





Article

Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms

K. R. Akshatha ¹, A. Kotegar Karunakar ^{2,*}, Satish B. Shenoy ³, Abhilash K. Pai ²
and Nikhil Hunjanal Nagaraj ⁴ and Sambhav Singh Rohatgi ⁴

¹ Department of Electronics and Communication Engineering, Centre for Avionics, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India; akshatha.kr@manipal.edu

² Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India; abhilash.pai@manipal.edu

³ Department of Aeronautical and Automobile Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India; satish.shenoy@manipal.edu

⁴ Department of Mechatronics Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India; nikhilnagaraj12@gmail.com (N.H.N.); sambhav300899@gmail.com (S.S.R.)

* Correspondence: karunakar.ak@manipal.edu

Abstract: The automatic detection of humans in aerial thermal imagery plays a significant role in various real-time applications, such as surveillance, search and rescue and border monitoring. Small target size, low resolution, occlusion, pose, and scale variations are the significant challenges in aerial thermal images that cause poor performance for various state-of-the-art object detection algorithms. Though many deep-learning-based object detection algorithms have shown impressive performance for generic object detection tasks, their ability to detect smaller objects in the aerial thermal images is analyzed through this study. This work carried out the performance evaluation of Faster R-CNN and single-shot multi-box detector (SSD) algorithms with different backbone networks to detect human targets in aerial view thermal images. For this purpose, two standard aerial thermal datasets having human objects of varying scale are considered with different backbone networks, such as ResNet50, Inception-v2, and MobileNet-v1. The evaluation results demonstrate that the Faster R-CNN model trained with the ResNet50 network architecture out-performed in terms of detection accuracy, with a mean average precision (mAP at 0.5 IoU) of 100% and 55.7% for the test data of the OSU thermal dataset and AAU PD T datasets, respectively. SSD with MobileNet-v1 achieved the highest detection speed of 44 frames per second (FPS) on the NVIDIA GeForce GTX 1080 GPU. Fine-tuning the anchor parameters of the Faster R-CNN ResNet50 and SSD Inception-v2 algorithms caused remarkable improvement in mAP by 10% and 3.5%, respectively, for the challenging AAU PD T dataset. The experimental results demonstrated the application of Faster R-CNN and SSD algorithms for human detection in aerial view thermal images, and the impact of varying backbone network and anchor parameters on the performance improvement of these algorithms.

Keywords: human detection; thermal camera; aerial images; convolutional neural network; object detection; Faster RCNN; SSD



Citation: Akshatha, K.R.; Karunakar, A.K.; Shenoy, S.B.; Pai, A.K.; Nagaraj, N.H.; Rohatgi, S.S. Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms. *Electronics* **2022**, *11*, 1151. <https://doi.org/10.3390/electronics11071151>

Academic Editor: George A. Tsihrintzis

Received: 18 February 2022

Accepted: 18 March 2022

Published: 6 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Object detection algorithms, in general, determine the category and location of various objects present in the images by combining both localization and classification tasks [1]. It helps to solve some of the advanced computer vision-related tasks, such as object tracking and image segmentation, and thus aids in solving the challenges faced by autonomous driving, automatic image captioning, intelligent video surveillance, and augmented reality applications. The traditional approaches of object detection generally require region of interest (RoI) generation, the extraction of features, and the classification of ROIs as its significant steps. Most of these approaches use handcrafted features, such as HoG [2],

LBP [3], and DPM [4], to describe the object properties. These features are not robust enough to handle the small target size, low resolution, occlusion, pose and scale variation challenges present in the aerial view thermal images. The selection of robust feature descriptors for such complex vision tasks needs a great deal of feature engineering to characterize the objects. The general object detection performance improved significantly after the evolution of convolutional neural networks for image classification [5,6]. Several deep learning-based frameworks have been evolved over the years for generic object detection and are broadly classified into two categories. The first one is the region proposal-based approach that follows the traditional pipeline of region proposal generation and classification. Since they use a two-stage approach, they are more accurate but require higher computation time. Some of the popular two-stage object detection approaches are R-CNN [7], Fast R-CNN [8], Faster R-CNN [9], feature pyramid networks [10], SPP-Net [11], Mask R-CNN [12]. The second one is a regression-based approach that uses a single-stage pipeline to perform the detection, and hence it is faster. Some of the popular single-stage approaches are YOLOv1-v4 [13–16], SSD [17], and DSSD [18]. Both region proposal and regression-based approaches use a pre-defined set of anchors for detection. Recently, few anchorless detectors, such as CenterNet [19], CornerNet [20], FCOS [21] and detection transformers (DETR) [22] were proposed to eliminate the complex anchor generation stage. In general, all these algorithms aimed to detect objects of various categories in the high-resolution PASCAL [23] or COCO [24] datasets. The objects in these datasets appear in large or medium sizes that occupy major portions of the image. As a result, many of these algorithms struggle to detect smaller objects, which occupy small areas of the images. This, together with the extra aerial view constraints, prompted us to look for a viable method for human detection in aerial thermal images [25].

Thermal images generated using thermal sensors are capable of sensing the thermal radiation emitted by various objects. Thermal cameras placed above the ground level (aerial view) find numerous applications in surveillance, border monitoring, search and rescue, etc. [26–28]. The automatic detection of humans in aerial view thermal imagery is a challenging task due to the small target size, low resolution, occlusion, human pose, and scale variations [29]. The sample thermal images from the OSU [30] and AAU PD T [31] aerial thermal datasets used in this study are shown in Figure 1.

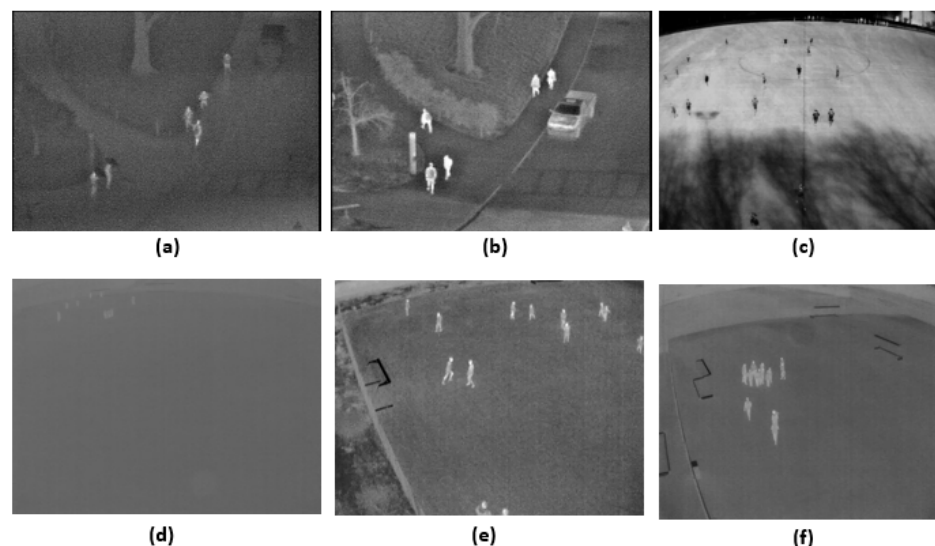


Figure 1. Sample thermal images from (a,b) OSU thermal dataset and (c–f) AAU PD T dataset.

