



## Article

# Small Object Detection and Tracking: A Comprehensive Review

Behzad Mirzaei <sup>1</sup>, Hossein Nezamabadi-pour <sup>1</sup>, Amir Raouf <sup>2</sup> and Reza Derakhshani <sup>2,3,\*</sup>

<sup>1</sup> Intelligent Data Processing Laboratory (IDPL), Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman 76169-13439, Iran

<sup>2</sup> Department of Earth Sciences, Utrecht University, 3584CB Utrecht, The Netherlands

<sup>3</sup> Department of Geology, Shahid Bahonar University of Kerman, Kerman 76169-13439, Iran

\* Correspondence: r.derakhshani@uu.nl

**Abstract:** Object detection and tracking are vital in computer vision and visual surveillance, allowing for the detection, recognition, and subsequent tracking of objects within images or video sequences. These tasks underpin surveillance systems, facilitating automatic video annotation, identification of significant events, and detection of abnormal activities. However, detecting and tracking small objects introduce significant challenges within computer vision due to their subtle appearance and limited distinguishing features, which results in a scarcity of crucial information. This deficit complicates the tracking process, often leading to diminished efficiency and accuracy. To shed light on the intricacies of small object detection and tracking, we undertook a comprehensive review of the existing methods in this area, categorizing them from various perspectives. We also presented an overview of available datasets specifically curated for small object detection and tracking, aiming to inform and benefit future research in this domain. We further delineated the most widely used evaluation metrics for assessing the performance of small object detection and tracking techniques. Finally, we examined the present challenges within this field and discussed prospective future trends. By tackling these issues and leveraging upcoming trends, we aim to push forward the boundaries in small object detection and tracking, thereby augmenting the functionality of surveillance systems and broadening their real-world applicability.

**Keywords:** small object; detection; tracking; computer vision; survey



**Citation:** Mirzaei, B.; Nezamabadi-pour, H.; Raouf, A.; Derakhshani, R. Small Object Detection and Tracking: A Comprehensive Review. *Sensors* **2023**, *23*, 6887. <https://doi.org/10.3390/s23156887>

Academic Editor: Yun Zhang

Received: 5 July 2023

Revised: 27 July 2023

Accepted: 1 August 2023

Published: 3 August 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Detecting and tracking moving objects in various forms of visual media is a critical aspect of numerous applications, from video surveillance [1] to intelligent traffic management [2] and digital city infrastructure [3]. The platforms implementing these tasks range from closed-circuit televisions (CCTVs) [4] and aircraft [5,6] and unmanned aerial vehicles (UAVs) [7] to the transport of nano-particles in soil for environmental protection [8].

Object detection involves identifying a target object within a single frame or image, whereas object tracking focuses on estimating or predicting the object's position throughout the video sequence, given its initial location [8–10]. These capabilities are essential in a wide array of computer vision tasks, including surveillance, autonomous navigation for vehicles and robots, and more. Object tracking typically occurs in two distinct contexts. Single object tracking (SOT) is focused on a singular target throughout the video, whereas multiple object tracking (MOT) or multiple target tracking (MTT) involves keeping track of numerous objects simultaneously [11]. For instance, the military may use tracking for national surveillance, including border patrols and base perimeter security, or for improving missile guidance systems. In the sporting domain, systems like Hawk-eye monitor ball movement, aiding line judges and player strategy development.

Detecting and tracking objects in infrared (IR) images is another critical application where the target might vary in size or approach at different velocities [12]. The tasks involved in detection and tracking largely depend on the granularity of information the user

seeks to extract [13]. Tracking can be quite intricate due to numerous factors. Challenges arise when objects move rapidly, experience occlusion, or become invisible. Difficulties also emerge in the presence of noise, non-rigid objects, rotation, scale changes, and moving cameras. Despite notable advancements, these complexities persist, especially when tracking small objects from a considerable distance [14].

A small object is one that appears tiny within the video frame, like a distant paraglider or soccer ball, often due to being tracked from afar. These objects, common in applications such as UAVs, remote sensing, and ball sports, may exhibit poor appearance attributes and are challenging to detect and track due to their size. They can often be mistaken for noise due to their minuscule size, negatively impacting tracking accuracy. The term “small objects” is commonly defined in two ways. First, it may refer to objects that are physically smaller in the real world. Alternatively, as per the MS-COCO [15] metric evaluation, small objects are those with an area of  $32 \times 32$  pixels or less, a threshold generally accepted for datasets involving common objects. Small object detection and tracking thus represent a specialized facet of object detection and tracking, necessitating distinct techniques for handling digital images and videos. For instance, aerial images often require advanced methods due to the inherent nature of small objects [16]. Figure 1 provides several examples of small objects.



**Figure 1.** Some representative instances of small objects.

Previous research primarily focused on the detection and tracking of larger objects within scenes, with lesser emphasis on smaller targets. This revealed a necessity for more nuanced and precise algorithms to address small object detection and tracking. Despite the prevalent usage and inherent challenges associated with small object detection and tracking, a comprehensive review focusing specifically on this subject has been noticeably absent. As such, the main purpose of this study was to thoroughly review the existing methods and provide a structured taxonomy.

The primary findings of this research can be outlined as follows:

- A detailed review of methods for detecting and tracking small objects, thereby addressing a significant research gap. To our knowledge, such a comprehensive review has not been previously conducted. The proposed taxonomies aim to provide researchers with a broader and more profound understanding of small object detection and tracking.
- Classification of existing methods into five categories: filter-based methods, search-based methods, background detection-based methods, classical computer-vision-based methods, and deep-learning-based methods.
- Segregation of public datasets used for small detection and object tracking into two categories: spectrum-based video datasets and source position-based video datasets. The datasets are introduced, and the most commonly employed evaluation metrics for detecting and tracking small objects are presented.
- A discussion on the primary challenges in this field, coupled with an analysis of the strengths and weaknesses of methods within each category. We also highlight potential future research trends.

In the ensuing sections of this paper, we propose a comprehensive taxonomy for small object detection and tracking methods. Section 2 undertakes a review of existing methods

from various perspectives. Section 3 provides details on the types of public datasets used for small object detection and tracking. Section 4 presents common evaluation metrics for the detection and tracking phases. In Section 5, we delve into the current challenges in the field and propose potential future trends. Section 6 concludes the paper.

Our goal in reviewing existing methods, addressing challenges, and suggesting future directions is to advance the field of small object detection and tracking. By doing so, we hope to facilitate improvements in various computer vision applications such as surveillance systems, autonomous navigation, and object tracking in dynamic environments.

## 2. Taxonomy and Review the Small Object Detection and Tracking Methods

Although many different approaches have been suggested in the field of detection and tracking of large objects, these methods are limited for detection and tracking of small objects which can be classified into two main groups such as unified track-and-detection methods and track-by-detection methods.

Unified track-and-detection methods are grouped into two categories as filter-based methods and search-based methods. Also, track-by-detection methods are divided as three other categories, including background detection-based methods, classical computer-vision-based methods, and deep-learning-based methods. This section will provide an overview of these methods.

### 2.1. Unified Track-and-Detection Methods

These types of methods perform the tracking and detection processes together in a singular framework without a distinct detector. These methods typically require manual initialization, and the number of objects in the initial frame is fixed. Subsequently, these objects are located and tracked throughout the succeeding frames. We review these methods in the subsections below.

#### 2.1.1. Filter-Based Methods

The methods under this category primarily employ Kalman and Particle filters for the tracking process. For instance, in Huang et al. [17], an adaptive particle filter coupled with an effective proposal distribution was proposed to handle cluttered backgrounds and occlusions efficiently. To facilitate a more diverse proposal distribution, an adaptive motion model was used. Additionally, the template correlation was integrated with motion continuity and trajectory smoothness in the observation likelihood, further eliminating visual distractions.

In the study by Habibi et al. [18], tracking was integrated with a super-resolution technique wherein a high-resolution image was created from multiple low-resolution images. Given that super-resolution enhanced the visual quality of small objects, the process provided more tracking information, thereby increasing precision. The tracking process was then conducted through an adaptive Particle filter, as proposed in Huang et al. [17].

Liu et al. [19] put forth an approach grounded in super-resolution, using convolutional neural network (CNN) to track small objects. A deep-learning network was deployed to enhance the visual quality of small objects, subsequently improving the tracking performance. A Particle filter was then employed for tracking [20].

In their research, Wu et al. [21] proposed an enhanced kernel correlation filter (KCF)-based approach for tracking small objects in satellite videos. Occlusion presents a significant hurdle in object tracking, especially apparent in satellite videos due to the minute size of objects, making them more susceptible to occlusion. The methodology in this study used the average peak correlation energy and the peak value of the response map to determine potential object occlusion. The object's subsequent location was forecasted employing a Kalman filter.

### 2.1.2. Search-Based Methods

Search-based methods attempt to discover the best tracks through extensive searching, with a particular focus on small object tracking. Notable among these methods is the algorithm proposed by Blostein et al. [22], dubbed Multiple Hypothesis Testing (MHT). This method operates under the assumption that the intensity values of background and noise are lower than the mean target intensity. In this approach, the track tree roots are chosen from a predetermined number of points with the highest intensity value. For each root, the algorithm selects neighboring points in the subsequent frame to construct a track tree. Within MHT, there are two thresholds— $T_1$  and  $T_2$ —against which each point on the track is compared. If the new point surpasses  $T_2$ , the algorithm records the track and proceeds to the next frame. If the point falls below  $T_1$ , the track is rejected. However, if the new point lies between  $T_1$  and  $T_2$ , the algorithm defers the decision to the next frame. Ultimately, the tree is pruned to yield a desired number of tracks. Nonetheless, this method faces challenges when tracking fast-moving small objects, as the search area increases exponentially. This makes current MHT algorithms computationally impractical for objects moving at speeds exceeding 1 pixel/frame. In response to this problem, Ahmadi et al. [23] utilized the Multi-Objective Particle Swarm Optimization algorithm (MOPSO) [23] to identify the most optimal track within each root.

