applied sciences

MDPI

Article

Deep-Learning Based Algorithm for Detecting Targets in Infrared Images

Lifeng Yang [1], Shengzong Liu [2,*] and Yiqi Zhao [3,*]

[1] School of Optical and Communication Engineering, Yunnan Open University, Kunming 650000, China; yanglifeng@ynou.edu.cn
[2] School of Information Technology and Management, Hunan University of Finance and Economics, Changsha 410000, China
[3] Computer Science and Engineering, Central South University, Changsha 410000, China
* Correspondence: lsz@hufe.edu.cn (S.L.); zhaoyiqi@csu.edu.cn (Y.Z.)

Abstract: Infrared image target detection technology has been one of the essential research topics in computer vision, which has promoted the development of automatic driving, infrared guidance, infrared surveillance, and other fields. However, traditional target detection algorithms for infrared images have difficulty adapting to the target's multiscale characteristics. In addition, the accuracy of the detection algorithm is significantly reduced when the target is occluded. The corresponding solutions are proposed in this paper to solve these two problems. The final experiments show that this paper's infrared image target detection model improves significantly.

Keywords: infrared image; deep learning; neural network; target detection; transfer learning; multiscale characteristics; context analysis

Citation: Yang, L.; Liu, S.; Zhao, Y. Deep-Learning Based Algorithm for Detecting Targets in Infrared Images. *Appl. Sci.* **2022**, *12*, 3322. https://doi.org/10.3390/app12073322

Academic Editor: Giacomo Fiumara

Received: 18 January 2022
Accepted: 23 March 2022
Published: 24 March 2022

1. Introduction

Infrared image target detection identifies and labels each target class from an infrared image containing multiple targets. Infrared images consist of information about the thermal radiation emitted by the target and are not susceptible to environmental influences. Therefore, infrared images have advantages over visible images in low-visibility environments, such as night scenes, haze, rain, snow, and dust. In recent years, IoT technologies such as nighttime intrusion warning systems have cited infrared images based on this advantage [1].

Target detection algorithms, in general, can be divided into two categories: traditional target detection algorithms based on image processing and machine learning and new target detection algorithms based on deep learning. Traditional infrared image target detection algorithms include edge detection, module matching, Hough transform, etc. Some target detection algorithms use edges, contours, and textures for target detection. Dalal et al. proposed using gradient direction histograms to detect HOG features of pedestrians [2]. They divided the image and obtained the directional histogram of the gradient edges of each pixel point in each region. The combined directional histogram was used as a feature representation for each area. Papageorgiou et al. proposed using Haar wavelet features for target detection, calculating the pixel values in adjacent rectangles obtained from the detection window and their differences and then using the differences to classify each region in the image [3]. Wu et al. proposed to detect pedestrians using Edgelet features and obtained high target detection performance [4]. Traditional target detection algorithms extract features manually for images. These features rely on a priori knowledge and have limited expressiveness, limiting the accuracy of target detection algorithms.

In recent years, with the rapid development of deep learning, many deep-learning algorithms have been applied to the field of computer vision. Deep-learning-based target detection algorithms have been proposed one after another. Compared with traditional

target detection algorithms that use manual feature extraction, deep-learning-based target detection algorithms can self-extract features, which do not require a priori knowledge and have the more expressive power of the extracted features. This is more beneficial to improve the performance of target detection models. In 2014, Girshick applied the regional convolutional neural network to target detection and proposed the R-CNN model [5]. This model is an essential milestone in deep-learning-based target detection algorithms. R-CNN first uses the Selective Search method to extract about 2000 candidate regions, then uses CNN to remove features from the stretched candidate regions, and finally uses support vector machine SVM to classify these features and box regression. In 2015, Girshick proposed a faster Fast R-CNN based on R-CNN [6]. Unlike the computational process of R-CNN, Fast R-CNN first convolves the whole image to get the feature map and then combines the two steps of candidate region classification and frame regression for training so that the computation speed is faster.

Neither R-CNN nor Fast R-CNN solves the problem of relying on the selective search algorithm in the candidate region generation phase, which causes a very time-consuming pain, so Ren et al. proposed the Faster R-CNN model. Faster R-CNN introduces a Region Proposal Network (RPN), which extracts candidate regions directly on the feature map output from the convolutional neural network, significantly improving the detection speed of the target detection model. Then, Bell proposed the ION model based on the Faster R-CNN model [7]. This model uses spatial recurrent neural networks to combine contextual features and the output of the features from different convolutional layers and uses them as multiscale features for target detection.

The above studies mainly focus on target detection in visible images. However, deep learning in target detection research of infrared images is not yet common. Inspired by the idea of transfer learning, this paper migrates the target detection algorithm on visible images to the infrared image target detection field. Firstly, we propose a target detection model CMF Net to solve the problem of the existence of target multiscale features. The CMF Net model is based on the VGG16 network (a convolutional neural network) and uses two multiscale feature extraction mechanisms for image feature extraction and fusion. This makes the final feature map input from the backbone network to the classification network contain low-level visual features that facilitate target localization and high-level semantic features that enable target recognition. Secondly, to solve the problem of low detection accuracy of the algorithm when the target is occluded, we propose the CMF-3DLSTM model. The model improves the classification network into a 3D long- and short-term memory network based on the CMF Net model. We use an attention mechanism to assign weights to the contextual features extracted in different dimensions. Finally, target detection features include multiscale features and contextual features to achieve the fusion of spatio-temporal features.

The rest of this paper is organized as follows: Section 2 summarizes the infrared image target detection algorithm-related work. Section 3 introduces the details of the CMF Net model. Section 4 introduces the structure and details of the CMF-3DLSTM model in detail. Section 5 describes the design and results of relevant experiments. Section 6 summarizes the work of this paper.

2. Related Work

2.1. Target Detection Framework Based on Deep Learning

Target detection aims to locate and identify each target instance using a bounding box. Traditional target detection algorithms include edge detection [8], module matching [9], Hough transform [10], etc. These target detection algorithms use edges, contours, and textures for target detection. These features rely on a priori knowledge and have limited expressiveness, limiting the accuracy of target detection algorithms. With the rapid development of deep learning, the field of computer vision has achieved remarkable success in target detection tasks using deep-learning algorithms [11]. Deep-learning-based target detection frameworks have also been proposed one after another [12,13].