The human targets in these datasets covering a very small portion of the entire image show the small target size and the large scale variations present in the dataset. The illumination variation, occlusion, and poor resolution are the additional challenges in these aerial thermal datasets that can cause general object detection algorithms to perform

poorly. It is challenging to decide the algorithm best suited for a specific application that demands small target detection. Speed and memory are the significant constraints for applications requiring deployment on mobile devices, and some other applications require high accuracy. A literature survey shows that two-stage detectors, such as Faster R-CNN, are more accurate, whereas single-stage detectors are faster and sometimes less accurate. However, SSD has demonstrated improved performance over Faster R-CNN in terms of accuracy and speed for the general object detection task. Hence, in this work we conduct an extensive experimental study to analyze the performance of highly accurate Faster R-CNN and SSD algorithms for human target detection that appear small in aerial thermal images. The choice of a backbone network for feature extraction in these algorithms is crucial, as the number and type of layers directly affect the detector's speed, memory, and performance. In recent years, deeper architectures, such as InceptionNets and ResNets, have shown improved accuracy and lighter MobileNet architectures have shown improved speed. In the proposed work, the Faster R-CNN and SSD algorithms' performance is evaluated using Inception-v2, ResNet-50, and MobileNet-v1 backbone networks on the two aerial thermal datasets comprising small targets. The general object detection algorithms' parameters are designed to detect medium- or large-sized objects, and hence they may not be suitable for small target detection. Therefore, we fine-tuned the original algorithms' default parameters to make them appropriate to detect smaller targets. The performance evaluation of Faster R-CNN and SSD algorithms with various backbone networks provides the rationale for choosing a suitable algorithm for various real-time applications aiming for small target detection. The contributions of this research work are as follows.

- The performance of the Faster R-CNN and SSD algorithms are experimentally evaluated on recently released challenging aerial thermal pedestrian datasets (AAU PD T and OSU thermal pedestrian datasets) for the detection of human targets that appear small in aerial view.
- The impact of varying the backbone network for Faster R-CNN and SSD algorithms concerning speed and accuracy in detecting small targets is evaluated. It will help to choose a suitable object detection model for a specific real-time application.
- The default anchor parameters of the original algorithms are fine-tuned to make the original algorithms more suitable for small target detection, which causes remarkable improvement in the detection performance.

The remainder of the paper is organized as follows. Section 2 overviews the related work, and Section 3 presents the method adopted in the paper. The experimental settings and results are reported in Section 4, and Section 5 discusses the results. Finally, Section 6 concludes the paper with remarks.

2. Related Work

In the past decade, there has been a significant contribution to the object detection domain. In this paper, the review is restricted only to human detection in thermal images and is categorized into traditional and deep-learning-based approaches.

Ma et al. [32] applied the traditional approach of blob extraction and blob classification to detect and track pedestrians. Region-wise gradient and geometric constraints filtering were performed for blob extraction, and HoG and DCT features were used with SVM for blob classification. Lahouli et al. [33] proposed an efficient framework for detecting and tracking pedestrians in thermal images. Saliency maps with contrast enhancement technique were used to extract the regions, and discrete Chebychev moments (DCM) were used as features for SVM classification. Younsi et al. [34] extracted the moving objects using GMM; the shape, appearance, spatial, and temporal based similarity function was used for detection. Teutsch et al. [35] extracted the hot spots using maximally stable extremal regions (MSER), classified using discrete cosine transform (DCT) and random naive Bayes classifier. Biswas et al. [36] used local steering kernel (LSK) as low-level feature descriptors for pedestrian detection in far infrared images. Oluyide et al. [37] proposed an approach

for candidate generation and ROI extraction using histogram specification and partitioning for pedestrian detection in IR surveillance videos.

Zhang et al. [38] presented an infrared-based video surveillance system that enhanced the resolution of data before applying a Faster R-CNN approach. Huda et al. [31] created a substantial diverse thermal dataset (AAU PD T) with variations in time of capture, weather, camera distance, varying body and background temperatures and shadows. They used the YOLOv3 detector to perform human detection. Chen and Shin [39] performed pedestrian detection in IR images using an attention-guided encoder–decoder convolutional neural network (AED-CNN), which generates multi-scale features. Tumas et al. [40] developed a 16-bit thermal pedestrian dataset (named ZUT), captured during severe weather conditions, and YOLOv3 was used to perform pedestrian detection. Huda et al. [41] analyzed the impact of using a pre-processed thermal dataset with the YOLOv3 object detector. The AAU data were enhanced using histogram stretching and the performance was compared with the data in original form. The best performance was obtained for the AAU data in their original form without using pre-processing techniques. Cioppa et al. [42] proposed a novel system to detect and count the players in a football field, using a network trained in a student–teacher distillation approach with custom augmentation and motion information. Farooq et al. [43] used the YOLOv5 framework for smart thermal perception systems using SGD and ADAM optimizers. Vasic et al. [44] proposed a novel method for person detection in aerial images captured using UAVs with the multimodal deep learning approach. It uses two different convolutional networks in region proposal and uses contextual information in the classification stage. Haider et al. [45] proposed fully convolutional regression network to perform human detection in thermal images. This network was designed to map the human heat signature in the thermal image to the spatial density maps. In addition, there are various recent approaches found in the literature that include the cascaded parsing network (CP-HOI) for multistage structured human object interaction recognition [46], and differentiable multi-granularity human representation learning [47], which can be adopted for these tasks due to their better performance in similar vision tasks.

From the literature review, it is found that very limited work has been done to perform small target detection specifically on aerial thermal images. As there are various object detection algorithms for generic object detection tasks, people may face ambiguity in choosing an algorithm for small target detection. The proposed research performs a performance analysis of Faster R-CNN and SSD algorithms to detect human targets with a small size in aerial view thermal images. Further, an attempt is made to fine-tune these algorithms to improve detection.

3. Methods

The evolution of Faster R-CNN and SSD algorithms is a significant milestone in the computer vision domain that has elevated object detection performance by a considerable amount. However, analyzing their performance for a specific task of small target detection is essential to make decisions while choosing the model to detect humans in aerial view thermal images. The performance analysis of the Faster R-CNN and SSD algorithms is carried out for small target detection in aerial view thermal images, using various backbone networks. A transfer learning approach for a diverse set of thermal images was adopted to build a human detection model. Figure 2 shows the approach adopted in this work to compare the two algorithms' performance for aerial thermal datasets.

