Depthwise Convolution For Compact Object Detector In nighttime Images

Heena Patel[1], Kalpesh Prajapati[1], Anjali Sarvaiya[1], Kishor Upla[1],
Kiran Raja[2], Raghavendra Ramachandra[2] and Christoph Busch[2]
[1]*Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India.*
[2]*Norwegian University of Science and Technology (NTNU), Gjøvik, Norway.*

Abstract

Despite thermal imaging primarily used for nighttime surveillance, uniform temperature of object and background makes it difficult to acquire details in the scene being observed and thereby object detection. Further, thermal images collected over long distances degrade the spatial resolution of the acquired objects and so do the moving objects leading to noisy features. We present a computationally efficient object detection approach using Depthwise Deep Convolutional Neural Network (DDCNN) for detecting and classifying objects in nighttime images under low resolution. The Depthwise Convolution (DC) employed in the proposed approach minimises the network's computational complexity resulting in the lowest number of training parameters (i.e., $3M$) as compared to the other existing state-of-the-art methods such as FRCNN ($52M$), SSD ($24M$) and YOLO-v3 ($61M$) parameters. Further, by introducing novel Tversky and Intersection over Union (IoU) loss functions into the compact architectural design, we improve nighttime object detection accuracy. The validity of the proposed model is assessed on numerous datasets such as FLIR, KAIST, MS, and our internal dataset having multiple objects in each image. The experimental results from the proposed method indicate both quantitative and qualitative improvements over the recent state-of-the-art methods for nighttime imaging. The proposed approach achieves a mean Average Precision (mAP) of 52.39% and a highest individual object detection accuracy of 72.70% accuracy for cars in nigh-time situations suggesting applications in real-time use cases.

1. Introduction

Object detection consists of finding and classifying areas of interest in the image and has been extensively studied in the past. Given an image, the object detection algorithm outputs one or more detection hypotheses with a probability score. An object detection algorithm not only finds the class of objects but also identifies the extent of the objects in the image. As objects can be placed anywhere in the image and can be of any size, object detectors are typically designed to detect multiple objects [17, 48, 49]. Such object detection approaches have a number of applications, especially in nighttime surveillance. The performance of the object detector in such applications decides the accuracy of the overall system. A robust and sophisticated object detector in night surveillance systems can therefore be asserted as a basic need for keeping society safe by detecting and classifying the objects of interest accurately in the area under surveillance.

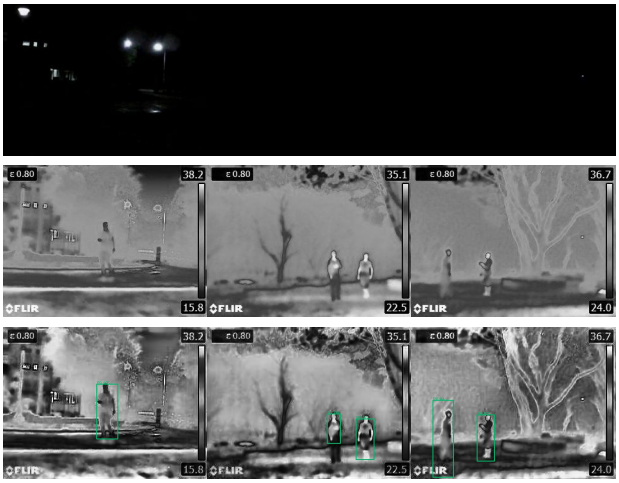


Figure 1. The examples of nighttime images. These images have been acquired with absence of artificial light source and in darker regions. The first and second row show the visible and its corresponding thermal images, respectively.The third row shows the detected objects obtained using the proposed method on thermal images.

Over the last few decades, research on automated visual systems has grown rapidly, and visible spectrum cameras have become the standard imaging device for acquiring those images for surveillance. Such cameras are usually equipped with regular Charge-Coupled Devices (CCDs), and hence, the visibility and color of the acquired objects

depend on the sources of illumination present during the acquisition process. Generally, these cameras have the advantage of high spatial resolution and are suitable for daytime and nighttime with proper lighting setup. However, the acquired objects are not clearly visible due to inherent hardware limitations in the absence of proper lighting sources. Hence, they tend to perform inadequately for no and/or poor lighting conditions for nighttime surveillance [58]. Furthermore, even in cases where the images can be captured during nighttime using visible spectrum cameras, the objects are often unclear due to the absence of proper illuminating sources [41].

Most common objects such as vehicles and pedestrians are difficult to be identified by humans during nighttime [19], and such an argument also holds for automated surveillance systems operating during the night. For instance, most automated night vision systems for monitoring intelligently moving objects assume that the input image has a clear view under lane light, but unfortunately, this does not always hold [58]. The quality of these images is affected by several atmospheric conditions that change the key characteristics of the light source due to scattering (i.e., intensity, color, polarization, coherence) [10,44]. It is therefore essential to obtain a better solution for major safety issues due to the collisions of vehicles in dark or in poor lighting conditions.

To deal with the limitations of CCD sensors during nighttime, a number of works have proposed using thermal cameras as an alternative [19, 27]. Thermal cameras, in particular, have the innate ability to capture images with/without the presence of light [33, 54]. Despite their advantage of capturing images in low lit conditions, object detection from thermal images remains a challenge [40,56]. As depicted in the Fig. 1, infrared cameras are rarely affected by changes in ambient lighting, and they can capture sufficient quality images in darkness, fog and other complex environments. This enables the proper and widespread use of thermal images in scenarios involving high-value targets, such as remote surveillance applications that monitor distant vehicles, pedestrians, or buildings [4]. However, object detection in the thermal images is still difficult due to its unique features [40] as depicted in Fig. 1. One of the challenges in detecting objects in thermal images is the perceived temperature of the object of interest being similar to the perceived temperature of the background. Such a challenge results in low contrast of the acquired image and adversely reduces automatic object detection/recognition performance in thermal images. It, therefore, manifests a serious obstacle to real-world applications where detecting objects is critical, such as in nighttime surveillance or autonomous driving.