Target detection frameworks based on deep learning mainly fall into two categories: two-level detection framework and single-level detection framework [14]. The two-level detection framework includes a pre-processing step for region recommendations. That is, candidate regions are selected and then classified. Such representatives include Faster R-CNN and Mask R-CNN [15], etc. They adopt the R-CNN proposed by Girshick et al. [16] as the target suggestion method [17], which significantly reduces the amount of calculation compared with the traditional method. The conventional method usually uses superpixel, edge [18], and shape to score Windows, containing objects to generate region suggestion boxes. Deepbox [19] used a lightweight ConvNet model for training, which rearranged the regional suggestion boxes generated by the Edge box. Compared with R-CNN, these traditional methods often require a more extensive calculation. The single-stage detection framework adopts the regression method to directly regress the position and type of the target from the feature graph. The grid method or convolution of different scales is used to operate the feature graph to obtain the position and classification information of the target directly. Such representatives include YOLO and SSD. Generally speaking, the two-stage detection frame has higher accuracy, and the single-pole detection frame is faster. In order to improve the performance of CMF-3DLSTM, we use an infrared image target detection model based on multiscale feature fusion and context analysis proposed in this paper, and we adopt the target detection framework of Faster R-CNN as its basic framework.

2.2. Transfer Learning

The main idea of transfer learning is to transfer labeled data or knowledge structures from related domains to accomplish or improve the learning of the target domain or task. One of the main assumptions in traditional machine learning algorithms is that training and test data must be in the same feature space and have the same distribution. Transfer learning relaxes the basic assumption that training and test data may be in different feature spaces or follow other data distribution [20]. Specifically for the target detection task of this paper, the task of the source domain is defined as the target detection based on the sizeable visible dataset ImageNet, and the task of the target domain is defined as the target detection based on the small infrared dataset FLIR. Since both infrared imaging and visible imaging are similar, they collect target information for imaging through optical systems. The migration learning approach can be used to initialize the parameters of the infrared image target detection model with the pre-trained model on the visible image dataset. Eventually, the model can be fine-tuned and trained using the infrared image dataset.

2.3. Cross-Layer Connection Mechanism

The cross-layer connection mechanism is a classical idea of direct routing from the lower to the higher, ignoring the middle layer. The specific details of the cross-layer connection method vary in different models. A cross-layer connection mechanism is proposed in this paper to solve the problem of multiscale feature detection in images. This method implements two multiscale feature extraction mechanisms and feature fusion mechanisms, which can adapt to the multiscale features of the target and improve the target detection performance of the model. The cross-layer connection mechanism used in this paper is closest to the pedestrian target detection method [21]. In contrast, the two multiscale feature extraction mechanisms proposed in this paper use parameter sharing to process the feature images output from the first, third, and fifth convolution layers in different ways. To keep the resolution consistent, we took the resolution of the feature graphs output by the third convolution layer and the fifth convolution layer as the benchmark, adjusted the resolution of the feature graphs output by other convolution layers, and finally realized the cross-layer connection.

3. CMF Net

This section introduces the CMF Net target detection model in detail. Its innovation lies in using multiscale feature extraction mechanisms of parameter sharing to extract multiscale feature information and carry out feature fusion.

3.1. Network Structure of CMF Net

The network structure of CMF Net is shown in Figure 1, which consists of four parts: backbone network, region proposal network RPN, ROI pooling layer, and classification network. The first part is the backbone network, which mainly adopts the migration learning method, two multiscale feature extraction mechanisms, and a feature fusion mechanism. The output feature map contains both low-level visual features and high-level semantic features. The second part is the region proposal network RPN [22], which is mainly used to extract the region proposal frames containing the target for the feature map output from the backbone network and filter out about 300 high-scoring regions proposal frames. The third part is the ROI pooling layer, mainly used to map ROI regions to convolutional regions and pool them into feature maps of fixed size. The fourth part is the classification network, which is used mainly for target location correction and classification recognition after mapping the ROI pooling layer and achieving target detection.

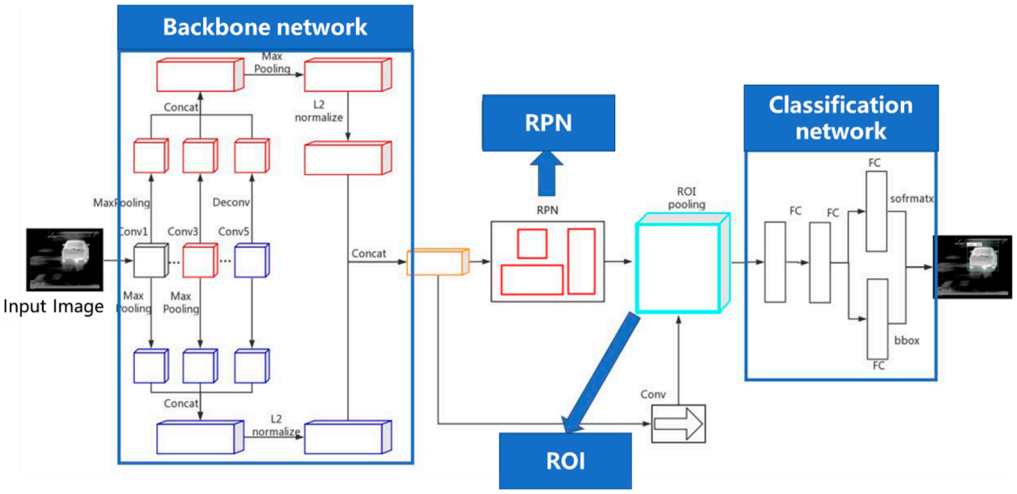

Figure 1. CMF Net architecture.

3.2. Multiscale Feature Extraction I

In the vgg16 network, the visual features extracted from the lower convolutional layer play an important role in the target location, while the semantic features extracted from the higher convolutional layer play an important role in target recognition. The multiscale feature fusion method can retain the low-level visual and high-level semantic features and avoid extracting redundant features from the two adjacent convolutional layers. The first multiscale feature fusion method is shown in Figure 2. Considering that the intermediate convolutional layer contains visual and semantic features, which combine the two, it is essential for target detection. Therefore, the feature map output by the third convolutional layer is retained completely, and the resolution of the feature map is taken as the benchmark. The feature map extracted from the first convolutional layer is divided into two pools, and the feature map extracted from the fifth convolutional layer is deconvoluted [23]. This can further study the feature map output of the first layer and the fifth volume layer, resolve the feature maps of the low-, middle-, and high-volume outputs that are adjusted to the same level, and, finally, connect them to achieve feature fusion.