The publicly available two standard aerial thermal datasets are collected, and the given annotations in the text format are converted into the required standard format (CSV or PASCAL VOC). The labeled training data are used to train and build deep-learning-based Faster R-CNN and SSD object detection models in a transfer learning approach, using COCO pre-trained object detection models. The number of classes is changed to one for detecting a single human category. The Inception-v2 is the common backbone network used for comparing the performance of two algorithms. Resnet50 with Faster R-CNN and Mobilenet-v1 with SSD are used to analyze the impact of varying backbone on accuracy and

speed, respectively. The performance of the trained model is evaluated for the validation and test data, using standard evaluation metrics, such as precision, recall, F-1 score, mean average precision, and the detection speed in terms of frames per second. Fine-tuning the model involves modifying the algorithm’s default anchor parameters to improve the detection performance for small target detection and hyperparameter tuning.

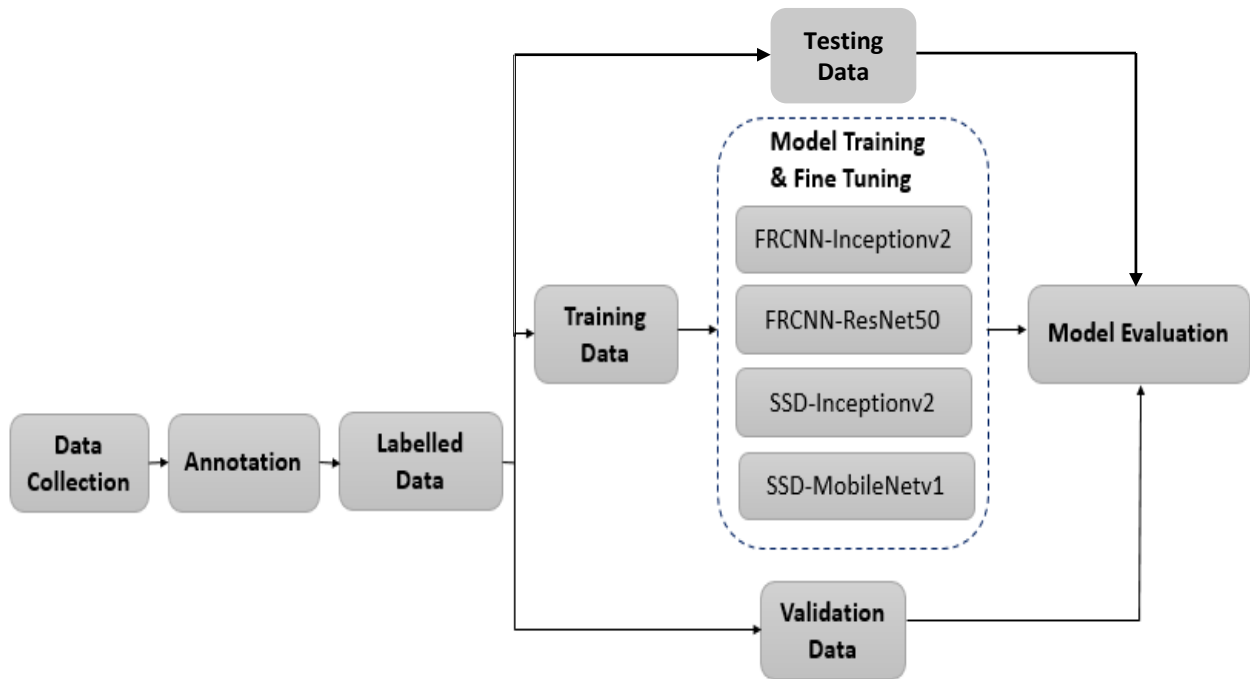


Figure 2. Proposed approach to compare various deep learning-based human detection models. Note: FRCNN stands for Faster R-CNN.

3.1. Faster R-CNN Algorithm

Faster R-CNN [9] is the state-of-the-art multiclass object detection algorithm that has shown good performance in various object detection tasks. Faster R-CNN employs a region proposal-based approach for object detection, having a region proposal network (RPN) and Fast R-CNN detector combined in a single framework as shown in Figure 3.

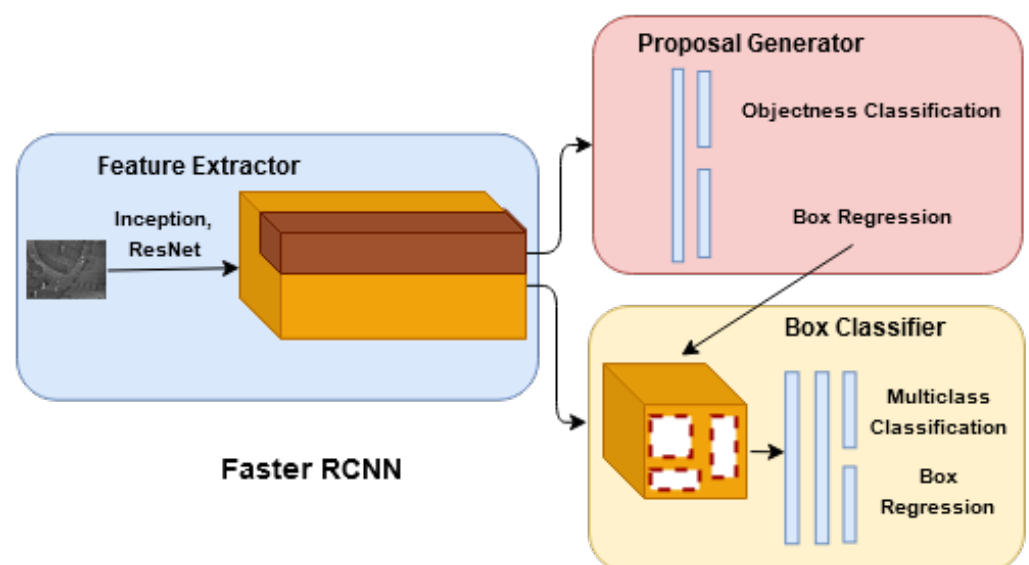


Figure 3. Faster R-CNN approach for object detection.

RPN generates a rectangular set of region proposals of different scales and aspect ratios to determine the regions containing the objects using k anchor boxes. The Fast R-CNN detector takes multiple regions generated from RPN and passes them through the ROI pooling layer and convolution layers to create a feature vector of fixed length. These feature vectors are passed through a set of fully connected layers consisting of the soft-max layer, which generates a probability estimate for each object class and a bounding box regressor to estimate the bounding box's coordinates localization. Non-max suppression (NMS) is performed to merge the region proposals with the highest intersection over union (IoU) with the ground truth and to retain the proposals that have the highest confidence scores.

