remote sensing

MDPI

Article

Multiple Ship Tracking in Remote Sensing Images Using Deep Learning

Jin Wu, Changqing Cao *, Yuedong Zhou, Xiaodong Zeng, Zhejun Feng, Qifan Wu and Ziqiang Huang

School of Physics and Optoelectronic Engineering, Xidian University, 2 South Taibai Road, Xi'an 710071, China; jinw9824@stu.xidian.edu.cn (J.W.); xyan_2@stu.xidian.edu.cn (Y.Z.); xdzeng@xidian.edu.cn (X.Z.); zhjfeng@mail.xidian.edu.cn (Z.F.); qfwu_1@stu.xidian.edu.cn (Q.W.); zqhuang_1@stu.xidian.edu.cn (Z.H.)
* Correspondence: chqcao@mail.xidian.edu.cn

Abstract: In remote sensing images, small target size and diverse background cause difficulty in locating targets accurately and quickly. To address the lack of accuracy and inefficient real-time performance of existing tracking algorithms, a multi-object tracking (MOT) algorithm for ships using deep learning was proposed in this study. The feature extraction capability of target detectors determines the performance of MOT algorithms. Therefore, you only look once (YOLO)-v3 model, which has better accuracy and speed than other algorithms, was selected as the target detection framework. The high similarity of ship targets will cause poor tracking results; therefore, we used the multiple granularity network (MGN) to extract richer target appearance information to improve the generalization ability of similar images. We compared the proposed algorithm with other state-of-the-art multi-object tracking algorithms. Results show that the tracking accuracy is improved by 2.23%, while the average running speed is close to 21 frames per second, meeting the needs of real-time tracking.

Keywords: multi-object tracking; remote sensing image; multiple granularity network (MGN); deep learning

Citation: Wu, J.; Cao, C.; Zhou, Y.; Zeng, X.; Feng, Z.; Wu, Q.; Huang, Z. Multiple Ship Tracking in Remote Sensing Images Using Deep Learning. *Remote Sens.* **2021**, *13*, 3601. https://doi.org/10.3390/rs13183601

Academic Editor: Józef Lisowski

Received: 6 August 2021
Accepted: 6 September 2021
Published: 9 September 2021

1. Introduction

With rapid developments in space science, remote sensing technology has greatly improved the small coverage of traditional ground detection and the lack of related data through the high-speed acquisition of omnidirectional and multi-view ground information. Ships are indispensable strategic resources and means of transportation in military and civilian fields. Therefore, ship target tracking is the focus of this study. Remote sensing images of ships are small in size and have complex backgrounds, which makes ship target tracking more challenging than multi-object tracking of pedestrians on the road.

Multi-object tracking technology combines the context information in the video sequence to perform location recognition, track maintenance, and ID recording for multiple targets of interest at the same time. The traditional MOT algorithm expresses the task as a data association problem. According to the use of historical frame information, it is divided into online association (e.g., probabilistic data association [1], Poisson multi-Bernoulli mixture [2], and joint probabilistic data association [3]) and offline association (e.g., multiple hypothesis tracking [4]). Early research mainly focused on proposing and solving better data association algorithms, but it is difficult to meet all possible actual situations due to the limitation of rules. Deep learning has greatly improved the performance of computer vision tasks due to its powerful feature extraction capabilities, but research related to multi-object tracking is minimal. Seyed et al. [5] first proposed an online multi-object tracking method based on a fully developed, end-to-end learning algorithm in 2016, which solved the problem of modeling target number changes and discrete data association. Since then, deep learning has been widely used in the field of multi-object tracking.

In existing multi-object tracking systems, the mainstream algorithm is the detection-based tracking (DBT) algorithm, including three stages, which are object detection, feature matching, and data association [6]. DBT algorithm is represented by single online and real-time tracking (SORT) [7] and DeepSORT algorithms [8]. Among them, the DeepSORT algorithm is proposed based on the SORT algorithm, and multi-object tracking is improved significantly by extracting deep feature information. Otherwise, Chu et al. [9] proposed a CNN-based framework for online MOT and used ROI pooling to track potential features for each target. Zhou et al. [10] replaced the object detection framework with CenterNet and applied the tracker to the previous frame. Wang et al. [11] presented a new instance of joint MOT based on graph neural networks (GNNs). Guo et al. [12] introduced a dynamic Siamese network through a fast transformation learning model, which enabled effective online learning of target. However, object tracking algorithms rely on the feature-extraction capability of target detectors, and the tracking speed largely depends on the target detection speed. Traditional target detection algorithms, such as edge detection algorithm [13], threshold segmentation method [14], and visual saliency detection [15], cannot meet real-time requirements.

Recently, with rapid developments in deep learning, the convolutional neural network (CNN) algorithm has become the preferred framework for target detection networks due to its powerful feature extraction and modeling capabilities. CNN-based target detection algorithms can be divided into two categories [16]: the first being the region-based target detection algorithm with the advantage of accuracy, forming two-stage algorithms represented by R-CNN [17–20]; the second being the regression-based target detection algorithm with the advantage of speed, forming a one-stage algorithm represented by YOLO [21–23], single-shot multi-box detector (SSD) [24], RetinaNet [25], CenterNet [26], and BorderDet [27]. Due to the superiority of CNN and the importance of ship targets, Sebastian et al. [28] proposed a benchmark for marine ship target detection using the Singapore Maritime Dataset. Dilip et al. [29] introduced the idea of using the proximity of the bottom edge of the box as a new evaluation criterion for marine target detection. Qiao et al. [30] constructed a multi-category and multi-angle ship target dataset, VesselID-539, and applied Resnet50 as the backbone network to complete a five-element network framework, improving the accuracy of ship reidentification. The abovementioned method provides a new way for improving the accuracy of the multi-object tracking algorithm, but it still cannot meet the real-time requirements.

YOLOv3 is significantly better than other algorithms in detecting accuracy and speed; thus, a multi-object tracking algorithm based on an improved YOLOv3 network is proposed herein. To better the detection ability, we first improved the YOLOv3 network model to a more suitable detection framework for this experiment. Concurrently, by introducing the MGN network to extract more accurate target appearance information for feature prediction and matching, the tracking effect of similar images was improved. The overall block diagram of the proposed multi-object tracking algorithm is shown in Figure 1.

