remote sensing

MDPI

Article

Visible-Thermal Image Object Detection via the Combination of Illumination Conditions and Temperature Information

Hang Zhou, Min Sun *, Xiang Ren and Xiuyuan Wang

Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China; hang.zhou@pku.edu.cn (H.Z.); ren.xiang@pku.edu.cn (X.R.); adhamwang@pku.edu.cn (X.W.)
* Correspondence: sunmin@pku.edu.cn

Abstract: Object detection plays an important role in autonomous driving, disaster rescue, robot navigation, intelligent video surveillance, and many other fields. Nonetheless, visible images are poor under weak illumination conditions, and thermal infrared images are noisy and have low resolution. Consequently, neither of these two data sources yields satisfactory results when used alone. While some scholars have combined visible and thermal images for object detection, most did not consider the illumination conditions and the different contributions of diverse data sources to the results. In addition, few studies have made use of the temperature characteristics of thermal images. Therefore, in the present study, visible and thermal images are utilized as the dataset, and RetinaNet is used as the baseline to fuse features from different data sources for object detection. Moreover, a dynamic weight fusion method, which is based on channel attention according to different illumination conditions, is used in the fusion component, and the channel attention and a priori temperature mask (CAPTM) module is proposed; the CAPTM can be applied to a deep learning network as a priori knowledge and maximizes the advantage of temperature information from thermal images. The main innovations of the present research include the following: (1) the consideration of different illumination conditions and the use of different fusion parameters for different conditions in the feature fusion of visible and thermal images; (2) the dynamic fusion of different data sources in the feature fusion of visible and thermal images; (3) the use of temperature information as a priori knowledge (CAPTM) in feature extraction. To a certain extent, the proposed methods improve the accuracy of object detection at night or under other weak illumination conditions and with a single data source. Compared with the state-of-the-art (SOTA) method, the proposed method is found to achieve superior detection accuracy with an overall mean average precision (mAP) improvement of 0.69%, including an AP improvement of 2.55% for the detection of the Person category. The results demonstrate the effectiveness of the research methods for object detection, especially temperature information-rich object detection.

Keywords: object detection; multi-spectral fusion; visible and thermal images; RetinaNet; illumination conditions; dynamic weight fusion; temperature information; a priori knowledge

Citation: Zhou, H.; Sun, M.; Ren, X.; Wang, X. Visible-Thermal Image Object Detection via the Combination of Illumination Conditions and Temperature Information. *Remote Sens.* **2021**, *13*, 3656. https://doi.org/10.3390/rs13183656

Academic Editor: Pedro Melo-Pinto

Received: 13 August 2021
Accepted: 9 September 2021
Published: 13 September 2021

1. Introduction

Object detection is a popular direction in computer vision and digital image processing, and is widely used in many fields such as autonomous driving, disaster rescue, robot navigation, intelligent video surveillance, etc. Object detection is also a fundamental type of algorithm in the field of pan-identity recognition, and plays a crucial role in subsequent tasks such as face recognition, gait recognition, crowd size assessment, and instance segmentation.

From the comprehensive perspective of international-related research, most object detection studies are based on visible images. However, for images taken at nighttime or those with insufficient illumination conditions, it is difficult for visible light-based object detection methods to achieve the expected results. While thermal imaging can significantly

compensate for the shortcomings of visible imaging, thermal images also have some defects such as high noise, low contrast, non-uniformity, and poor spatial resolution [1], so the results are not ideal when applied to target detection alone. Moreover, because individual objects are represented differently in thermal infrared images, the temperature information contained in thermal infrared images can be applied as a priori knowledge for target detection.

Therefore, in the present study, the channel attention method is applied to the dynamic weighted fusion of multiple data sources. Furthermore, the temperature information of thermal infrared images is also applied to a deep learning model to improve the detection accuracy.

2. Related Works

In early research, Choi et al. [2] used low-level image fusion to fuse the best features of two types of sensor to achieve superior performance in human detection. They proposed a new joint bilateral filter that is useful for the fusion of the edge information in a visible image and the white region in a thermal image. While the results of experiments demonstrated the effective and easy human detection by the proposed algorithm, its detection accuracy was found to be relatively low.

In recent years, the development of deep learning, models that fuse visible and thermal images have become a popular research direction, which has led to significant improvements in the accuracy of object detection. According to the location of fusion in the model, related studies can be divided into those conducted on data-level, feature-level, and decision-level fusion [3], among which the most studied methods are feature-level and decision-level fusion. Wagner et al. [4] was the first to conduct research on convolutional neural network (CNN)-based multispectral pedestrian detection, and he evaluated the results of feature-level-based early and late fusion, which showed that late fusion achieved superior accuracy. Chen et al. [5] proposed a feature-level-based novel multi-layer fused CNN (MLF-CNN) for the detection of pedestrians under adverse illumination conditions, and they used the sum fusion method for the integration of the two convolutional layers. The detection accuracy of the proposed method was found to be 28.62% better than the baseline and 11.35% better than the well-known faster R-CNN halfway fusion [6] in the KAIST multispectral pedestrian dataset.

These priori studies were all based on feature-level fusion without the consideration of the illumination conditions. In fact, the contribution proportions of visible and thermal images under different illumination conditions should be different; visible light contributes more under bright conditions, while thermal infrared light contributes more under dark conditions.

To consider the ratio of the contributions of different source images to the detection results, Li et al. [7] designed an illumination-aware Faster R-CNN-based [8] structure; they adaptively merged color and thermal sub-networks via a gate function defined over the illumination value. Furthermore, Guan et al. [9] proposed a novel illumination-aware weighting mechanism that incorporates illumination information into two-stream deep CNNs to learn multispectral human-related features under different illumination conditions (daytime and nighttime) to accurately depict the illumination condition of a scene. These studied both improved detection accuracy of pedestrians, but the fusion strategies are decision-level, i.e., the detection results are obtained separately using two branched networks, and the results are then weighted and fused using the coefficients determined by the illumination conditions. However, both methods demand more network parameters, higher computational intensity, and higher platform requirements; in addition, the determination of the image illumination conditions is not specific.

The methods proposed in these studies all deal with thermal images in the same way as visible images, which means that only the semantic information of the visual layer is utilized, whereas the temperature information associated with the objects is not. In reality, the temperature of pedestrians on the road is distinct from that of the background; thus, in

theory, the a priori knowledge of temperature can be fully exploited to further improve the object detection accuracy.

Zhao et al. [10] constructed an image-temperature transformation formula based on infrared image formation theory, converted infrared images into corresponding temperature maps, and trained a temperature network for detection. However, the method first uses a deep learning network to obtain the pedestrian frame with the highest confidence, after which the original image is converted to the temperature map based on the pedestrian frame, and the converted temperature map is finally fed into the model for relearning. Thus, via this method, the model trained twice, making the process repetitive and time-consuming.

To this end, the present study explores a feature-level-based fusion method. First, the spatial attention method is used to dynamically learn different channel weights at the feature level to improve the rationality of the fusion process. Then, different fusion parameters are selected according to different illumination scenarios. Furthermore, the a priori knowledge of temperature is applied to the preliminary feature extraction, instead of spatial attention, to achieve guided a priori judgment, thereby achieving the reduction of the computational volume and the improvement of the detection accuracy. The contributions of this study are as follows:

1 Different fusion parameters are provided for different scenes in the visible-thermal image feature fusion process, which allows the effects caused by different illumination conditions to be taken into account;
2 The channel attention mechanism is used when fusing feature maps from different data sources to allow for the dynamic learning of the fusion weights;
3 Temperature information is used as a priori knowledge to accelerate model convergence and improve the detection accuracy;
4 As compared to previous similar studies (mainly refers to decision-level-based methods such as [7,9]), fewer parameters and simpler models are used;
5 The proposed temperature information module is plug-and-play, and can be applied to all temperature-related target detection applications.

The remainder of this paper is organized as follows. Section 3 introduces the research methodology, and Section 4 presents the experiments. Then, Sections 5 and 6 analyze and discuss the experimental results. Finally, Section 7 presents the conclusions of this research.