3.2. Single Shot Multi-Box Detector (SSD)

SSD [17] is the single shot multi-box detector algorithm that uses multi-scale features and default anchor boxes to detect multiple-sized objects present within the scene in a single step as shown in Figure 4.

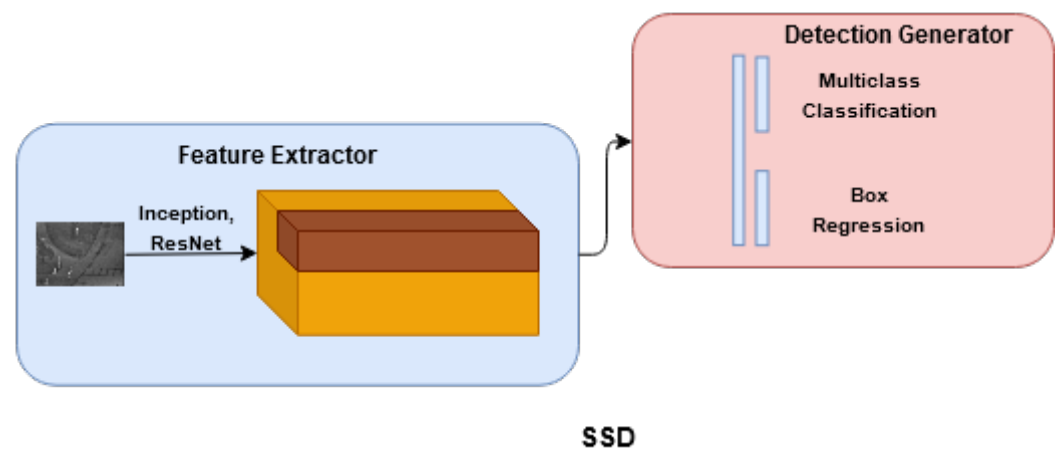


Figure 4. SSD approach for object detection.

SSD does not involve a region proposal network, and hence it is a much faster algorithm than the two-stage detectors. SSD uses a feed-forward convolutional network to produce bounding boxes of fixed size and scores for each box. The convolution layer is added to the base network, and feature maps from multiple layers are used to detect the objects. Features extracted from shallow layers help detect smaller objects, and the deeper layer features are responsible for detecting larger objects. Each feature map cell uses a set of default bounding boxes used to estimate class scores and bounding box locations. End-to-end training is performed using the backpropagation algorithm, and the loss function is composed of the weighted sum of localization and classification losses. Finally, NMS and IoU are conducted to produce the most appropriate bounding boxes.

3.3. Model Fine-Tuning

The choice of a backbone network for feature extraction has a significant impact on the memory, speed, and the detector's performance. Additionally, the original algorithms' default anchor parameters are meant for detecting objects of medium or large sizes in a generic object detection task. They may not be suitable for detecting small targets present in aerial images. Building a deep learning model always needs fine-tuning of the hyperparameters during training to choose the best performing model. Overall, model fine-tuning in this work involves varying the backbone network, fine-tuning the Faster R-CNN and SSD default anchor parameters, and involves hyperparameter tuning to improve the detection performance.

3.3.1. Choosing the Network Architecture

Object detectors mainly consist of two modules, i.e., feature extractor and feature classifier. Initially, when the Faster R-CNN and SSD algorithms were introduced, VGG-Net [48] and ZF-Net [49] were adopted for feature extraction and classification, employing fully connected layers for classification. Deeper and broader convolutional networks are used in the recent object detection algorithms due to the improved performance of fully convolutional networks for classification. Each network architecture has its strength and weakness. VGG-Net is the deeper architecture developed with a fixed-sized convolutional kernel of 3×3 , max pool kernels with size 2×2 , and a stride of 2 to reduce the training time and the number of parameters. Having a fixed size kernel is not always sufficient to recognize variable-sized features in applications, such as object detection. Inception networks [50] aim to detect objects well at different scales, and they are made to increase the depth and width of the network without increasing the computation budget. However, using deeper architectures sometimes causes a vanishing gradient problem as the gradient is back-propagated to lower layers during training. It results in saturation or degradation in the performance when the network becomes deeper. ResNet architectures [51] are developed to eliminate the vanishing gradient problem by introducing residual blocks with skipped connection, which has shown significant improvement in the performance over the earlier architectures. Such architectures cannot be useful for some real-time applications, where speed and memory are the constraints due to their substantial computational costs. MobileNet [52] is the network that is trained to minimize the computational resources. MobileNets use fewer trainable parameters and hence are faster and used for real-time applications. Ren et al. [53] experimentally demonstrated the performance improvement of the object detection algorithms by using deeper architectures, such as ResNets, InceptionNets, and GoogleNet. Based on this study, we have chosen the Inception-v2 backbone for both Faster R-CNN and SSD algorithms for comparing the performance of the two algorithms. ResNet-50 is another backbone network used with Faster R-CNN to check the performance improvement over the Inception-v2 model. Mobilenet-v1 is the lighter network used with the SSD algorithm for speed improvement for use with mobile devices. All these combinations are used to select the best performing model for applications in which speed is the critical requirement.

3.3.2. Modifying the Faster R-CNN and SSD Parameters

The region proposal network of the Faster R-CNN algorithm predicts the region proposals based on the pre-defined set of anchors with three different scales and aspect ratios. They used the scales [0.5, 1, 2] representing the box areas of 128^2 , 256^2 , and 512^2 pixels with aspect ratios 1:1, 1:2, and 2:1 to detect varying sized objects. These values are chosen efficiently to detect objects with a wide range of scales, and they performed very well on PASCAL VOC and COCO object detection tasks. However, humans in thermal images occupying very few pixels may not be detected by using these default anchor parameters. The default anchor boxes used in the RPN of the Faster R-CNN algorithm are too large to detect the smaller targets present in the aerial thermal images [54]. Hence the appropriate anchor parameters for the Faster R-CNN algorithm are experimentally selected to detect small targets. Another parameter, 'stride', decides the step size over the feature map, which is usually set to 16 in the original Faster R-CNN algorithm. The stride value of 16 or 32 is too large to miss detecting smaller human targets lying within stride steps. Based on these observations, we experimented by varying the anchors' scale and stride values and observed the impact on detection performance.

The SSD architecture adopted in this work resizes the input images into 300×300 before detection. The experiment is conducted by increasing the input resolution to 600×600 , and the impact on the performance is observed. The SSD algorithm's anchor size is decided based on the minimum and maximum scale values of the algorithm. The default minimum value is kept at 0.2, which may be large to detect the small objects. The minimum scale of

the anchor parameter is reduced further during the experimentation to observe the effect on performance.

3.3.3. Hyperparameter Tuning

Tuning of the hyperparameters involves changing the learning rate and batch size to make the model learn effectively with the available resources and to avoid over-fitting. Suitable values for these parameters are chosen during the training process by carefully observing the training progress based on the loss function.