Salari et al. [24] presented an effective algorithm for tracking dim targets within digital image sequences. The algorithm operates in two stages: noise removal and tracking. Initially, the Total Variation (TV) filtering technique is employed to improve the Signal Noise Ratio (SNR) and eliminate the image's noise. Subsequently, to detect and track dim tiny targets, a genetic algorithm with associated genetic operators and encoding is used. In the study by Shaik et al. [25], Bayesian techniques were deployed for the detection and tracking of targets in infrared (IR) images. The algorithm begins by applying preprocessing to incoming IR targets to reduce noise and segmentation. The initial position of the object is ascertained utilizing ground truth (GT) data. Subsequently, a grid composed of segments around the target's position in the ensuing frame is chosen, and regions with high-intensity within this segment are highlighted. Employing Bayesian probabilistic methodologies, the likelihood of the object shifting its position from the current frame to any high-intensity location within this grid is then calculated. The position suggesting the highest probability is chosen, and the object's position in the following frame is established. Given that an object's intensity may not necessarily be the highest in a frame, the position and intensity of the object in the previous frame are considered in the Bayesian probabilistic equation to determine its position in the next frame.

## 2.2. Track-by-Detection Methods

In track-by-detection methods, the detector autonomously identifies the desired object or objects in each frame. Subsequently, the tracking process is conducted to associate the detected objects across successive frames. The objects are initially detected, and then their trajectories are linked. This category of methods is capable of handling frames containing a variable number of objects. Thus, if a new object enters the scene, unlike methods in the previous category, these methods will not face any complications. This flexibility has made them more popular, and the majority of studies presented in the field of small object detection and tracking fall within this category. However, it should be noted that the efficacy of these methods largely depends on the accuracy of the detector they utilize. In the following sections, we will investigate and review such methods.

### 2.2.1. Background Detection-Based Methods

Methods utilizing frame differencing and background subtraction gained considerable attention due to their simplicity and robust performance in real-time applications. For instance, the study outlined by Archana et al. [26] employed this technique to detect and track a tennis ball and players within video frames. Initially, the input images are smoothed, and background images are accumulated to create an average background

model. To execute detection in the current frame, the difference between the frame and its predecessor is calculated, and the logical AND operation is executed between the difference image and the derived background. Subsequently, the ball and player candidates are determined based on the size of the identified contour. Ultimately, tracking is executed separately for the ball and players using the centroid of the detected contour. In another study, [27], the authors proposed an adaptive background subtraction algorithm for the detection and tracking of moving objects within a video sequence. They extracted the video's background image and applied a median filter to denoise the video sequence. The objects were then identified utilizing an adaptive background subtraction algorithm along with mathematical morphology. The tracking process employed the objects' centers. Importantly, this method incorporates an update of the background at each stage.

An alternative methodology was introduced by Srivastav et al. [28], which incorporated three-frame differencing and background subtraction for detecting moving objects in videos. The procedure commences with the selection of three successive frames from the image sequence. Subsequently, the difference between the first and second frames is computed, denoted as  $D_1$ . Similarly, the outcome of the difference between the second and third frames is labeled as  $D_2$ . If  $DB$  signifies the result of subtracting the background from the current frame, moving objects are detected by implementing a pixel-wise logical OR operation on  $D_1$ ,  $D_2$ , and  $DB$ . Finally, background noise is eliminated by utilizing a median filter.

Zhu et al. [29] incorporated three-frame differencing and operations such as "AND" and "XOR" for swift detection of moving objects. The difference image,  $p_1$ , is obtained by calculating the difference between the initial two frames, and  $p_2$  is obtained from the difference between the second and third frames. Subsequently, a new image,  $p_3$ , is created by performing  $p_1$  AND  $p_2$ . The next step involves obtaining  $p_2$  XOR  $p_3$ , resulting in a new image,  $p_4$ . Ultimately, the detection image is derived from  $p_1$  AND  $p_4$ . Following detection, noise is mitigated using post-processing algorithms. In their research, Yin et al. [30] proposed an algorithm known as Motion Modeling Baseline (MMB), designed to detect and track small, densely clustered moving objects in satellite videos. The process commences with the extraction of candidate slow-moving pixels and region of interest proposals using accumulative multi-frame differencing (AMFD). The full targets are then efficiently detected using low-rank matrix completion (LRMC). Lastly, the motion trajectory-based false alarm filter mitigates false alarms by compiling the trajectory over time, underlining that authentic moving targets are more likely to exhibit continuous trajectories.

Zhou et al. [31] presented a study that utilized an efficient and unsupervised approach, employing background subtraction for object delineation in Wide Area Motion Imagery (WAMI). Initially, background subtraction is used to detect low contrast and small objects, leading to the extraction of objects of interest. Following this, a convolutional neural network (CNN) is trained to reduce false alarms by considering both temporal and spatial data. Another CNN is subsequently trained to forecast the positions of several moving targets within a specified area, thus reducing the complexity of the necessary multi-target tracker. A Gaussian Mixture-Probability Hypothesis Density (GM-PHD) filter is finally employed to correlate detections over time.

Teutsch et al. [32], proposed an algorithm for detecting moving vehicles in Wide Area Motion Imagery that enhanced object detection by utilizing two-frame differencing along with a model of the vehicle's appearance. The algorithm amalgamates robust vehicle detection with the management of splitting and merging, and applies an appearance-based similarity measure to estimate assignment likelihoods among object hypotheses in consecutive frames.

Aguilar et al. [33] proposed a multi-object tracking (MOT) technique for tracking small moving objects in satellite videos. They used a patch-based CNN object detector with a three-frame difference algorithm to concentrate on specific regions and detect adjacent small targets. To improve object location accuracy, they applied the Faster Region-based convolutional neural network (Faster R-CNN) [34] since the three-frame difference algo-

rithm neither regularizes targets by area nor captures slow-moving targets. Furthermore, they applied a direct MOT data-association approach facilitated by an improved GM-PHD filter for multi-target tracking.

This approach was advanced by Aguilar et al. [35], where the performance of Faster R-CNN's object detection was significantly boosted by merging motion and appearance data on extracted patches. The new approach comprises two steps: initially obtaining rough target locations using a lightweight motion detection operator and, then, to enhance the detection results, combining this information with a CNN. An online track-by-detection methodology is also applied during the tracking process to convert detections into tracks based on the Probability Hypothesis Density (PHD) filter.

In the research conducted by Lyu et al. [36], a real-time tracking algorithm was introduced, specifically designed for ball-shaped, fast-moving objects, leveraging frame difference and multi-feature fusion. The process initiates by applying frame difference between two consecutive frames, after which the resulting differential image is segmented into smaller contours. A multi-feature-based algorithm is then used to determine if these are moving areas with ball-shaped objects.

Hongshan et al. [37] proposed a wiener filter-based infrared tiny object detection and tracking technique that optimizes filtering under stable conditions based on the least mean square error metrics. Given that the background is distributed in the image's low-frequency part and the high-frequency part primarily encompasses small objects, an adaptive background suppression algorithm is performed, taking advantage of the low-pass Wiener filter's characteristics. Appropriate segmentation then reveals potential targets. The relationship between multiple frames, including the continuity and regularity of target motion, is utilized for detection and tracking.

In the research conducted by Deshpande et al. [38], they applied max-mean and max-median filters on a series of infrared images for the detection of small objects. The initial step involves applying either the max-mean or max-median filter to the unprocessed image. Subsequently, the filtered image is subtracted from the original one to highlight potential targets. A thresholding step, which is guided by the image's statistical characteristics, limits the quantity of potential target pixels. Finally, the output images are cumulatively processed to track the target. The post-processing algorithm is equipped to detect the continuous trajectory of the moving target.

### 2.2.2. Classical Computer-Vision-Based Methods

These types of methodologies use algorithms and techniques that are based on mathematical models and heuristics to execute detection and tracking processes. For example, a study conducted in 2016 used frequency and spatial domain information to track small dim objects in infrared image sequences [39]. This method consists of three stages: initially, each frame produces six high-frequency sub-bands using the Dual-Tree Complex Wavelet Transform (DT-CWT). The potential targets in these high-frequency sub-bands are then detected by the Constant False Alarm Rate (CFAR) detection module. Finally, the potential targets are refined using an SVM classifier.

In their research, Ahmadi et al. [40] proposed an algorithm to identify and monitor small and dim infrared targets by integrating three mechanisms of the Human Visual System (HVS). The procedure involves four stages. Initially, a multi-scale Difference of Gaussians (DOG) filter is applied to the current image frame to create a series of feature maps at different scales. These maps are subsequently consolidated to form a saliency map. In the next stage, a Gaussian window is established at a location near the target, referred to as the visual attention point, on the saliency map. After normalizing the entire image, the target's position within the current frame is determined. Finally, the Proportional-Integral-Derivative (PID) algorithm is used to forecast the location of the visual attention point in the following frame.

Dong et al. [41] also introduced a method for detecting small moving targets in infrared images. Initially, the points of interest, including moving targets, are extracted using DOG

filters. These interest points are tracked across multiple frames, and the relationship between these points in the first and last frames is established. Ultimately, based on the relationships, these interest points are divided into two clusters: target points and background points.