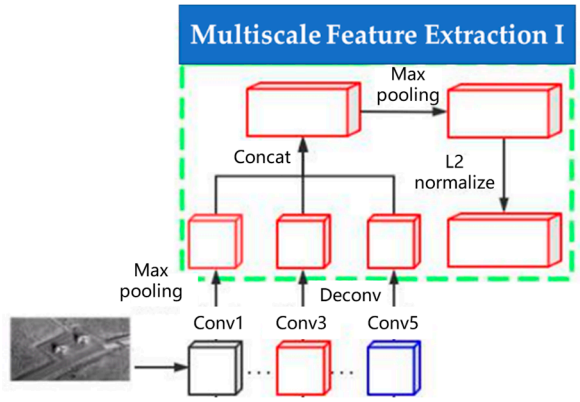

Figure 2. Multiscale feature extraction I.

3.3. Multiscale Feature Extraction II

In the first multiscale feature extraction mechanism, the feature graph output from the third convolution layer is retained completely. Then, based on the resolution of the feature images output from the third layer, the feature images output from the first convolution layer are pooled, and the feature images output from the fifth convolution layer is deconvolution processed. This makes the feature images output by the first, third, and fifth convolution layers maintain the exact resolution. However, this processing method loses the high-level semantic features learned by the fifth convolution layer to some extent, which affects the accuracy of the target detection model. Therefore, a second multiscale feature extraction method is proposed, whose structure is shown in Figure 3. The feature map output from the fifth convolutional layer is retained completely, and the resolution of the feature map is taken as the benchmark. The feature map extracted from the first convolutional layer is pooled twice. The feature map extracted from the third convolutional layer is pooled once. The resolution of the feature map output from the first, third, and fifth layers is adjusted to be the same and connected to achieve feature fusion.

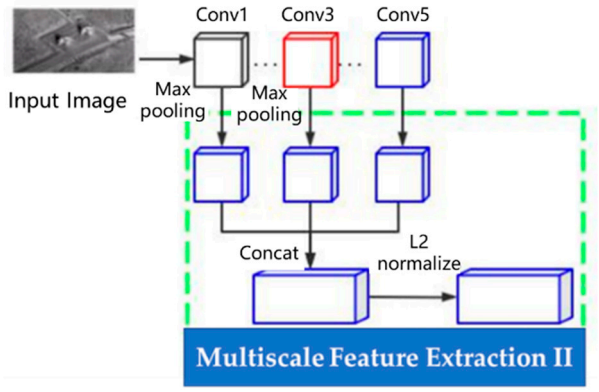

Figure 3. Multiscale feature extraction II.

3.4. Feature Fusion Strategy

Two multiscale feature extraction methods use different strategies to extract multiscale feature maps. The final output of the feature map has its unique advantages. The first multiscale feature extraction method ultimately retains the mixed features of target location and target recognition extracted from the middle convolutional layer. Still, it loses some high-level semantic features and affects target recognition. The second multiscale feature extraction method preserves the semantic features of a high-level convolutional layer but loses some mixed features, which affects the target location. Therefore, it is necessary to fuse the two kinds of feature maps so that the final output feature map simultaneously contains rich mixed features and semantic features.

The final feature maps obtained by the multiscale feature extraction mechanism suffer from inconsistent resolution and inconsistent amplitude of feature values. Therefore, these feature maps cannot be directly fused with features. The feature fusion method proposed in this paper makes the feature maps output by the first multiscale feature extraction mechanism consistent with the resolution of the feature maps output by the second multiscale feature extraction mechanism through a pooling process. Secondly, L2 normalization is performed in the feature maps outputted by the two feature extraction methods, so that the amplitudes of feature values in the two feature maps are consistent. Then, we connect two feature maps to get a feature map containing rich visual features, mixed features, and semantic features. Finally, we input it to RPN to extract ROI information.

We assign a binary class label to each box (including the target or excluding the target). We set a binary class label (include target or not) to each box. We assign a positive title to a box with an IoU threshold higher than 0.7 with any ground truth box and then assign a negative label to a box with an IoU threshold lower than 0.3 with all ground truth boxes. Our goal is to minimize a multitask loss function.

$$L(k, k^*, t, t^*) = L_{cls}(k, k^*) + \lambda L_{reg}(t, t^*) \tag{1}$$

L_{cls} is the classification loss, L_{reg} is the coordinate regression loss of the box with a positive label assigned. k^* and k are true to label and predicted labels separately, respectively. $L_{reg}(t, t^*) = R(t - t^*)$ where R is the smoothed loss function defined in Faster R-CNN. We express the coordinates of the positive box as $t = (t_x, t_y, t_w, t_h)$ and the coordinates of the predicted box as $t^* = \left(t_x^*, t_y^*, t_w^*, t_h^*\right)$.

$$\begin{array}{c} t_x = (G_x - P_x)/P_w \cdot t_y = \left(G_y - P_y\right)/P_h \\ t_w = \log(G_w - P_w) \cdot t_h = \log(G_h - P_h) \end{array} \tag{2}$$

where $P^i = \left(P_x, P_y, P_w, P_h\right)$ specifies the coordinates of the center point of the predicted box. G^i specifies the coordinates of the center point of the positive box.

The RPN module first resamples the unbalanced sample set of positive and negative samples, using the oversampling method in random sampling to obtain more sample first data balance by randomly repeating examples from a small number of class sample sets. Then, the gradient descent method is used for training, and the classification loss error L_{cls} and regression loss error L_{reg} are back-propagated to update the model parameters until the RPN module converges. The parameters for the training of the RPN module are set as shown in Table 1.

Table 1. RPN training parameters list.

Description	Value
The Anchor scale	32, 64, 128
MiniBatch Quantity	256
PRN foreground–background ratio	1:1
IOU threshold used by NMS for RPN training	0.3
IOU thresholds used by NMS for RPN prediction	0.7

4. CMF-3DLSTM

The infrared image target detection model CMF Net based on multiscale feature fusion proposed in Section 3 adopts two multiscale feature extraction mechanisms and feature fusion methods for the final output feature graph of the backbone network. It adapts to the multiscale characteristics of the target. It can be regarded as the feature fusion of spatial dimension, which is of great help to improve the performance of infrared image target detection. CMF Net can achieve better target detection performance when the background environment is the relatively simple spacing between targets. However, the problem with

CMF Net is that it is easy to misjudge when multiple targets are close together, overlapping, and semantically confusing, as shown in Figure 4.