The main contributions of the paper can be summarized as follows:

(1) We chose YOLOv3 as the target detection framework and improved it to enhance the detection ability of the model. By removing the 52×52 prediction scale, the network depth is reduced, and the training time is greatly saved. In order to fit the target characteristics of the ship, we linearly stretched the anchor box after the K-means algorithm. We also adjusted the loss of the model to solve the imbalance of positive and negative samples.

(2) To solve the problem of insufficient data, we selected ship targets under different complex backgrounds to make datasets and improve the performance evaluation of the network. MGN was used to extract more detailed target appearance information to facilitate the formation of a complete motion tracking trajectory.

The remainder of this paper is organized as follows: Section 2 introduces the related work. Section 3 gives a detailed description of our method. In Section 4, we introduce the results of the experiment. Finally, in Section 5, conclusions are drawn.

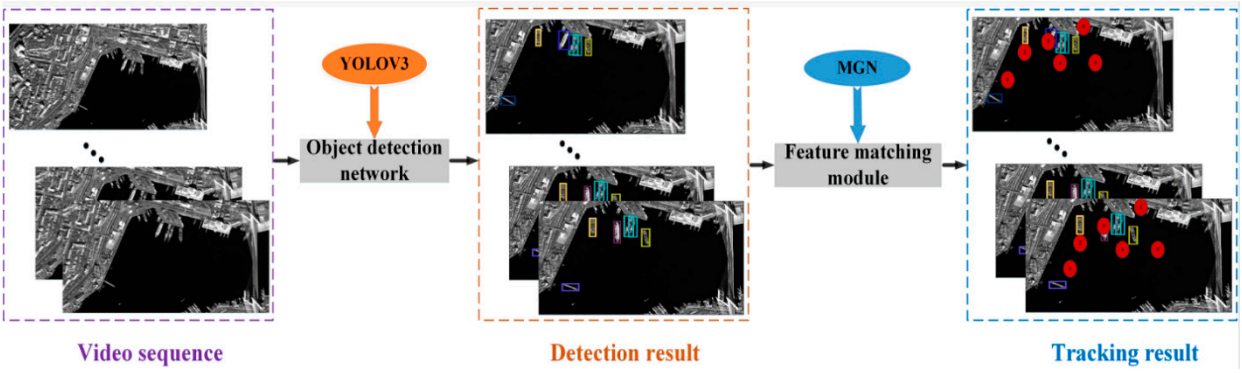


Figure 1. Detection and segmentation effect on our proposed method.

2. Related Work

Target tracking models the motion and appearance characteristics of the target by analyzing the context information of the image sequence and then predicts the target motion state and calibrates the target position. Compared with single object tracking (SOT), which focuses on locating targets in subsequent frames, the MOT algorithm focuses more on target information matching and ID maintenance. Considering the requirements of real-time target tracking, the current DBT tracking mode has become mainstream in this field. The DBT algorithm is mainly composed of three parts: target detection, feature matching, and data association.

2.1. Object Detection

The target detection system analyzes the input image and predicts the location of the target of interest. As an important part of the multi-target tracking algorithm, the performance of the target detector will directly affect the effect of multi-object tracking. In the process of target detection, missed or false detection will cause frequent switching of the target ID and disorder of the target's movement trajectory. For the above reasons, the reasonable choice of target detection network is crucial. The traditional target detection algorithms based on image processing technology are mainly divided into optical flow method [31], inter-frame difference method [32], background difference method [33], and template matching method [34]. The abovementioned algorithms complete the detection task by detecting the strong contrast between the target and the image background, which is difficult to meet the target detection under the complex background. To further meet real-time requirements, research based on deep learning has been widely used in the field of target detection. Feature extraction is the core step of the target detection algorithm. By judging whether to manually extract feature information, it can be divided into machine learning-based and deep learning-based target detection. The former is mostly manual extraction of features, such as haar feature [35], histogram of oriented gradient [36], and local binary pattern [37]. The latter learns features by itself through convolution operations and is mainly composed of one-stage algorithms (e.g., YOLO [21–23], SSD [24], RetinaNet [25], CenterNet [26], and BorderDet [27]) and two-stage algorithms (e.g., R-CNN [17–20]). Currently, target detection based on deep learning has become the mainstream model, which has promoted the rapid development of MOT technology.

2.2. Feature Matching

Taking into account the misalignment when the uncertainty of object state estimation is low, we usually measure the matching degree of features by combining the motion information and appearance information of the target.

2.2.1. Motion Model

To achieve continuous tracking of the target, the motion model predicts the position of the target in the video frame by frame and transforms the tracking box of the previous

frame to the current frame to match the detection box. In most motive scenes, the camera is fixed, and the interval between two adjacent frames is short, and it can default to a uniform linear motion. Therefore, the Kalman filter [2] model is usually used to predict the motion of two adjacent frames. Subsequently, more and more scholars carry out research on the motion characteristics of the target. Martin et al. [38] used correlation filters (CF) to replace the Kalman filter in SORT to improve the tracking effect. Bochinski et al. [39] proposed the use of kernelized correlation filters (KCFs) to expand high-speed intersection over union (IOU) tracking, which solved the problem of trajectory clutter and frequent ID switching. Zhao et al. [40] introduced a new type of compressed CNN based on correlation filtering, which has the ability to reidentify when the target is lost. With the rise of deep learning, optical flow estimation is directly applied to neural networks. Fischer et al. [41] designed an optical flow network (Flownet) to directly predict optical flow with a trained decoder. Liu et al. [42] applied PWC-Net [43] to alleviate camera motion defects and regarded the optical flow network as an auxiliary tracker to solve the problem of missed target detection. In the process of short-term target tracking, it is more accurate to use motion features to match the target.

2.2.2. Appearance Model

The appearance model aims to learn the external characteristics of the target so that the same target feature in each frame of the image sequence is more similar than the different target features. In the early research, traditional artificial features (e.g., color histograms, and gradient features) were often used to characterize the appearance of the target, but the recognition effect was poor on complex target tracking problems. The rapid development of deep learning has made the appearance features of targets extracted by neural networks widely used in the field of target recognition. Laura et al. [44] trained the Siamese net to learn descriptors encoding local spatiotemporal structures between the two input image patches. Schroff et al. [45] used a novel online triplet mining method to realize tasks such as face recognition, verification, and clustering. On this basis, Hermans et al. [46] designed a variant of the triplet loss to perform end-to-end deep metric learning. Xiao et al. [47] proposed margin sample mining loss (MSML), which can achieve better learning losses. Son et al. [48] presented the quadruplet network model to learn more target appearance features. The more detailed the appearance features are, the higher the tracking efficiency of similar images is. When the target is occluded, it is better to use the appearance information to match the target.