As a result, the vision community has recently started showing a great interest in developing efficient object detection techniques that can perform satisfactorily on night vision images, such as thermal images. Along the same lines, we introduce a computationally compact object detection algorithm for thermal images to address the above limitations. Our key aim is to develop an efficient and computationally compact algorithm for object detection in nighttime thermal images. The proposed object detection approach for thermal images is based on Depthwise Convolution (DC) in a DCNN to reduce the total number of parameters. Further, the use of loss functions such as Tversky and Intersection over Union (IoU) helps to improve the detection accuracy of the proposed method. In addition, we provide an extensive set of experiments to demonstrate the performance of our proposed approach on three publicly available datasets such as (1) FLIR validation [55], (2) KAIST testing dataset [26] and (3) MSOD dataset [57]. We further create a new dataset consisting of 998 images acquired in fully dark conditions to test the proposed approach's generalizability. We present both quantitative and qualitative results to support the applicability of our proposed approach. The object detection output obtained using the proposed method on sample images is depicted in Fig. 1, where one can inspect that the proposed detector can identify the objects in a thermal image. Therefore, the contributions of this work are highlighted as follows:

- We propose a computationally efficient object detection framework for nighttime situations. The compactness of the proposed object detector module is attributed to Depthwise Convolution (DC), resulting in approximately $3M$ parameters. The proposed detector is computationally efficient as compared to the other existing detector methods both in terms of parameters and inference time (see Fig. 7).
- Further, the generalizability of the proposed network is verified by conducting experiments on disjoint datasets unseen during training. The results are demonstrated on three publicly available datasets on which the proposed approach obtains better results than state-of-the-art methods.
- Owing to limited datasets captured using thermal cameras at nighttime, a new dataset has been constructed with 998 images, and this dataset has been specifically used to study the effectiveness and generalizability of the proposed approach on images acquired under darker conditions (sample images are shown in Fig. 1).

The rest of the paper is organized as follows. The next section thoroughly reviews the literature related to object detection approaches for thermal images. Then, in Section 3, we elaborate on the proposed approach for object detection for nighttime thermal images. Next, we report an extensive set of experiments performed to evaluate the effectiveness of the proposed approach in Section 4, and in Section 5, we conclude the work with a discussion on our

contributions.

2. Related works

Object detection is difficult in many common situations, such as nighttime illumination and bad weather conditions due to fog, rain, and dust [29, 34]. In such situations, most object detectors fail as they are trained on visible images which do not capture all variations [5, 39, 63]. As a result, object detection in thermal images for nighttime has rarely been investigated in the past [1, 6, 12].

With the breakthrough of CNNs in object detection, alternative and better object detection approaches [6, 35, 59] have started utilizing CNN-based approaches [13]. These methods can be categorized as region proposal-based and Single-Shot Detectors (SSDs), depending on the network forwarding pipeline [20, 38, 48, 50]. Fast R-CNN [20] applies selective search [60] to obtain region proposals, while Faster R-CNN [50] proposes to learn a Region Proposal Network (RPN) to accelerate the proposal generation process. In the most common region-based object detection methods [14, 20, 36, 50], category-related region proposals are assumed in the first phase, and then those proposals are refined and classified based on the CNN features using Region of Interest (RoI) pooling or align layer. The approaches mentioned above provide high detection accuracy, but the inference speed is usually slow due to the two-step mechanism. To speed up the discovery pipeline, region proposal generation is discarded in the region-free framework [38, 47–49]. To further reduce the computational need for proposal generation, single-shot approaches [38, 47–49] deploy a fixed set of predefined anchor boxes as proposals that directly predicts the category and offsets for each anchor box. Although these methods achieve state-of-the-art performance, such success hinges on the substantial amount of labelled training data, which requires a high labour cost. These methods can further overfit the training domain, making it difficult to generalize the approaches to many real-world scenarios. Despite real-time processing speed, detection accuracy is also sacrificed compared to best in class region-based approaches. Recently, Deconvolutional Single Shot Detector (DSSD) [18] and RetinaNet [36] have been proposed that provide competitive detection performance compared to the top of the region-based methods. Unfortunately, a very deep feature extractor (ResNet-101 [24]) is used in those methods, and due to the additional layers, these tend to have a computational overload.

The use of a deep learning model in object detection helps to obtain substantial detection efficacy on thermal images in terms of the various detection metrics (such as precision, recall, f_1 score, etc.) over the traditional methods; however, there are many limitations associated with those methods. Therefore, some key challenges based on reviewing the aforementioned existing object detection methods can be listed as provided here:

- As mentioned earlier, the performance of object detection on thermal image using deep learning methods is better over the traditional methods. Hence, many of aforementioned existing works [15, 16, 22, 37] achieve superior performance; however, they are limited to the images captured under the presence of sufficient natural or artificial lighting conditions. In the case of poor lighting sources and/or darker regions, they show inferior detection accuracy (see Fig. 6).
- Nighttime images often have multiple objects similar to daytime images and require multiple object detection. However, most of the works on night time images focus on detecting single object alone (i.e., people or car) [1–3, 6, 7, 12, 19, 21, 25, 27, 30, 33, 35, 37, 52] which limits their use in many applications for surveillance systems. In the proposed object detection approach, we perform multiple object detection simultaneously on thermal images such as person, car and bicycle (see Fig. 3, Fig. 5-Fig. 6).
- Most of the existing works [11, 28, 31, 32, 42, 53, 62] utilize pre-trained object detection models such as FRCNN [50], YOLO-v3 [48] and SSD [38] which is proven to be inefficient in the case of more diverse data (see Fig. 3-Fig. 6).
- Finally, the existing state-of-the-art methods for object detection [8, 9, 16, 18, 22, 22, 49, 50] employ a complex architecture that could be difficult to deploy in real-time applications. In the proposed method, we present an architecture which is computationally cheaper (utilizes only $3M$ training parameters) when compared to other existing object detection algorithms (see Fig. 7).

