International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.13, No.1 (May-23)

http://dx.doi.org/10.12785/ijcds/1301101

Pedestrian detection in thermal and color images using a new combination of saliency network and Faster R-CNN

Amlan Jyoti Das[1], Simantika Choudhury[2], Navajit Saikia[3] and Subhash Chandra Rajbongshi[2]

[1]*Obaforta India Pvt. Ltd., Guwahati, India*
[2]*Department of Electronics and Communication Engineering, Gauhati University, Guwahati, India*
[3]*Department of Electronics and Telecommunication Engineering, Assam Engineering College, Guwahati, India*

Received 29 May. 2022, Revised 9 May. 2023, Accepted 14 May. 2023, Published 30 May. 2023

Abstract: Pedestrian detection has been an important topic of research due to its increasing demand in the surveillance based applications. Thermal and color images are used to detect pedestrian under different illumination conditions. Recently people have used saliency maps to augment the images as an attention mechanism. This work employs different saliency based networks to evaluate their performances when used for augmentation and to determine the kind of saliency networks which derive better results in combination with Faster R-CNN. It also proposes an enhanced version of the KAIST multispectral dataset with corrected and extended set of annotations for both color and thermal channels separately. Pixel-level annotations for saliency networks are also proposed for thermal and color channels separately by using a subset of KAIST dataset. A detailed analysis of the saliency network performance is presented in terms of precision, recall, F-measure and mean absolute error. A new metric "region-level F-measure" is introduced to study the efficacy of saliency networks while used for augmentation. This work also presents the best combinations of saliency network and Faster R-CNN detector for both thermal and color channels maintaining a trade-off between detection performance and computation speed. The proposed detectors outperform existing detectors of similar type.

Keywords: Pedestrian detection, Saliency network, KAIST multispectral dataset, Faster R-CNN, PoolNet

1. Introduction

Pedestrian detection is one of the popular subjects for research because of its significant applications in the area of computer vision. It is an essential task for an intelligent system to add useful information for better understanding of the images or video footages. Even though there have been remarkable improvements in accuracy, pedestrian detection still has difficulties associated with it that need to be addressed. In many security and surveillance applications, one important requirement is that these intelligent systems perform optimally under different illumination conditions. It is often observed that the detector performance degrades for color images in night conditions and for thermal images in daylight conditions due to lack of sufficient information. This unmatched potential of color and thermal images has inspired researchers to work on combined information from color and thermal image pairs [1] or to use fusion architectures [2]. But, the drawback of these approaches is that the set up required for image acquisition is quite expensive. Also, combination of color-thermal pair is done by image registration which needs to be accurate. A slight misalignment in the image pairs may degrade the detector performance. To eliminate these limitations, people are also working on pedestrian detectors using individual channels in multi-spectral dataset [3]. This motivates us to use color and thermal images separately for detecting pedestrians under different illumination conditions.

To overcome the challenge of improving the performance of detecting pedestrians in color images during night and thermal images during day, the usage of saliency maps may be considered. Saliency object detection is a process to detect the salient objects in an image and to segment the exact shapes at the locations of the objects. The resultant binary mask is called the saliency map. These saliency maps are mixed with the thermal channel in implementing pedestrian detector for the first time in [3] recently. Authors in [3] hypothesize that thermal images augmented with their saliency maps may improve the performance of pedestrian detector. Two deep neural networks, R^3-Net and PiCANet are used in [3] to produce saliency maps. Driven by intuition, we also feel that color and thermal images combined with the respective saliency maps might improve performance of the detector. We therefore explore and analyse the performances of detectors by using saliency maps generated by different deep salient networks for color and thermal images separately. It is crucial to select a good performing deep salient network which produces good

E-mail address: amlan@spinaanalytica.com (*), csimantika@gauhati.ac.in (*), navajit.ete@aec.ac.in and subhash73@gauhati.ac.in,
**Authors contributed equally*

quality saliency maps with more true positive and less false positive regions. The quality of the saliency maps may in turn affect the final detection performance. To the best of our knowledge, there are no such literatures analysing the type of saliency networks that need to be used for augmentation in improving the overall detection performance. Here, it is intended to propose an efficient way of evaluation which estimates the quality of saliency maps created by these networks prior to its use in augmenting the data employed for training the final pedestrian detector.