2.3. Data Association

Data association refers to ID matching of multiple targets according to the similarity measurement results. The similarity measurement measures the similarity between the detection box and the tracking box by calculating the position distance and the feature distance. Commonly used measurement methods, such as IOU measurement, Mahalanobis distance, and cosine distance [3], cannot form a complete trajectory. As a key step in the DBT task, data association directly determines the matching efficiency and the final tracking effect of the tested targets. In the process of data association, it is difficult to distinguish the target characteristics, which will lead to frequent ID switching. After the target is occluded for a short time, how to continue to accurately recover the target information is also the focus of the current research in the field of multi-object tracking. The most common data association algorithm in MOT is the Hungarian algorithm [2], which is to obtain the maximum matching value of the bipartite graph through the augmented path. In addition, algorithms such as recurrent neural network (RNN) [49] and deep multi-layer perceptron (MLP) [50] are gradually being applied at the data association stage.

3. Methods

The MOT algorithm simultaneously performs location recognition, track maintenance, and ID recording for multiple targets in videos, which is difficult to track. As illustrated

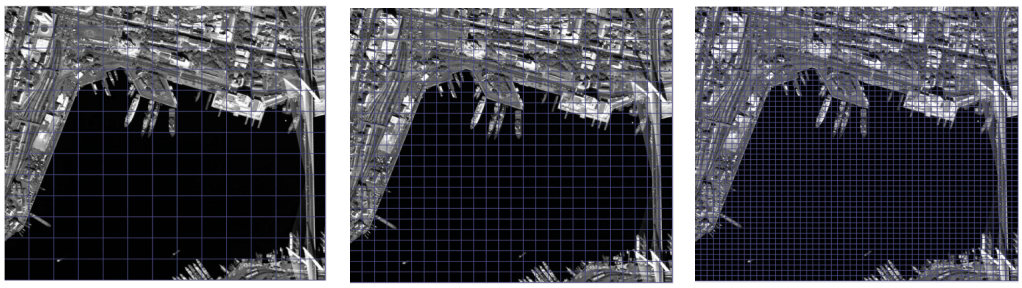
in Figure 1, we used the improved YOLOv3 algorithm for target detection and formed a complete trajectory of ship targets according to the multi-object tracking algorithm with MGN. To further improve the accuracy and speed of object tracking tasks, we made the following optimizations.

3.1. Improvement of YOLOv3

The existing YOLOv3 algorithm is only suitable for target detection with high contrast in natural scenes. Therefore, we improved the network model using the characteristics of ship targets.

3.1.1. Detection Scale

The YOLOv3 network effectively combines deep semantic information and shallow feature information through the introduction of a multi-scale prediction structure, improving the accuracy of target detection. The feature pyramid was constructed with three sizes of feature maps for detecting different sizes of targets. The receptive fields at three scales are shown in Figure 2.



(**a**) (**b**) (**c**)

Figure 2. Receptive field of the same aerial remote sensing image at different scales: (**a**) 13 × 13; (**b**) 26 × 26; (**c**) 52 × 52.

However, our target detection dataset was derived from remote sensing images. From Figure 2, the ship target studied is more suitable for the prediction of the two scales of 13 × 13 and 26 × 26. Thus, the detection scale of 52 × 52 was omitted from this paper to avoid wasted time caused by model complexity. This reduces the network model depth and number of parameters and saves the detection time of the target. The improved network model graph is shown in Figure 3.

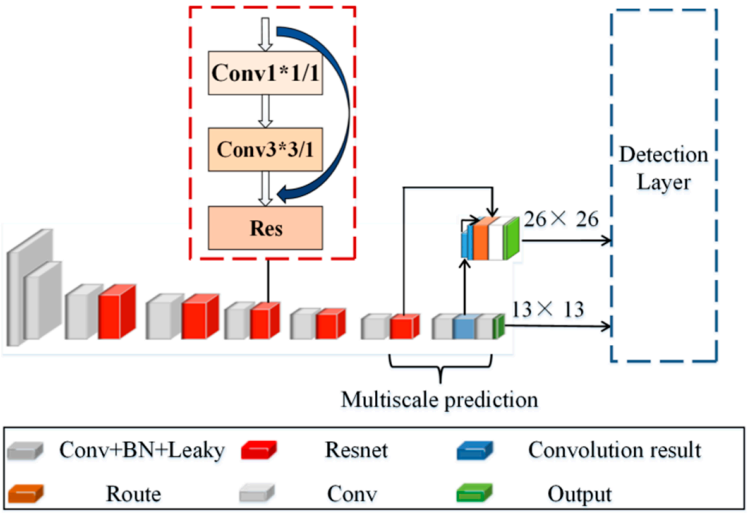


Figure 3. Structure of the improved network model.

3.1.2. Anchor Box

The ship target in the remote sensing image is elongated and different from the fixed anchor box shape originally set by YOLOv3. It was therefore necessary to readjust the number and size of the anchor boxes when performing object detection. The K-means algorithm [51] was used for clustering analysis to determine the prior anchor box more suitable for the existing dataset. We set the range of K from 1 to 12, and the IOU values corresponding to different cluster numbers of the dataset are shown in Figure 4.

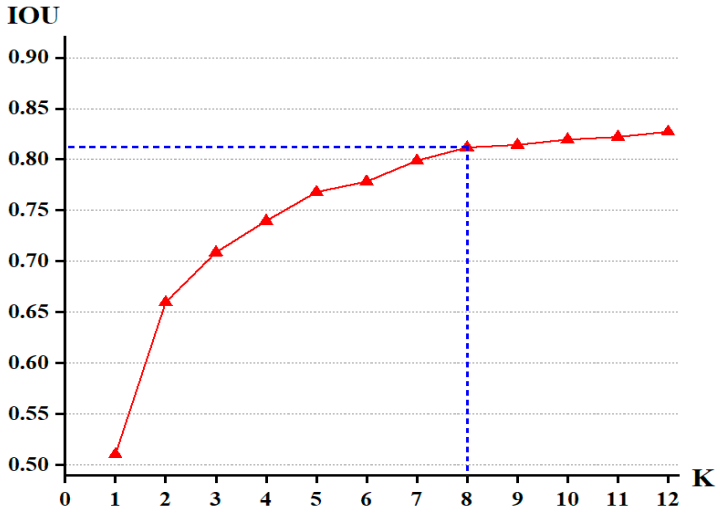


Figure 4. The clustering results of our datasets.

