for object detection. *Pattern Recognit.* **2020**, *100*, 107149. [CrossRef]
214. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12595–12604.
215. Yun, J.; Jiang, D.; Liu, Y.; Sun, Y.; Tao, B.; Kong, J.; Tian, J.; Tong, X.; Xu, M.; Fang, Z. Real-time target detection method based on lightweight convolutional neural network. *Front. Bioeng. Biotechnol.* **2022**, *10*, 861286. [CrossRef] [PubMed]
216. Heo, S.; Cho, S.; Kim, Y.; Kim, H. Real-time object detection system with multi-path neural networks. In Proceedings of the 2020 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), Sydney, Australia, 21–24 April 2020; IEEE: New York, NY, USA, 2020; pp. 174–187.
217. Chandana, R.; Ramachandra, A. Real time object detection system with YOLO and CNN models: A review. *arXiv* **2022**, arXiv2208.
218. Bansal, A.; Rambhatla, S.S.; Shrivastava, A.; Chellappa, R. Detecting human-object interactions via functional generalization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10460–10469.
219. Antoun, M.; Asmar, D. Human object interaction detection: Design and survey. *Image Vis. Comput.* **2023**, *130*, 104617. [CrossRef]
220. Wang, T.; Yang, T.; Danelljan, M.; Khan, F.S.; Zhang, X.; Sun, J. Learning human-object interaction detection using interaction points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4116–4125.
221. Kim, S.; Jung, D.; Cho, M. Relational context learning for human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2925–2934.
222. Krišto, M.; Ivasic-Kos, M.; Pobar, M. Thermal object detection in difficult weather conditions using YOLO. *IEEE Access* **2020**, *8*, 125459–125476. [CrossRef]
223. Jiang, C.; Ren, H.; Ye, X.; Zhu, J.; Zeng, H.; Nan, Y.; Sun, M.; Ren, X.; Huo, H. Object detection from UAV thermal infrared images and videos using YOLO models. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102912. [CrossRef]
224. Ivasic-Kos, M.; Kristo, M.; Pobar, M. Person Detection in thermal videos using YOLO. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys), London, UK, 5–6 September 2019*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 2, pp. 254–267.

225. Mushahar, M.F.A.; Zaini, N. Human body temperature detection based on thermal imaging and screening using YOLO Person Detection. In Proceedings of the 2021 11th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 27–28 August 2021; IEEE: New York, NY, USA, 2021; pp. 222–227.
226. Ippalapally, R.; Mudumba, S.H.; Adkay, M.; HR, N.V. Object detection using thermal imaging. In Proceedings of the 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 10–13 December 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
227. Hsu, H.K.; Yao, C.H.; Tsai, Y.H.; Hung, W.C.; Tseng, H.Y.; Singh, M.; Yang, M.H. Progressive domain adaptation for object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Aspen, CO, USA, 1–5 March 2020; pp. 749–757.
228. Zhang, C.; Li, Z.; Liu, J.; Peng, P.; Ye, Q.; Lu, S.; Huang, T.; Tian, Y. Self-guided adaptation: Progressive representation alignment for domain adaptive object detection. *IEEE Trans. Multimed.* **2021**, *24*, 2246–2258. [CrossRef]
229. Li, W.; Li, L.; Yang, H. Progressive cross-domain knowledge distillation for efficient unsupervised domain adaptive object detection. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105774. [CrossRef]
230. Li, X.; Xiong, H.; Li, X.; Wu, X.; Zhang, X.; Liu, J.; Bian, J.; Dou, D. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* **2022**, *64*, 3197–3234. [CrossRef]
231. Xie, E.; Wang, W.; Wang, W.; Ding, M.; Shen, C.; Luo, P. Segmenting transparent objects in the wild. In *Computer Vision–ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 696–711.