In the study by Zhang et al. [42], an algorithm was presented to detect and track dim moving points with low Signal-to-Noise Ratio (SNR) in IR image sequences. This algorithm first applies a temperature non-linear elimination and Top-hat operator to preprocess the original images. Then, a composite frame is obtained by reducing the three-dimensional (3D) spatio-temporal scanning of an object to 2D spatial hunting. The tracking process, conducted in the final step, finds the object trajectory using the condition of a constant false alarm probability.

Lastly, Shaik et al. [43] proposed a strategy based on object preprocessing, tracking, and classification for infrared target sequences. Initially, preprocessing stages comprising normalization, rank order filtering to enhance the object features in the frame, and morphological operations to remove noise from frames are conducted in sequence. For tracking multiple objects, a correlation matrix is utilized, and the objects' positions in the frames are determined. Finally, a self-organizing map (SOM) is employed for classification based on various features, such as statistics and the object's shape.

An aggregation signature was utilized by Liu et al. [44] for small object tracking. This method suppresses the background through the aggregation signature, achieving highly distinctive features for small objects. The approach also generates a saliency map for frames and performs tracking. The tracker employs both the target's prior information and the context data to relocate the tracked target.

In the study by Tzannes et al. [45], temporal filters were applied to detect point targets in IR imagery. The central concept involves exploiting the difference in temporal profiles between target and clutter pixels. The "temporal profile" refers to the changing pixel value over a short period in an IR sequence. The pixels of a point target exhibit a pulse-like temporal profile. In this study, the implementation of the continuous wavelet transform (CWT) was explored for these temporal pixel profiles.

Bae et al. [46] introduced a technique for detecting small targets in sequences of infrared images, grounded in the cross-product of temporal pixels, relying on temporal profiles. Small targets display temporal characteristics that are distinguishable from various backgrounds. Consequently, the approach discriminates target and background pixels based on the cross-product of pixel values within the temporal profile. The method aggregates temporal background pixels to forecast the temporal background in the target region. Small target detection is then executed by measuring the absolute difference between the original temporal profile and the predicted background temporal profile, postulating that the absolute difference in target pixels exhibits a higher grey level. Ultimately, the target's trajectory is determined using a threshold that validates the presence or absence of a target based on each pixel's value in the calculated difference.

Bae [47] proposed a spatial and temporal bilateral filter (BF) intended for the identification of small trajectories in his study. This approach functions through the extraction of spatial target details via a spatial BF, and temporal target data via a temporal BF. The detection of small targets is realized by subtracting the anticipated spatial and temporal background profiles from the original infrared image and the original temporal profile, respectively. Similar to previous methods, this technique employs a threshold to ascertain a small target trajectory.

In the work presented by Choudhary [48], an automated process for the identification and tracking of targets within sequences of infrared images was achieved through the application of morphological connected operators. It involved two stages: intraframe and interframe. In the intraframe stage, the background is minimized to enhance the targets' visibility, and a binary image is created using adaptive double thresholding. In the interframe stage, the algorithm first assigns labels to the binary detections from the

previous stage, ensuring that detections associated with the same target receive the same label. Finally, targets not consistently detected in the sequence are eliminated.

In the research conducted by Son et al. [49], a model for tracking minuscule drones was put forth, featuring a predictor based on the Kalman filter and multiple trackers. This model is composed of two distinct trackers, a predictor, and a refining operation. The first tracker employs motion flow to identify a moving target, while the second utilizes histogram features to establish the region of interest. The trajectory of the target is then forecasted using the Kalman filter. Lastly, the refinement operation is employed to settle on the precise location of the target.

### 2.2.3. Deep-Learning-Based Methods

In this category, deep learning is used to detect small objects and, afterwards, their tracking. Deep networks can learn features and classifiers for detection and tracking tasks. For example, convolutional neural networks (CNNs) can learn intricate and distinctive features from massive datasets, deal with various kinds of objects and scenes, adjust to changing situations and appearance changes, and transfer to new settings and domains. They have several benefits, such as high precision, reliability, and flexibility. However, CNN-based methods also have some limitations, such as high computational expense, large memory usage, vulnerability to scale and occlusion, and challenge in handling small objects [35].

In the study conducted by Zhang et al. [50], three conventional convolutional neural networks, namely Faster R-CNN, YOLOv3 [51], and YOLOv3 tiny, were employed for the real-time detection and tracking of a golf ball within video sequences. A discrete Kalman filter was utilized during the tracking phase, predicting the golf ball's position based on previous observations. To improve accuracy and processing speed, image patches, rather than full images, were employed for detection purposes. Aktaş et al. [52] employed aerial imaging for small object detection and tracking. Given the unique challenges of aerial imagery due to wide-field images and tiny target objects, the standard Faster R-CNN model is modified. The goal is to use both spatial and temporal information from the image sequence, as appearance information alone is insufficient. The method adjusts the anchors in the Region Proposal Network (RPN) stage and optimizes the intersection over union (IoU) for small objects. After improving detection performance, the Deep SORT algorithm [53] is applied for small object tracking.

Behrendt et al. [54] introduced a system to detect, track, and perform 3D localization of traffic lights for automated vehicles using deep learning. A trained neural network detects traffic lights (as small as  $3 \times 10$  pixels), while the tracker triangulates the traffic lights' locations in the 3D space using stereo imagery and vehicle odometry information. A neural network is then used to correct the location estimate. In the study by Hurault et al. [55], a self-supervised soccer player detector and tracker robust to small players and suitable for wide-angle video games was presented. The detection framework involves sequential training of a teacher and student for domain adaptation and knowledge distillation, followed by player trajectory tracking using spatial and visual consistency. Zhu et al. [56] proposed a Multilevel Knowledge Distillation Network (MKDNet), which enhances feature representation, localization abilities, and discrimination for tiny object tracking. MKDNet performs three levels of knowledge distillation: score, feature, and IoU-level distillation. In the study by Liu et al. [57], transformers were employed for aerial small object tracking. Transformer, a popular network structure in deep learning, uses the attention module to provide a global response between the template frame and the search frame, effectively supplementing contextual information for small objects. Therefore, it can address the issue of poor feature expression capability in small objects. Also, a template update strategy is applied to manage template updates under occlusion, out-of-view, and drift conditions.

Huang et al. [58] proposed a deep autoencoder architecture called TrackNet for tracking tennis balls from broadcast videos. TrackNet generates a detection heatmap from either a single frame or several consecutive frames to locate the ball and learn flying patterns.



Yoshihashi et al. [59] presented the Recurrent Correlational Network (RCN) for joint detection and tracking of small flying objects, where detection and tracking are jointly carried out on a multi-frame representation learned through an end-to-end, trainable, and single network. A convolutional long short-term memory network learns informative appearance change in the detection process, enabling correlation-based tracking over its output. Another approach was proposed by Marvasti-Zadeh et al. [60], which used a two-stream multitask network and an offline proposal approach to track aerial small objects. The network, aided by the proposed proposal generation strategy, leverages context information to learn a generalized target model and handle viewpoint changes and occlusions. The reviewed methods are subsequently categorized based on the above taxonomy in Table 1.

**Table 1.** Categorization of the applied methods in the literature.

Types of Methods	Category	References
Unified track-and-detection	Filter-based	[17–19,21]
	Search-based	[22–25]
Track-by-detection	Background-detection-based	[26–38]
	Classical computer-vision-based	[39–49]
	Deep-learning-based	[50,52,54–60]

### 3. Types of Public Data Sets Used for Small Object Detection and Tracking

The datasets used for small object detection and tracking are essential for evaluating and benchmarking the performance of various algorithms in this field. These datasets can be broadly classified into two groups: spectrum-based video datasets and source position-based video datasets. Spectrum-based datasets include infrared (IR) videos whereas source position-based datasets include aerial videos, satellites videos, and normal videos. Note that videos captured by satellites or Unmanned Aerial Vehicles (UAVs) present more challenges due to larger dimensions, smaller objects, and the presence of many more objects compared to normal videos. The difficulty is especially pronounced in satellite videos due to the higher altitude of the target object. The following are some popular and publicly available datasets in the field.

#### 3.1. *Small90 and Small112*

Small90 is one of the most comprehensive datasets in the field of small object tracking in normal videos. It includes 90 sequences of annotated small-sized objects. The dataset was created from existing tracking datasets by selecting videos with an object area size ratio smaller than 0.01 of the entire image. Additional challenges, such as target drifting and low resolution, were also considered. Small112 extends Small90 by adding 22 more challenging sequences. Each sequence in these datasets is classified using 11 attributes for a comprehensive study of tracking approaches [44].

#### 3.2. *Large-Scale Tiny Object Tracking (LaTOT)*

LaTOT stands for Large-scale Tracking of Tiny Objects, and it was developed as a response to the limitations observed in the Small90 dataset, particularly its small-scale size and limited challenges. The LaTOT dataset consists of 434 video sequences, encompassing more than 217,000 frames, all captured in real-world scenarios. Each frame in the dataset is meticulously annotated with high-quality bounding boxes. Additionally, the dataset incorporates 12 challenging tracking attributes aimed at encompassing a wide range of viewpoints and scene complexities [56].

#### 3.3. *Video Satellite Objects (VISO)*

VISO is a large-scale dataset designed for the purpose of detecting and tracking moving objects in satellite videos. It comprises 47 high-quality satellite videos, which were

acquired using Jilin-1 satellite platforms. Within this dataset, there are 3711 object tracking trajectories, and it contains over 1.6 million instances of interest specifically for object detection tasks. Each image in the dataset possesses a resolution of  $5000 \times 12,000$  pixels, and it encompasses a diverse array of objects varying in scale. The dataset encompasses various types of moving objects, including trains, cars, ships, and planes, and it is formulated to incorporate seven key challenges commonly encountered in object tracking research [30].