3. Methodology

The flowchart of this research based on RetinaNet [11] is presented in Figure 1. First, the dataset is used to train an illumination condition classification model. Second, two branch networks are used to extract the feature information of visible and thermal images, respectively. Moreover, the channel attention and a priori temperature mask (CAPTM) module is proposed to facilitate feature extraction by using the temperature information as a priori knowledge. Third, based on the classification results, the features are fused before ResNet [12] and followed by a feature pyramid network (FPN [13]). The detailed fusion structure is presented in Section 3.4. Different from the traditional sum fusion method, a channel attention mechanism is introduced and different channel weights are used for different illumination conditions to fully exploit the visible or thermal image information. Finally, the fused feature map is ultimately used to obtain the object detection results.

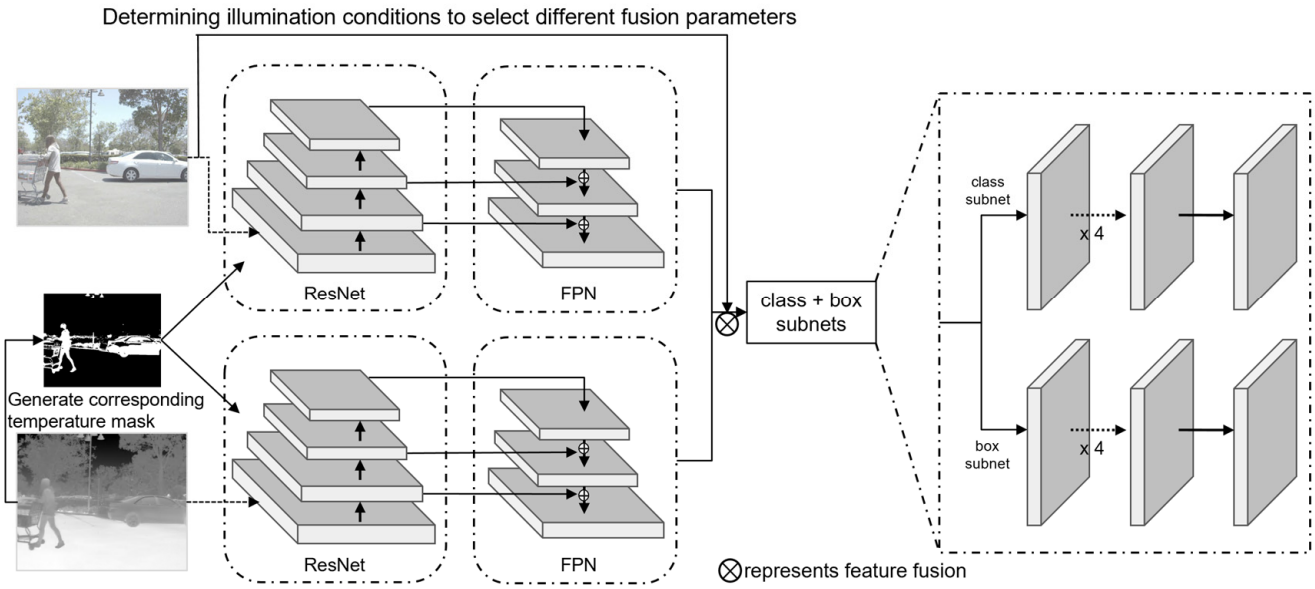


Figure 1. The flowchart of the research framework.

3.1. Illumination Conditions Discrimination

To choose the most suitable fusion parameters, the illumination conditions (day or night) of the image must be determined in advance.

In this study, the original FLIR dataset [14] was used for model training. The original FLIR dataset [14] captured vehicles and pedestrians on the streets and highways of Santa Barbara, California, during daytime (60%) and nighttime (40%) from November to May. Given the computational intensity, the dataset was divided into only two categories—daytime and nighttime—according to the illumination scenario, thereby making full use of the visible and thermal information. For example, during the daytime, both visible and thermal images are clear, and can be complementary; in contrast, thermal images taken during nighttime have more semantic information. The examples of different illumination scenes in the FLIR dataset are shown in Table 1.

The dichotomous classification of image scenarios is a simple and common classification task. In this study, ResNet50 [12], in which only visible images were input, was used to accomplish the task. The classification details are described in Section 4.2.1.

3.2. Fusing Visible and Thermal Feature Maps

The fusion operation occurs at the intersection of the two branch networks, which forms a composite feature map with the semantic information of both visible and thermal images via certain operations. The most common fusion operations include the concatenate, sum, max, and mean operations. Pei et al. [15] explored the effects of these methods on fusion, and found that sum fusion is superior. The function of the summation operation can be expressed as Equation (1):

$$y_{b,c,h,w}^{sum} = f^{sum}(V, T) = v_{b,c,h,w} + t_{b,c,h,w}, \quad (1)$$

where $v_{b,c,h,w}$ and $t_{b,c,h,w}$ denote the values of visible and thermal infrared eigenmaps at (h, w, c, b), respectively, and b, c, h, and w are the batch size, number of channels, height, and width of the feature maps. In this equation, the coefficients of the variables are the same, without distinction between different illumination conditions. However, strictly speaking, different data sources should be given different weights considering their diverse performance in different scenes, e.g., in dark nighttime situations, thermal images are the mainstay of object detection.

Table 1. Examples of different illumination scenes in the FLIR dataset [14].

	Visible Image	Thermal Image
Day		
Night		

For the model to automatically learn the appropriate weights under different lighting conditions, three preconceived schemes were considered:

1 The first scheme was the improvement of a sum fusion method to a weighted summation, as given in Equation (2):

$$y_{b,c,h,w}^{weighted\ sum} = \lambda \cdot v_{b,c,h,w} + (1-\lambda) \cdot i_{b,c,h,w}, \tag{2}$$

where λ is a learnable parameter that can be learned in the network. The method is straightforward and has only one parameter, but the results of an experiment indicated that λ was updated slowly.

2 The second scheme was convolving the concatenated feature map to a new one. While previous studies (e.g., [16]) have used this type of method, it has a large number of parameters, which would substantially increase the model complexity and training time.

3 The third scheme was the use of the channel attention mechanism to score individual channels, i.e., to enhance the important channels and suppress the unimportant channels. This method has a moderate number of parameters and balances the weights not only of the data sources, but also the feature map channels.

Therefore, based on the comprehensive consideration of these schemes, scheme 3 was adopted in this work.

There are many available methods for channel attention. In this study, the squeeze-and-excitation (SE) block [17] was used for feature fusion due to its fewer parameters and good performance. The SE block consists of three main components, namely the squeeze, excitation, and scale components. First, the feature map is compressed using a global pooling layer, after which the corresponding channel weights are learned using two fully connected (FC) layers, and, finally, the weights are mapped to the original feature map to enhance and suppress different channels.

Hu et al. [17] embedded the SE block into the residual block, which enabled ResNet [12] to learn the channel attention of the feature map continuously during extracting features.

By contrast with this method, in the model proposed in the present study, the SE block is applied to feature map fusion to learn the weights of the visible and thermal feature map channels, as shown in Figure 2. Two visible and thermal feature maps are first concatenated, the SE block is used once for the concatenated feature map, and the weighted feature map is finally split for the sum operation to obtain the required fused feature map.

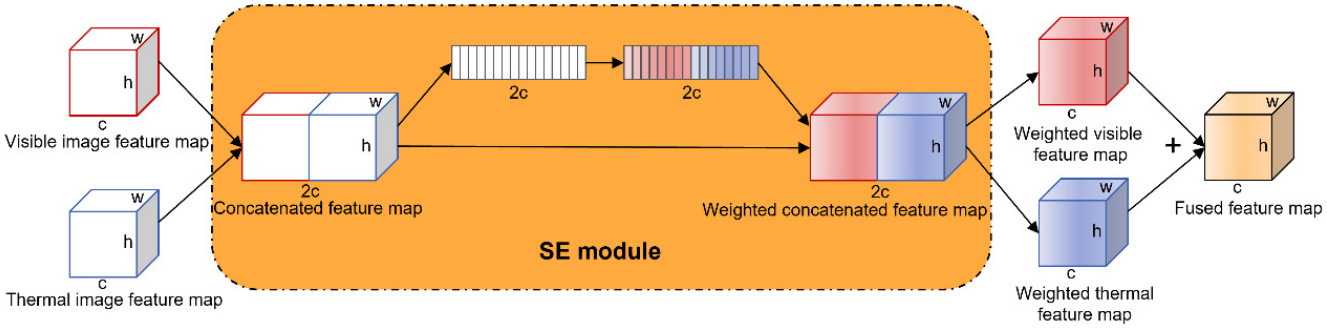


Figure 2. The use of a SE block to merge visible and thermal feature maps (the SE block is indicated in orange). Two branch feature maps are merged into one after an SE block.