When K = 8, the curve is stable. Thus, we selected eight sets of anchor frames. As the ship target is elongated and large, to fit its characteristics more closely, we linearly lifted the prior box; the specific formula is as follows:

$$\begin{cases} x_1' = \alpha x_1 \\ x_8' = \beta x_8 \\ x_i' = \frac{(x_i - x_1)}{(x_8 - x_1)}(x_8' - x_1') + x_1' \\ y_i' = x_1' \frac{y_i}{x_i} \end{cases} \quad (1)$$

where x_i and y_i are the horizontal and vertical dimensions of the i-th anchor box after linear scaling, respectively; α and β are scaling factors with values of 0.8 and 1.2, respectively. The final sizes of the anchor boxes after linear scaling are (89 × 217), (96 × 134), (142 × 184), (181 × 121), (183 × 243), (222 × 182), (293 × 195), and (334 × 147), respectively.

From the clustering results, it can be observed that the size of the detection dataset elongates in agreement with the actual situation. Further, choosing the appropriate size and number of corresponding anchor boxes can improve the model's precision in positioning and detection of specific targets.

3.1.3. Loss Function

For the one-stage algorithm, many easy negative examples in the dataset exist, that is, samples belonging to the background. In the target detection process, these samples dominate the updated direction of the gradient and hinder the reduction of the loss function, thereby seriously affecting the detection accuracy of the target. Focal loss [25] solves the unbalanced sample allocation by optimizing the cross-entropy function. The expression of the cross-entropy loss function is as follows:

$$CE(p, y) = \begin{cases} -\log(p), y = 1 \\ -\log(1 - p), y \neq 1 \end{cases} \quad (2)$$

where p is the output value of the activation function. The following transformation is performed on p to obtain the cross-entropy loss function:

$$p_t = \begin{cases} p, y = 1 \\ 1 - p, y \neq 1 \end{cases} \tag{3}$$

$$CE(p, y) = CE(p_t) = -\log(p_t) \tag{4}$$

In the iterative process of many simple negative samples, the loss function drops slowly and never attains the optimum value. The weight factor α_t is introduced to balance the uneven proportions of positive and negative samples. It is expressed as the proportion of the opposite class, that is, the more negative the sample is, the smaller the weight is. For the balance of simple and difficult samples, a dynamically changing weight $(1 - p_t)^\gamma$ is introduced, making the network pay more attention to difficult and misclassified samples. The final loss function is as follows:

$$Loss_{FL} = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{5}$$

where γ is the modulation factor used for reducing the loss of easy-to-classify samples, p_t is the prediction probability, and α_t is the weight factor whose value is in the range of [0, 1].

3.2. Improvement of Appearance Model

Remote sensing images become easily occluded by cloud interference. Appearance features, as highly distinguishable attributes in the target detection field, have advantages in identifying targets with similar shapes or occluded. Following the rise of deep learning, appearance features of targets extracted using neural networks are widely used in the fields of target detection and tracking. The multi-granularity network [52] based on deep feature extraction was proposed for person reidentification in 2018. Therefore, we selected the MGN network to extract richer and fine-grained information to complete the feature matching of the targets. The structure diagram is shown in Figure 5.

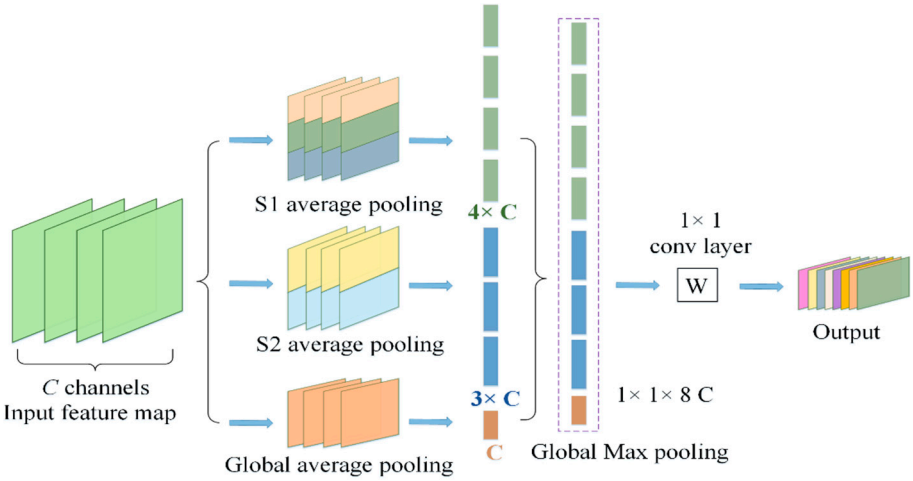


Figure 5. Schematic diagram of MGN structure.

From Figure 5, this network acquires different levels of multi-granularity feature information by extracting the global and local features of the target separately. We extracted the information from the target using the MGN module to obtain a more integral appearance feature. As the granularity of segmentation increases, the network learns more detailed information, thereby improving the association of target information in the long-term target tracking process. The multi-dimensional extraction of global and local features of the input feature map obtains more comprehensive feature information. The improvement

of appearance features strengthens the efficiency of tracking similar pictures and reduces the number of ID switches.

4. Experiments

4.1. Experimental Design

Remote sensing images have various scales and severe background interference, making data acquisition extremely difficult. Training using convolutional neural networks requires a large number of data samples, hence the difficulty to directly detect and track targets from existing datasets.