### 3.4. UAV123

UAV123 is a dataset for object tracking in UAV aerial imagery, comprising over 110K frames and 123 video sequences. The sequences are high-resolution, fully annotated, and captured from an aerial perspective at low altitudes. The dataset has three sections: high-quality sequences taken at diverse heights, noisy and lower-quality sequences, and synthetic sequences [61].

### 3.5. Dataset of Object deTectioN in Aerial Images (DOTA)

DOTA, which stands for detection in aerial images, is a valuable resource that facilitates object detection tasks in Earth Vision. This dataset contains a substantial collection of 1,793,658 instances distributed across 11,268 images, encompassing 18 commonly encountered categories. Each object within the dataset is annotated with a bounding box, which may be either horizontal or oriented. Despite the dataset containing a significant number of small objects, it was observed that these objects are predominantly concentrated in a limited number of categories, specifically within the “small-vehicle” category. This concentration can be attributed to the high diversity of orientations observed in overhead view images and the substantial variations in large-scale amongst the instances [62].

### 3.6. Vision Drone (VisDrone)

VisDrone is an extensive dataset collected through the use of drones, spanning various urban and suburban regions across 14 distinct cities in China. This dataset is curated to facilitate four essential computer vision tasks, namely image object detection, video object detection, single object tracking, and multi-object tracking. For the image object detection track, VisDrone consists of 10,209 images, each having a resolution of  $2000 \times 1500$  pixels, and the dataset encompasses a total of 542,000 instances representing 10 typical object types commonly found in traffic scenes. The imagery in VisDrone was captured using drones from multiple viewpoints within diverse urban settings, leading to variations in viewpoints and significant occlusion. Consequently, the dataset includes a substantial number of small objects [63].

### 3.7. Tsinghua-Tencent 100K (TT100K)

The TT100K dataset is utilized for realistic traffic sign detection in normal videos, encompassing 30,000 instances of traffic signs distributed across 100,000 images. These images represent 45 distinct classes of typical Chinese traffic signs. Each traffic sign within the TT100K dataset is meticulously annotated, providing detailed bounding boxes and instance-level masks. The images in TT100K were captured using Tencent Street Views, which offers a wide range of weather conditions and illumination settings, adding to the dataset’s realism. Notably, TT100K exhibits a considerable number of small instances, leading to a long-tail distribution, with approximately 80% of instances occupying less than 0.1% of the entire image area [64].

## 4. Performance Evaluation Metrics

In the realm of small object detection and tracking, the majority of methods employ a track-by-detection approach, commonly evaluating their detection and tracking modules individually. This section elaborates on several evaluation metrics utilized in the quantitative appraisal of algorithms during these two stages. During the detection stage, primary metrics encompass Precision, Recall, F1-score, the Precision-Recall (PR) curve, Average

Precision (AP), and mean Average Precision (mAP). Regarding the tracking stage, evaluation measures encompass Overlap Success Rate (OSR), Distance Precision Rate (DPR), Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and Completeness.

#### 4.1. Metrics for Detection

##### - Precision and Recall:

Before introducing these metrics, it is necessary to explain the following concepts:

- True Positive (TP): A ground-truth bounding box that is detected correctly;
- False Positive (FP): Mistakenly detecting an object that does not exist or detecting an object that does exist in the wrong place;
- False Negative (FN): This term refers to an instance where a ground-truth bounding box goes undetected.

The ability to distinguish between True Positive (TP) and False Positive (FP) is critical for object detection [65]. Additionally, False Negative (FN) is the name given to missed true targets. Precision is defined as the ratio of true positives to the detected targets as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

A detector's capability to capture the targets is calculated using Recall [65]. This metric is defined as the ratio of TP to the number of all existing true targets:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

##### - F1-score

A traditional criterion for categorizing objects into either being targets or not is the F1-score that is equivalent to Precision and Recall's harmonic mean [30,35,66], i.e.,

$$\text{F1 - score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

Despite the widespread usage of Precision, Recall, and F1-score as evaluative measures for object detection algorithms, these metrics have inherent limitations. It is crucial to acknowledge their potential for unfairness before application and to discern chance or base case levels of the statistic. In certain cases, a system with objectively lower quality, in terms of informedness, may appear to possess superior quality when judged using these popular metrics [67].

##### - Precision-Recall (PR) curve

The Precision-Recall (PR) curve serves as a graphical representation depicting the trade-off between precision and recall at different threshold levels. A high precision value indicates a low false-positive rate, whereas high recall corresponds to a low false-negative rate. A large area under the PR curve signifies both high recall and precision, reflecting accurate classifier results (high precision) and a substantial proportion of true positives (high recall) [30]. It is important to note that the shape of this curve is influenced by the threshold chosen, which, in turn, impacts the trade-off between precision and recall.

##### - Average-Precision (AP)

The AP metric, which is defined as the area under the Precision-Recall curve, is a widely accepted measure of detection model accuracy [65]. However, calculating an accurate Area Under the Curve (AUC) can be challenging in practice, primarily due to the characteristic zigzag shape of the precision-recall plot. To mitigate this issue, the zigzag behavior is eliminated by processing the precision-recall curve prior to estimating the AUC. The two most common techniques for this processing are the 11-point interpolation and

all-point interpolation. The 11-point interpolation method aims to capture the essence of the precision-recall curve's shape by calculating the average maximum precision values at a series of 11 evenly spaced recall levels [0, 0.1, 0.2, ..., 1]. The shape of the precision  $\times$  recall curve is summarized by the 11-point interpolation as:

$$AP_{11} = \frac{1}{11} \sum_{R \in \{0, 0.1, \dots, 0.9, 1\}} P_{\text{interp}}(R) \quad (4)$$

$$P_{\text{interp}}(R) = \max_{\tilde{R}: \tilde{R} \geq R} P(\tilde{R}) \quad (5)$$

Rather than using the observed precision  $P@R$  at each recall level  $R$ , as specified in this definition, one can determine the Average Precision (AP) by considering the maximum precision  $P_{\text{interp}}(R)$  with a recall value equal to or greater than  $R$ .

In the all-points interpolation method, instead of interpolating only at 11 evenly spaced points, one can choose to interpolate at all points as follows:

$$AP_{\text{all}} = \sum_n (R_{n+1} - R_n) P_{\text{interp}}(R_{n+1}) \quad (6)$$

$$P_{\text{interp}}(R_{n+1}) = \max_{\tilde{R}: \tilde{R} \geq R_{n+1}} P(\tilde{R}) \quad (7)$$

In this case, the maximum precision with a recall value equal to or greater than  $R_{n+1}$  is used to interpolate the precision at each level. This provides a more accurate representation of AP than merely relying on the precision observed at a limited number of points. Average precision has some limitations. First, it assumes that precision is a continuous function of recall, which is not true in practice. Moreover, this measure ignores the size and shape of the objects, and it requires binarizing the output for multi-class or multi-label classification.

- mean Average-Precision (mAP)

Within a specific dataset, the accuracy of object detectors is evaluated across all classes using the mAP, which is simply the average of the AP scores across all classes as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (8)$$

where  $N$  indicates how many classes are being evaluated in total, and  $AP_i$  is the AP for the  $i$ th class [65].

#### 4.2. Metrics for Tracking

Evaluation criteria for the tracking stage are divided into two categories: single object tracking and multiple object tracking, detailed as follows.

##### 4.2.1. Single Object Tracking

Key metrics employed within this category involve the Overlap Success Rate (OSR) and the Distance Precision Rate (DPR). The Distance Precision Rate (DPR) is determined by computing the ratio of frames where the center location error, denoted as the Euclidean distance between the centers of the predicted box ( $B_p$ ) and the ground truth box ( $B_g$ ), falls below a designated threshold  $\alpha$ . On the other hand, the Overlap Success Rate (OSR) is defined as the proportion of frames where the overlap ratios with the ground-truth box surpass a specified threshold  $\beta$ . The overlap ratio is mathematically expressed in Equation (9):

$$S = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \quad (9)$$

where  $|\cdot|$  indicates the number of pixels in an area  $\cap$  and  $\cup$  show the intersection and union of two areas, respectively [68].

OSR and DPR rely on the selection of thresholds for determining the overlap rate and the distance mistake, which can influence the sensitivity and specificity of the evaluation. Furthermore, these measures neglect the scale variation of the target, which is a frequent difficulty in tracking [69]. These are two important limitations for these metrics.

#### 4.2.2. Multiple Object Tracking

##### - Multiple Object Tracking Accuracy (MOTA)

MOTA is a widely used metric for assessing the overall performance of a multiple-object tracker. It evaluates the quality of the recovered tracks by considering identity switches, missed targets, and false positives as follows:

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad (10)$$

Here,  $g_t$  represents the number of ground-truth objects at frame index  $t$ . The terms  $m_t$ ,  $fp_t$ , and  $mme_t$  denote missed targets, false positives, and ID switches at the  $t$ -th frame, respectively. MOTA provides an aggregate count of tracking errors and ranges from  $(-\infty, 1]$  with negative values indicating poor performance and one representing the highest possible score [70].

##### - Multiple Object Tracking Precision (MOTP)

MOTP considers the distance between the detected objects and the ground truth objects. It is a representative measure of a tracker's precision performance and is defined by the following equation:

$$\text{MOTP} = \frac{\sum_{t,i} d_t^i}{\sum_t c_t} \quad (11)$$

where  $d_t^i$  is the distance between the object  $o_i$  and its corresponding hypothesis and  $c_t$  is the total number of matches in frame  $t$  among the true targets and the hypothesized objects. The MOTP score falls in the interval  $[0, \infty)$ , where a score of 0 indicates good performance, and higher values indicate poorer performance [71].

It should be noted that MOTA and MOTP are not appropriate for real-time tracking applications, because they need the ground truth information of all the frames in a video sequence, which is not accessible in real-time situations. Moreover, they are not adaptable to multi-class tracking problems, because they suppose that all the objects are in the same class and have the same appearance model [72].