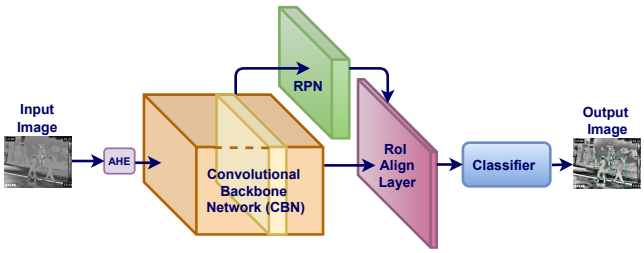


Figure 2. The block schematic of the proposed object detection algorithm.

3. Proposed Methodology

The proposed approach aims to implement a computationally compact and efficient object detection algorithm that can work with different illumination conditions at nighttime. The schematic representation of the proposed method is provided in Fig. 2. It mainly consists of five modules to perform specific tasks:

- Adaptive Histogram Equalization (AHE),
- Convolutional Backbone Network (CBN),
- Region Proposal Network (RPN),

Table 1. A detailed description of convolutional layers utilized in each module of the proposed object detection model. Here, DC(·) and G represent depthwise convolution and number of groups considered in DC, respectively. f, s and FC denote output feature maps, stride value and fully connected layer, respectively.

Modules	Layers	Discription
CBN	Conv_1	7×7, f = 32, s = 2
	Conv_2	DC(3×3, f = 32, G = 32) DC(3×3, f = 32, G = 32) DC(1×1, f = 64, G = 32) $\times 3$
	Conv_3	DC(3×3, f = 64, G = 32) DC(3×3, f = 64, G = 32) DC(1×1, f = 128, G = 32) $\times 3$
	Conv_4	DC(3×3, f = 128, G = 32) DC(3×3, f = 128, G = 32) DC(1×1, f = 256, G = 32) $\times 3$
	Conv_5	DC(3×3, f = 256, G = 32) DC(3×3, f = 256, G = 32) DC(1×1, f = 512, G = 32) $\times 3$
RPN	Conv_1	DC(3×3, f = 256, G = 128)
	Conv_2	conv(1×1, 6)
	Conv_3	conv(1×1, 12)
RoI	Conv_1	conv(1×1, 128) DC(3×3, f = 128, G = 64)
	Conv_2	conv(1×1, f = 128) DC(3×3, f = 128, G = 64)
	Conv_3	conv(1×1, f = 128) DC(3×3, f = 128, G = 64)
	Conv_4	conv(1×1, f = 128) DC(3×3, f = 128, G = 64)
Classifier	Conv_1	conv(7×7, 512 nodes)
	FC	linear, 256 nodes
	FC	linear, 3 nodes
Functions		Activation = ELU, Optimizer = Adam, Loss Function = Tversky_IoU

- Region of Interest (RoI) Align Layer, and
- Classifier (object detection).

Initially, to enhance the details present in the thermal image, we pass it through Adaptive Histogram Equalization (AHE) block. The salient features in the enhanced thermal images are then extracted using the Convolutional Backbone Network (CBN). Next, a shallow Region Proposal Network (RPN) is employed to propose the bounding boxes of objects. Further, the alignment of the features of interest available at the Region of Interest (RoI) is performed through the alignment layer, and then the classification and bounding-box regression are carried out as depicted in Fig. 2. A detailed description of each module employed in the proposed object detection model is elaborated below, and its architecture is tabulated in Table 1.

3.1. Adaptive Histogram Equalization (AHE)

In low-contrast imagery, usually, the important regions occupy a small portion of gray-level intensities, while uninteresting regions such as background and noise occupy the majority of gray-level. Thus, a large number of pixels and hence large peaks in the histogram correspond to those uninteresting regions. Adaptive Histogram Equalisation (AHE) in such a case can improve the contrast using local image data. The basic idea of AHE is to partition the given image into a grid of rectangular contextual sections and apply standard histogram equalisation to each of them. The number of contextual regions and their sizes are determined by the type of input image, with 8×8 (pixels) being the most frequent region size [43, 45, 46].

3.2. Convolutional Backbone Network (CBN)

The most popular algorithms such as FRCNN [50], MR-CNN [23], YOLO-v3 [49] and SSD [18] employ the ResNet and VGG structures for high-level feature extraction. Since these deep networks are composed of many hidden layers, the models above are computationally intensive (i.e., consume more than 40-50 M number of training parameters), and hence, they increase the overall complexity of the network. To overcome this deficiency, we present a computationally cheaper CBN module with approximately 85-90% lesser parameters than the abovementioned algorithms. This module takes an image and extracts high-level features from it. The number of layers and structure of this module are mentioned in Table 1. The use of Depthwise Convolution (DC) in the proposed CBN module further helps to reduce the overall computational complexity of the network. It divides the channels into G-groups and performs the convolution independently for each group. If the number of groups equals the number of channels, the grouped convolutions are reduced to depth-wise convolutions. The CBN module consists of five convolutional layers in which the size of the kernel and number of output channels generated from each layer are represented as $m \times n$, f in Table 1. All convolutional layers except the first layer utilize three DC layers. Here, a group of three DC layers is considered as one block and each block is recursively repeated a number of times. We repeat it three times (i.e., $\times 3$) to extract meaningful features available in the thermal image in the proposed CBN network.