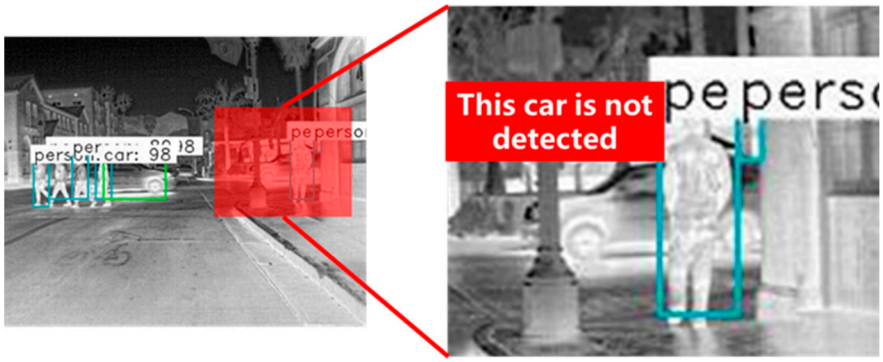

Figure 4. CMF Net target detection results.

This section proposes CMF-3DLSTM, an infrared image target detection model based on spatio-temporal feature fusion and attention mechanism. This model first inherits the multiscale feature fusion strategy of CMF Net to achieve feature fusion in the spatial dimension. Then, the model is based on 3DLSTM, which extracts contextual information along with the positive and negative directions of each dimension from the length, width, and height dimensions of the 3D feature map. Meanwhile, the model uses an attention mechanism to assign weights to the contextual features extracted in various dimensions and directions. CMF-3DLSTM effectively improves target detection performance in complex situations such as multiple targets approaching each other, overlapping each other, and semantic confusion.

4.1. Network Structure of CMF-3DLSTM

CMF-3DLSTM target detection model includes four modules, namely trunk network, regional proposal network, ROI pooling layer, and classification network, as shown in Figure 5.

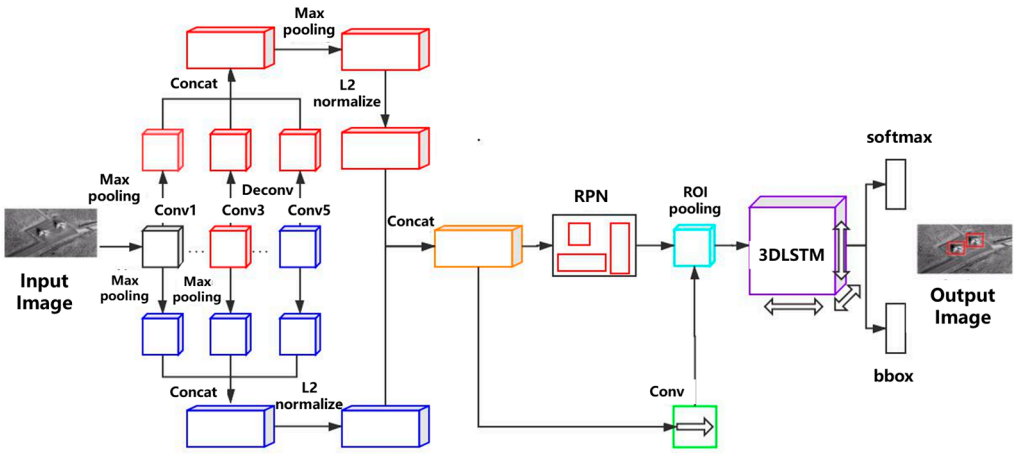

Figure 5. Structure diagram of CMF-3DLSTM model.

The trunk network still adopts a 16-layer VGG16 network. The model file saved after pre-training the target detection model Faster R-CNN on the visible light domain dataset Image Net is used for parameter initialization in CMF-3DLSTM utilizing the idea of transfer learning, and two multiscale feature extraction mechanisms are still used for multiscale feature extraction and feature fusion. The process of generating candidate boxes for regional proposal network RPN is unchanged. It still extracts and screens regional proposal boxes that may contain targets from the feature graph output by the trunk network, and about

300 high-scoring regional proposal boxes are screened. The ROI pooling layer mainly generates a fixed-size feature map based on region proposal mapping of candidate box generated by RPN network recommendation for subsequent classification and regression. The classification network primarily uses the multiscale feature map processed by ROI pooling layer mapping and uses the 3D six-way long- and short-term memory network 3DLSTM, constructed based on BI-LSTM, to extract context information. At the same time, the attention mechanism is used to assign different weights to the context features extracted from other dimensions and directions to achieve the fusion of spatio-temporal features. Finally, the input is given to the classification and location layers for target classification recognition and position correction.

4.2. Context Information Extraction Network

Figure 6 shows a 3D long- and short-term memory network, which can extract context information. The 3DLSTM network firstly transforms the 3D feature image into the 2D feature image. Then, each row in the two-dimensional feature graph is regarded as a vector or a sequence, and each column in the two-dimensional feature graph is considered to be a time step. Finally, the context information of feature map extraction is transformed into the extraction of vector or sequence relations.

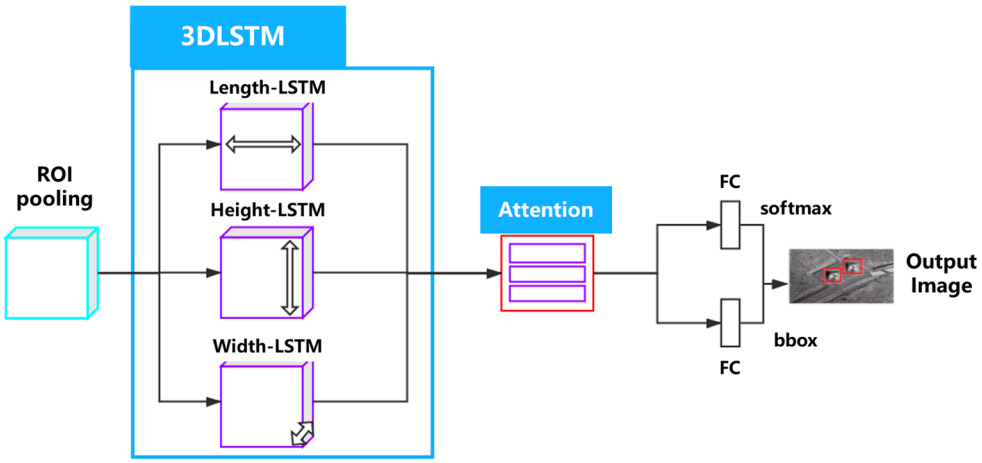

Figure 6. 3DLSTM network structure diagram.