In this work, we examine the detector performances using five deep saliency networks (BASNet [4], PFANet [5], R^3-Net [6], PiCANet [7] and PoolNet [8]) to augment thermal and color images separately. Faster R-CNN detectors trained using only thermal and color images separately are used as our baseline detectors. The KAIST multispectral pedestrian dataset [1] contains huge number of thermal-color image pairs and is widely used. We also use this dataset for experimentation after enhancing the annotations.

The rest of the paper is organized as follows: Section 2 presents a brief literature review on the related pedestrian detectors and salient object detectors. Section 3 presents the proposed methodology. The enhanced annotations and proposed pixel-level annotations of the dataset are introduced in Section 4. In Section 5, the implementation details regarding the saliency networks and Faster R-CNN are discussed. Section 6 presents the experimental analysis and performance comparison. Section 7 concludes the work.

2. Background

This section presents a brief overview of the existing literatures related to the present work.

Pedestrian detection: Traditional pedestrian detectors use hand-crafted features like HOG [9], Haar-like [10], LBP [11], ICF [12], ACF [13], etc. and classifiers like SVM [14], AdaBoost [15], etc. In recent times, deep learning based detectors have outperformed many of these traditional detectors. In deep learning based category, Zhang et al. [16] adopt Faster R-CNN for pedestrian detection. The Faster R-CNN is one of the popular object detection algorithms that basically contain two modules, region proposal network (RPN) and fast R-CNN network [17]. YOLO (You Only Look Once) [18] is another popular one-stage architecture using CNN which is Faster than RPN based architectures. In [19], a Faster R-CNN based multi-spectral deep neural network fuses the identical information from color and thermal images. Wagner et al. [2] analyze the potential of deep learning for multispectral pedestrian detection. In [20], authors investigate transformer based backbone architectures in pedestrian detection and concludes that CNN based backbones perform better in terms of large scale datasets and generalization. This paper proposes a fine-tuning approach for improvement in generalization for the later generations of pedestrian detectors. The use of multispectral images in deep learning based methods increased with the release of multispectral datasets containing color and thermal image pairs. These pairs have complementary potential for different illumination conditions. In [1], authors introduce KAIST multispectral pedestrian dataset which is large enough to be used in deep learning based techniques. Authors in [3] use thermal images from KAIST dataset and combine saliency masks to illuminate pedestrians for better performance of Faster R-CNN detector. However, the suitability of a saliency network for augmentation is not properly analysed in [3]. In [21], a network is proposed which introduces instance-level segmentation in pedestrian detection using thermal images. Authors in [22] proposes a two-stage method for improving detection in pedestrian. The first stage uses a generative augmentation technique and then a YOLOv3 based pedestrian detector is used for detection in thermal domain.

Salient object detection (SOD): Saliency detection methods can be categorized into traditional and deep learning based models. The traditional models are mostly based on low-level heuristics like color, texture, global and local contrast, background prior, etc. [23][24] [25][26]. Although these models allows performance in real time, they suffer in challenging environments like occlusion, cluttered environment, variation in illumination, etc. [27]. Authors in [27] present a comprehensive review on deep SOD networks. Deep learning based methods, which often use CNN, can overcome these limitations. U-Net based networks like BASNet, PiCANet, PoolNet, etc. have drawn attention in the domain of saliency object detection due to their capacity to construct detailed rich feature maps by building top-down pathways upon classification networks. In DSS [28], several short connections are added from the deeper-side outputs of CNN to the shallow ones which helps in locating salient regions. In RAS [29], reverse attention blocks are embedded to perform salient residual learning. The non-salient areas are emphasized by these blocks using deeper-level outputs' complement. In [30], InSPyReNet is proposed which is based on image pyramid. This network produces a saliency map with a strict image pyramid framework. Authors in [31] proposes TRACER which incorporates tracing modules to produce saliency maps with explicit edges. The tracing module comprises of a fast fourier transform which propagates fine edge information to the feature extraction. Another deep network C2S-Net [32] extracts saliency maps from contours and outputs fine predictions. In PFANet, both low-level and high-level features are used for the refinement of saliency maps. It also preserves edges to guide the network. R^3-Net is another network which uses residual refinement block (RRB) to refine saliency map by learning the residuals between coarse saliency map and ground truth.