3.4. Metrics Used for Model Evaluation

The general evaluation criteria used for object detection algorithms are detection speed in terms of frames per second and mean average precision obtained using precision and recall. The F-1 score is another single metric used for evaluating the object detection algorithms that combine both precision and recall. The confidence score and IoU values decide true positives (TP) and false positives (FP) for the prediction. The precision–recall curve (PR curve) is another metric that shows the association between precision and recall values based on the confidence threshold selected. The model's detection speed is measured based on FPS, which indicates the maximum number of frames that are predicted for objects in one second. It demonstrates how well the algorithm is suitable for applications demanding speed as the primary requirement. In this work, the performance of the trained model is evaluated using precision, recall, F-1 score, PR curve, mean average precision, and the detection speed in terms of frames per second. The confidence score threshold of 0.6 and IoU threshold of 0.5 are used in this experiment for evaluation.

4. Experiments and Results

The experiment was conducted using an HP Elite Desk 800 G4 workstation, with 32 GB of RAM with Intel Core i7-8700K CPU. It has an NVIDIA GeForce GTX 1080 graphic processing unit (GPU) with 8 GB of RAM. The algorithm was implemented in Python 3.7 using the Tensor Flow-1 object detection API framework.

4.1. Dataset and Experiments

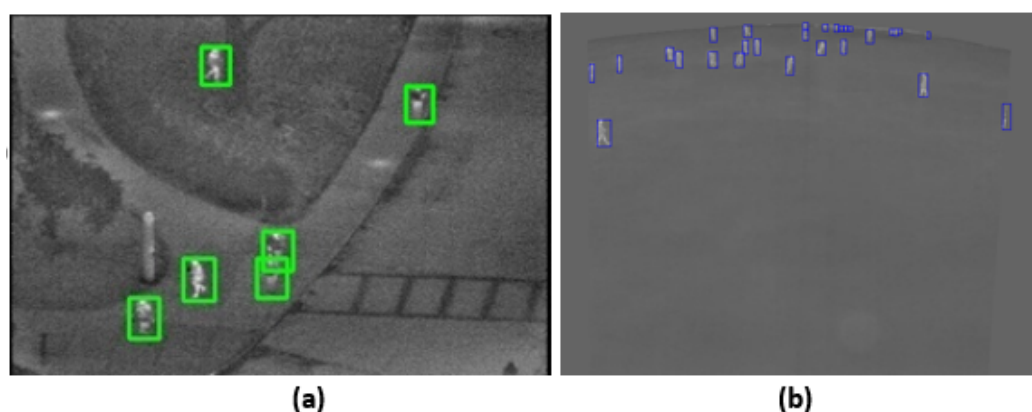
In this work, we used images from OSU and AAU aerial thermal pedestrian datasets for experimentation. The sample images of the dataset with the given annotations are shown in Figure 5. The OSU thermal pedestrian database is the standard dataset used for testing and evaluating the computer vision models, whereas the AAU PD T is the newly released challenging thermal dataset acquired using a aerial thermal camera in a complex outdoor environment. It includes people with different scales, pose variations, interactions/occlusions between people with fast and erratic motion. There exists a lot of variability in the scenes pertaining to the change in weather and lighting conditions, varying body and background temperatures, shadows, wind and snow. In addition, human objects in the AAU PD T dataset are very small and hence it is very challenging dataset when compared to the OSU thermal dataset. The detailed analysis and comparison of the two datasets are illustrated in Table 1.

The OSU and AAU datasets have been split into training, validation and test sets. The training set combines 237 images of OSU dataset and 706 images from the AAU dataset, which is used to train and build various object detection models. The combined validation set, comprising 24 images from OSU and 108 images from the AAU dataset, is used to validate the performance of the model during training. Individual test sets, comprising 24 OSU test images and 130 AAU test images, are used to evaluate the performance of the trained models for two datasets separately.

The performance of the various object detection models based on mAP and the speed they achieved for the COCO object detection task is available in the Tensorflow object detection API. Faster R-CNN with ResNet50 and Inception-v2, SSD with Inception-v2, and MobileNet-v1 are chosen in this work for experimentation.

Table 1. Details regarding the dataset.

Parameter	OSU Dataset	AAU Dataset
Resolution	360 × 240	640 × 480
Number of images	285	944
Camera Position	Rooftop	Light pole
Number of objects	984	7809
Average number of objects per image	3	8
Median of object size (in sq. pixels)	540	256
Median of overlap area between the object and the image	0.63%	0.08%

**Figure 5.** Sample annotated images from (a) OSU dataset (b) AAU PD T dataset.

4.2. Experimental Results

The results of various deep-learning-based human detection models for the validation set and two test sets of OSU and AAU thermal datasets are presented in this section. The Faster RCNN algorithm is trained with the Inception-v2 and ResNet50 backbone architectures, separately. For the resizing of the images, the ‘keep aspect ratio resizer’ function with dimension 600 × 1024 is used. This function resizes the smaller image length to 600 pixels, and if the larger side is greater than 1024, the images are resized to make the long edge 1024. The aspect ratio of the images is preserved. The best performance is obtained for the following set of hyperparameters.

- Learning Rate = 0.0002;
- Batch size=1;
- Augmentation is performed using random horizontal flip and contrast adjustment;
- Momentum optimizer value = 0.9;
- Training steps = 20,000.

The batch size is set to 1 to fit into the available GPU memory. The training is stopped at 20,000 steps, as no further performance improvement is seen during training.

The SSD algorithm is trained using Inception-v2 and MobileNet-v1 network architectures with the following hyperparameter settings. Fixed-sized resizer of dimension 300 × 300 is used prior to training.

- Batch size = 32;
- Learning Rate = 0.0002;
- Augmentation using random horizontal flip and contrast adjustment;
- Momentum 0.9;
- Training steps 20,000.

The default learning rate and momentum optimizer values are selected for both the algorithms while training the model. The training is stopped after 20,000 steps, as no further improvement in the performance is seen during training. The maximum batch size

of 32 is used with the available GPU memory. The results obtained for Faster R-CNN and SSD models using various backbone networks for the combined validation dataset and individual test sets of OSU and AAU datasets are shown in Tables 2–4, respectively.

Table 2. The performance of various object detection models for the validation dataset.

Model Name	Precision	Recall	F-1 Score	mAP (0.5 IoU)	Speed in FPS
FRCNN Inception-v2	75.47%	75.04%	75.25%	55.1%	8
FRCNN ResNet50	77.32%	75.61%	76.45%	56.7%	6
SSD Inception-v2	93.18%	62.68%	74.94%	26.4%	33
SSD MobileNet-v1	96.07%	58.1%	72.41%	24.5%	44

Note: FRCNN stands for Faster R-CNN.