We fix the feature map's length, width, and height separately and stretch the other two directions. In this way, the shape of the 3D feature map can be transformed into a 2D feature map. This two-bit information is then input into Bi-LSTM to extract contextual information along the fixed direction of the original feature map. Taking the length direction of the fixed feature map as an example, the 3D feature map becomes a 2D feature map (length, width $\times$ height) after transformation, and the specific transformation process is shown in Figure 7.

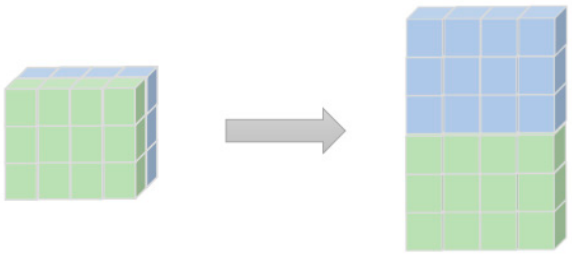

Figure 7. 3D feature-length expansion diagram.

Finally, the features generated are connected. At this time, the feature graph input to the classification regression network contains the context information extracted from the

length, width, and height of the original feature graph, so the network is called 3DLSTM. Compared with RNN, LSTM, and BI-LSTM, which extract 2D context information, this network is more conducive to improving the performance of the target detection model.

4.3. Attentional Mechanism

Attention mechanisms are generally used in natural language processing. As the length of text sequence increases in practical applications, the more advanced information in the series is lost more seriously, leading to a significant decline in model performance. A common solution is to input text sequences in both sequential and reverse order or LSTM. Although the two methods can improve the model performance to a certain extent, it is still difficult to effectively solve the problem of a too-long sequence.

The final output size of the feature graph of 3DLSTM is 3256, which adopts bi-LSTM to output vectors with the length of 256 from the three dimensions of length, width, and height of the 3D feature graph, respectively. Each element in the vector represents the neuron's output under a time step. Such a feature map size has the problem of a too-long sequence. Moreover, the weights of the three vectors are different, and the weight of each element in each vector should also be different. Therefore, an attention mechanism is adopted that can selectively screen out a small amount of important information from a large amount of information and focus on this vital information.

A source in attention consists of a series of key–value pairs. The weight coefficient of each key corresponding to value is obtained by calculating the similarity of each key in input vector query and source. Formula (3) is the calculation of the similarity between the query and the key. Formula (4) determines the weight coefficient of each key corresponding to value by the $Soft_{max}$ function.

$$Sim_i(Query, Key_i) = \frac{Query \cdot Key_i}{\|Query\| \cdot \|Key_i\|} \tag{3}$$

$$a_i = Soft_{max}(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{N} e^{Sim_j}} \tag{4}$$

The weighted sum of values obtains the final attention value according to these weight coefficients. The calculation formula of attention value is as follows:

$$Attention(Query, Source) = \sum_{i=1}^{N} a_i \cdot Value_i \tag{5}$$

Using the attention mechanism, you can assign different weights to the context information collected by 3DLSTM in different directions and to the different elements in the context information in each direction. This enables the target detection model CMF-3DLSTM to pay more attention to the salient features of the target, thus improving the target detection performance of CMF-3DLSTM.

4.4. Model Training Strategy

CMF-3DLSTM uses the same training methods of pre-training migration and model fine-tuning as CMF Net, but the training strategy of the classification module is different. In the classification module, 3DLSTM is based on three bidirectional long- and short-term memory networks, while the network structure of Bi-LSTM is based on LSTM. Therefore, the ultimate goal of 3DLSTM is the same as that of LSTM, which is to minimize a loss function $L(t)$. The specific calculation formula is as follows:

$$L = \sum_{t=1}^{T} l(t) \tag{6}$$

where t represents the current moment, T represents the total time step, and $l(t)$ represents the loss function at the current moment. The calculation formula of $l(t)$ is as follows:

$$l(t) = f(h(t), y(t)) = \|h(t) - y(t)\|^2 \tag{7}$$

$h(t)$ represents the hidden layer output at the current time t, and $y(t)$ represents the output layer output at the present time t. To minimize $l(t)$ loss function, the 3DLSTM network is trained by the gradient descent method. When the error is propagated back, the chain derivative method is used to update the model weight parameters. The specific calculation formula is as follows:

$$\frac{\partial L}{\partial w} = \sum_{t=1}^{T} \sum_{i=1}^{M} \frac{\partial L}{\partial h_i(t)} \cdot \frac{\partial h_i(t)}{\partial w} \tag{8}$$

i represents the memory unit of the hidden layer. M is the number of memory units. w represents the model weight parameter. $h_i(t)$ represents the output of the memory unit in the hidden layer at the current time t. After calculating the gradient of weight parameter w of all models, 3DLSTM uses the gradient descent method to update the parameters iteratively. Finally, it minimizes the loss function L to achieve the purpose of training the classification module.

For infrared image target detection model CMF-3DLSTM, a joint optimization 10-step training process is designed in this paper, as shown in Algorithm 1.

Algorithm 1: CMF-3DLSTM training process

Input: Infrared image dataset.
Output: Target detection model CMF-3DLSTM.
Step 1: Initialize the network parameters in Step2 and Step3 using the pre-training model on the VOC2007 dataset.
Step 2: Use the first multiscale feature extraction mechanism to extract feature information.
Step 3: Using the second multiscale feature extraction mechanism to extract feature information.
Step 4: CMF Net is used to carry out feature fusion for the feature information extracted by Step 2 and Step 3.
Step 5: Train the RPN network to generate the proposals using the characteristic information obtained from Step 4.
Step 6: Implement ROI Pooling of Step 5 and adjust them to the same size.
Step 7: The 3DLSTM network is used to extract the context information of ROI in Step 6.
Step 8: The attention mechanism is used to assign weight to the output of the features by Step 7.
Step 9: Classification layer and regression layer are used for target detection for the output of the features by Ste p8.
Step10: The unified network of Step 5 and Step 9 joint training is taken as the final model.