4.1.1. Datasets Creation

The particularity of remote sensing images raises the requirements of the network model for target tracking accuracy. Among the commonly known datasets of remote sensing images, the DOTA dataset has the best image quality presently, and the image size is around 4000–5000 pixels. The DOTA dataset was used herein to complete the production of the multi-object tracking dataset. The main process was as follows:

(1) We manually selected 40 pictures from the DOTA dataset to ensure that each picture contains approximately 10 ship targets and cropped them to 1024 × 1024 size images.
(2) Regarding ships on the sea, the speed of ordinary cargo ships is 22–27 km/h and the speed of large container ships is 36–52 km/h. We divided the size of the ship target in the selected picture according to the pixel value: targets larger than 150 × 150 are considered large targets; otherwise, the targets are considered as small targets. Thus, the small target translates forward by 5 pixels per frame and the large target by 3 pixels.
(3) We coded to implement operations, such as translation and rotation of the ship target in each picture, to obtain the required dataset.
(4) Repeat the above steps until the production of 40 video sequences (the MOT dataset) was completed.

One picture was taken from the 40 video sequences every five frames, and 4026 pictures were obtained. To avoid overfitting, the dataset was amplified through rotation angle and mirror flip. Finally, 20,000 images in the target detection dataset required for the experiment were obtained. Concurrently, all pictures were divided into a training set, a validation set, and a test set in the ratio of 3:1:1, which were used for training and evaluating the target detection network. In addition, because the detection and tracking datasets had partially overlapped pictures, we divided the 40 video sequences in the object tracking dataset into training and validation sets at a ratio of 7:3 and used these to retrain the object detection network and verify the proposed model presented in this study.

4.1.2. Evaluation and Implementation

In this paper, we chose the average precision (AP) and output tensors of our model as the evaluation indexes for object detection. Further, the multi-object tracking precision (MOTP), the multi-object tracking accuracy (MOTA), the identity switches (IDsw), and the frames per second (FPS) were selected to evaluate the tracking capability of the network model. All experiments were based on a computer with an Intel Core i7-7700K CPU at 4.20 GHz and NVIDIA GTX 1060Ti GPU. There were 12,000 training samples, 4000 validation samples, and 4000 test samples in our self-made datasets. In the process of model training, the self-made dataset was used for preliminary training of the network, and the validation set was used to further verify the effect of the improved model. The preset training parameters were as follows: the momentum was 0.9, the weight attenuation was 0.0005, the initial learning rate was 0.001, and the maximum number of iterations was 20,000. The learning rate drops to 0.0001 after 7500 iterations, and to 0.00001 after 15,000 iterations. We set the batch of the improved model to 32; the training epoch is 50, with 20 rounds in the first stage and 30 rounds in the second stage. The training loss function diagram of the improved model on the dataset is shown in Figure 6.

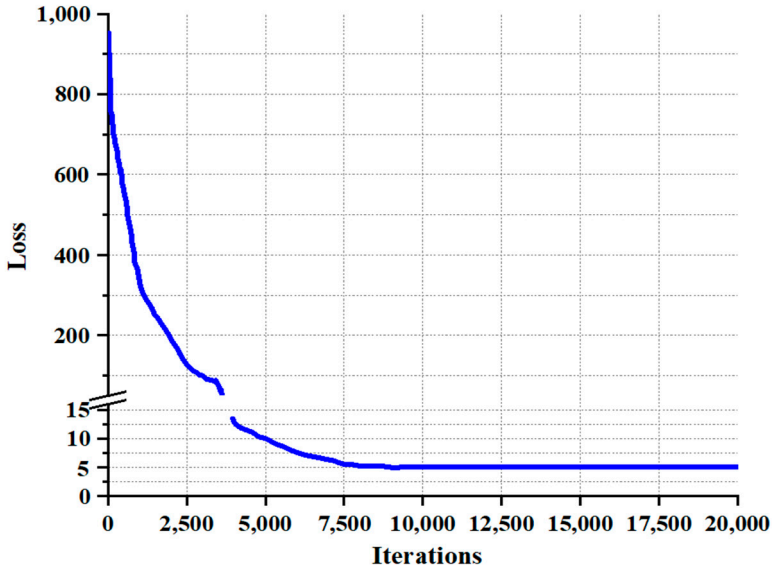


Figure 6. The training loss function diagram of the improved model on the dataset.

Figure 6 shows that the network model after the first 20 rounds of training was still underfitting, resulting in a large loss function. In the subsequent 30 rounds, the learning rate of the network model was adjusted in time, which effectively avoided the recurrence of overfitting. The training loss value after the final iteration was 4.7382.

4.2. Experiments and Analysis

This study was based on the improved YOLOv3 network along with the DeepSORT algorithm for multi-object tracking of remote sensing ship images. To test the effectiveness of the improved model, we used the mainstream detection and multi-object tracking algorithms to compare experiments with the improved network, comprehensively considering the detection and tracking capabilities of the network model from the detection and tracking result graph and multiple evaluation indicators.

4.2.1. Object Detection

We conducted experimental comparisons on the performance of each part of the improved YOLOv3 model.

From Table 1, YOLOv3(1) is the original YOLOv3 network with a detection accuracy of 91.32%, and the output tensor is 6.4×10^4. YOLOv3(2) is the network model after adjusting the prior anchor, and its detection accuracy is increased by 1.45%. YOLOv3(3) model deleted the 52×52 detection scale to reduce the complexity of the model and replaced the original loss function with focal loss to balance the uneven distribution. The final detection accuracy is slightly improved, but the output tensor is significantly reduced. Finally, our network model increased the detection accuracy to 93.55%, while reducing the number of output prediction tensors to a quarter of the original.

Table 1. Performance comparison of each part of the improved model.

Network Model	52×52 Scale	Loss_{FL}	Anchor Setting	AP/%	Tensors
YOLOv3(1)	✓			91.32	6.4×10^4
YOLOv3(2)	✓		✓	92.77	6.2×10^4
YOLOv3(3)		✓		92.09	1.7×10^4
ours		✓	✓	93.55	1.6×10^4

From the experimental results, it can be observed that the evaluation index of the improved model is considerably better than that of the YOLOv3 algorithm; moreover, the

AP value increased by 2.23%. The number of output prediction tensors of the improved model is only a quarter of that of YOLOv3, which greatly reduces the parameter-processing scale of the network. The results of the algorithm mentioned in this paper are shown in Figure 7.