From the above discussion, we can draw the following motivations:

i) Due to the challenges associated with multispectral dataset, performance of pedestrian detector using image pairs may degrade as discussed in Section 1. This motivates us to look into detectors using the individual channels of

multispectral dataset. When the channels are used separately to achieve better detection performance in different illumination conditions, improvement in the annotations may be an aspect of interest.

ii) The performance of deep saliency networks used for augmentation is not studied yet. This motivates us to analyze the suitability of different available saliency networks for augmentation and to conclude their suitability in pedestrian detection.

iii) Thermal images are less informative during daytime. In [3], only thermal images are used for detection by using saliency masks for augmentation. The performance of the detectors may also be analyzed for the color channel.

iv) Faster R-CNN and YOLO are two popular object detectors. The former performs better when objects are close to each other and smaller in size [18]. However, it is slower than YOLO which is a single stage detector. Due to better detection performance, Faster R-CNN is often considered in pedestrian detection.

This work considers five state-of-art deep saliency networks for analysis in object detection framework. These networks are selected based on their performances on ECSSD dataset [33] which is popular in the domain of SOD. Here, we consider three U-Net based networks which includes BASNet, PiCANet and PoolNet. PFANet is another network which is considered due to its good performance in SOD. Moreover, we select R^3-Net for the purpose of comparing the present work with existing literature. Based on above, the contributions of this work are as follows:

i) We propose an enhanced version of annotations separately for color and thermal images in the KAIST multispectral pedestrian dataset. The reasons for the proposed modifications are discussed in detail in Section 4. A set of pixel-level annotations for salient pedestrian dataset of color images is also proposed. The annotations for salient pedestrian dataset of thermal images introduced in [3] are also improved.

ii) To the best of our knowledge, no work has been reported to analyze which type of deep saliency networks can improve performance of a pedestrian detector. This analysis is important as the performance of the saliency network largely impacts the performance of Faster R-CNN. We present a detailed analysis by using the above said five saliency networks with Faster R-CNN for thermal images.

iii) The next contribution is the use of the combination of saliency maps and color images for data augmentation. To the best of our knowledge, this is the first work in this direction. Similar set of experimentations are performed in case of color images as mentioned in clause (ii) above.

iv) The quality of saliency maps are proposed to be enhanced by dense CRF technique. A new metric called "region-level F-measure" is also proposed to check the quality of a saliency network which in turn helps to select a suitable saliency network for augmentation without testing them with the final Faster R-CNN detector.

v) A new combination of deep saliency network and Faster R-CNN is proposed for thermal images which outperforms the existing similar detectors. We also propose a new combination for color images which performs the best based on experimentations.

3. Proposed methodology

The proposed methodology uses the color and thermal images separately for training and testing. Figure 1 shows the various steps involved which are discussed below.

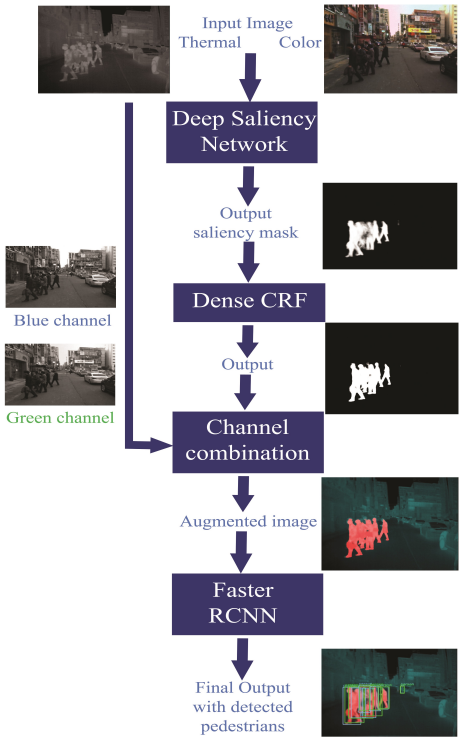


Figure 1. System block diagram

Saliency mask creation: Saliency maps are used to assist the final detector (Faster R-CNN here) with an attention mechanism and hence it may not be required to cover the whole pedestrian by the mask. We, therefore, hypothesize that it is more important for a saliency detector not to provide the Faster R-CNN with misclassified regions than to provide masks with proper shapes. Given an input image, it is fed to the trained deep saliency network to derive a saliency map containing binary masks of the pedestrians in the image.