The training parameters of CMF-3DLSTM, CMF Net, and Faster R-CNN are shown in Table 2, in which Faster R-CNN has 136,708,989 training parameters, and CMF Net has 152,048,765 training parameters. The number of training parameters of CMF-3DLSTMA is 40,761,469, which drops to the level of 10 million and dramatically reduces the space complexity of the algorithm. However, the 3DLSTM network and attention mechanism introduced in the classification module of CMF-3DLSTM is more complex than the fully connected layer in the classification module of CMF Net, thus causing an increase in time complexity. The CMF Net model performs target detection at a speed of about 0.87 s/pc on a machine with a graphics card configuration of GeForce GTX 1080. In the same experimental environment, the Faster R-CNN performs target detection at about 0.75 s/pc, and the CMF-3DLSTMA has a reduced target detection speed of about three s/pc.

Table 2. Experimental environment.

Model	Backbone Network	RPN	ROI Pooling	Classification Network	Total
Faster R-CNN	14,714,688	2,382,893	0	11,961,1408	136,708,989
CMF Net	17,077,824	7,691,309	0	127,279,632	152,048,765
CMF-3DLSTM	17,077,824	7,691,309	0	15,992,336	40,761,469

5. Experiment and Analysis of Experimental Results

5.1. Description of Dataset

We have trained and evaluated our model on FLIR, a public infrared driving image dataset, and achieved excellent results. The introduction of the dataset is shown in Table 3.

Table 3. Dataset specifications.

Content	Synced annotated thermal imagery and non-annotated RGB imagery for reference. Camera centerlines approximately 2 inches apart and collimated to minimize parallax
Images	>10 K from short video segments and random image samples.
Image Capture Refresh Rate	Recorded at 30Hz. Dataset sequences sampled at 2 frames/s or 1 frame/s. Video annotations were performed at 30 frames/s recording.
Frame Annotation Label Totals	10,228 total frames and 9214 frames with bounding boxes. 1. Person (28,151); 2. Car (46,692); 3. Bicycle (4457); 4. Dog (240); 5. Other vehicle (2228).
Driving Conditions	Day (60%) and night (40%) driving on Santa Barbara, CA area streets and highways from November to May with clear to overcast weather.
Dataset File Format	1. Thermal—14-bit TIFF (no AGC); 2. Thermal—8-bit JPEG (AGC applied) w/o bounding boxes embedded in images; 3. Thermal—8-bit JPEG (AGC applied) with bounding boxes embedded in images for viewing purposes; 4. RGB—8-bit JPEG; 5. Annotations: JSON (MSCOCO format).

5.2. Description of Evaluation

In this paper, the target detection performance of the method on the FLIR infrared driving image dataset is evaluated from mAP (mean average precision), which is widely used as a standard measure in previous target detection research. The calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision\left(R_{jk}\right) \tag{10}$$

where TP indicates that the prediction is true and the label is true, FP indicates that the prediction is true and the label is false. Q is the set of target categories to be detected, m_j is the number of pictures of all categories corresponding to Q_j, R_{jk} is the set of all pictures in the returned result until picture k is found. That is to say, the corresponding precision is calculated in this set.

5.3. Experimental Analysis

The training set of Infrared image dataset FLIR contains 7860 IR images, and the test set has 1360 Infrared images. To facilitate the experiment, we converted the image

annotation files of the Infrared image from JSON format (MSCOCO format) to XML format (VOC2007 format).

We conducted a total of four experiments. The first set of experiments tested the performance of the Faster R-CNN target detection model under various network layer combinations. The second group of experiments analyzed the performance comparison between CMF Net (Faster R-CNN model that adopts two multiscale feature extraction mechanisms and carries out feature fusion) and those using two multiscale feature extraction mechanisms alone. The third group of experiments analyzed the performance comparison between CMF Net and other target detection models. The fourth group of experiments analyzed the performance comparison between CMF-3DLSTM (using 3DLSTM network to replace the full connection layer in CMF Net) and other target detection models.

(1) *Experiment I*: The performance of the target detection model depends mainly on whether the feature map contains rich features or not. To investigate which network layers and network layer combinations can make the model the best performance, we conduct seven sets of tests based on the Faster R-CNN target detection model. The final target detection performance of the feature maps output by convolutional layer 1 (single 1), convolutional layer 3 (single 2), and convolutional layer 5 (single 3) are first tested separately. Then, the target detection is performed for the feature maps output by the convolutional layer combination 1+2+3 (Group 1) and 3+4+5 (Group 2), respectively. Finally, the target detection is performed for the feature maps output by the convolutional layer combination 1+3+5 with two different multiscale feature extraction mechanisms (Group 3 and Group 4). The experimental results are shown in Table 4.

Table 4. Results of combining different convolutional layers.

Layers	Single 1	Single 2	Single 3	Group 1	Group 2	Group 3	Group 4
mAP	0.514	0.605	0.583	0.567	0.618	**0.636**	**0.661**

Experimental results show that the target detection model of convolution layer combination 1+3+5 with two different multiscale feature extraction mechanisms (Group 3 and Group 4) has better detection performance on FLIR. Therefore, the proposed infrared image target detection model uses a 1+3+5 convolution layer combination.

(2) *Experiment II*: To verify the performance of the two multiscale feature extraction mechanisms and CMF Net, we carried out three experiments, and the experimental results are shown in Figure 8. We found that our target detection model CMF Net has a great improvement.

The mAP of CMF Net improved about 6.8% and 4.4% compared to the first multiscale feature extraction mechanism and the second multiscale feature extraction mechanism, respectively. Although the target detection accuracy of CMF Net decreased in the bicycle category, it improved by 6.1% and 24% in the car and person categories, respectively, compared to the first multiscale feature extraction mechanism, and 13.2% and 9.3%, respectively, compared to the second multiscale feature extraction mechanism, CMF Net's accuracy only in the bicycle target. The accuracy of CMF Net is reduced by 9.6% compared to the first multiscale feature extraction mechanism and 9.3% compared to the second multiscale feature extraction mechanism. This experimental result illustrates the importance of feature fusion based on two multiscale feature extraction mechanisms compared to one multiscale feature extraction mechanism alone to improve the performance of the target detection model.

(3) *Experiment III*: To fully prove the correctness of our multiscale feature extraction strategy, we still adopted the idea of transfer learning to migrate the pre-training networks of Faster R-CNN, YOLO, and SSD, which are currently popular in the

visible light domain, to FLIR infrared driving image dataset, continue training until the model converges.