Table 3. The performance of various object detection models for OSU thermal test dataset.

Model Name	Precision	Recall	F-1 Score	mAP (0.5 IoU)	Speed in FPS
FRCNN Inception-v2	95.28%	100%	97.58%	100%	8
FRCNN ResNet50	98.06%	100%	99.02%	100%	6
SSD Inception-v2	100%	87.13%	93.12%	98.6%	33
SSD MobileNet-v1	64%	95.05%	76.49%	81.9%	44

Note: FRCNN stands for Faster R-CNN.

Table 4. The performance of various object detection models for AAU PD T test dataset.

Model Name	Precision	Recall	F-1 Score	mAP (0.5 IoU)	Speed in FPS
FRCNN Inception-v2	74.43%	74.1%	74.27%	54.5%	8
FRCNN ResNet50	78.76%	77.69%	78.62%	55.7%	6
SSD Inception-v2	92.91%	62.41%	74.66%	25.9%	33
SSD MobileNet-v1	94.09%	58.72%	72.31%	23.7%	44

Note: FRCNN stands for Faster R-CNN.

Faster R-CNN with the Resnet50 architecture out-performed the Inception-v2 in terms of precision, recall, and F-1 score. Both Faster R-CNN networks detected humans with high precision and recall for the OSU dataset; however, for the AAU PD T dataset, performance was relatively low due to the tiny size of the targets and the additional challenges in the dataset as highlighted in Table 1.

SSD with Inception-v2 showed improved detection accuracy when compared to the MobileNet-v1 architecture. From Tables 2–4, we can infer that though SSD models are precise in detecting the human objects in all the datasets, the low recall value indicates the difficulty of these models in detecting tiny objects in low-resolution images. The performance comparison of Faster R-CNN and SSD models with different backbone networks in terms of mAP and FPS is illustrated in Tables 2–4. The PR curve comparison between various models for the validation set is shown in Figure 6.

From Figure 6, we can observe that, though both Faster R-CNN models performed equally well, the ResNet50 model out-performed them in terms of mAP. The special residual blocks of the ResNet architecture performed well in extracting rich feature representations from the poor-quality images. SSD MobileNet-v1 achieved the highest speed due to fewer trainable parameters involved in the architecture.

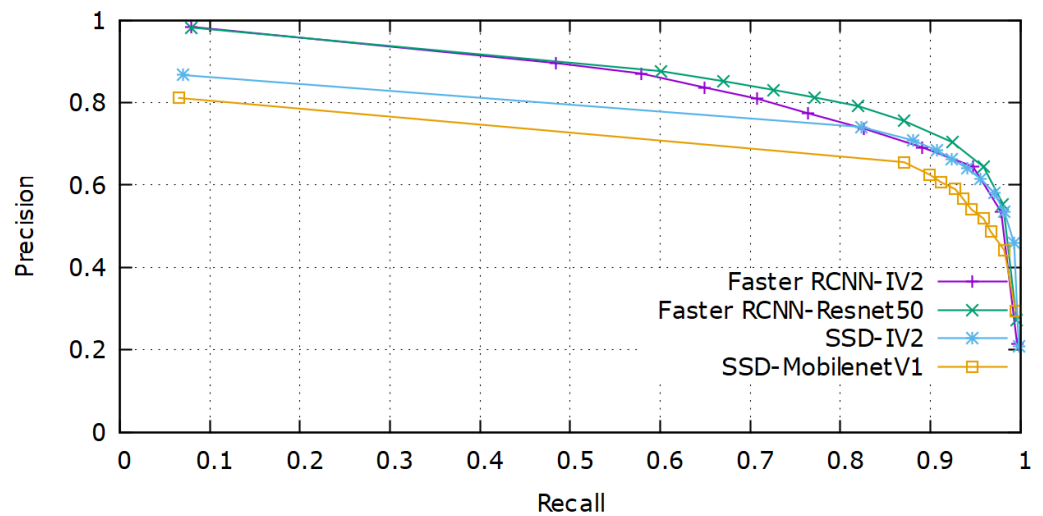


Figure 6. Precision–recall curve comparison of various models for validation set. Note: Faster RCNN-IV2 stands for Faster R-CNN with Inception V2, SSD-IV2 stands for SSD with Inception-V2.

4.2.1. Model Fine-Tuning Results

Faster R-CNN with ResNet50 being the best performing model, as observed in Tables 2–4, is considered for further analysis. Anchor scales of the Faster R-CNN ResNet50 model are changed to [0.5, 0.25, 0.125, 0.0625] and stride values are reduced to 8, and the results compared with the original algorithm are tabulated in Table 5.

Table 5. Performance of Faster R-CNN ResNet50 model before and after fine-tuning.

Dataset	Precision	Original Model			Fine-Tuned Model			
		Recall	F-1 Score	mAP	Precision	Recall	F-1 Score	mAP
Validation	77.32%	75.61%	76.45%	56.7%	94%	75.69%	83.86%	65.4%
OSU test set	98.06%	100%	99.02%	100%	100%	100%	100%	100%
AAU PD-T	78.76%	77.69%	78.62%	55.7%	90.93%	77.52%	83.69%	66.7%

Fine-tuning of the Faster R-CNN anchor parameters caused significant improvement in the detection performance. The performance of the original Faster R-CNN and fine-tuned model is compared using a PR curve for the validation set as shown in Figure 7.

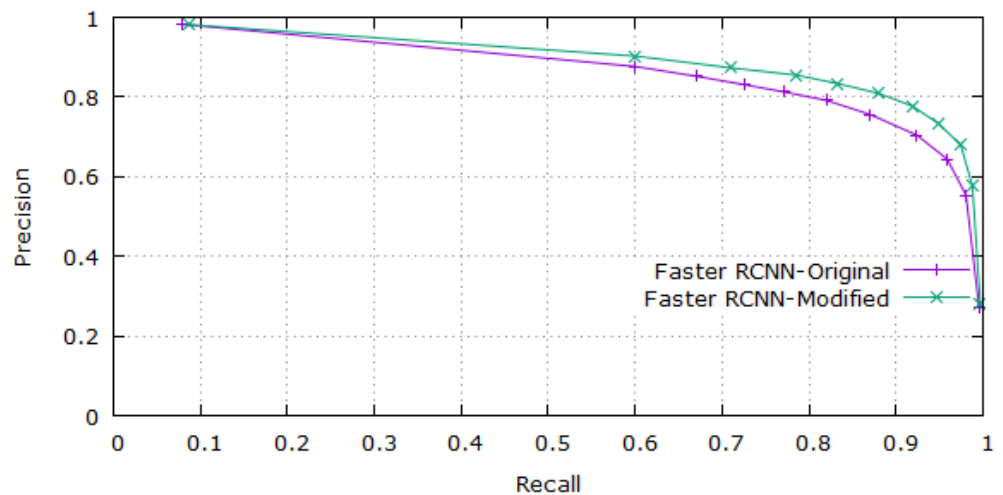


Figure 7. Performance comparison between original faster R-CNN ResNet50 model and its modified version on validation dataset.