3.3. Region Proposal Network (RPN)

This module predicts the presence of objects based on the region of the feature map from CBN. A small convolutional network is slid over the feature maps output by the last shared convolutional layer to generate region propos-

als. This small network takes as input of $n \times n$ spatial window of the input convolutional feature map. Each sliding window is mapped to a lower-dimensional feature. Hence, it works by passing a sliding window over the CNN feature map, and each window outputs l potential bounding boxes and scores for how good each of that boxes is expected to be. Then, the available features are fed into two sibling fully-connected layers, i.e., a box-regression layer (reg) and a box-classification layer (CLS) (see Table 1). Intuitively, objects in an image must fit certain common aspect ratios and sizes. For that, it requires some rectangular boxes that resemble the shapes of objects. Hence, it creates l such common aspect ratios as anchor boxes. Each such anchor box outputs one bounding box and score per position in the image. Hence, the classification layer has $2\times$ times the Number of classes (i.e., 6) output parameters that represent class name along with a score of each class, and the regression layer has $4\times$ Number of classes (i.e., 12) output parameters that describe the coordinates of the bounding box of each class.

3.4. Region of Interest (RoI) Alignment Layer

At this stage, another neural network with four convolutional layers is used, taking the proposed regions from the previous stage and fitting them into several specific areas in a feature map. A single block of convolutional layer consists of pair of simple and Depthwise Convolutions (DC). It scans those areas, assigns bounding boxes, and then predicts the score for each object in the feature map. The regions of the feature map selected by the RoI layer are slightly misaligned from the regions of the original image. Hence, it is adjusted and precisely aligned using the RoI Align layer that converts all the regions to the same shape.

3.5. Classifier

The fixed size feature maps obtained by RoI layer are passed through two Fully Connected (FC) layers followed by a convolutional layer Hence, 7×7 convolution with 512 nodes is applied to the backbone feature map of RoI to provide a feature map of single vector dimension. Here, first and second FC layers have 256 and 3 nodes, respectively. It is then fed to the Softmax classifier module to predict the class labels and bounding boxes.

3.6. Loss Functions

Further, the proposed method is trained on the combination of Tversky and Intersection over Union (IoU) losses inspired by earlier work [50]. The Tversky loss [51] adds weight to False Positive (FP) and False Negative (FN) with the help of a constant coefficient, and it calculates the similarity between two objects. Our choice of Tversky loss is further motivated to mitigate the false detection arising out of low lit images. Mathematically, the Tversky loss ($\ell_{tversky}$) can be formulated as,

$$\ell_{tversky} = \frac{|P \cap T|}{(|P| \cap |T|) + \alpha((1-T) * P) + \beta((1-P) * T)}. \quad (1)$$

Here, P and T represent predicted and target output images. α and β coefficient values are considered 0.7 and 0.3, respectively as suggested in [51]. Additionally, the IoU loss (ℓ_{IoU}) is calculated as the ratio between the overlap of the positive instances between two sets, and their mutual combined values which can be calculated as [61],

$$\ell_{IoU} = \frac{|P \cap T|}{|P| + |T| - |P \cap T|}. \quad (2)$$

Finally, the proposed object detection module is trained with ℓ_s loss function which is the combination of Tversky and IoU losses by exploiting the benefits of both approaches, and same can be represented as,

$$\ell_s = \ell_{tversky} + \ell_{IoU}. \quad (3)$$

4. Experimental Analysis

Numerous experiments have been performed on different datasets of night-vision thermal images to evaluate the performance of the proposed compact object detection model and the detailed description associated to this is depicted here. All experiments have been performed on a computer with Intel Xeon(R) $CPUE5-2620$ v4 processor @$2.10GHz \times 32$ running on a 128GB RAM and two $NVIDIA$ Quadro P5000 with $16GB\ GPUs$. Further, the proposed method is implemented in the PyTorch library.

The proposed object detection method is trained on FLIR dataset [55] which consists of visible and its corresponding thermal images with annotation of three different classes such as person, car and bicycle. Hence, it consists ground-truth of the objects present in the thermal images. It has 10,228 images and same is divided for training (8,862 images) and validation (1,366 images) purposes. Additionally, the training images are also augmented with random rotation of 0 or 90, random horizontal flipping and random cropping operations in order to avoid the problem of underfitting.

The proposed method is trained upto 4×10^5 number of iteration with batch size of 8 and it is optimized using Adam optimizer. Additionally, the learning rate and IoU threshold is set to 2×10^{-3} and 0.9, respectively.

4.1. Testing Details

The quantitative and qualitative evaluations of the proposed and other existing methods are presented by testing it on four different datasets: (1) FLIR validation [55] (2) KAIST testing dataset [26], and (3) MSOD dataset [57] and (4) our internal dataset.

The FLIR validation dataset contains 1366 thermal images utilized for testing in our work. It consists of annotations of car, bicycle and person classes. Additionally, the KAIST testing dataset consists of 32,770 number of image pairs, from that 2000 number of nighttime images are selected for testing purposes. It consists of details of one class: person. Further, the MSOD dataset contains 3,772 nighttime images; out of these, 300 images are randomly selected for testing. Similar to the FLIR dataset, it has annotations of those three classes. Moreover, to check the generalizability of the proposed method on real-time night situations, we have prepared our dataset and performed the testing of the proposed method on that dataset. This dataset includes 998 images that have been acquired under total darkness at nighttime without any artificial and/or natural light around the premises of interest. The FLIR E8-XT camera is used to capture the images under different weather conditions during the nighttime. This camera covers the spectral range of 7.5-13 μm with a spatial resolution of 320×240 pixels. In the internal dataset, there are two classes: person and car. Thus, the potential of the proposed method is validated on numerous datasets of having multiple classes on nighttime scenarios.

4.2. Ablation study

To show the effectiveness of various modules used in the proposed method, many experiments have been conducted in ablation study. It includes utilization of AHE, role of activation functions (i.e., ReLU, Leaky ReLU (LReLU), Parametric ReLU (PReLU) and Exponential Linear Unit (ELU)), importance of proposed loss function, optimizer and depthwise convolution in the architectural design. The detail description of each module is elaborated in the *supplementary material*.