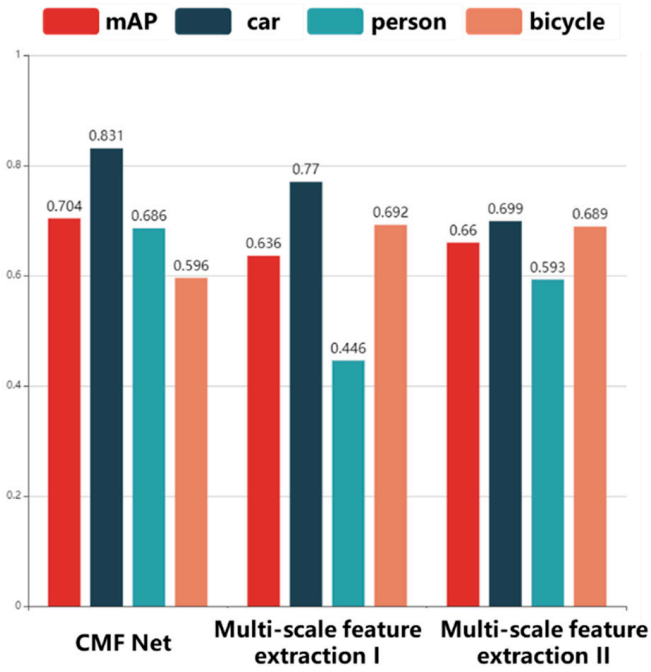

Figure 8. Performance comparison between CMF Net and two multiscale feature extraction mechanisms.

The experimental results are shown in Figure 9 We evaluate the performance of CMF Net on the test set of FLIR infrared driving image dataset. Using the above methods, we get about 71% of mAP by CMF Net, 58% by Faster R-CNN, 65% by YOLO, and 54% by SSD.

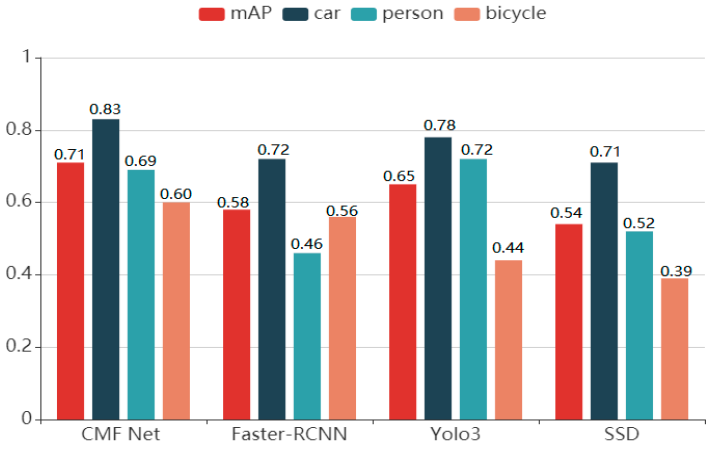

Figure 9. Performance comparison between CMF Net and other target detection models.

Compared with other common target detection model's mAP, our multiscale feature fusion model CMF Net achieved significant improvement in accuracy, about 13 percentage points higher than Faster R-CNN's mAP, about 6 percentage points higher than YOLO's mAP, and about 17 percentage points higher than SSD's mAP.

In FLIR infrared driving image dataset, due to the different shooting distances, the size of the car, person, and bicycle targets in the infrared image is different, which has significant multiscale characteristics. Our model adopts two multiscale feature extraction mechanisms and two-level feature fusion methods, which makes the final output of the backbone network contains rich visual features and semantic features, so the detection

accuracy of car, person, and bicycle is far higher than the other three networks. The accuracy of CMF Net on car target is 11%, 5%, and 12% higher than Faster R-CNN, YOLO, and SSD, respectively; the accuracy on person target is 23% and 17% higher than Faster R-CNN and SSD, respectively, and the accuracy on bicycle target is 4%, 16%, and 21% higher than Faster R-CNN, YOLO, and SSD respectively. Compared with YOLO, the accuracy of the personal target is reduced by 3%. The importance of the combination of two multiscale feature extraction mechanisms and two-level feature fusion methods is fully proved.

As shown in Figure 10, the target detection result of CMF Net is on the left, and the target detection result of Faster R-CNN is on the right. The scenes on the left and right are the same, with vehicles and pedestrians appearing on the street at different scales. Faster R-CNN detected most targets in the image well but failed to detect pedestrians appearing at a small scale in the middle of the image. CMF Net can adapt to the multiscale characteristics of the target because it adopts two multiscale feature extraction mechanisms. Therefore, the pedestrians on a small scale can be successfully identified and positioned correctly.

Figure 10. Comparison of target detection results between CMF Net and Faster R-CNN.

(4) *Experiment IV*: CMF Net, a target detection model based on multiscale feature fusion, has a problem: it is easy to cause misjudgment in the complex situation of multi-target detection. In particular, it is challenging to detect CMF Net effectively when multiple targets are close to or even overlapping each other. It is imperative to use the contextual information around the target effectively. This paper proposes an infrared image target detection model CMF-3DLSTM based on multiscale feature fusion and context analysis. CMF-3DLSTM is inherited from CMF Net. The difference between CMF-3DLSTM and CMF Net is that it replaces the complete connection layer of the classification regression network with a 3D long- and short-term memory network. Context information can be extracted based on multiscale feature fusion. CMF-3DLSTM improved target detection performance by about 2.9% on the infrared image dataset FLIR compared to CMF Net's mAP.

The experimental results are shown in Figure 11. The target detection model CMF-3DLSTM and other target detection models are evaluated on the test set of FLIR. Using the above methods, CMF-3DLSTM obtained about 73.3% mAP, while the mAP on CMF Net was about 70.4%, the mAP on Faster R-CNN was about 68.7%, the mAP on YOLO3 was approximately 64.8%, and the mAP on SSD was about 60.8%.

Although the CMF-3DLSTM model does not achieve optimal detection results for car, person, and bicycle alone, the average detection accuracy is more important in complex situations where multiple targets are nearby or even overlap or obscure each other. The target detection results of CMF-3DLSTM on the infrared image dataset FLIR are illustrated in Figure 12.