3.3. Utilizing Temperature Information as a Priori Knowledge

As mentioned previously, thermal images provide temperature information in addition to texture features. Due to the different characteristics of various objects, the objects in a thermal image have different forms. By using this unique feature, the detection target is separated from the background by setting a certain pixel threshold, and this can be added to the model as a priori knowledge.

According to the detection objects considered in this study, the temperature of the human body is relatively stable; thus, it is easy to segment humans from the background. A car has a high temperature when the engine is running, but cars are also easy to segment when the engine is off due to the high reflectivity of the metal or glass material on their surface. Bicycles are the most challenging objects to differentiate due to their small size and surface material. According to the lighting scenes considered in this study, thermal images taken at nighttime are primarily sensitive to the heat emitted by the target itself, so the segmentation of a person and car is more accurate in nighttime scenes than in daytime scenes. However, it should be noted that the thresholds selected for segmentation would be different in the daytime due to the different lighting conditions.

Therefore, the dataset was roughly divided into three scenes, namely sunny day, cloudy day, and nighttime scenes, and different thresholds were set used to segment these different scenes. For example, suppose the pixel values of a person, bicycle, and car based on a thermal infrared sensor under the night condition are respectively 7400–7500, 7200–7400, and 7500–8000; the threshold value of 7200–8000 can be set to segment the targets from the background. It was experimentally proven that the background noise is not a concern because the temperature knowledge just provides a general direction for the detection model.

The part within the threshold was set as 1 and the part beyond the threshold was set as 0. The resulting visualization of the final temperature mask map is presented in Figure 3. From top to bottom, the three rows respectively indicate segmentation under the scenes of a sunny day, cloudy day, and night. From the figure, it can be seen that during the daytime, people, cars, and bicycles were segmented well but with higher noise; at night, even a person in the distance was segmented accurately and with less noise. Taken together, these segmentation maps are relatively noisy, e.g., the street light in the night scene in the third row of the figure looks like a person in the masked image. However, the proposed model does not rely entirely on the temperature mask; it also relies on other semantic information in the original visible-thermal infrared image. Therefore, most of the noise would not be misidentified.

(**a**) Temperature mask (**b**) Visible image (**c**) Thermal image

Figure 3. The temperature masks in different illumination conditions (sunny day, cloudy day and night).

Spatial attention, i.e., focusing on local information in the spatial domain, aims to identify the regions on the feature map that deserve attention, and to ultimately achieve better detection results. Ordinary spatial attention is obtained computationally and is a posteriori knowledge. In contrast, the temperature mask obtained from thermal infrared images can be treated as parameter-free and a priori knowledge-based spatial attention. Therefore, the CAPTM module, which is based on the convolution block attention module (CBAM) [18] and fuses the a priori knowledge of temperature, is proposed in this work. As shown in Figure 4, the spatial attention part of the original module is replaced with the temperature mask, thereby reducing the number of parameters and simplifying the complexity of the original CBAM. The channel attention module of the CAPTM module is identical to that of the original CBAM. However, the spatial attention module of the CAPTM omits the parameter operations and instead directly uses the weight matrix formed by the temperature mask. The vector obtained by the channel attention module is then element-wise multiplied with the weight matrix generated by the temperature mask to obtain the final output feature map. As shown in Figure 5, the CAPTM module was inserted into ResNet [12] for feature extraction in practical applications.

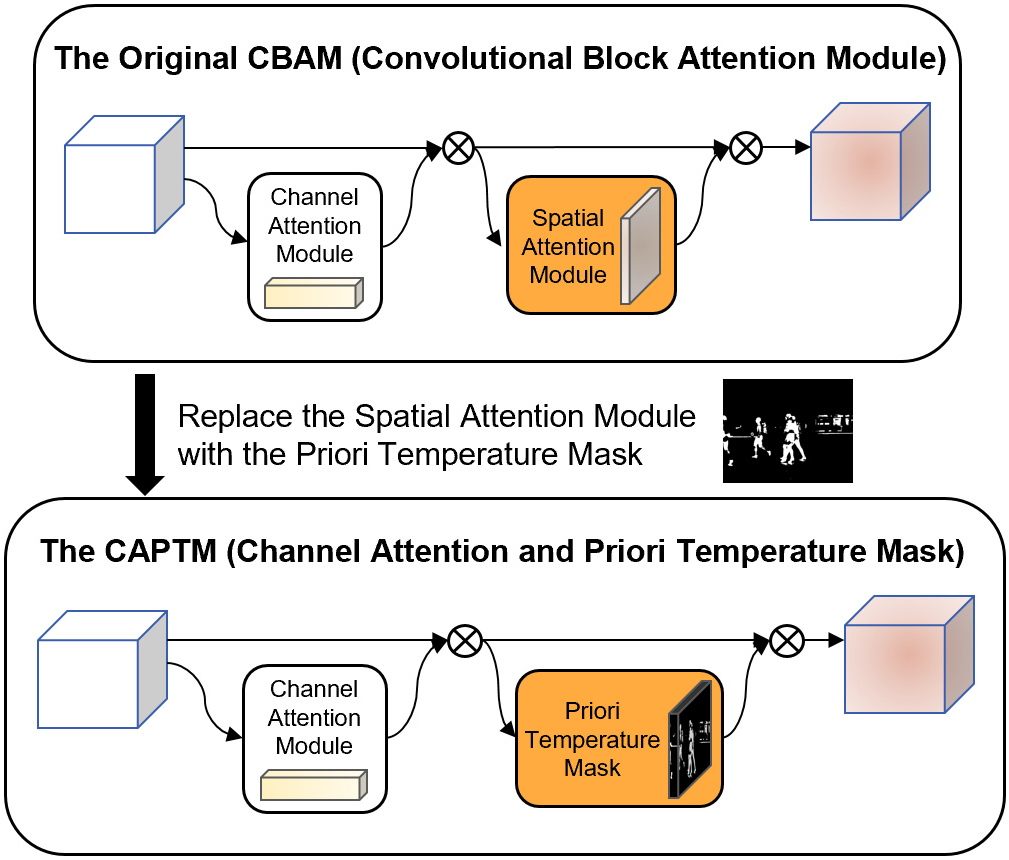


Figure 4. The convolution block attention module (CBAM) and channel attention and a priori temperature mask (CAPTM) module.

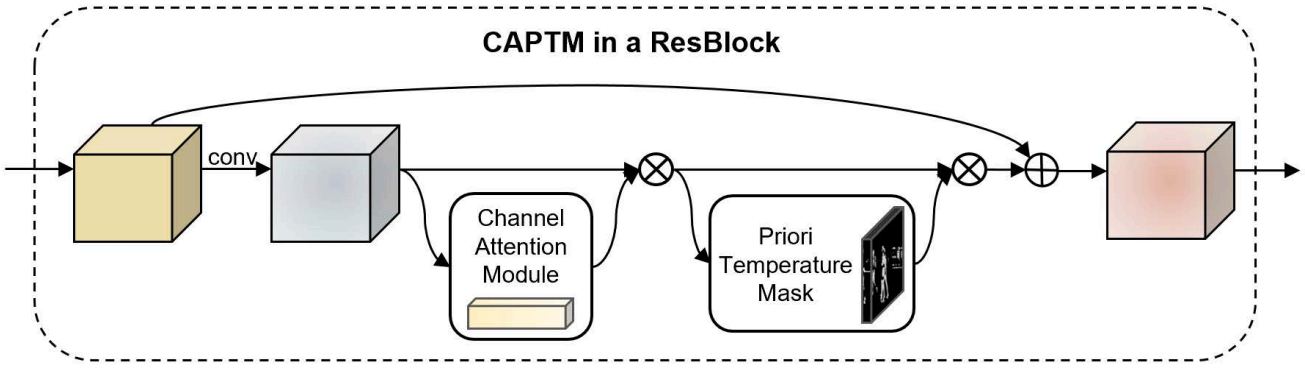


Figure 5. Embedding the CAPTM module into the residual block in ResNet [12].