4.3. Comparison with State-of-the-art Methods

To verify the efficacy of the proposed approach, the qualitative and quantitative assessments have been conducted and their detailed demonstration is presented in this section.

4.3.1 Quantitative fidelity

The proposed method is evaluated quantitatively in terms of various detection metrics such as Average Precision (AP), recall, f_1 score and mean of AP (mAP). The higher value of these measures represents better detection accuracy. Additionally, we also add the inference time taken by each method to obtain the detection results which must be as minimum as possible. Here, in order to understand the effectiveness of the proposed object detection module over the different state-of-the-art detector methods, we add its comparison in Table 2. Here, the different state-of-the-art detectors such as YOLO-v3 [49], SSD [18] and FRCNN [50] are used to show the effectiveness of the proposed object detection module. All the above mentioned existing methods have been re-trained on the training protocol of the proposed method and their detection scores are obtained and tabulated in Table 2.

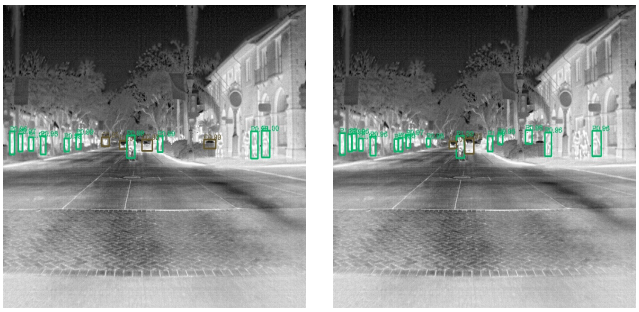
Table 2. The quantitative comparison of the proposed object detection model. The red color fonts indicate the highest value among all.

Metrics/Methods	YOLO-v3 [49]			SSD [18]			FRCNN [50]			Proposed		
	person	bicycle	car	person	bicycle	car	person	bicycle	car	person	bicycle	car
FLIR Validation Dataset : Person, Bicycle and Car Detection												
AP	19.44	32.91	36.29	25.36	18.54	58.44	14.41	0.21	44.05	43.72	40.77	72.70
recall	12.41	42.13	26.50	40.98	39.65	28.40	7.92	0.21	25.22	35.83	40.33	47.87
f1 score	21.66	41.69	39.21	24.48	41.04	32.08	14.62	0.42	39.72	40.16	46.12	62.31
mAP		29.55			34.11			19.56			52.39	
Inference Time/image (in sec.)		0.778132			0.231047			0.018554			0.0000013	
KAIST Dataset : Person Detection												
AP	28.12	-	-	30.56	-	-	27.65	-	-	59.43	-	-
recall	15.85	-	-	14.22	-	-	13.89	-	-	42.91	-	-
f1 score	27.05	-	-	21.08	-	-	20.74	-	-	55.88	-	-
mAP		28.12			30.56			27.65			59.43	
Inference Time/image (in sec.)		0.860113			0.239660			0.029388			0.0000016	
MSOD dataset : Person, Bicycle and Car Detection												
AP	6.94	0.15	3.75	11.50	0.27	9.64	26.84	-	8.46	63.94	0.74	4.98
recall	10.30	0.25	10.80	18.34	0.19	6.65	15.51	-	5.37	44.03	0.74	3.75
f1 score	13.89	0.47	9.01	22.62	1.46	10.89	26.55	-	9.97	57.70	1.47	7.09
mAP		3.62			7.29			11.76			23.22	
Inference Time/image (in sec.)		0.559803			0.229065			0.037591			0.00000037	
Our Internal Dataset: Person and Car detection												
AP	1.36	-	18.59	5.22	-	11.79	0.33	-	24.78	13.61	-	33.59
recall	16.91	-	32.54	32.48	-	21.71	2.77	-	21.91	53.20	-	35.49
f1 score	4.89	-	26.84	13.09	-	24.71	2.07	-	32.13	21.24	-	42.14
mAP		9.97			8.50			12.56			23.60	
Inference Time/image (in sec.)		0.473345			0.223567			0.013533			0.0000000750	

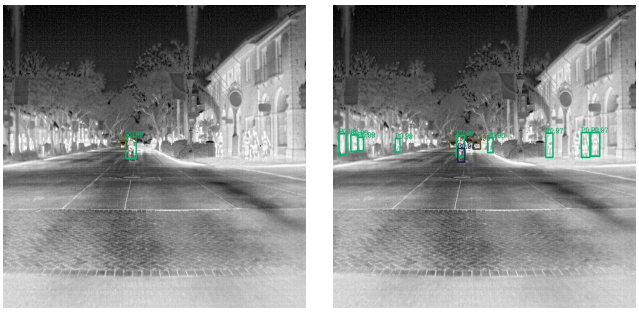
The FLIR validation and MSOD datasets contain annotations of common objects such as person, bicycle and car; hence, performance of the proposed algorithm along with the existing methods is verified by evaluating metrics on these three objects. While KAIST dataset includes annotation of person only. Therefore, its quantitative performance is measured on object of person. Further, we have prepared our internal dataset by annotating person and car. Thus, the detection accuracy of these two objects are measured in terms of different detection metrics. By inspecting Table 2, one can observe that the proposed object detection model performs better than the existing state-of-the-art object detection methods such as FRCNN, YOLO-v3 and SSD for most of the datasets. In case of FLIR validation dataset, recall scores for person and bicycle classes are better for SSD and YOLO-v3 models, respectively. Further, the car class for MSOD dataset is detected accurately by SSD method. Additionally, one can compare the inference time required by the proposed method which is optimum when compared to the other existing methods.