232. Huff, D.T.; Weisman, A.J.; Jeraj, R. Interpretation and visualization techniques for deep learning models in medical imaging. *Phys. Med. Biol.* **2021**, *66*, 04TR01. [CrossRef] [PubMed]
233. Mi, J.X.; Li, A.D.; Zhou, L.F. Review study of interpretation methods for future interpretable machine learning. *IEEE Access* **2020**, *8*, 191969–191985. [CrossRef]
234. Zhang, F.; Pan, T.; Wang, B. Semi-supervised object detection with adaptive class-rebalancing self-training. In Proceedings of the AAAI conference on artificial intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 3252–3261.
235. Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; Luo, P. Detco: Unsupervised contrastive learning for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 8392–8401.
236. Elezi, I.; Yu, Z.; Anandkumar, A.; Leal-Taixe, L.; Alvarez, J.M. Towards reducing labeling cost in deep object detection. *arXiv* **2021**, arXiv:2106.11921.
237. Liu, N.; Xu, X.; Gao, Y.; Zhao, Y.; Li, H.C. Semi-supervised object detection with uncurated unlabeled data for remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *129*, 103814. [CrossRef]
238. Waelen, R.A. The ethics of computer vision: An overview in terms of power. *AI Ethics* **2024**, *4*, 353–362. [CrossRef]
239. Fabbrizzi, S.; Papadopoulos, S.; Ntoutsi, E.; Kompatsiaris, I. A survey on bias in visual datasets. *Comput. Vis. Image Underst.* **2022**, *223*, 103552. [CrossRef]
240. Wang, Z.; Qinami, K.; Karakozis, I.C.; Genova, K.; Nair, P.; Hata, K.; Russakovsky, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8919–8928.
241. Le, N.; Rathour, V.S.; Yamazaki, K.; Luu, K.; Savvides, M. Deep reinforcement learning in computer vision: A comprehensive survey. *Artif. Intell. Rev.* **2022**, *55*, 2733–2819. [CrossRef]
242. Cheng, M.; Wang, H.; Long, Y. Meta-learning-based incremental few-shot object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2158–2169. [CrossRef]
243. Zhou, D.; Fang, J.; Song, X.; Liu, L.; Yin, J.; Dai, Y.; Li, H.; Yang, R. Joint 3d instance segmentation and object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1839–1849.
244. Marella, S.T.; Parisa, H.S.K. Introduction to quantum computing. In *Quantum Computing and Communications*; Intech Open: London, UK, 2020.
245. Wang, Y.K.; Wang, S.E.; Wu, P.H. Spike-Event Object Detection for Neuromorphic Vision. *IEEE Access* **2023**, *11*, 5215–5230. [CrossRef]
246. Yao, X.; Farha, F.; Li, R.; Psychoula, I.; Chen, L.; Ning, H. Security and privacy issues of physical objects in the IoT: Challenges and opportunities. *Digit. Commun. Netw.* **2021**, *7*, 373–384. [CrossRef]
247. Liu, X.; Xie, L.; Wang, Y.; Zou, J.; Xiong, J.; Ying, Z.; Vasilakos, A.V. Privacy and security issues in deep learning: A survey. *IEEE Access* **2020**, *9*, 4566–4593. [CrossRef]
248. Yang, J.; Wang, C.; Jiang, B.; Song, H.; Meng, Q. Visual perception enabled industry intelligence: State of the art, challenges and prospects. *IEEE Trans. Ind. Inform.* **2020**, *17*, 2204–2219. [CrossRef]
249. Xiao, B.; Kang, S.C. Development of an image data set of construction machines for deep learning object detection. *J. Comput. Civ. Eng.* **2021**, *35*, 05020005. [CrossRef]

250. Zhou, L.; Zhang, L.; Konz, N. Computer vision techniques in manufacturing. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *53*, 105–117. [CrossRef]
251. Paneru, S.; Jeelani, I. Computer vision applications in construction: Current state, opportunities & challenges. *Autom. Constr.* **2021**, *132*, 103940.
252. Reja, V.K.; Varghese, K.; Ha, Q.P. Computer vision-based construction progress monitoring. *Autom. Constr.* **2022**, *138*, 104245. [CrossRef]