##### - Completeness

Completeness metrics evaluate the overall continuity of the tracking trajectory. Specifically, trajectories resulting from combined tracker outputs can be classified into three categories: Mostly Tracked (MT), Partially Tracked (PT), and Mostly Lost (ML). The MT category suggests that the tracker output covers more than 80% of the ground truth trajectory's length, whereas the ML category signifies it covers less than 20%. The PT category applies to all other instances. Therefore, an optimal tracker should ideally produce a larger proportion of sequences classified as MT and a smaller proportion as ML [71].

Completeness metrics have some limitations. In this way, they do not measure the accuracy of the location of the tracked objects, which can influence the quality of the tracking results. Furthermore, these criteria do not consider the identity switches of the tracked objects, which can lead to confusion and mistakes [73].

## 5. Challenges and Future Trends

In this section, we initially contrast the methods reviewed, focusing on their respective strengths and shortcomings, before delving into a detailed analysis. We also discuss the

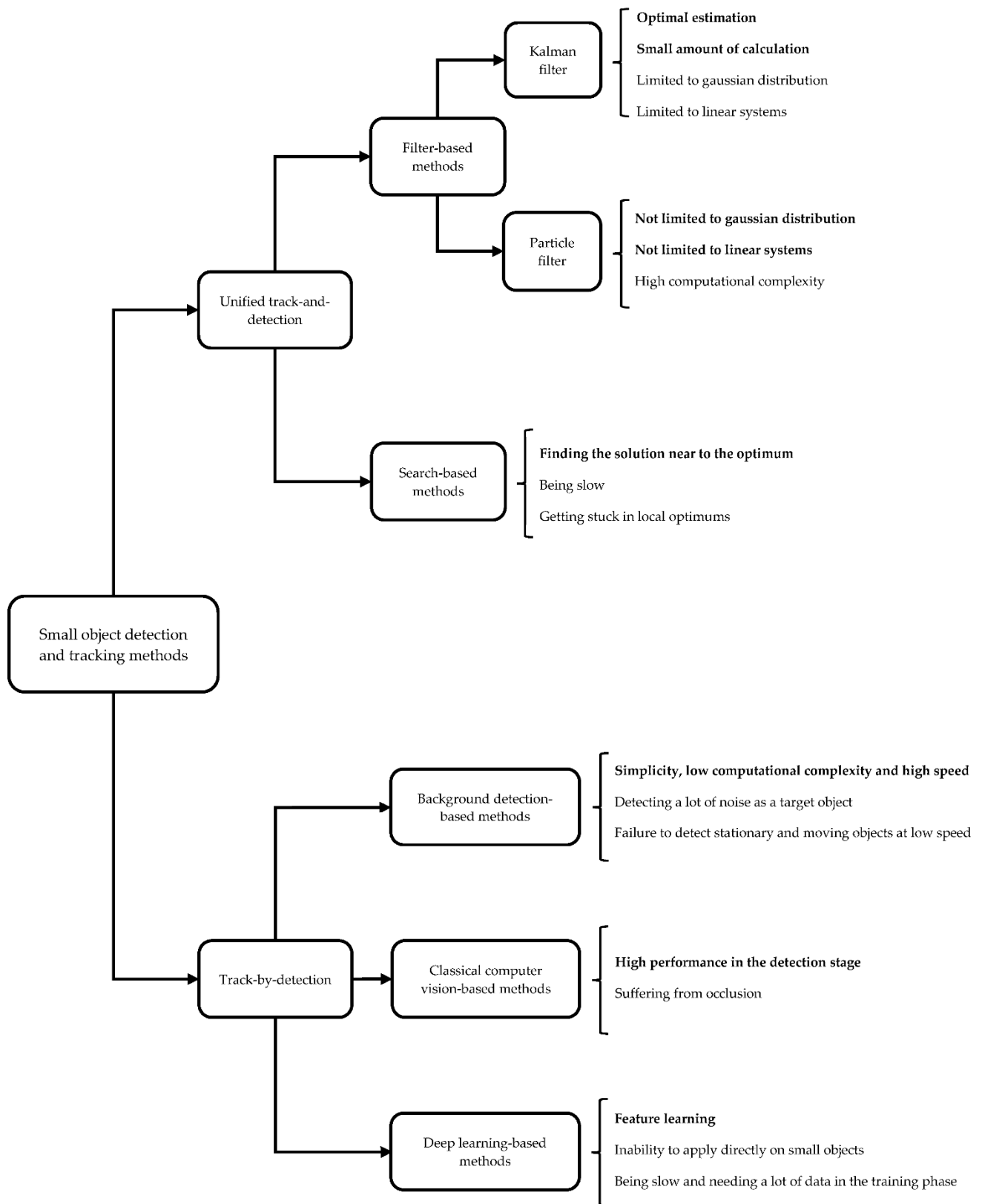
challenges faced by existing object detection and tracking methods and explore potential future trends to overcome these limitations. We also introduce a sensor fusion-based approach that incorporates 3D LiDAR into the object detection and tracking processes. Table 2 compares unified track-and-detection methods with tracking-by-detection methods based on their advantages and shortcomings. Additionally, Figure 2 individually delineates the taxonomy of small object detection and tracking methods along with the merits and pitfalls of existing methods across each category. Note that the merits of methods across each category are highlighted in boldface style.

**Table 2.** Comparison of unified track-and-detection methods versus track-by-detection methods.

Types of Methods	Advantages	Shortcomings
Unified track-and-detection	Free of object detector	Manual initialization, Fixed number of objects to be tracked
Track-by-detection	Capable of managing a varying number of objects in video, Automatic initialization	Performance depends on object detection

According to Table 2, it can be concluded that the advantage of unified track-and-detection methods is that these types of methods perform detection and tracking processes in a unified manner and without a detector. In other words, these methods are free of object detectors. However, their shortcomings are that they require a manual initialization with a fixed number of objects in the first frame such that these objects are located and tracked in the next frames. This means that when new objects enter the scene, these methods are not able to manage them and so have limited applications. On the other hand, track-by-detection methods are more popular and have more applications due to performing a detection stage and automatically detecting objects in each frame by detection algorithms. The advantages of these methods are automatic initialization and managing the scene with a varying number of objects due to the detection stage. However, the performance of these methods is highly dependent on the performance of the utilized object detector which is their main shortcoming. This means that poor performance in the detection process can adversely affect the performance of the tracking process. In brief, tracking-by-detection methods are more widely used within the field of small object detection and tracking because of discovering new objects and terminating disappearing objects automatically. That is while unified track-and-detection methods cannot deal with these challenges.

As Figure 2 elucidates, in the context of Filter-based methods, the Kalman filter generally enhances the tracking process's efficiency through optimal estimations. Its simplicity and minimal computational requirements make it apt for tasks necessitating high real-time responsiveness. Nonetheless, the Kalman filter's assumptions of Gaussian distribution for state variables and its suitability only for linear systems are limitations. The particle filter resolves these constraints, but its elevated computational cost renders it unsuitable for real-time applications. It should be mentioned that there are a lot of Kalman filter variations, such as the Extended Kalman Filter (EKF) and Unscented Kalman Filter (UKF), which are the famous heuristic variations of the Kalman filter for nonlinear systems with Gaussian distribution [74,75]. Search-based methods, on the other hand, suffer from slow speeds, high computational complexity due to the search process, and also the potential to become trapped in local optima. Hence, they are unsuitable for real-time applications. Yet, these methods' utility lies in their ability to deliver near-optimal solutions with acceptable error margins.



**Figure 2.** The taxonomy of detection and tracking methods of small objects along with their advantages and drawbacks.

Background-based methods boast simplicity, low computational complexity, and high speed. As a result, these methods have good performance in real-time applications.

However, they tend to generate considerable noise as the target object, necessitating post-processing methods for noise removal. Moreover, they might fail to detect certain objects, particularly those that are stationary or moving slowly. Classical computer-vision-based methods exhibit excellent performance in object detection, effectively identifying objects within video frames using mathematical and heuristics techniques. Unfortunately, these methods often overlook the challenge of occlusion during the tracking stage, a common and critical issue in object tracking. This occurs when objects appear to blend or merge due to close proximity or background integration, potentially causing the tracker to lose or erroneously track objects after overlapping [76]. In other words, occlusion means that an object is partially or completely hidden by other objects, which reduces the appearance and location information of the object. This issue is particularly prevalent with small objects due to their diminutive size.

To counter occlusion, trackers typically employ data association. However, there are various methods for small object detection and tracking under occlusion that can modify accuracy, speed, robustness, and generality. Some of these methods are: using a large number of anchor points to cover a wider search space, using multiple-frame receptive networks for small object tracking, using motion as a strong cue for small object detection in videos, using correlation filter to maintain the identity and trajectory of objects over time, and using reconstruction methods to recover the appearance and location of the object after occlusion [77,78]. For instance, Wang et al. [79] addressed occlusion by calculating the correlation between different frames and learning motion information. In the study by Wojke et al. [53], the Deep SORT algorithm integrated motion and appearance information to solve this issue by utilizing Mahalanobis distance and a convolutional neural network (CNN) trained on a large-scale person re-identification dataset. It should be noted that the real-time performance and computational complexity of classical computer-vision-based methods depend on several factors, such as the size and quality of the input images, and the type and complexity of that methods.