4.3.2 Visual fidelity

The visual fidelity of the proposed method has been quantified by comparing it with the existing state-of-the-art methods. In Fig. 3 - Fig. 6, we show the bounding boxes obtained on FLIR, KAIST, MSOD and our internal datasets,

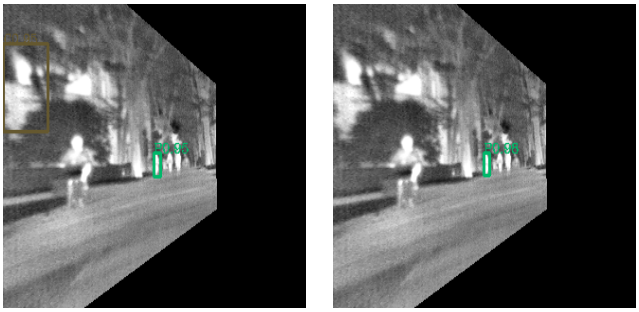


(a) YOLO-v3 [49] (b) SSD [18]

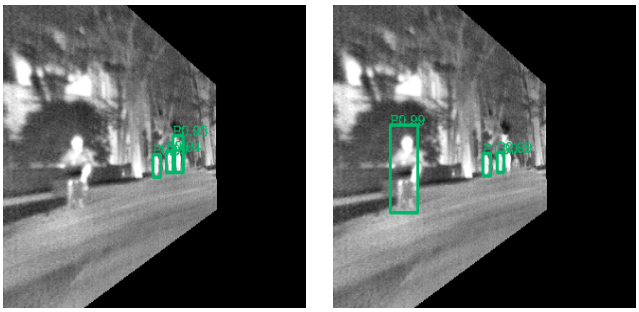


(c) FRCNN [50] (d) Proposed

Figure 3. The qualitative comparison of proposed method with existing state-of-the-art methods on FLIR dataset (zoom it for better visualization).

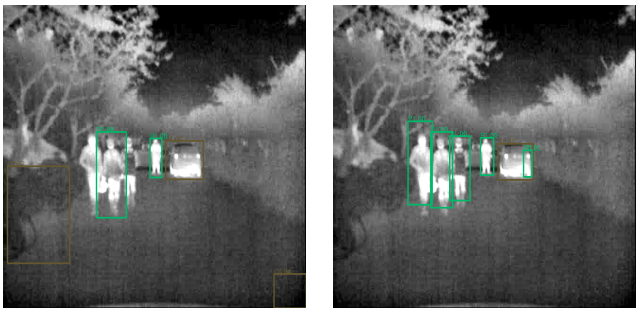


(a) YOLO-v3 [49] (b) SSD [18]

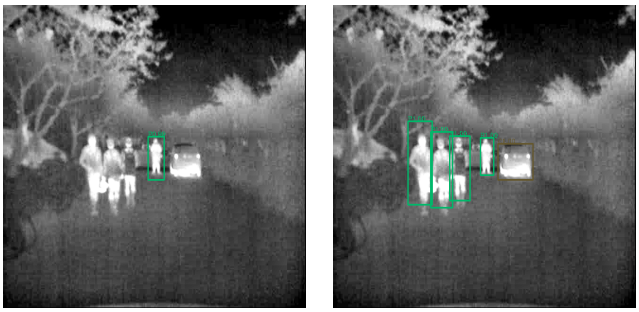


(c) FRCNN [50] (d) Proposed

Figure 5. The qualitative comparison of proposed method with existing state-of-the-art methods on MSOD dataset (zoom it for better visualization).

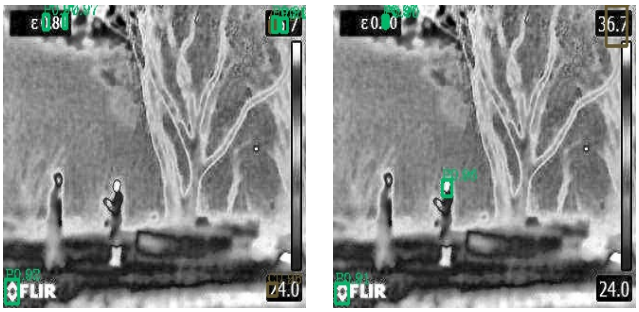


(a) YOLO-v3 [49] (b) SSD [18]

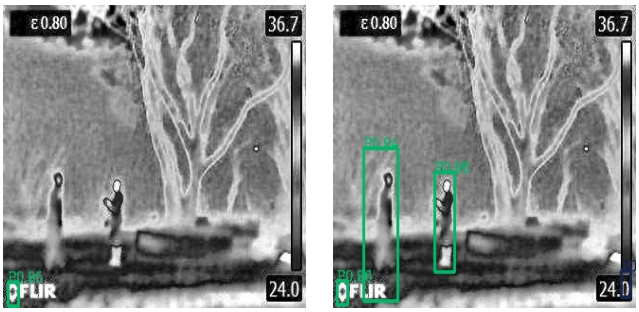


(c) FRCNN [50] (d) Proposed

Figure 4. The qualitative comparison of proposed method with existing state-of-the-art methods on KAIST dataset (zoom it for better visualization).



(a) YOLO-v3 [49] (b) SSD [18]



(c) FRCNN [50] (d) Proposed

Figure 6. The qualitative comparison of proposed method with existing state-of-the-art methods on our internal dataset (zoom it for better visualization).

respectively. The qualitative comparison of proposed object detection model (i.e., Fig. 3(d)-Fig. 6(d)) are compared with YOLO-v3 [49] (i.e., Fig. 3(a)-Fig. 6(a)), SSD [18] (i.e., Fig. 3(b)-Fig. 6(b)), and FRCNN [50] (i.e., Fig. 3(c)-Fig. 6(c)). For better visualisation, Fig. 3-Fig. 6 can be zoomed to see the differences among all these results.