For the SSD algorithm with Inception-v2, we tested by decreasing the anchor scale values. Changing the minimum scale value to 0.15 from a default value of 0.2 caused the mAP to improve from 26.4% to 29.9% for the validation dataset, which is not the significant improvement compared to Faster R-CNN. Increasing the input resolution size of the SSD algorithm to 600×600 from the 300×300 input size resulted in an mAP of 30.5% on the validation set. This is because increasing the input resolution improves the performance of the small target detection.

4.2.2. Qualitative Results

The images from different test datasets are given as input to the trained model for prediction, and the results are visualized. The sample qualitative results of the detection obtained by various models are shown in Figure 8.

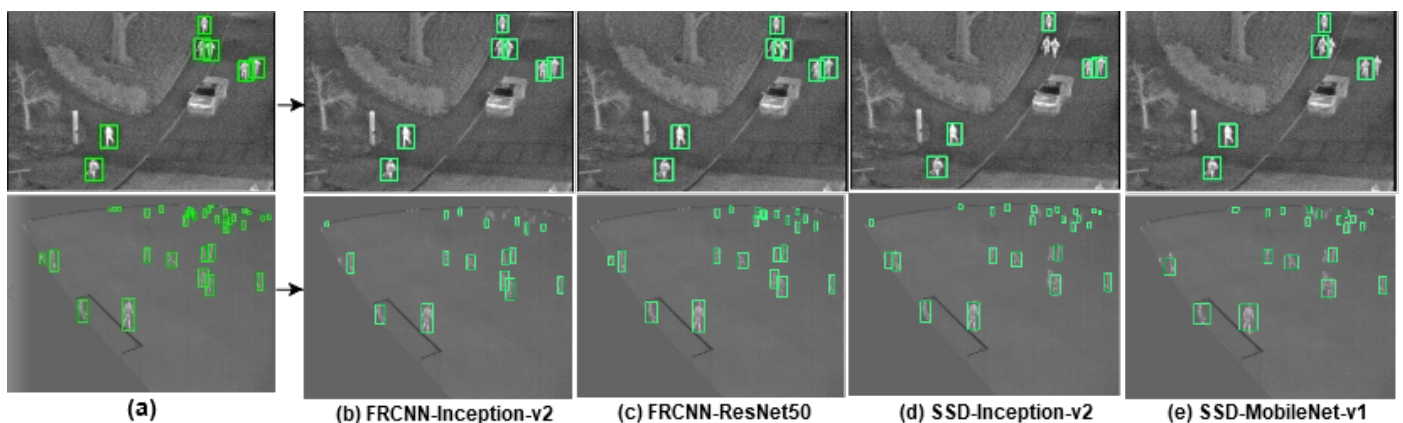


Figure 8. Qualitative results of various models (a) ground truth (b–e) predictions.

5. Discussion

The performance comparison of the Faster R-CNN and SSD algorithms with different backbone network architectures was analyzed through this study in order to detect humans in the aerial view thermal images. For all the datasets used in the experimentation, the Faster R-CNN algorithm gave more accurate results in detecting human instances with small target sizes. The higher input resolution size and the additional region proposal network helped to produce more accurate results. In contrast, the simple single-stage architecture and absence of a region proposal stage in SSD helped achieve the highest detection speed. This study enabled us to choose the specific algorithm depending on the application.

The impact of choosing a specific network architecture on detection accuracy and speed is observed after performing the experiments. As different backbone networks vary in their architecture, layers and type of filters, they showed significant variation in the performance from one another. For Faster R-CNN, both the Inception-v2 and ResNet50 gave better results, whereas the ResNet50 model out-performed due to the ResNet architecture's specialty to deal with the gradient vanishing problem. Though in the PASCAL VOC object detection task, SSD performed better than Faster R-CNN in terms of accuracy and speed, it failed to detect the smaller objects with high accuracy. SSD MobileNet-v1 being a lighter model and having fewer trainable parameters achieved the highest detection speed of 44 FPS. Hence SSD models can be used for real-time applications, where speed is the primary requirement. It is observed that there is a huge difference in the performance in terms of mAP and speed when both algorithms use a common Inception-v2 backbone network. The difference in the performance is less when the backbone network is changed for each algorithm. This study demonstrates that the choice of algorithm plays a significant role as compared to the choice of the backbone network.

An attempt made to fine-tune the default anchor parameters of both Faster R-CNN and SSD algorithms to detect smaller targets caused improved performance. The scale

and stride parameters for anchor generation in the Faster R-CNN algorithm were found to impact the performance of the algorithm significantly. The original algorithm had scale values of [0.5, 1, 2], and stride values were 16 pixels in both horizontal and vertical directions. These scale values are large to detect very small-scale objects, and the stride steps are so large that they can miss many smaller objects. The original Faster R-CNN ResNet50 algorithm with default anchor parameters resulted in a mAP of 56.7%, whereas changing the scales to [0.5, 0.25, 0.125, 0.0625] did not cause any noticeable improvement in the performance. Changing the stride parameter from 16 to 8 improved the mAP to 59.5%. However, reducing the stride to 8 and decreasing the scales to [0.5, 0.25, 0.125, and 0.0625] gave us the highest mAP of 65.4%, which is significantly improved compared to the default parameters.

Similar experimentation was done for the SSD algorithm with Inception-v2, wherein we changed the minimum scale value to 0.15 from a default value of 0.2. It caused the mAP to improve from 26.4% to 29.9%. When images are resized to 300×300 size in the SSD algorithm before applying detection, it causes small human objects in the original image to appear further smaller. Thus SSD struggles to detect many tiny objects, such as humans present in the dataset. Increasing the input resolution size of the SSD algorithm to 600×600 resulted in a mAP of 30.5%. The best-performing hyperparameter values for learning rate and batch size were chosen based on the experimentation. Having a large batch size is always better; hence the maximum possible batch size is selected to fit into the available GPU memory constraints.

6. Conclusions

Performing human detection in aerial thermal images is a useful task for various applications in surveillance, security, search and rescue, border monitoring. Hence, performance analysis of Faster R-CNN and SSD algorithms to detect humans in thermal images is performed. The Faster R-CNN ResNet50 model achieved the highest mAP, and SSD Mobilenet-v1 achieved the highest detection speed. This study suggests that Faster R-CNN algorithms are better suited in applications where accuracy is the main criterion, and SSD algorithms are suitable for real-time applications where speed is the primary requirement. The fine-tuning of Faster R-CNN anchor parameters has shown improvement in mAP by 10%, whereas SSD showed 3.5% improvement compared to the original parameters. Though Faster R-CNN has shown better accuracy when compared to SSD, there is scope for further improvement. This study can be further extended to other recent deep learning-based object detectors. Additionally, the use of pre-processing techniques to enhance the thermal images' quality may lead to better detection performance with the cost of increased computation load.