Deep-learning-based methods offer the advantage of feature learning, facilitating end-to-end functionality. However, these methods face challenges when directly applied to small objects as these possess limited and poor features, which may get lost during deep network processing, because these networks have many layers and, as a result, include many processes. This challenge means that small objects have fewer and less distinguishable features than larger objects due to their small shape size. This makes it difficult for deep networks to learn and extract suitable features for small object detection and tracking. To solve this problem, various methods have been proposed to enhance and improve the feature information of small objects that can improve accuracy, speed, robustness, and generality. Some of these methods are: using feature enhancement modules (FEM) that can use feature aggregation structures (FAS) and attention generation structures (AGS) to reduce the interference of background by extracting multi-scale contextual information and combining a coordinate attention mechanism, thus improving the perception of small objects, using generative adversarial networks (GAN) to generate new high-quality images that contain small objects, using self-supervised methods to learn features using motion, and using supervised methods to learn features using annotations [80,81].

On the other hand, deep networks require substantial data for the training phase to circumvent overfitting. The cause is that these networks need to learn intricate and distinctive features from massive data, which are normally not accessible for small objects. In addition, small objects usually have a lot of variety and change in their appearance, shape, pose, colour, orientation, and location, which require very diverse and balanced data [82]. Generally, deep networks are slow but there are different methods for small object detection and tracking based on deep learning. As a result, each method has its own advantages and limitations in terms of real-time performance and computational complexity. Hence, it is impossible to find one optimal method for small object detection and tracking that can achieve both high real-time performance and low computational complexity. Various methods may work better for different situations and tasks based



on the needs and limitations. Some potential directions for future research are creating more efficient network structures, developing more effective loss functions, using more contextual information, and adding more existing knowledge.

Multiple object tracking (MOT) is another vital consideration. Few works address multiple small object tracking, primarily employing GM-PHD filters for this task. In addition to creating sophisticated appearance and/or motion models to address challenges like scale changes, out-of-plane rotations, and illumination variances (central to single object tracking), MOT involves determining the variable number of objects and maintaining their identities [11]. Traditionally, the track-by-detection paradigm has been employed for MOT, leveraging the inherent separation of detection and data association tasks [83]. Thus, the tracking performance is significantly influenced by the detection results [84,85]. The MOT problem can be viewed as a data association task, aiming to connect detections across video sequence frames [86]. In addition to the common challenges encountered in both SOT and MOT, MOT faces unique concerns such as (1) initializing and terminating tracks, (2) managing interactions among multiple objects, (3) dealing with frequent occlusions, and (4) handling similar appearances. A similar appearance complicates the tracking process as the object's appearance information is insufficient for differentiation. The study by Dicle et al. [87] applied motion dynamics as a cue to distinguish and track multiple objects with similar appearances. This issue is addressed by formulating the problem as a Generalized Linear Assignment (GLA) of tracklets that are dynamically similar, incrementally associated into longer trajectories.

Scale variation is another important issue in this domain which means that the relative size of objects to the image or camera is changing, which affects the appearance and location information of the object. This can cause algorithms to fail to detect or track the object correctly. To solve this problem, various methods have been proposed for small object detection and tracking under scale variation in which accuracy, speed, robustness, and generality are improved. Some of these methods are: using image pyramid to provide images with different levels of details, using scale filter to estimate the target scale, using multi-scale feature fusion to reduce scale mismatch, and using multiple-frame receptive networks for small object tracking [56,88,89].

### 5.1. Sensor Fusion-Based Approach

Sensor fusion is a popular approach in modern perception systems, where information from multiple sensors is combined to enhance the overall perception and understanding of the environment. In the context of object detection and tracking, integrating 3D LiDAR data with other sensors, such as cameras, can provide more accurate and comprehensive information about the surrounding objects [90,91].

#### 5.1.1. Hydro-3D: Hybrid Object Detection and Tracking for Cooperative Perception Using 3D LiDAR

One noteworthy system that utilizes this sensor fusion-based approach is the “hydro-3D” framework, a hybrid object detection and tracking system for cooperative perception using 3D LiDAR. By leveraging the high-resolution 3D point cloud data provided by LiDAR, hydro-3D improves the accuracy and robustness of object detection and tracking tasks. The 3D information allows the system to handle various challenges, such as occlusion and object detection in complex scenarios [92].

#### 5.1.2. Automated Driving Systems Data Acquisition and Analytics Platform

The development of advanced object detection and tracking systems, like the hydro-3D, is further facilitated by automated driving systems data acquisition and analytics platforms. These platforms provide a large-scale collection of diverse real-world driving data, enabling researchers and developers to train and validate their algorithms using a vast range of scenarios. The availability of such datasets accelerates progress in the field and fosters the creation of more robust and reliable object detection and tracking systems [93].

### 5.1.3. YOLOv5-Tassel: Detecting Tassels in RGB UAV Imagery with Improved YOLOv5 Based on Transfer Learning

In recent years, transfer learning has become a widely adopted technique in deep learning-based object detection tasks. YOLOv5-Tassel is an example of a transfer learning-based approach specifically designed for detecting tassels in RGB UAV imagery. By leveraging pre-trained models and fine-tuning on tassel-specific datasets, YOLOv5-Tassel demonstrates enhanced performance and efficiency in detecting small and intricate objects, like tassels, in aerial images [94].

### 5.1.4. Comparison and Future Trends

In contrast to the methods reviewed earlier, the sensor-fusion-based approach using 3D LiDAR, exemplified by hydro-3D, provides a more comprehensive understanding of the environment. By combining 3D LiDAR data with other sensor inputs, like cameras, the system can address challenges related to occlusion, object detection, and tracking in dynamic and complex scenarios.

Furthermore, the use of automated driving systems data acquisition and analytics platforms facilitates the development and testing of advanced object detection and tracking algorithms, enabling researchers to train and validate their models on diverse real-world data.

For deep-learning-based methods, transfer learning, as demonstrated by YOLOv5-Tassel, becomes crucial in improving the performance of object detection on small objects or objects with limited features. By utilizing pre-trained models and fine-tuning on specific datasets, transfer learning allows the network to learn and generalize better, even with limited training data.

Future trends in object detection and tracking may involve even more sophisticated sensor fusion techniques, integrating data from various sensors, such as LiDAR, cameras, radars, and more. Additionally, the continued advancement of deep learning and transfer learning approaches is likely to enhance the accuracy and efficiency of small object detection and tracking tasks.

In conclusion, the field of object detection and tracking is continually evolving, with sensor-fusion-based approaches and transfer learning playing pivotal roles in overcoming challenges and improving performance. Automated driving systems data acquisition platforms contribute significantly to advancing research and development in this domain. As technology continues to progress, we can expect further breakthroughs that will refine and expand the capabilities of object detection and tracking systems, making them more reliable and applicable across various real-world scenarios.

## 6. Conclusions

This paper reviewed different methods for detecting and tracking small objects, which are important aspects of image and video analysis, mainly due to the limited and inferior features of these objects. We introduced two main categories: unified track-and-detection and track-by-detection methods. The former was split into filter-based and search-based methods, while the latter was divided into background-based, classical computer-vision-based, and deep-learning-based methods. This review also categorized public datasets for small object detection and tracking, dividing them into spectrum-based video datasets and source position-based video datasets. This classification helps researchers to do experimental work. Moreover, we explained the usual evaluation metrics for detection, single object tracking, and multiple object tracking, providing a basis for validating and comparing results.

A significant portion of this review was dedicated to discussing the prevalent challenges in small object detection and tracking. These challenges, encompassing issues such as occlusion, scale variation, speed, computational cost, and the application of deep-learning techniques for small objects, show the complexities in this domain. Moreover, sensor-fusion-based approach with 3D LiDAR can help small object detection and tracking tasks owing to providing more accurate and comprehensive information about the surrounding objects. Concurrently, we assessed the strengths and limitations different methods in each category,

providing an evaluation of current approaches and their potential suitability for different scenarios. Furthermore, we outlined potential future research directions, emphasizing the need for better performance and accuracy, especially in dealing with multiple object tracking (MOT). Existing works in MOT are sparse, and more research is needed to handle extra tasks, such as changing numbers of objects, keeping their identities, and managing frequent occlusions or similar appearances.

In conclusion, this comprehensive review is a helpful guide for researchers within the field of small object detection and tracking. Research could focus on dealing with multiple object tracking and improving the performance and accuracy of small object detection and tracking. By outlining the current challenges and future trends, this review gives a roadmap to guide future research in this crucial area of image and video analysis.

**Author Contributions:** Conceptualization, H.N.-p., A.R. and R.D.; methodology, B.M.; software, B.M.; validation, H.N.-p.; formal analysis, B.M.; investigation, B.M.; resources, A.R. and R.D.; data curation, B.M., A.R. and R.D.; writing—original draft preparation, B.M. and H.N.-p.; writing—review and editing, B.M., H.N.-p., A.R. and R.D.; visualization, H.N.-p. and B.M.; supervision, H.N.-p. and A.R.; project administration, H.N.-p. and R.D.; funding acquisition, R.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is contained within the article.