The Fig. 3 shows the comparison of the proposed method along with other existing methods on FLIR dataset. It can be observed that the bicycle is not detected in most of the cases of existing methods (see Fig. 3(a-c)). However, the bicycle object is easily identified along with persons in the proposed method (see Fig. 3(d) by zooming it). It can be noted that the existing methods are unable to perform well on night-vision images. Further, the proposed method shows its superiority among the state-of-the-art methods. Hence, it can be deduced that the proposed method is helpful to obtain superior detection efficiency (i.e., more persons are detected with better accuracy).

The visual comparison of the proposed method along with the existing methods on the KAIST dataset is depicted in Fig. 4. This dataset contains very noisy images; hence, it is difficult to extract salient features. However, while comparing the above mentioned existing object detection methods, the proposed object detection model works well on noisy data (See Fig. 4(d)). The SSD method also detects all persons successfully (see Fig. 4(b)); However, it also results in the wrong detection of the front boundary of a car as a person. The proposed model improves the detection accuracy by providing accurate bounding boxes surrounding the object (i.e., person) that can be observed by zooming Fig. 4(d). Further, Fig. 5 displays the detection results on the MSOD dataset. Similar to the earlier dataset, the images in this dataset are also very noisy and low resolution. It can be observed that the state-of-the-art object detection methods such as YOLO-v3 (i.e., Fig. 5(a)), SSD (i.e., Fig. 5(b)), and FRCNN (i.e., Fig. 5(c)) have challenges in detecting objects in noisy and low-resolution images when compared to the proposed object detection model (i.e., Fig. 5(d)).

In addition to that, our internal dataset contains images captured in total darker situations, which are depicted in Fig. 6. Here, it is worth noting that the thermal cameras also provide insufficient information about objects due to nighttime illumination. Thus, it is difficult to retain sufficient information about the objects. Hence, most of the existing detection techniques fail to obtain better accuracy on this dataset (see Fig. 6(a-c)). However, the proposed method detects objects in nigh-time situations, and its performance proves its efficiency for real-time darker region scenarios. The additional detection results obtained using the proposed method on this dataset are also displayed in Fig. 1.

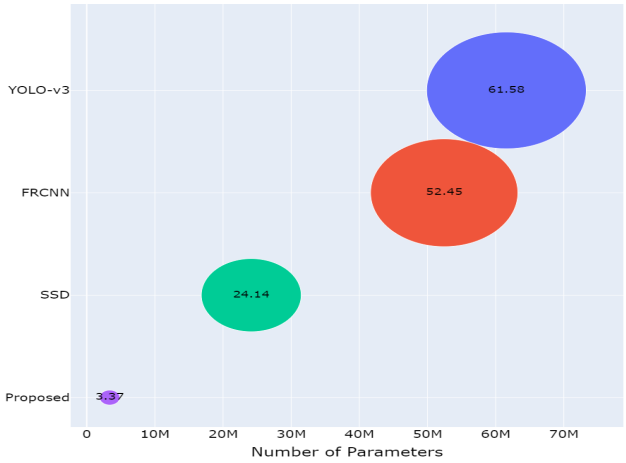


Figure 7. The computational complexity of different existing object detection methods in terms of number of parameters required to train their models.

4.3.3 Computational Complexity

The compactness of the proposed object detection module is attributed to $3M$ parameters only. The computational complexity of the different methods along with the proposed module is depicted in Fig. 7. One can deduce that the proposed object detection module needs the lowest number of training parameters (i.e., $3M$) as compared to the other existing state-of-the-art methods such as FRCNN [50], SSD [38], and YOLO-v3 [49] which are computationally intensive as they need $52M$, $24M$ and $61M$ number of parameters, respectively. The computational complexity of the proposed object detection module is, therefore, considerably less than the existing state-of-the-art methods without the loss of performance.

5. Conclusion

An efficient object detection approach using Depthwise DCNN for nighttime images was proposed in this work. The proposed network architecture utilizes AHE prior to the object detection task, which enhances the feature of the thermal input data. The design of the CBN module builds the proposed model efficient to extract features where the use of depthwise convolution makes it computationally inexpensive. The efficacy of the proposed method is validated on four different diverse datasets on night images. The qualitative and quantitative evaluations show that the proposed approach is generalizable and superior to the existing state-of-the-art object detection methods for nighttime scenarios. The proposed object detection model also reduces the computational complexity significantly compared to the other existing methods.

References

[1] Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification. *Pattern Recognition*, 48(6):1947 – 1960, 2015. 3

[2] Ala A. Alsanabani, Mohammed A. Ahmed, and Ahmad M. Al Smadi. Vehicle counting using detecting-tracking combinations: A comparative analysis. In *2020 The 4th International Conference on Video and Image Processing*, pages 48–54, 2020. 3

[3] Vinicius F Arruda, Thiago M Paixão, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. pages 1–8. IEEE, 2019. 3

[4] Jeonghyun Baek, Sungjun Hong, Jisu Kim, and Euntai Kim. Efficient pedestrian detection at nighttime using a thermal camera. *Sensors*, 17(8):1850, 2017. 2

[5] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*, pages 613–627. Springer, 2014. 3

[6] Sujoy Kumar Biswas and Peyman Milanfar. Linear support tensor machine with lsk channels: Pedestrian detection in thermal infrared images. *IEEE transactions on image processing*, 26(9):4229–4242, 2017. 3

[7] Raluca Brehar and Sergiu Nedevschi. Pedestrian detection in infrared images using hog, lbp, gradient magnitude and intensity feature channels. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1669–1674. IEEE, 2014. 3

[8] Philippe Burlina. Mrcnn: A stateful fast r-cnn. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3518–3523. IEEE, 2016. 3

[9] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. pages 8869–8878, 2020. 3

[10] Chuan Chong Chen. Attenuation of electromagnetic radiation by haze, fog, clouds, and rain. Technical report, RAND CORP SANTA MONICA CA, 1975. 2