3.4. The Full Structure of the Model Framework

RetinaNet [11] was used as the main framework in this study, and was chosen over other deep learning models for three reasons. First, RetinaNet is a one-stage detection method, which runs faster than two-stage detection methods. Second, RetinaNet has a simple structure but excellent performance, and the structure is clear when exploring different fusion locations. Finally, RetinaNet contains the FPN [13] structure; thus, it is more effective for the FLIR dataset [14], which is characterized by a large range of target sizes.

RetinaNet consists of three components; the first component is the feature extraction part composed of ResNet, the second component is the multi-scale feature extraction part composed of the FPN, and the third component includes the subnets used to classify and regress the prediction frame. Therefore, to explore the position of the fusion operation, it is explicit to locate the position before ResNet, before ResNet after FPN, and after FPN, which can be referred to as early fusion, middle fusion, and late fusion, respectively. Pei et al. [15] found that middle fusion is superior; thus, middle fusion was used as the base framework in the present study for further exploration.

Middle fusion occurs after ResNet and before the FPN, as shown in Figure 6. After determining the image illumination conditions, the visible and thermal images first respectively pass through two convolutional layers and ReLU layers with the same settings. Then, they pass through their respective branches of ResNet (where the CAPTM module is applied). The features extracted in the last three layers of ResNet are then fused layer-by-layer according to the illumination conditions. At this point, the two branch networks converge into one (Pei et al. [15] used sum fusion at this stage, whereas the proposed model uses a dynamic fusion method based on spatial attention). Finally, the fused features pass through the class and box subnets to obtain the final result. It should be noted that the parameters of the branch networks used to extract the image features are not shared.

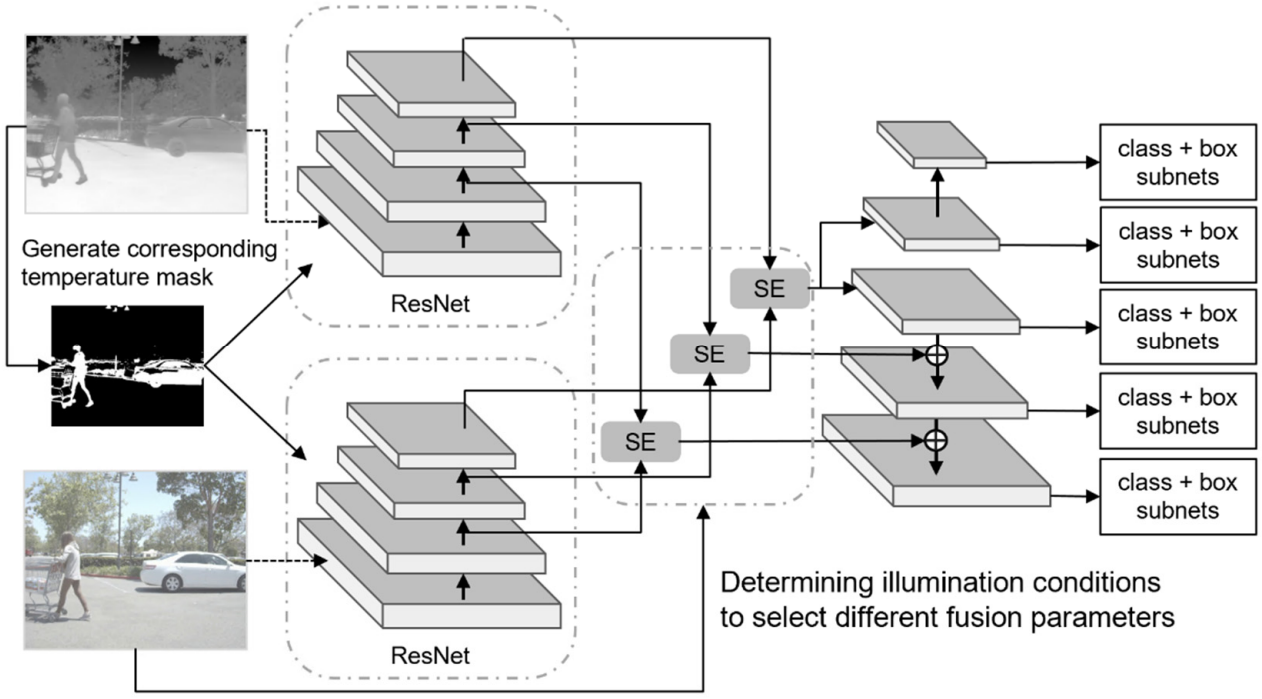


Figure 6. The schematic diagram of the model framework of the proposed model.

Lin et al. proposed an improved cross-entropy (CE loss) loss function called focal loss [11]; in this function, the original CE loss is multiplied by an exponent that weakens the contribution of the easily detectable targets. This enables focal loss to successfully overcome the problem of common loss functions being swayed by a large number of negative samples under the condition of an extreme imbalance of positive and negative samples. This expression is given by Equation (3). However, the final focal loss [11] also includes the introduction of the coefficient α so that it can balance the difficult and easy samples, and its expression is given by Equation (4). The model proposed in this paper employs the smooth $L1$ as the prediction frame regression loss function, focal loss as the category classification loss function, and the sum of the two as the total loss function, as given by Equation (5).

$$FL(p_t) = -(1-p_t)^{\gamma} log(p_t), \tag{3}$$

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}log(p_t) \quad (4)$$

$$L = smooth\_L1 + FL \quad (5)$$

4. Experiments

4.1. Experimental Platform and Data

All experiments conducted in this study were based on a Dell PowerEdge T640 tower server, the configuration of which is described in Table 2. The training and evaluation of the deep network in this study were performed on this platform using PyTorch.

Table 2. The configuration of the experimental platform.

Configuration Items	Configuration Content
Central Processing Unit (CPU)	Intel Xeon Silver 4116 @ 2.10 GHz × 4
Random Access Memory (RAM)	128 G
Hard Disk Drive (HDD)	4T RAID6
Graphics Processing Unit (GPU)	NVIDIA GTX 1080 Ti 11 GB × 4

The FLIR dataset [14] provides the annotated thermal imaging dataset and corresponding unannotated RGB images for training (85%, 8353) and validation (15%, 1267). It (the original FLIR dataset) captured vehicles and pedestrians on the streets and highways of Santa Barbara, California, during daytime (60%, 6190) and nighttime (40%, 3430) from November to May. However, the visible and thermal images in this dataset are not aligned. Therefore, the aligned dataset from Zhang et al. [19] was used, in which the unaligned image pairs were eliminated from the FLIR dataset [14] and 4129 pairs of training data and 1013 pairs of test data were ultimately retained. The images in the aligned dataset have a resolution of 640 × 512 and were captured by an FLIR Tau2 Camera. Approximately 80% (4130) of the images were captured during the daytime, and 20% (1012) images were captured during the nighttime.

The aligned FLIR dataset [19] includes four categories of objects, namely Car, Person, Bicycle, and Dog. The statistics are reported in Table 3, from which it can be seen that the number of Dog objects is too small; thus, only Car, Person, and Bicycle objects were considered in this study.

Table 3. The statistics on the number of classes in the aligned FLIR dataset.

Class Name	Number of Each Class
Car	24,732
Person	13,094
Bicycle	2926
Dog	108

4.2. Parameter Configuration

4.2.1. Illumination Discrimination Network

ResNet50 [12] was used as the illumination discrimination network. The number of neurons in the final fully connected layer was modified to two (day and night). The model was initialized with the parameters of the ResNet50 model pre-trained on ImageNet, and was fine-tuned with the original FLIR dataset [14] for 10 epochs. The reason for choosing the original FLIR dataset for training rather than the aligned FLIR dataset is that the former has a closer sample size of 1:1 for daytime and nighttime. Then, a classification accuracy of 99.22% was reached.