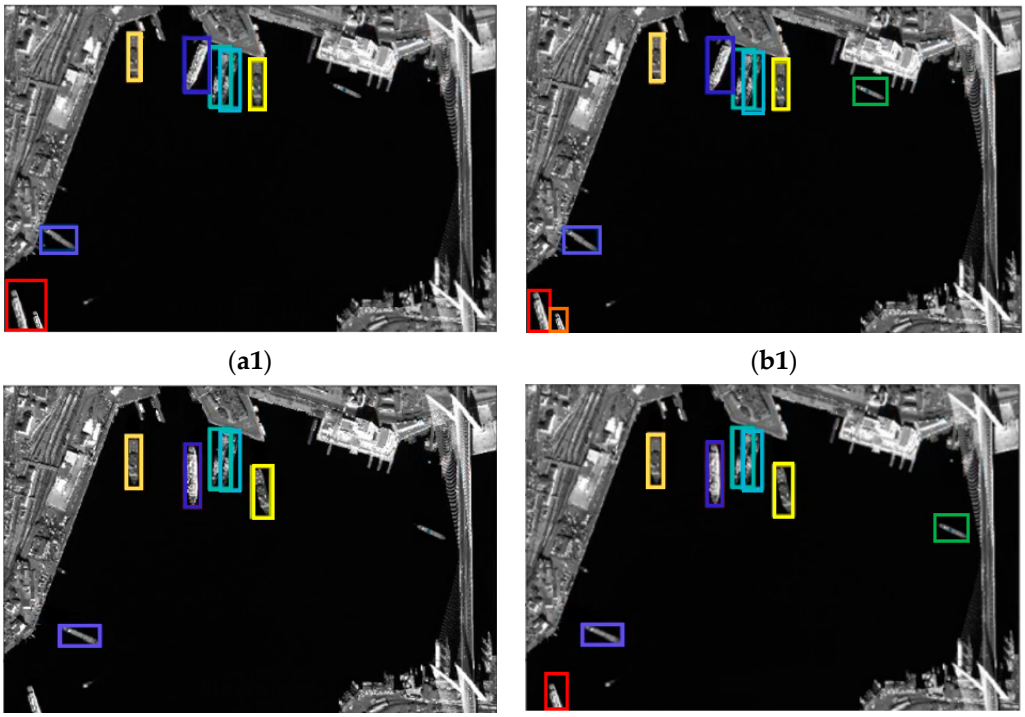


Figure 7. The results showing the detection of different algorithms on the dataset: (**a1**,**a2**) the YOLOv3 model; (**b1**,**b2**) the proposed model in this paper.

In Figure 7, the YOLOv3 model has a poor ability to detect the edge of the image; however, the improved algorithm makes up for this defect. Comprehensive performance indicators and detection results prove that the proposed algorithm considers both speed and accuracy and is suitable for remote sensing ship target detection.

4.2.2. Multi-Object Tracking

To verify the performance of the proposed algorithm in the multi-object tracking of ships, we conducted tests on the target tracking dataset and compared the experimental results with those for the SORT and DeepSORT algorithms. The results are shown in Table 2.

Table 2. Comparison of multi-object tracking effects of different algorithms.

Method	MOTA/%	MOTP/%	IDsw	FPS
SORT	59.7	73.5	122	13
DeepSORT	62.2	72.5	65	9
DeepSORT + MGN	63.1	72.3	54	9
ours	64.5	71.8	50	21

From Table 2, it can be observed that the multi-object tracking accuracy rate of the DeepSORT algorithm with the MGN network is slightly improved, compared with the previous two algorithms. However, the MGN network obtains more complete appearance characteristics by extracting the deep-level information of the target, which helps to improve the association of the target information in the long-term tracking process to form

a correct trajectory. Therefore, the addition of the MGN network is effective in reducing the number of IDsw. In addition, the multi-object tracking accuracy rate of our method is 64.5%, which is 4.8% and 2.3% higher than that of the SORT and DeepSORT algorithms, respectively. Concurrently, the target ID switching times of the improved algorithm are 15 less than those of the DeepSORT algorithm. In particular, after the introduction of the YOLOv3 detection model, the average running speed is close to 21 frames per second to meet the real-time requirements. The tracking results of the algorithm proposed in this paper are shown in Figure 8.