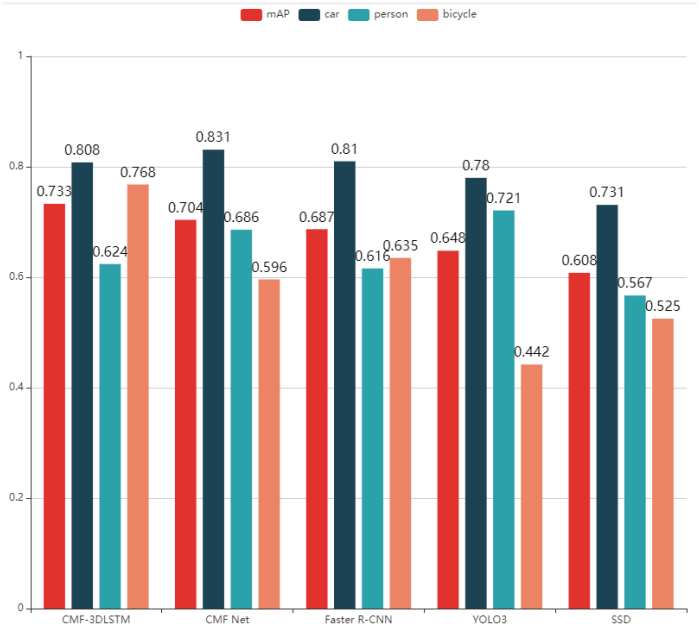

Figure 11. Performance comparison between CMF-3DLSTM and other target detection models.

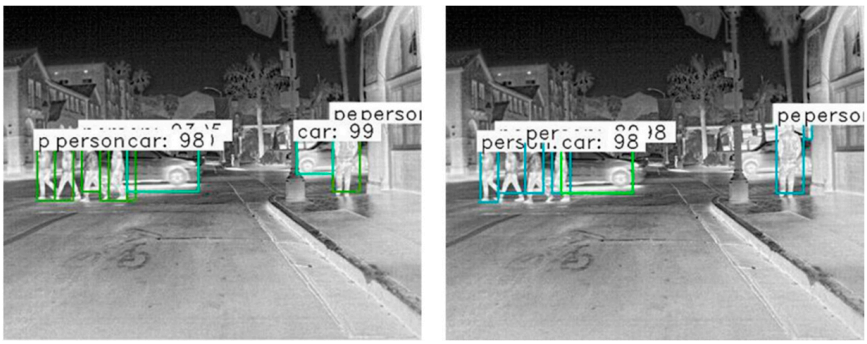

Figure 12. Comparison of target detection results between CMF-3DLSTM and CMF Net.

6. Conclusions

This paper transfers the target detection model in the visible light domain to the infrared environment by transfer learning. In our work, we creatively proposed the method of feature fusion and multiscale feature extraction with two shared features. We obtained the network architecture of CMF Net that makes full use of multiscale feature fusion information for target detection. Through multiscale feature fusion, rich visual and semantic features can be obtained to improve the accuracy of target detection and adapt to the multiscale characteristics of the target to be detected. To improve the target detection performance of the model in complex scenes such as mutual occlusion and overlapping of multiple targets, we constructed a 3D long- and short-term memory network based on CMF Net to extract context information and finally realized the CMF-3DLSTM model. Compared with Faster R-CNN, YOLO3, and SSD, CMF-3DLSTM achieves higher target detection performance on infrared image dataset FLIR. This proves the importance of constructing an infrared image target detection model based on multiscale feature fusion and context analysis.

However, we still need to make further improvements work. The target detection model proposed in this paper improves the accuracy, reduces the number of parameters, and decreases the spatial complexity compared with models such as Faster-RCNN, but it increases the time complexity, making it challenging to meet the requirements of real-time detection. We need to optimize the network structure further to improve the model's

real-time detection capability. At the same time, we need to train and evaluate the model on more infrared image datasets with different scenes to meet the requirements of other scenes in practical applications.

Author Contributions: Data curation, L.Y.; Investigation, S.L.; Methodology, L.Y. and Y.Z.; Resources, S.L.; Supervision, S.L.; Validation, Y.Z.; Writing—review & editing, L.Y., S.L. and Y.Z. 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.

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

References

1. Tan, P.; Mao, K.; Zhou, S. Image Target Detection Algorithm of Smart City Management Cases. *IEEE Access* **2020**, *8*, 163357–163364. [CrossRef]
2. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 886–893. [CrossRef]
3. Papageorgiou, C.; Poggio, T. A Trainable System for Object Detection. *Int. J. Comput. Vis.* **2000**, *38*, 15–33. [CrossRef]
4. Wu, B.; Nevatia, R. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'2005), Beijing, China, 17–21 October 2005; Volume 1. [CrossRef]
5. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
6. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015.
7. Bell, S.; Lawrence Zitnick, C.; Bala, K.; Girshick, R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2874–2883.
8. Enkelmann, W.; Struck, G.; Geisler, J. ROMA—A system for model-based analysis of road markings. In Proceedings of the Intelligent Vehicles '95. Symposium, Detroit, MI, USA, 25–28 September 1995. [CrossRef]
9. Otsuka, Y.; Muramatsu, S.; Takenaga, H.; Kobayashi, Y.; Monj, T. Multitype lane markers recognition using local edge direction. In Proceedings of the Intelligent Vehicle Symposium, Versaille, France, 17–21 June 2002. [CrossRef]
10. Kluge, K.; Lakshmanan, S. A deformable-template approach to lane detection. In Proceedings of the Intelligent Vehicles '95. Symposium, Detroit, MI, USA, 25–26 September 2002. [CrossRef]
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2012**, *60*, 84–90. [CrossRef]
12. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
13. Zhang, Y.; Sohn, K.; Villegas, R.; Pan, G.; Lee, H. Improving object detection with deep convolutional networks via Bayesian optimization and structured prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 249–258. [CrossRef]
14. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comp. Vis.* **2020**, *128*, 261–318. [CrossRef]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
16. Cheng, M.M.; Zhang, Z.; Lin, W.Y.; Torr, P. BING: Binarized normed gradients for objectness estimation at 300 fps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
17. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef] [PubMed]
18. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.
19. Kuo, W.; Hariharan, B.; Malik, J. Deepbox: Learning objectness with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015.
20. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *arXiv* **2014**, arXiv:1411.1792.
21. Sermanet, P.; Kavukcuoglu, K.; Chintala, S.; LeCun, Y. Pedestrian detection with unsupervised multi-stage feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.

22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
23. Wang, C.; Shi, J.; Yang, X.; Zhou, Y.; Wei, S.; Li, L.; Zhang, X. Geospatial object detection via deconvolutional region proposal network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2019**, *12*, 3014–3027. [CrossRef]