4.2.2. Temperature Mask Extraction

The FLIR dataset [14] includes original thermal TIFF files in addition to RGB images, which is a feature unavailable in other visible-thermal infrared datasets. Because the

thermal RGB images of a common dataset are obtained by stretching and converting TIFF files, the same pixel value of different images does not represent the same radiation value of the original TIFF images. Consequently, the original TIFF images must be used if the temperature information of the thermal infrared images is to be extracted, which is why the FLIR dataset was chosen for use in the present study.

However, when collecting thermal infrared images, the sensor receives not only the radiation of the object itself, but also environmental radiation. Therefore, the pixel values of the same object imaged in different environments will be different. Generally, if the dataset contains weather information, a TIFF file pixel value-temperature model can be constructed to obtain the temperature image, after which a single temperature threshold can be set for mask extraction. The FLIR dataset is not labeled with weather conditions, so a ResNet classifier (designed to pre-process the dataset) was used to classify the original dataset into three categories according to the illumination conditions (sunny day, cloudy day, and night), and different thresholds were set under the three illumination conditions, namely 7500–7700, 7300–7500, and 7200–8000 respectively. Similar to Section 4.2.1, the number of neurons in the final fully connected layer of this classifier was modified to three (sunny day, cloudy day, and night). The model was initialized with the parameters of the ResNet50 model pre-trained on ImageNet, and was fine-tuned with the aligned FLIR dataset [19] for 20 epochs. Then, the best classification accuracy of 94.07% was reached at epoch 13.

As shown in Figure 7, after obtaining the temperature mask, the temperature mask map must be converted into the weight map with weights within the threshold set to 1 and weights outside the threshold set to w. In the figure, white represents the in-threshold part and black represents the out-of-threshold part. Via the use of the equation $img = img * (1 - w) + w$, the transformed weighted map can be obtained, in which the threshold w should be less than 1 for suppression. Then, as described in Section 3.3, the CAPTM module is fed into the feature extraction network.

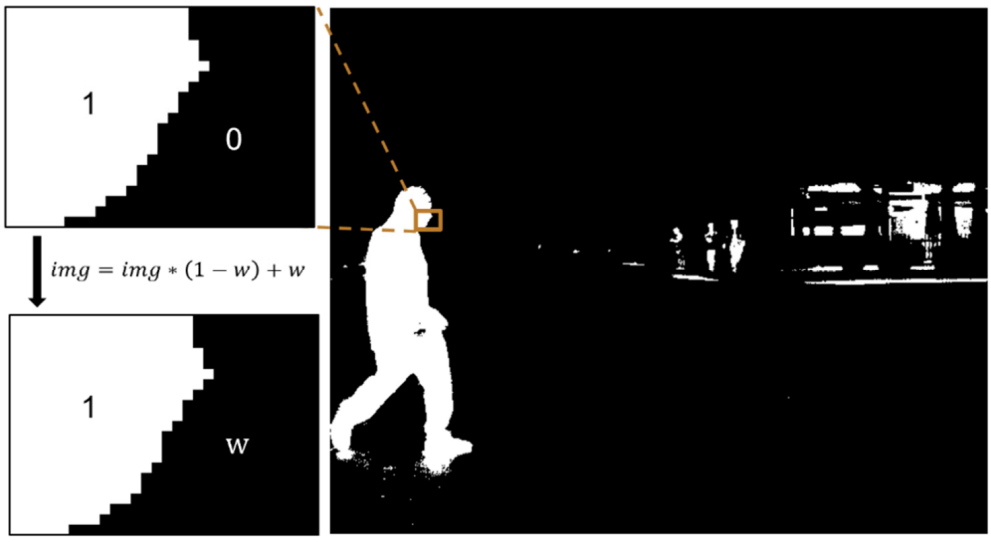


Figure 7. An example of converting the temperature mask map to a weighted map.

4.2.3. Parameter Setting of Anchor

Based on the sizes of the input images and output feature map, anchor_areas $= [16 \times 16,\ 32 \times 32,\ 64 \times 64,\ 128 \times 128,\ 256 \times 256]$ and scale_ratios $= \left[1,\ 2^{\frac{1}{3}},\ 2^{\frac{2}{3}}\right]$ were set. In addition, the aspect ratios of different categories of target objects in the real sample were counted to better set the pre-defined anchor frame aspect ratios, and the statistical results are exhibited in Figure 8. Based on the results, aspect_ratios $= \left[\frac{1}{3},\ \frac{1}{2},\ 1\right]$ was set.

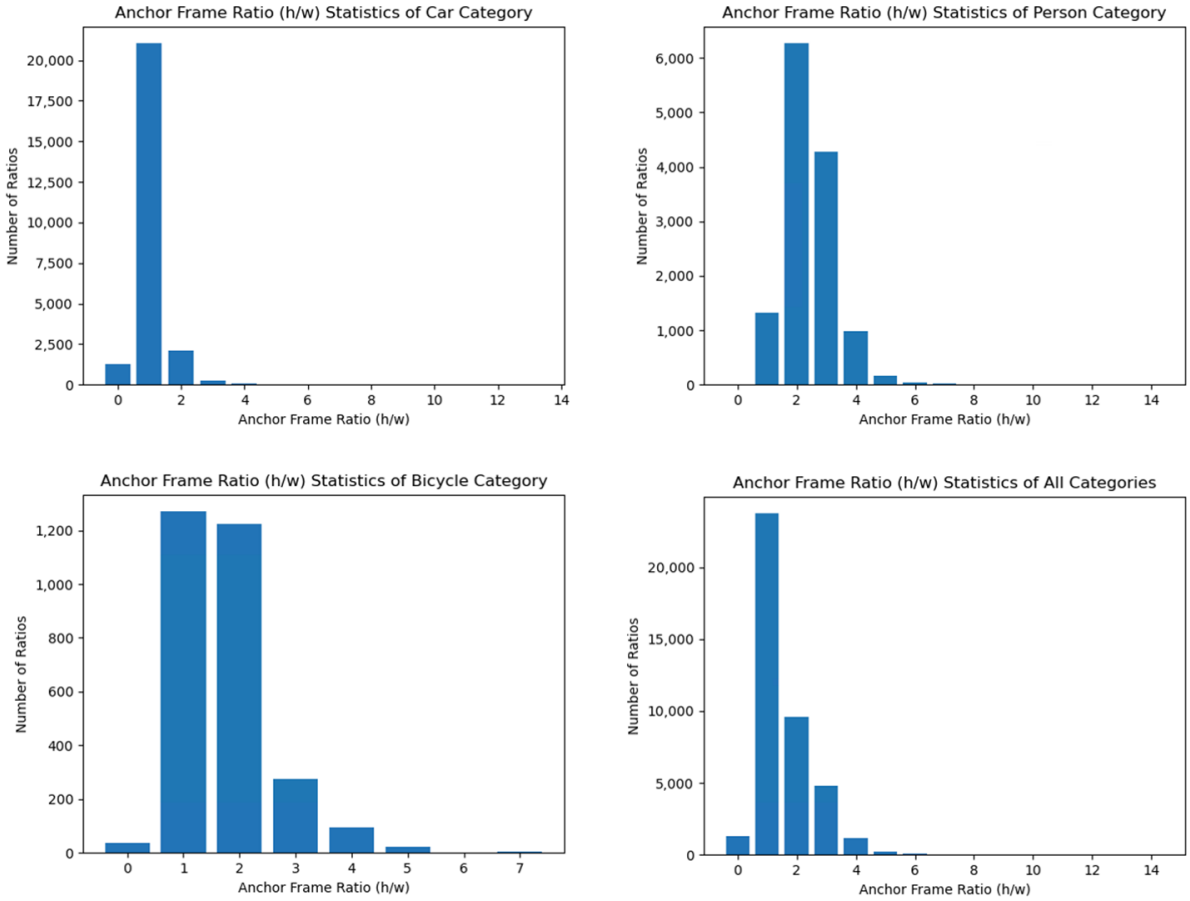


Figure 8. The anchor frame ratio statistics. The horizontal coordinate represents the inverse of the aspect ratio and the vertical coordinate represents the number of different ratios.