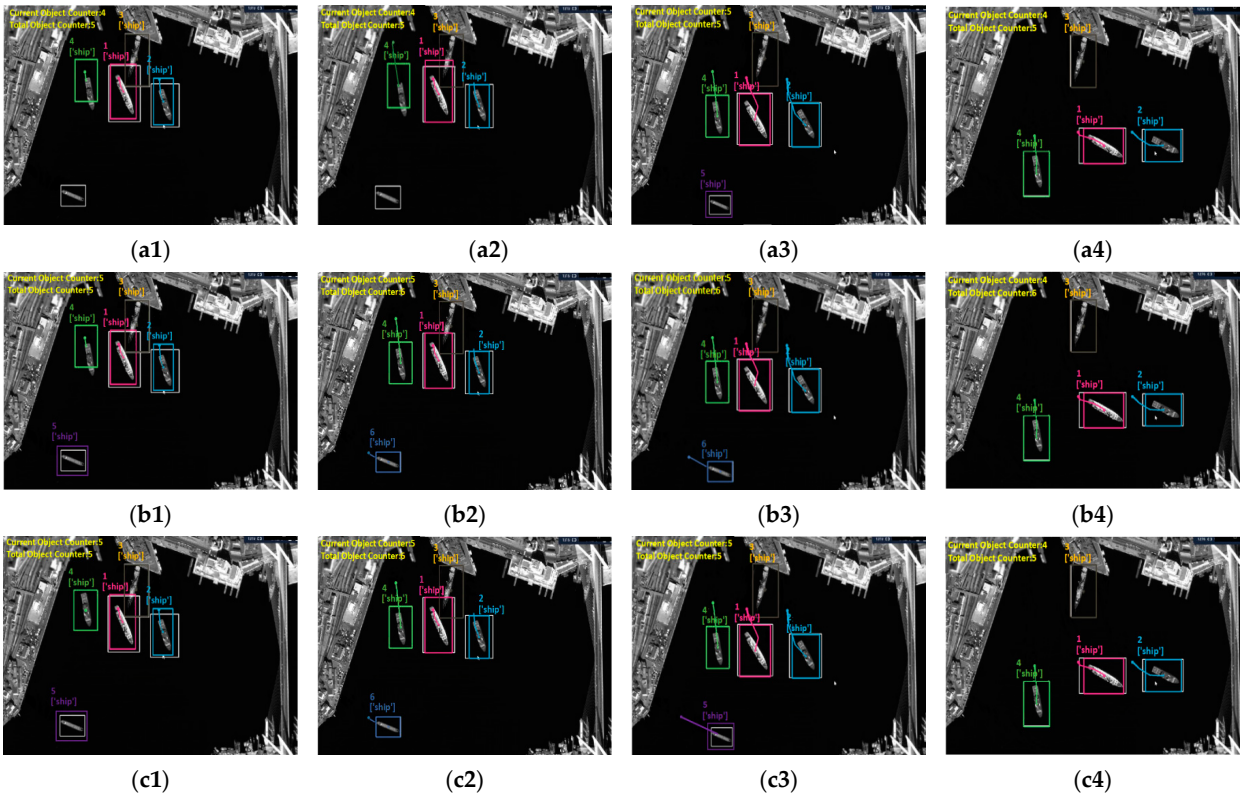


Figure 8. Tracking results of one frame on the dataset using different algorithms: (**a1**–**c1**): the 12th frames; (**a2**–**c2**): the 24th frames; (**a3**–**c3**): the 53rd frames; (**a4**–**c4**): the 62nd frames; (**a1**–**a4**) the tracking results of the SORT model; (**b1**–**b4**) the tracking results of the DeepSORT model; (**c1**–**c4**) the tracking results of the model proposed in this paper.

In Figure 8, the SORT algorithm missed detecting the 12th and 24th frames; that is, the fifth target was not detected. Both DeepSORT and improved algorithms mistakenly regarded the fifth target as a new target at frame 24; however, the improved algorithm re-identified it using matching appearance features at frame 53, while the DeepSORT algorithm still had errors in detection. In terms of tracking speed, the proposed algorithm strongly outperforms the other two algorithms.

The comprehensive network-tracking capability of the proposed model improved in terms of the index parameters or the results of the actual tracking diagram. Due to the particularity of remote sensing images, comprehensive accuracy, and speed, the proposed algorithm is more suitable for tracking remotely sensed ship targets, compared to current mainstream algorithms.

5. Conclusions

As an important strategic resource and transportation tool, ships have practical value that cannot be ignored in the research on remote sensing images. The background of remote sensing images is complex, and the scales are changeable; therefore, traditional multi-object tracking algorithms are difficult for extracting target features and accurately

locate ships. We proposed a multi-object tracking algorithm using a combination of the improved YOLOv3 network and the DeepSORT algorithm. By selecting the YOLOv3 network as the target detection framework and making corresponding adjustments to the target ship to enhance the detection ability, the MGN network was introduced to extract more complete target appearance information for feature matching. To obtain a more accurate motion trajectory, we set the matching priority relative to the frequency of target occurrences to enhance the data association. The experimental results show that the accuracy of the proposed model increased by 2.23%, and the FPS is more than twice the other algorithms, thereby improving the ability of the network. This study fully evaluated the degree of optimization of the improved algorithm using ship targets in complex backgrounds, which has practical significance for the research of remote sensing image multi-object tracking technology.

Author Contributions: Conceptualization, C.C. and Q.W.; methodology, Y.Z.; software, C.C.; validation, Y.Z., X.Z. and Z.H.; formal analysis, Q.W. and X.Z.; investigation, J.W.; resources, C.C.; data curation, J.W. and Z.F.; writing—original draft preparation, J.W.; writing—review and editing, Z.F. and C.C.; visualization, C.C.; supervision, Y.Z.; project administration, J.W.; funding acquisition, J.W. 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: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: The authors thank the team of optical sensing and measurement of Xidian University for their help. This research was supported by the National Natural Science Foundation of Shaanxi Province (Grant No. 2020JM-206), the National Defense Basic Research Foundation (Grant No. 61428060201), and 111 Project (B17035).

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

References

1. Bar-Shalom, Y.; Daum, F.; Huang, J. The probabilistic data association filter. *IEEE Control. Syst.* **2009**, *29*, 82–100.
2. Samuel, S.; Joachim, B.; Emil, R.; Amrit, K.; Karl, G. Mono-Camera 3D Multi-Object Tracking Using Deep Learning Detections and PMBM Filtering. In Proceedings of the IEEE IV, Suzhou, China, 26–30 June 2018; pp. 433–440.
3. Seyed, H.R.; Anton, M.; Zhang, Z.; Shi, Q.F.; Anthony, D.; Ian, R. Joint probabilistic data association revisited. In Proceedings of the IEEE ICCV, Santiago, Chile, 11–18 December 2015; pp. 3047–3055.
4. Kim, C.; Li, F.; Ciptadi, A.; Rehg, J.M. Multiple hypothesis tracking revisited. In Proceedings of the IEEE ICCV, Santiago, Chile, 11–18 December 2015; pp. 4696–4704.
5. Seyed, H.R.; Milan, A.; Zhang, Z.; Shi, Q.; Dick, A.; Reid, I. Joint Probabilistic Matching Using m-Best Solutions. In Proceedings of the IEEE CVPR, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 136–145.
6. Zhang, Y.; Mu, H.; Jiang, Y.; Ding, C.; Wang, Y. Moving Target Tracking Based on Improved GMPHD Filter in Circular SAR System. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 559–563. [CrossRef]
7. Bewley, A.; Ge, Z.Y.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the IEEE ICIP, Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
8. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the IEEE ICIP, Beijing, China, 17–20 September 2017; pp. 3645–3649.
9. Chu, Q.; Ouyang, W.; Li, H.; Wang, X.; Liu, B.; Yu, N. Online Multi-Object Tracking Using CNN-based Single Object Tracker with Spatial-Temporal Attention Mechanism. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 4836–4845.
10. Zhou, X.-Y.; Koltun, V.; Krähenbühl, P. Tracking objects as points. *arXiv* **2020**, arXiv:2004.01177.
11. Wang, Y.-X.; Weng, X.; Kitani, K. Joint Detection and Multi-Object Tracking with Graph Neural Networks. *arXiv* **2020**, arXiv:2006.13164.
12. Guo, Q.; Feng, W.; Zhou, C.; Huang, R.; Wan, L.; Wang, S. Learning Dynamic Siamese Network for Visual Object Tracking. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 1763–1771.
13. Mittal, M.; Verma, A.; Kaur, I.; Kaur, B.; Sharma, M.; Goyal, L.M.; Roy, S.; Kim, T.H. An Efficient Edge Detection Approach to Provide Better Edge Connectivity for Image Analysis. *IEEE Access* **2019**, *7*, 33240–33255. [CrossRef]