**Acknowledgments:** We are grateful to the anonymous reviewers and editors, who provided thorough and insightful reviews that helped improve the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Zhou, J.T.; Du, J.; Zhu, H.; Peng, X.; Liu, Y.; Goh, R.S.M. AnomalyNet: An Anomaly Detection Network for Video Surveillance. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2537–2550. [[CrossRef](#)]
2. Zhu, L.; Yu, F.R.; Wang, Y.; Ning, B.; Tang, T. Big Data Analytics in Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 383–398. [[CrossRef](#)]
3. Hua, S.; Kapoor, M.; Anastasiu, D.C. Vehicle Tracking and Speed Estimation from Traffic Videos. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; Volume 2018.
4. Hagiwara, T.; Ota, Y.; Kaneda, Y.; Nagata, Y.; Araki, K. Method of Processing Closed-Circuit Television Digital Images for Poor Visibility Identification. *Transp. Res. Rec.* **2006**, *1973*, 95–104. [[CrossRef](#)]
5. Crocker, R.I.; Maslanik, J.A.; Adler, J.J.; Palo, S.E.; Herzfeld, U.C.; Emery, W.J. A Sensor Package for Ice Surface Observations Using Small Unmanned Aircraft Systems. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1033–1047. [[CrossRef](#)]
6. Zhang, F.; Du, B.; Zhang, L.; Xu, M. Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5553–5563. [[CrossRef](#)]
7. Zhou, H.; Wei, L.; Lim, C.P.; Creighton, D.; Nahavandi, S. Robust Vehicle Detection in Aerial Images Using Bag-of-Words and Orientation Aware Scanning. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7074–7085. [[CrossRef](#)]
8. de Vries, E.T.; Tang, Q.; Faez, S.; Raouf, A. Fluid Flow and Colloid Transport Experiment in Single-Porosity Sample; Tracking of Colloid Transport Behavior in a Saturated Micromodel. *Adv. Water Resour.* **2022**, *159*, 104086. [[CrossRef](#)]
9. Delibaşoğlu, İ. Moving Object Detection Method with Motion Regions Tracking in Background Subtraction. *Signal Image Video Process.* **2023**, *17*, 2415–2423. [[CrossRef](#)]
10. Tsai, C.Y.; Shen, G.Y.; Nisar, H. Swin-JDE: Joint Detection and Embedding Multi-Object Tracking in Crowded Scenes Based on Swin-Transformer. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105770. [[CrossRef](#)]
11. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Kim, T.K. Multiple Object Tracking: A Literature Review. *Artif. Intell.* **2021**, *293*, 103448. [[CrossRef](#)]
12. Desai, U.B.; Merchant, S.N.; Zaveri, M.; Ajishna, G.; Purohit, M.; Phanish, H.S. Small Object Detection and Tracking: Algorithm, Analysis and Application. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3776. [[CrossRef](#)]
13. Rout, R.K. *A Survey on Object Detection and Tracking Algorithms*; National Institute of Technology Rourkela: Rourkela, India, 2013.
14. Yilmaz, A.; Javed, O.; Shah, M. Object Tracking: A Survey. *Acm Comput. Surv. (CSUR)* **2006**, *38*, 13-es. [[CrossRef](#)]

15. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693.
16. Naik, B.T.; Hashmi, M.d.F. YOLOv3-SORT: Detection and Tracking Player/Ball in Soccer Sport. *J. Electron. Imaging* **2022**, *32*, 011003. [[CrossRef](#)]
17. Huang, Y.; Llach, J. Tracking the Small Object through Clutter with Adaptive Particle Filter. In Proceedings of the ICALIP 2008—2008 International Conference on Audio, Language and Image Processing, Shanghai, China, 7–9 July 2008.
18. Habibi, Y.; Sulistyaningrum, D.R.; Setiyono, B. A New Algorithm for Small Object Tracking Based on Super-Resolution Technique. In *Proceedings of the AIP Conference Proceedings*; AIP Publishing: Long Island, NY, USA, 2017; Volume 1867.
19. Liu, W.; Tang, X.; Ren, X. A Novel Method for Small Object Tracking Based on Super-Resolution Convolutional Neural Network. In Proceedings of the 2019 2nd International Conference on Information Systems and Computer Aided Education, ICISCAE 2019, Dalian, China, 28–30 September 2019.
20. Mahmoodi, J.; Nezamabadi-pour, H.; Abbasi-Moghadam, D. Violence Detection in Videos Using Interest Frame Extraction and 3D Convolutional Neural Network. *Multimed. Tools Appl.* **2022**, *81*, 20945–20961. [[CrossRef](#)]
21. Wu, D.; Song, H.; Yuan, H.; Fan, C. A Small Object Tracking Method in Satellite Videos Based on Improved Kernel Correlation Filter. In Proceedings of the 2022 14th International Conference on Communication Software and Networks, ICCSN 2022, Chongqing, China, 10–12 June 2022.
22. Blostein, S.D.; Huang, T.S. Detecting Small, Moving Objects in Image Sequences Using Sequential Hypothesis Testing. *IEEE Trans. Signal Process.* **1991**, *39*, 1611–1629. [[CrossRef](#)]
23. Ahmadi, K.; Salari, E. Small Dim Object Tracking Using a Multi Objective Particle Swarm Optimisation Technique. *IET Image Process.* **2015**, *9*, 820–826. [[CrossRef](#)]
24. Salari, E.; Li, M. Dim Target Tracking with Total Variation and Genetic Algorithm. In Proceedings of the IEEE International Conference on Electro Information Technology, Milwaukee, WI, USA, 5–7 June 2014.
25. Shaik, J.; Iftekharruddin, K.M. Detection and Tracking of Targets in Infrared Images Using Bayesian Techniques. *Opt. Laser Technol.* **2009**, *41*, 832–842. [[CrossRef](#)]
26. Archana, M.; Geetha, M.K. Object Detection and Tracking Based on Trajectory in Broadcast Tennis Video. *Procedia Comput. Sci.* **2015**, *58*, 225–232. [[CrossRef](#)]
27. Zhang, R.; Ding, J. Object Tracking and Detecting Based on Adaptive Background Subtraction. *Procedia Eng.* **2012**, *29*, 1351–1355. [[CrossRef](#)]
28. Srivastav, N.; Agrwal, S.L.; Gupta, S.K.; Srivastava, S.R.; Chacko, B.; Sharma, H. Hybrid Object Detection Using Improved Three Frame Differencing and Background Subtraction. In Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering, Noida, India, 12–13 January 2017.
29. Zhu, M.; Wang, H. Fast Detection of Moving Object Based on Improved Frame-Difference Method. In Proceedings of the 2017 6th International Conference on Computer Science and Network Technology, ICCSNT, Dalian, China, 21–22 October 2017; Volume 2018, pp. 299–303. [[CrossRef](#)]
30. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [[CrossRef](#)]
31. Zhou, Y.; Maskell, S. Detecting and Tracking Small Moving Objects in Wide Area Motion Imagery (WAMI) Using Convolutional Neural Networks (CNNs). In Proceedings of the FUSION 2019 22nd International Conference on Information Fusion, Ottawa, ON, Canada, 2–5 July 2019.
32. Teutsch, M.; Grinberg, M. Robust Detection of Moving Vehicles in Wide Area Motion Imagery. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June 2016–1 July 2016.
33. Aguilar, C.; Ortner, M.; Zerubia, J. Small Moving Target MOT Tracking with GM-PHD Filter and Attention-Based CNN. In Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, MLSP, Gold Coast, Australia, 25–28 October 2021; pp. 1–6. [[CrossRef](#)]
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
35. Aguilar, C.; Ortner, M.; Zerubia, J. Small Object Detection and Tracking in Satellite Videos With Motion Informed-CNN and GM-PHD Filter. *Front. Signal Process.* **2022**, *2*, 827160. [[CrossRef](#)]
36. Lyu, C.; Liu, Y.; Li, B.; Chen, H. Multi-Feature Based High-Speed Ball Shape Target Tracking. In Proceedings of the 2015 IEEE International Conference on Information and Automation, ICIA 2015—In conjunction with 2015 IEEE International Conference on Automation and Logistics, Lijiang, China, 8–10 August 2015.
37. Hongshan, N.; Zhijian, H.; Jietao, D.; Jing, C.; Haijun, L.; Qiang, L. A Wiener Filter Based Infrared Small Target Detecting and Tracking Method. In Proceedings of the 2010 International Conference on Intelligent System Design and Engineering Application, ISDEA 2010, Changsha, China, 13–14 October 2010; Volume 1.
38. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. Max-Mean and Max-Median Filters for Detection of Small Targets. In *Signal and Data Processing of Small Targets 1999*; SPIE: Cergy, Germany, 1999; Volume 3809, pp. 74–83.
39. Ahmadi, K.; Salari, E. Small Dim Object Tracking Using Frequency and Spatial Domain Information. *Pattern Recognit.* **2016**, *58*, 227–234. [[CrossRef](#)]