[11] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. pages 3339–3348, 2018. 3

[12] Yunfan Chen, Han Xie, and Hyunchul Shin. Multi-layer fusion techniques using a cnn for multispectral pedestrian detection. *IET Computer Vision*, 12(8):1179–1187, 2018. 3

[13] Y. Choi, N. Kim, S. Hwang, and I. S. Kweon. Thermal image enhancement using convolutional neural network. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 223–230, Oct 2016. 3

[14] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 3

[15] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51(3):1244–1261, 2021. 3

[16] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[17] Christian Eggert, Stephan Brehm, Anton Winschel, Dan Zecha, and Rainer Lienhart. A closer look: Small object detection in faster r-cnn. In *2017 IEEE international conference on multimedia and expo (ICME)*, pages 421–426. IEEE, 2017. 1

[18] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 3, 4, 6, 7, 8

[19] Junfeng Ge, Yupin Luo, and Gyomei Tei. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):283–298, 2009. 2, 3

[20] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[21] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 3

[22] Tiantong Guo, Cong Phuoc Huynh, and Mashhour Solh. Domain-adaptive pedestrian detection in thermal images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1660–1664. IEEE, 2019. 3

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[25] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. Cnn-based thermal infrared person detection by domain adaptation. *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, 10643, 2018. 3

[26] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 2, 5

[27] Vijay John, Seiichi Mita, Zheng Liu, and Bin Qi. Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks. In *2015 14th IAPR international conference on machine vision applications (MVA)*, pages 246–249. IEEE, 2015. 2, 3

[28] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. pages 480–490, 2019. 3

[29] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. Domain adaptation for privacy-preserving pedestrian detection in thermal imagery. In *International*

Conference on Image Analysis and Processing, pages 203–213. Springer, 2019. 3

[30] Jong Hyun Kim, Ganbayar Batchuluun, and Kang Ryoung Park. Pedestrian detection based on faster r-cnn in nighttime by fusing deep convolutional features of successive images. *Expert Systems with Applications*, 114:15–33, 2018. 3

[31] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. pages 6092–6101, 2019. 3

[32] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. pages 12456–12465, 2019. 3

[33] Taehwan Kim and Sungho Kim. Pedestrian detection at night time in fir domain: Comprehensive study about temperature and brightness and new benchmark. *Pattern Recognition*, 79:44–54, 2018. 2, 3

[34] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*, 2018. 3

[35] Jianfu Li, Weiguo Gong, Weihong Li, and Xiaoying Liu. Robust pedestrian detection in thermal infrared imagery using the wavelet transform. *Infrared Physics & Technology*, 53(4):267–273, 2010. 3

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[37] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 3

[38] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3, 8

[39] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2019. 3

[40] William Maddern and Stephen Vidas. Towards robust night and day place recognition using visible and thermal imaging. In M Devy, A Kelly, T Peynot, and S Monteiro, editors, *Proceedings of the RSS 2012 Workshop: Beyond laser and vision: Alternative sensing techniques for robotic perception*, pages 1–6. University of Sydney, Australia, 2012. 2

[41] Sascha Mahlke, Diana Rösler, Katharina Seifert, Josef F Krems, and Manfred Thüring. Evaluation of six night vision enhancement systems: Qualitative and quantitative support for intelligent image processing. *Human Factors*, 49(3):518–531, 2007. 2

[42] Fatemeh Mirrashed, Vlad I Morariu, Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Domain adaptive object detection. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 323–330. IEEE, 2013. 3

[43] Shivangi Mishra, Ashish Dwivedi, and Badal Soni. Foggy image enhancement using improved histogram equalization and guided filter. In *Smart and Intelligent Systems*, pages 443–452. Springer, 2022. 4

[44] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. *International journal of computer vision*, 48(3):233–254, 2002. 2

[45] Monal Patel, Arvind Yadav, and Carlos Valderrama. Image enhancement and object recognition for night vision traffic surveillance. In *Soft Computing for Security Applications*, pages 733–748. Springer, 2022. 4

[46] Stephen M Pizer, E Philip Amburn, John D Austin, et al. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987. 4

[47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[48] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1, 3

[49] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 3, 4, 6, 7, 8

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3, 4, 5, 6, 7, 8

[51] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. pages 379–387. Springer, 2017. 5

[52] Marcel Sheeny, Andrew Wallace, Mehryar Emambakhsh, Sen Wang, and Barry Connor. Pol-lwir vehicle detection: Convolutional neural networks meet polarised infrared sensors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1247–1253, 2018. 3

[53] Zhiqiang Shen, Harsh Maheshwari, Weichen Yao, and Marios Savvides. Scl: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses. *arXiv preprint arXiv:1911.02559*, 2019. 3

[54] Tomasz Sosnowski, Grzegorz Bieszczad, and Henryk Madura. Image processing in thermal cameras. In *Advanced technologies in practical applications for national security*, pages 35–57. Springer, 2018. 2

[55] Arthur Stout, Kedar Madineni, Louis Tremblay, and Zachary Tane. The development of synthetic thermal image generation tools and training data at flir. In *Automatic Target Recognition XXIX*, volume 10988, page 1098814. International Society for Optics and Photonics, 2019. 2, 5

[56] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Learning to colorize infrared images. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 164–172. Springer, 2017. 2

[57] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 35–43, 2017. 2, 5

[58] Robby T Tan. Visibility in bad weather from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[59] Michael Teutsch, Thomas Muller, Marco Huber, and Jurgen Beyerer. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 209–216, 2014. 3

[60] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 3

[61] Floris van Beers, Arvid Lindström, Emmanuel Okafor, and Marco A Wiering. Deep neural networks with intersection over union loss for binary image segmentation. pages 438–445, 2019. 5

[62] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. pages 11724–11733, 2020. 3

[63] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European conference on computer vision*, pages 443–457. Springer, 2016. 3