Author Contributions: Conceptualization, K.R.A. and N.H.N.; methodology, K.R.A., N.H.N. and S.S.R.; validation, K.R.A., A.K.K. and A.K.P.; formal analysis, K.R.A., A.K.K. and S.B.S.; investigation, K.R.A. and A.K.P.; resources, A.K.K.; data curation, N.H.N. and S.S.R.; writing—original draft preparation, K.R.A. and A.K.K.; writing—review and editing, K.R.A., A.K.K., S.B.S. and A.K.P.; visualization, K.R.A., N.H.N. and S.S.R.; supervision, A.K.K. and S.B.S.; project administration, K.R.A., A.K.K. and S.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
3. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
4. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef]
5. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
6. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
8. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
10. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
12. 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; pp. 2961–2969.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; Volume 2016, pp. 779–788. [CrossRef]
14. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
15. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. 2018. Available online: <http://xxx.lanl.gov/abs/1804.02767> (accessed on 21 January 2021).
16. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
18. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
19. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6569–6578.
20. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
21. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. 2019. Available online: <http://xxx.lanl.gov/abs/1904.01355> (accessed on 21 January 2021).
22. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
23. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
24. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
25. Nguyen, N.D.; Do, T.; Ngo, T.D.; Le, D.D. An Evaluation of Deep Learning Methods for Small Object Detection. *J. Electr. Comput. Eng.* **2020**, *2020*, 3189691. [CrossRef]
26. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2014**, *25*, 245–262. s00138-013-0570-5. [CrossRef]
27. Berg, A.; Ahlberg, J.; Felsberg, M. A thermal object tracking benchmark. In Proceedings of the 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Karlsruhe, Germany, 25–28 August 2015; pp. 1–6.

28. Sambolek, S.; Ivasic-Kos, M. Automatic Person Detection in Search and Rescue Operations Using Deep CNN Detectors. *IEEE Access* **2021**, *9*, 37905–37922. [CrossRef]
29. Sumit, S.S.; Rambli, D.R.A.; Mirjalili, S. Vision-Based Human Detection Techniques: A Descriptive Review. *IEEE Access* **2021**, *9*, 42724–42761. [CrossRef]
30. Davis, J. OSU Thermal Pedestrian Database. Available online: <http://vcipl-okstate.org/pbvs/bench/> (accessed on 21 March 2020).
31. Huda, N.U.; Hansen, B.D.; Gade, R.; Moeslund, T.B. The effect of a diverse dataset for transfer learning in thermal person detection. *Sensors* **2020**, *20*, 1982. [CrossRef]
32. Ma, Y.; Wu, X.; Yu, G.; Xu, Y.; Wang, Y. Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery. *Sensors* **2016**, *16*, 446. [CrossRef]
33. Lahouli, I.; Haelterman, R.; Chtourou, Z.; De Cubber, G.; Attia, R. Pedestrian detection and tracking in thermal images from aerial MPEG videos. In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018), Funchal, Portugal, 27–29 January 2018; Volume 5, pp. 487–495. [CrossRef]
34. Younsi, M.; Diaf, M.; Siarry, P. Automatic multiple moving humans detection and tracking in image sequences taken from a stationary thermal infrared camera. *Expert Syst. Appl.* **2020**, *146*, 113171. [CrossRef]
35. Teutsch, M.; Mueller, T.; Huber, M.; Beyerer, J. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 209–216. [CrossRef]
36. Biswas, S.K.; Milanfar, P. Linear Support Tensor Machine with LSK Channels: Pedestrian Detection in Thermal Infrared Images. *IEEE Trans. Image Process.* **2017**, *26*, 4229–4242. [CrossRef]
37. Oluyide, O.M.; Tapamo, J.R.; Walingo, T.M. Automatic Dynamic Range Adjustment for Pedestrian Detection in Thermal (Infrared) Surveillance Videos. *Sensors* **2022**, *22*, 1728. [CrossRef]
38. Zhang, H.; Luo, C.; Wang, Q.; Kitchin, M.; Parmley, A.; Monge-Alvarez, J.; Casaseca-de-la Higuera, P. A novel infrared video surveillance system using deep learning based techniques. *Multimed. Tools Appl.* **2018**, *77*, 26657–26676. [CrossRef]
39. Chen, Y.; Shin, H. Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network. *Appl. Sci.* **2020**, *10*, 809. [CrossRef]
40. Tumas, P.; Nowosielski, A.; Serackis, A. Pedestrian Detection in Severe Weather Conditions. *IEEE Access* **2020**, *8*, 62775–62784. [CrossRef]
41. Ul Huda, N.; Gade, R.; Moeslund, T.B. Effects of Pre-processing on the Performance of Transfer Learning Based Person Detection in Thermal Images. In Proceedings of the 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, 16–18 July 2021; pp. 86–91. [CrossRef]
42. Cioppa, A.; Deliege, A.; Huda, N.U.; Gade, R.; Van Droogenbroeck, M.; Moeslund, T.B. Multimodal and multiview distillation for real-time player detection on a football field. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 880–881.
43. Farooq, M.A.; Corcoran, P.; Rotariu, C.; Shariff, W. Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS). *IEEE Access* **2021**, *9*, 156465–156481. [CrossRef]
44. Kundid Vasić, M.; Papić, V. Multimodal deep learning for person detection in aerial images. *Electronics* **2020**, *9*, 1459. [CrossRef]
45. Haider, A.; Shaikat, F.; Mir, J. Human detection in aerial thermal imaging using a fully convolutional regression network. *Infrared Phys. Technol.* **2021**, *116*, 103796. [CrossRef]
46. Zhou, T.; Wang, W.; Liu, S.; Yang, Y.; Van Gool, L. Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1622–1631.
47. Zhou, T.; Qi, S.; Wang, W.; Shen, J.; Zhu, S.C. Cascaded parsing of human-object interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
48. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
49. Matthew, D.; Fergus, R. Visualizing and understanding convolutional neural networks. In Proceedings of the 13th European Conference Computer Vision and Pattern Recognition, Zurich, Switzerland, 6–12 September 2014; pp. 6–12.
50. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
51. 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.
52. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
53. Ren, Y.; Zhu, C.; Xiao, S. Object detection based on fast/faster RCNN employing fully convolutional architectures. *Math. Probl. Eng.* **2018**, *2018*, 3598316. [CrossRef]
54. Ren, Y.; Zhu, C.; Xiao, S. Small object detection in optical remote sensing images via modified faster R-CNN. *Appl. Sci.* **2018**, *8*, 813. [CrossRef]