14. Tang, Z.; Wu, Y. One image segmentation method based on Otsu and fuzzy theory seeking image segment threshold. In Proceedings of the ICECC, Ningbo, China, 9–11 September 2011; pp. 2170–2173.
15. Cong, R.; Lei, J.; Fu, H.; Cheng, M.M.; Lin, W.; Huang, Q. Review of Visual Saliency Detection with Comprehensive Information. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2941–2959. [CrossRef]
16. Cao, C.; Wu, J.; Zeng, X.; Feng, Z.; Wang, T.; Yan, X.; Wu, Z.Y.; Wu, Q.F.; Huang, Z.Q. Research on Airplane and Ship Detection of Aerial Remote Sensing Images Based on Convolutional Neural Network. *Sensors* **2020**, *20*, 4696. [CrossRef]
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE CVPR, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. Girshick, R. Fast R-CNN. In Proceedings of the IEEE ICCV, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
20. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE CVPR, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
22. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
23. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. In Proceedings of the IEEE CVPR, Salt Lake City, UT, USA, 18–22 June 2018; arXiv:1804.02767. Available online: https://arxiv.org/abs/1804.02767 (accessed on 5 July 2021).
24. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot multibox Detector. In *Computer Vision—ECCV 2016*; Springer: Cham, Switzerland, 2016; pp. 21–37.
25. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 318–327.
26. Zhou, X.Y.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
27. Qiu, H.; Ma, Y.; Li, Z.; Liu, S.; Sun, J. Borderdet: Border Feature for Dense Object Detection. *arXiv* **2020**, arXiv:2007.11056.
28. Moosbauer, S.; Konig, D.; Jakel, J.; Teutsch, M. A Benchmark for Deep Learning Based Object Detection in Maritime Environments. In Proceedings of the IEEE CVPR, Long Beach, CA, USA, 16–20 June 2019; pp. 916–925.
29. Prasad, D.K.; Dong, H.; Rajan, D.; Quek, C. Are Object Detection Assessment Criteria Ready for Maritime Computer Vision. *IEEE Trans. Intell. Trans. Syst.* **2020**, *21*, 5295–5304. [CrossRef]
30. Qiao, D.; Liu, G.; Dong, F.; Jiang, S.X.; Dai, L. Marine Vessel Re-Identification: A Large-Scale Dataset and Global-and-Local Fusion-Based Discriminative Feature Learning. *IEEE Access* **2020**, *8*, 27744–27756. [CrossRef]
31. Horn, B.K.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203. [CrossRef]
32. Bruhn, A.; Weickert, J.; Schnörr, C. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *Int. J. Comput. Vis.* **2005**, *61*, 211–231. [CrossRef]
33. Goyal, K.; Singhai, J. Review of background subtraction methods using Gaussian mixture model for video surveillance systems. *Artif. Intell.* **2018**, *50*, 241–259. [CrossRef]
34. Omachi, S.; Omachi, M. Fast template matching with polynomials. *IEEE Trans. Image Process.* **2007**, *16*, 2139–2149. [CrossRef] [PubMed]
35. Lienhart, R.; Maydt, J. An Extended Set of Haar-like Features for Rapid Object Detection. In Proceedings of the IEEE ICIP, Santa Clara, CA, USA, 22–25 September 2002; pp. 900–903.
36. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE CVPR, San Diego, CA, USA, 20–25 June 2005.
37. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
38. Martin, D.; Häger, G.; Khan, F.; Felsberg, M. Accurate Scale Estimation for Robust Visual Tracking. In Proceedings of the BMCV, Nottingham, UK, 1–5 September 2014; pp. 74–81.
39. Bochinski, E.; Eiselein, V.; Sikora, T. High-Speed Tracking-by-Detection Without Using Image Information. In Proceedings of the IEEE AVSS, Lecce, Italy, 29 August–1 September 2017.
40. Zhao, D.W.; Fu, H.; Xiao, L.; Wu, T.; Dai, B. Multi-Object Tracking with Correlation Filter for Autonomous. *Sensors* **2018**, *18*, 2004. [CrossRef] [PubMed]
41. Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Smagt, P.; Cremers, D.; Brox, T. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE ICCV, Santiago, Chile, 11–18 December 2015; pp. 2758–2766.
42. Liu, W.Q.; Mu, J.; Liu, G. Multiple Object Tracking with Motion and Appearance Cues. In Proceedings of the IEEE CVPR, Long Beach, CA, USA, 16–20 June 2019; pp. 161–169.
43. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In Proceedings of the IEEE CVPR, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8934–8943.
44. Laura, L.T.; Canton-Ferrer, C.; Schindler, K. Learning by Tracking: Siamese CNN for Robust Target Association. In Proceedings of the IEEE CVPR, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 418–425.

45. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE CVPR, Boston, MA, USA, 7–12 June 2015; pp. 815–821.
46. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 1526–1535.
47. Xiao, Q.Q.; Luo, H.; Zhang, C. Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017.
48. Son, J.; Baek, M.; Cho, M.; Han, B. Multi-Object Tracking with Quadruplet Convolutional Neural Network. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 3786–3795.
49. Ma, C.; Yang, C.; Yang, F.; Zhuang, Y.; Zhang, Z.; Jia, H.; Xie, X. Trajectory factory: Tracklet cleaving and re-connection by deep Siamese bi-gru for multiple object tracking. In Proceedings of the IEEE ICME, San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
50. Kieritz, H.; Hubner, W.; Arens, M. Joint detection and online multi-object tracking. In Proceedings of the IEEE CVPR, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1540–1548.
51. Yi, F.; Moon, I. Extended K-Means Algorithm. In Proceedings of the IHMSC, Hangzhou, China, 26–27 August 2013; pp. 263–266.
52. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 15 October 2018; pp. 274–282.