4.2.4. Configuration of the Remaining Parameters

In this study, ResNet50 was used as the feature extraction backbone and the input image size was set to 640 × 512. The ResNet model pre-trained on ImageNet was used to initialize the network, and the remaining layers were initialized using Xavir [20]. Furthermore, random clipping and flipping were conducted for data enhancement, and stochastic gradient descent (SGD) was used for backward propagation, for which the momentum parameter $v = 0.9$ and weight decay $\lambda = 0.0001$. Moreover, gradient clipping was used to crop the parameter gradients when their L2 norm was greater than 100. Finally, the learning rate was set as 0.001, and the branch network parameters used to extract visible and thermal infrared image features were not shared.

5. Results

The performance of the CAPTM module when used for different layers of ResNet was evaluated, and the results are reported in Table 4. As revealed by the table, the mean average precision (mAP) was the highest when the CAPTM was applied to the first three layers of ResNet; however, the accuracy decreased after the CAPTM was applied to all layers. It is posited that this is because the deeper layers in ResNet extract deeper semantic information, while the temperature mask is more effective for only primary object detection, which is prone to false recognition if applied to deeper layers. As shown in Figure 9, the effects of different values of w of the temperature mask on the results were also evaluated. It can be seen from the figure that the best results were achieved when $w = 0.8$. Thus, in the subsequent experiments, the relevant models were set with $w = 0.8$.

Table 4. The results of the CAPTM module after temperature information was inserted into different layers. Note: √ indicates the use of the CAPTM module, △ indicates the use of the CBAM module, and CAPTM_x indicates the use of the CAPTM module for the first x layers in ResNet.

Name	ResNet				mAP	AP		
	Layer1	Layer2	Layer3	Layer4		Car	Person	Bicycle
CAPTM_1	√	△	△	△	70.16%	83.29%	74.81%	52.38%
CAPTM _2	√	√	△	△	73.01%	84.28%	76.35%	**58.40%**
CAPTM _3	√	√	√	△	**73.15%**	**84.61%**	**77.04%**	57.79%
CAPTM _4	√	√	√	√	70.87%	84.02%	75.81%	52.79%

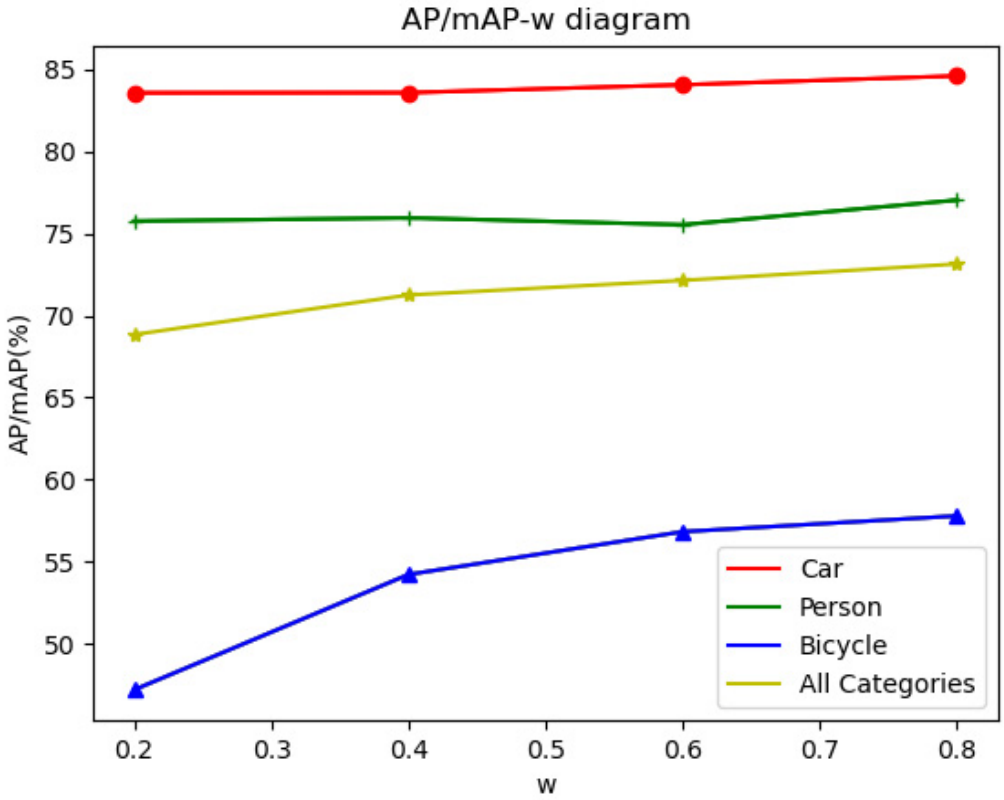
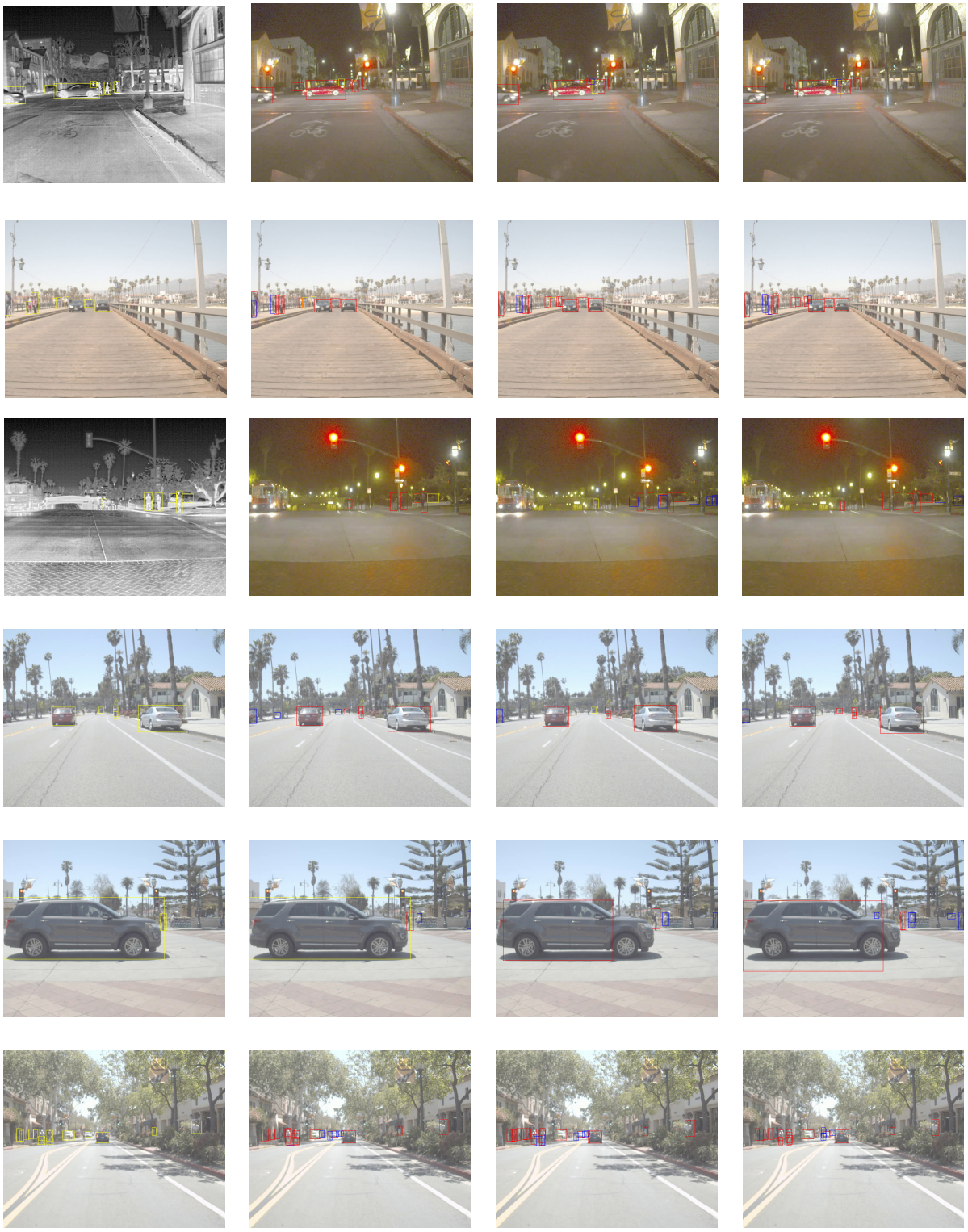


Figure 9. The AP/mAP-w diagram.