40. Dong, X.; Huang, X.; Zheng, Y.; Shen, L.; Bai, S. Infrared Dim and Small Target Detecting and Tracking Method Inspired by Human Visual System. *Infrared Phys. Technol.* **2014**, *62*, 100–109. [[CrossRef](#)]
41. Dong, X.; Huang, X.; Zheng, Y.; Bai, S.; Xu, W. A Novel Infrared Small Moving Target Detection Method Based on Tracking Interest Points under Complicated Background. *Infrared Phys. Technol.* **2014**, *65*, 36–42. [[CrossRef](#)]
42. Zhang, F.; Li, C.; Shi, L. Detecting and Tracking Dim Moving Point Target in IR Image Sequence. *Infrared Phys. Technol.* **2005**, *46*, 323–328. [[CrossRef](#)]
43. Shaik, J.S.; Iftexharuddin, K.M. Automated Tracking and Classification of Infrared Images. In Proceedings of the International Joint Conference on Neural Networks, Portland, OR, USA, 20–24 July 2003; Volume 2.
44. Liu, C.; Ding, W.; Yang, J.; Murino, V.; Zhang, B.; Han, J.; Guo, G. Aggregation Signature for Small Object Tracking. *IEEE Trans. Image Process.* **2020**, *29*, 1738–1747. [[CrossRef](#)] [[PubMed](#)]
45. Tzannes, A.P.; Brooks, D.H. Temporal Filters for Point Target Detection in IR Imagery. In Proceedings of the Infrared Technology and Applications XXIII, Orlando, FL, USA, 20–25 April 1997; Volume 3061.
46. Bae, T.W.; Kim, B.I.; Kim, Y.C.; Sohng, K.I. Small Target Detection Using Cross Product Based on Temporal Profile in Infrared Image Sequences. *Comput. Electr. Eng.* **2010**, *36*, 1156–1164. [[CrossRef](#)]
47. Bae, T.W. Spatial and Temporal Bilateral Filter for Infrared Small Target Enhancement. *Infrared Phys. Technol.* **2014**, *63*, 42–53. [[CrossRef](#)]
48. Choudhary, M. Automatic Target Detection and Tracking in Forward-Looking Infrared Image Sequences Using Morphological Connected Operators. *J. Electron. Imaging* **2004**, *13*, 802–813. [[CrossRef](#)]
49. Son, S.; Kwon, J.; Kim, H.Y.; Choi, H. Tiny Drone Tracking Framework Using Multiple Trackers and Kalman-Based Predictor. *J. Web Eng.* **2021**, *20*, 2391–2412. [[CrossRef](#)]
50. Zhang, X.; Zhang, T.; Yang, Y.; Wang, Z.; Wang, G. Real-Time Golf Ball Detection and Tracking Based on Convolutional Neural Networks. In Proceedings of the Conference Proceedings—IEEE International Conference on Systems, Man and Cybernetics, Prague, Czech Republic, 9–12 October 2020; Volume 2020.
51. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
52. Aktaş, M.; Ateş, H.F. Small Object Detection and Tracking from Aerial Imagery. In Proceedings of the 6th International Conference on Computer Science and Engineering, UBMK, Ankara, Turkey, 15–17 September 2021; pp. 688–693. [[CrossRef](#)]
53. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. In Proceedings of the Proceedings—International Conference on Image Processing, ICIP, Athens, Greece, 7–10 October 2018; Volume 2017.
54. Behrendt, K.; Novak, L.; Botros, R. A Deep Learning Approach to Traffic Lights: Detection, Tracking, and Classification. In Proceedings of the Proceedings—IEEE International Conference on Robotics and Automation, Singapore, 9 May 2017–3 June 2017.
55. Hurault, S.; Ballester, C.; Haro, G. Self-Supervised Small Soccer Player Detection and Tracking. In Proceedings of the MMSports 2020—Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports, Seattle, WA, USA, 16 October 2020.
56. Zhu, Y.; Li, C.; Liu, Y.; Wang, X.; Tang, J.; Luo, B.; Huang, Z. Tiny Object Tracking: A Large-Scale Dataset and a Baseline. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–15. [[CrossRef](#)] [[PubMed](#)]
57. Liu, C.; Xu, S.; Zhang, B. Aerial Small Object Tracking with Transformers. In Proceedings of the 2021 IEEE International Conference on Unmanned Systems, ICUS 2021, Beijing, China, 15–17 October 2021.
58. Huang, Y.C.; Liao, I.N.; Chen, C.H.; Ik, T.U.; Peng, W.C. TrackNet: A Deep Learning Network for Tracking High-Speed and Tiny Objects in Sports Applications. In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2019, Taipei, Taiwan, 18–21 September 2019.
59. Yoshihashi, R.; Trinh, T.T.; Kawakami, R.; You, S.; Iida, M.; Naemura, T. Differentiating Objects by Motion: Joint Detection and Tracking of Small Flying Objects. *arXiv* **2017**, arXiv:1709.04666.
60. Marvasti-Zadeh, S.M.; Khaghani, J.; Ghanei-Yakhdan, H.; Kasaei, S.; Cheng, L. COMET: Context-Aware IoU-Guided Network for Small Object Tracking. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12623.
61. Mueller, M.; Smith, N.; Ghanem, B. A Benchmark and Simulator for UAV Tracking. In *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9905.
62. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [[CrossRef](#)]
63. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and Tracking Meet Drones Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7380–7399. [[CrossRef](#)] [[PubMed](#)]
64. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
65. Padilla, R.; Netto, S.L.; Da Silva, E.A.B. A Survey on Performance Metrics for Object-Detection Algorithms. In Proceedings of the International Conference on Systems, Signals, and Image Processing, Niterói, Brazil, 1–3 July 2020; Volume 2020.
66. Mirzaei, B.; Rahmati, F.; Nezamabadi-pour, H. A Score-Based Preprocessing Technique for Class Imbalance Problems. *Pattern Anal. Appl.* **2022**, *25*, 913–931. [[CrossRef](#)]
67. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *arXiv* **2020**, arXiv:2010.16061.

68. Hu, Y.; Xiao, M.; Li, S.; Yang, Y. Aerial Infrared Target Tracking Based on a Siamese Network and Traditional Features. *Infrared Phys. Technol.* **2020**, *111*, 103505. [[CrossRef](#)]
69. Yuan, D.; Zhang, X.; Liu, J.; Li, D. A Multiple Feature Fused Model for Visual Object Tracking via Correlation Filters. *Multimed. Tools Appl.* **2019**, *78*, 27271–27290. [[CrossRef](#)]
70. Bernardin, K.; Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The Clear Mot Metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [[CrossRef](#)]
71. Dendorfer, P.; Osep, A.; Milan, A.; Schindler, K.; Cremers, D.; Reid, I.; Roth, S.; Leal-Taixé, L. MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 845–881. [[CrossRef](#)]
72. Nalawade, R.; Mane, P.; Haribhakta, Y.; Yedke, R. Multiclass Multiple Object Tracking. In *Lecture Notes in Networks and Systems*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 119.
73. Wu, Y.; Lim, J.; Yang, M.H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [[CrossRef](#)]
74. Urrea, C.; Agramonte, R. Kalman Filter: Historical Overview and Review of Its Use in Robotics 60 Years after Its Creation. *J. Sens.* **2021**, *2021*, 9674015. [[CrossRef](#)]
75. Pei, Y.; Biswas, S.; Fussell, D.S.; Pingali, K. An Elementary Introduction to Kalman Filtering. *Commun. ACM* **2019**, *62*, 122–133. [[CrossRef](#)]
76. Cheong, Y.Z.; Chew, W.J. The Application of Image Processing to Solve Occlusion Issue in Object Tracking. In *MATEC Web of Conferences*; EDP Sciences: Les Ulis, France, 2018; Volume 152.
77. Yuan, Y.; Chu, J.; Leng, L.; Miao, J.; Kim, B.G. A Scale-Adaptive Object-Tracking Algorithm with Occlusion Detection. *EURASIP J. Image Video Process.* **2020**, *2020*, 7. [[CrossRef](#)]
78. Cui, Y.; Hou, B.; Wu, Q.; Ren, B.; Wang, S.; Jiao, L. Remote Sensing Object Tracking with Deep Reinforcement Learning under Occlusion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [[CrossRef](#)]
79. Wang, Q.; Zheng, Y.; Pan, P.; Xu, Y. Multiple Object Tracking with Correlation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3876–3886.
80. Luo, H.; Wang, P.; Chen, H.; Kowelo, V.P. Small Object Detection Network Based on Feature Information Enhancement. *Comput. Intell. Neurosci.* **2022**, *2022*, 6394823. [[CrossRef](#)] [[PubMed](#)]
81. Shi, T.; Gong, J.; Hu, J.; Zhi, X.; Zhang, W.; Zhang, Y.; Zhang, P.; Bao, G. Feature-Enhanced CenterNet for Small Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5488. [[CrossRef](#)]
82. Ajaz, A.; Salar, A.; Jamal, T.; Khan, A.U. Small Object Detection Using Deep Learning. *arXiv* **2022**, arXiv:2201.03243.
83. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Detect to Track and Track to Detect. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Volume 2017.
84. Yu, F.; Li, W.; Li, Q.; Liu, Y.; Shi, X.; Yan, J. Poi: Multiple Object Tracking with High Performance Detection and Appearance Feature. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; Proceedings, Part II 14. Springer: Berlin/Heidelberg, Germany, 2016; pp. 36–42.
85. Sowah, N.L.; Wu, Q.; Meng, F. A Classification and Clustering Method for Tracking Multiple Objects. In Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Reno, NV, USA, 8–10 January 2018; pp. 537–544.
86. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple Online and Realtime Tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, Arizona, USA, 25–28 September 2016; pp. 3464–3468.
87. Dicle, C.; Camps, O.I.; Sznai, M. The Way They Move: Tracking Multiple Targets with Similar Appearance. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013.
88. Gong, Z.; Li, D. Towards Better Object Detection in Scale Variation with Adaptive Feature Selection. *arXiv* **2020**, arXiv:2012.03265.
89. Singh, B.; Davis, L.S. An Analysis of Scale Invariance in Object Detection—SNIP. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
90. Chavez-Garcia, R.O.; Aycard, O. Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 525–534. [[CrossRef](#)]
91. Deo, A.; Palade, V.; Huda, M.N. Centralised and Decentralised Sensor Fusion-based Emergency Brake Assist. *Sensors* **2021**, *21*, 5422. [[CrossRef](#)] [[PubMed](#)]
92. Meng, Z.; Xia, X.; Xu, R.; Liu, W.; Ma, J. HYDRO-3D: Hybrid Object Detection and Tracking for Cooperative Perception Using 3D LiDAR. *IEEE Trans. Intell. Veh.* **2023**, 1–13. [[CrossRef](#)]
93. Xia, X.; Meng, Z.; Han, X.; Li, H.; Tsukiji, T.; Xu, R.; Zheng, Z.; Ma, J. An Automated Driving Systems Data Acquisition and Analytics Platform. *Transp. Res. Part. C Emerg. Technol.* **2023**, *151*, 104120. [[CrossRef](#)]
94. Liu, W.; Quijano, K.; Crawford, M.M. YOLOv5-Tassel: Detecting Tassels in RGB UAV Imagery With Improved YOLOv5 Based on Transfer Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8085–8094. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.