Some representative images were selected for visualization, and the results are shown in Figure 10. It is evident from the figure that the results improved from left to right, and the proposed model using SE and CAPTM_3 was the best. The detection results of the proposed model using SE and CAPTM_3 were more accurate, and more small objects at farther distances were detected; moreover, many objects that were not labeled by the ground truth were detected. These findings prove the important role of the CAPTM module.



(**a**) Ground truth (**b**) Sum fusion method (**c**) Ours (+SE) (**d**) Ours (+SE, +CAPTM_3)

Figure 10. The detection results of different models. From left to right, each column represents the ground truth, sum fusion, the proposed model with only SE, and the proposed model with both SE and the CAPTM module. Furthermore, red, yellow, and blue respectively represent true-positive, false-negative, and false-positive results.

The floating point operations (FLOPs), the number of parameters and the average detection time per image of different models are shown in Table 5. As can be seen from the table, the FLOPs, the number of parameters and the average detection time of all the models show a trend from rise to decline, reaching the maximum at Model 4. Compared with Model 4, the proposed model with SE and CAPTM_3 decreases in each metric and performs better, which indicates that using the temperature mask instead of the spatial attention method can both improve the accuracy and reduce the model complexity.

Table 5. The floating point operations (FLOPs), the number of parameters and the average detection time per image of different models.

Model Name	Flops (G)(Input Size=224×224)	Number of Parameters (M)	Average Detecting Time per Image (ms)
Model 1 RetinaNet-sum-fusion	29.058	59.620	76.130
Model 3 Ours (+SE)	29.067	62.402	83.098
Model 4 Ours (+SE) + CBAM	29.114	67.435	157.708
Model 5 Ours (+SE, +CAPTM_3)	29.109	67.433	105.178

The proposed model was compared with other models, and the results are reported in Table 6. Due to the misalignment in the original dataset, almost no studies have used the FLIR dataset for multispectral object detection. Zhang et al. [19] published the aligned FLIR dataset and performed multispectral object detection; thus, it was considered as the current state-of-the-art (SOTA) method in this study. As presented in Table 5, Models 3 and 1 were compared to effectively prove the effectiveness of the proposed fusion method based on the illumination conditions and the channel attention mechanism; Models 5 (the proposed model) and 4 were compared to prove the effectiveness of the proposed CAPTM module; Models 5 and 2 (SOTA) were compared to prove the validity of all the proposed methods. It can be seen that the proposed model (Model 5) achieved a great improvement in accuracy as compared with Models 1, 3, and 4, and the accuracy on the Person object category exhibited a greater improvement as compared with that of Model 2, which proves that the proposed model is more suitable for the detection of objects with richer temperature information. However, compared with Models 2 and 3, the mAP of the Car object category was reduced, which is likely due to the existence of many unlabeled objects in the ground truth; as shown in Figure 11, many vehicles parked on the roadside were detected by the proposed model while they were not labeled in the ground truth. These objects can increase the number of false-positive results and consequently reduce the mAP value.

Table 6. Results of different models.

Model Name	mAP	AP		
		Car	Person	Bicycl
Model 1 RetinaNet-sum-fusion	67.80%	82.27%	72.09%	49.04%
Model 2 FSSD-CFR_3 (SOTA)	72.39%	84.91%	74.49%	57.77%
Model 3 Ours (+SE)	72.45%	**85.60%**	75.10%	56.65%
Model 4 Ours (+SE) + CBAM	71.06%	84.00%	74.88%	54.30%
Model 5 Ours (+SE, +CAPTM_3)	**73.15%**	84.61%	**77.04%**	**57.79%**

(**a**) Ground truth (**b**) Ours (+SE, +CAPTM_3)

Figure 11. Many vehicles parked on the roadside were not labeled. Red, yellow, and blue respectively represent true-positive, false-negative, and false-positive results.

6. Discussion

6.1. Whether the Aligned FLIR Dataset Has a Class Imbalance Problem

As described in Section 4.1, the number ratio of daytime images to nighttime images in the aligned FLIR dataset is not close to 1:1 (80% of daytime images and 20% of nighttime images), so there is a high probability of class imbalance. To examine whether this ratio causes class imbalance, the daytime images and nighttime images in the test set were split for accuracy verification, and the results are shown in Table 7. From the table, it can be seen that the mAP in the nighttime scene is similar to that in the daytime scene, which indicates that although the number of images in the nighttime is less than that in the daytime, it is sufficient for training to achieve the desired accuracy. Actually, the sample ratio of the class imbalance problem is generally 10:1 or higher.

Table 7. Detection accuracy comparisons in daytime and nighttime scenes.

Model Name	Daytime				Nighttime			
	mAP	AP			mAP	AP		
		Car	Person	Bicycl		Car	Person	Bicycle
Model 1 RetinaNet-sum-fusion	68.24%	82.62%	68.75%	53.35%	66.44%	85.42%	79.91%	34.00%
Model 3 Ours (+SE)	71.92%	83.83%	72.43%	59.50%	72.68%	89.34%	81.43%	47.28%
Model 4 Ours (+SE) + CBAM	70.67%	82.64%	72.18%	57.18%	70.07%	86.91%	81.13%	42.16%
Model 5 Ours (+SE, +CAPTM_3)	72.95%	83.05%	73.97%	61.82%	72.13%	88.16%	84.10%	44.14%

From the horizontal comparison point of view, the accuracy of both Car and Person categories is higher at night than during the day; the accuracy of Bicycle category is lower at night than during the day. It is speculated that it is due to that both Car and Person categories have richer temperature information at night, while Bicycle category has weaker temperature information and can hardly be distinguished by thermal infrared sensors at night. From the longitudinal comparison point of view, the proposed model (Model 5) has higher accuracy in the Person category and reaches the highest accuracy in the nighttime scene; the Car category has higher accuracy than Models 1 and 4, but lower accuracy than Model 3. It is speculated that it is due to the Person category has the richest temperature information, in addition to the inaccurate labeling of the Car category in the dataset as

mentioned above. Comprehensive results demonstrate the effectiveness of the proposed method for objects with rich temperature information.

6.2. Threshold Setting in Temperature Mask Extraction

One of the objectives of the experimental design of this study is to maximize the exploitation of thermal infrared images. The temperature information used in the CAPTM model proposed in this paper is a priori and known, which means that the environmental information of the application scene when using the model can be readily accessible, and then the appropriate threshold for extracting the corresponding temperature mask file can be set.

For this study, the most ideal dataset should have the following conditions:

1 Having visible and thermal infrared images in the same place and same time for feature fusion;
2 Having raw thermal infrared data (TIFF file) instead of stretched and processed thermal infrared images (jpg/jpeg file), for extracting temperature mask information;
3 Having environmental information, such as light, temperature, humidity, atmospheric pressure, etc., for extracting the temperature mask by constructing a TIFF file pixel value-temperature model.

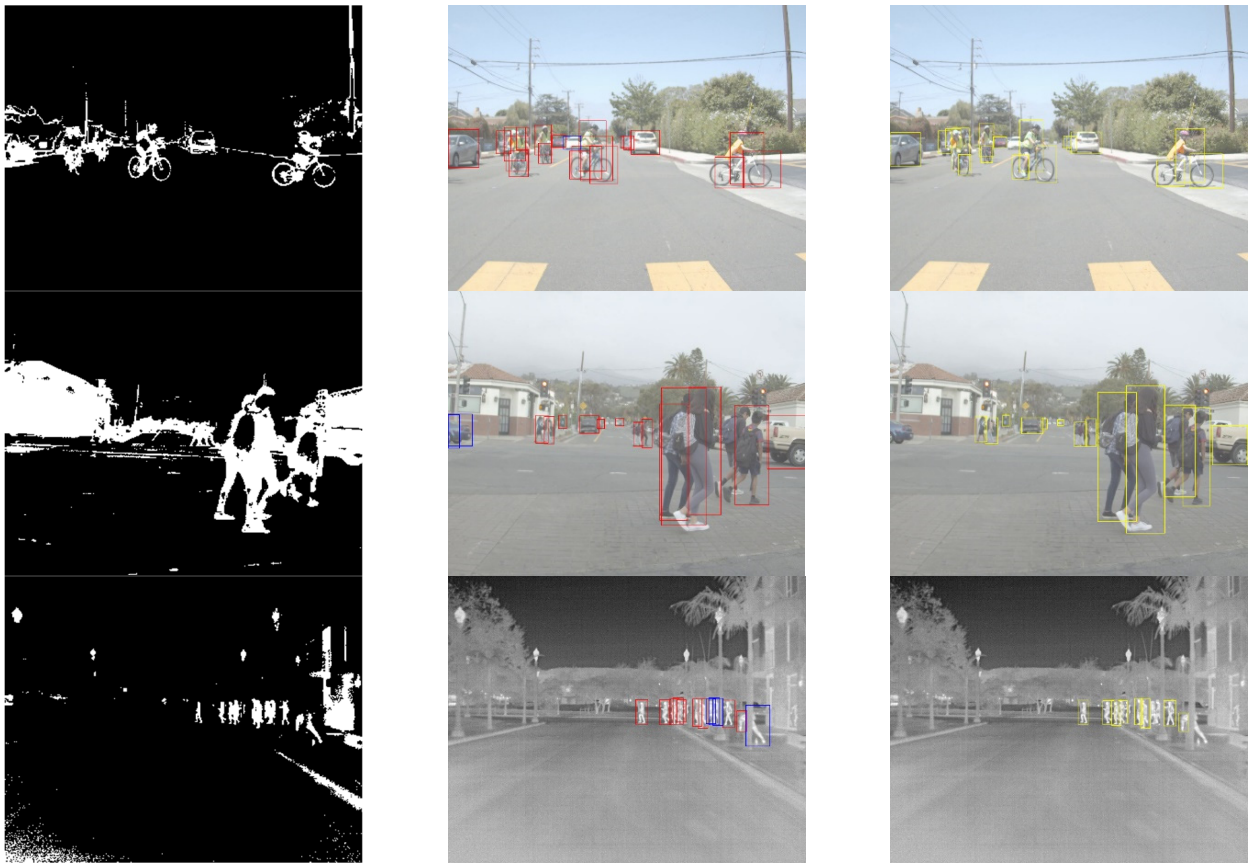
If a dataset satisfies conditions 1 and 2 along with condition 3, then a TIFF file pixel value-temperature model can be constructed and the temperature mask can be extracted by setting a fixed temperature threshold, e.g., if the human body temperature is kept between 20–40 °C, then a threshold of 20–40 can be set for extracting the Person category.

However, there is no publicly available dataset that fulfills all the conditions, and only the FLIR dataset satisfies conditions 1 and 2. Therefore, a ResNet50 was used to classify the data illumination scenes and the thresholds with good results were used for mask extraction (any classifier can be used as long as good results can be obtained).

6.3. Noise in the Temperature Mask Files

The background of the temperature mask file extracted by setting the temperature threshold generates lots of noise, and shadows, buildings, street lights, etc. may be extracted (as shown in Figures 3a and 12a). However, this study only applies the CAPTM module to the shallower part of the feature extraction network (the first three layers of ResNet) to obtain the shallow semantic information. Moreover, in this study, the temperature mask as a priori information only provides a general direction for the object detection and does not directly interfere with the final detection results. The graph of the detection results in Figure 12b can effectively prove the above arguments, and even the street light in the third row, which has a very person-like silhouette, was not falsely detected.

Since the noise of the temperature mask does not unduly affect the final detection results, the threshold values for extracting the mask file do not need to be too precise.



(**a**) Temperature mask (**b**) Detection results by our model (+SE, +CAPTM_3) (**c**) Ground truth

Figure 12. The detection results in (**b**) do not contain the noise (such as shadows, buildings, and street lights, etc.) in the background of (**a**). Even the street light in the third row, which has a very person-like silhouette, was not falsely detected. (Red, yellow, and blue respectively represent true-positive, false-negative, and false-positive results).

7. Conclusions

Based on previous studies, this research proposed a multi-data-source feature fusion method with a channel attention mechanism. Moreover, for the first time, the temperature information contained in thermal infrared images was used as a priori knowledge to improve the accuracy of object detection. First, the FLIR dataset [14] was used to train an illumination condition classification model. Second, two branch networks were respectively used to extract the feature information of visible and thermal images with the proposed CAPTM module. Then, the features were fused based on the classification results with a channel attention method before ResNet followed by the FPN. Finally, the object detection results were obtained. The results of experiments demonstrate that the proposed method is very effective. Compared with the existing SOTA method, the overall mAP of the proposed model was found to be improved by 0.69%, and the AP of the Person object category was improved by 2.55%, which is because the temperature information of people is richer in any environment. The results also validate the effectiveness of the proposed methods and indicate their importance for many fields, such as autonomous driving, disaster rescue, robot navigation, and intelligent video surveillance.

However, the proposed method only classifies, and does not quantify, the illumination conditions, which is one direction that could be improved in the future. Secondly, because the FLIR dataset does not include specific environmental information (e.g., temperature, humidity, time, etc.), different thresholds were only set for temperature mask extraction according to different scenes. Furthermore, these experiments were not tested on other

datasets. In future studies, the original images and corresponding environmental data can be collected to construct a pixel value-temperature model of TIFF files so that a single temperature threshold can be set for mask extraction, and the proposed methods can be tested on a bigger dataset.

Author Contributions: Conceptualization, H.Z. and M.S.; methodology, H.Z.; validation, H.Z., M.S. and X.R.; formal analysis, H.Z.; investigation, H.Z., X.R. and X.W.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z., M.S. and X.R.; visualization, H.Z.; supervision, M.S., X.R.; project administration, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Department of Sciences and Technology of the Xinjiang Production and Construction Corps, China, under Grant 2017DB005, in part by the Key Technologies Research and Development Program of China under Grant 2016YFD0300601, and supported by the High-Performance Computing Platform of Peking University.

Data Availability Statement: The original FLIR dataset is openly available, access through link: https://www.flir.com/oem/adas/adas-dataset-form (accessed on 30 March 2021). The aligned FLIR dataset was published by Zhang et al. [19] (accessed on 30 March 2021).

Acknowledgments: We would like to thank Zhang et al. [19] for the publicly available well-aligned FLIR dataset, which laid the foundation for our experiments.

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

References

1. Qian, C. The Status and Development Trend of Infrared Image Processing Technology. *Infrared Technol.* **2013**, *35*, 311–318.
2. Choi, E.J.; Park, D.J. Human detection using image fusion of thermal and visible image with new joint bilateral filter. In Proceedings of the 5th International Conference on Computer Sciences and Convergence Information Technology, Seoul, Korea, 30 November–2 December 2010; pp. 882–885. [CrossRef]
3. Zhang, X.-W.; Zhang, Y.-N.; Guo, Z.; Zhao, J.; Tong, X.-M. Advances and perspective on motion detection fusion in visual and thermal framework. *J. Infrared Millim. Waves* **2011**, *30*, 354–360. [CrossRef]
4. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In Proceedings of the ESANN, Bruges, Belgium, 27–29 April 2016; pp. 509–514.
5. Chen, Y.; Xie, H.; Shin, H. Multi-layer fusion techniques using a CNN for multispectral pedestrian detection. *IET Comput. Vis.* **2018**, *12*, 1179–1187. [CrossRef]
6. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.
7. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
9. Guan, D.; Cao, Y.; Yang, J.; Cao, Y.; Yang, M.Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Inf. Fusion* **2019**, *50*, 148–157. [CrossRef]
10. Zhao, Y.; Cheng, J.; Zhou, W.; Zhang, C.; Pan, X. Infrared pedestrian detection with converted temperature map. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 2025–2031.
11. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
13. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
14. Free Flir Thermal Dataset for Algorithm Training. Available online: https://www.flir.com/oem/adas/adas-dataset-form/ (accessed on 30 March 2021).
15. Pei, D.; Jing, M.; Liu, H.; Sun, F.; Jiang, L. A fast RetinaNet fusion framework for multi-spectral pedestrian detection. *Infrared Phys. Technol.* **2020**, *105*, 103178. [CrossRef]
16. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully convolutional region proposal networks for multispectral person detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 49–56.

17. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
18. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
19. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 276–280.
20. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010; pp. 249–256.