Dense CRF: The post-processing process, dense/ fully-connected conditional random field (CRF) [34], helps to enhance the spatial coherence of the binary masks derived

from the deep saliency network. The output binary masks are then up-sampled to the size of the original images.

Channel combination: The enhanced saliency map is used to augment the input image by using two channels of the input image (here, green and blue channels) and replacing the third (red) channel with the saliency mask. This augmented image contains texture information as two channels in the input image are retained and also illuminates the detected saliency regions by the deep saliency network. The channel combination helps the Faster R-CNN detector not to solely depend on the masked regions but also to learn from texture information.

Faster R-CNN: The final stage of the proposed methodology is the Faster R-CNN network which detects the pedestrians in the augmented image.

4. Proposed annotations for KAIST multispectral pedestrian dataset

This section provides the details of the proposed modifications in the KAIST multispectral dataset annotations. It also introduces the proposed salient pedestrian dataset for color and thermal images. These proposed sets of annotations address the limitations in the previous annotation sets and also consider more challenging conditions like occlusion, etc.

A. Annotations for pedestrian detector

We filter the original KAIST multispectral pedestrian dataset as done in [1] to extract 7601 training images containing pedestrians and 2252 testing images out of 95,328 frames. In the original annotations, the pedestrians are annotated in such a way that each pedestrian is accommodated within the same bounding box region for both color and thermal channels. In [35], it is observed that problems of imprecise localization, misaligned regions and misclassification are present in these annotations and accordingly a new annotated "sanitized dataset" is introduced. However, these annotations are labelled with the objective to use with models that apply multispectral image pairs. Hence, there is scope to improve the annotations in the sanitized dataset for designing detection systems for color and thermal images separately. We, therefore, address the following issues which also motivates us to present a set of "enhanced annotations". We use the LabelImg annotation tool [36] to label the pedestrians.

i) In many of the original KAIST annotations, bounding box is large as compared to the pedestrian and the region of interest is not well-centred in the bounding box, especially at the corners of the images as shown in Figure 2a and Figure 2e. The unwanted regions other than pedestrians are also covered in the bounding box which may lead to misclassification. Although these problems are addressed in the sanitized annotations, but, total correction could not be achieved due to the intended use of the image pairs with fusion architectures. Hence, we attempt to create well-centred and fitted bounding boxes around the pedestrians to present separate sets of annotations for color and thermal channels.

TABLE I. Comparison of pedestrian counts in different annotation versions of the KAIST multispectral dataset

Annotations of KAIST dataset		Original [1]	Sanitized [35]	Our version
Thermal training set	**Day**	12265	14317	14571
	Night	9172	9992	10147
	Total	21392	24309	24718
Color training set	**Day**	12265	14317	14491
	Night	9172	9992	10037
	Total	21392	24309	24528
Thermal testing set	**Day**	2003	2003	2688
	Night	754	754	1009
	Total	2757	2757	3697
Color testing set	**Day**	2003	2003	2325
	Night	754	754	952
	Total	2757	2757	3277

ii) There are many missing annotations which are clearly visible in the image pairs. Some of these are due to absence of corresponding pedestrian in one image of the image pair due to the effect of parallax. As the improved version provides same annotations for both the channels, these pedestrians are also ignored there. These conditions are shown in Figure 2b and Figure 2f. Since, we aim to provide separate annotations for the two channels, those pedestrians are also considered for annotation in the individual channels.

iii) In the sanitized annotations, the cases of misaligned bounding boxes in the image pairs are handled by evaluating the Intersection of Union (IoU) between the bounding boxes separately created for thermal and color image pair. When IoU<0.5, the pedestrian is ignored during training resulting in omission of a significant amount of annotations. The count of instances in these annotations are presented in Table I. As this work attempts separate annotations for color and thermal channels, these annotations are considered individually.

iv) It is observed that in some images a significant part of a few pedestrians are not covered by the bounding boxes as in Figure 2c and Figure 2g. These bounding boxes are also appropriately fitted in this work.

v) There are also image pairs in the dataset where a person behind a speeding car is visible in one channel but not visible in the other due to mismatch in temporal information. This problem is shown in Figure 2d and Figure 2h. Such pedestrians are also annotated here separately for color and thermal images.

B. Annotations for salient pedestrian detector

SOD requires pixel-level annotations of salient objects as binary masks. The popular salient object datasets contain

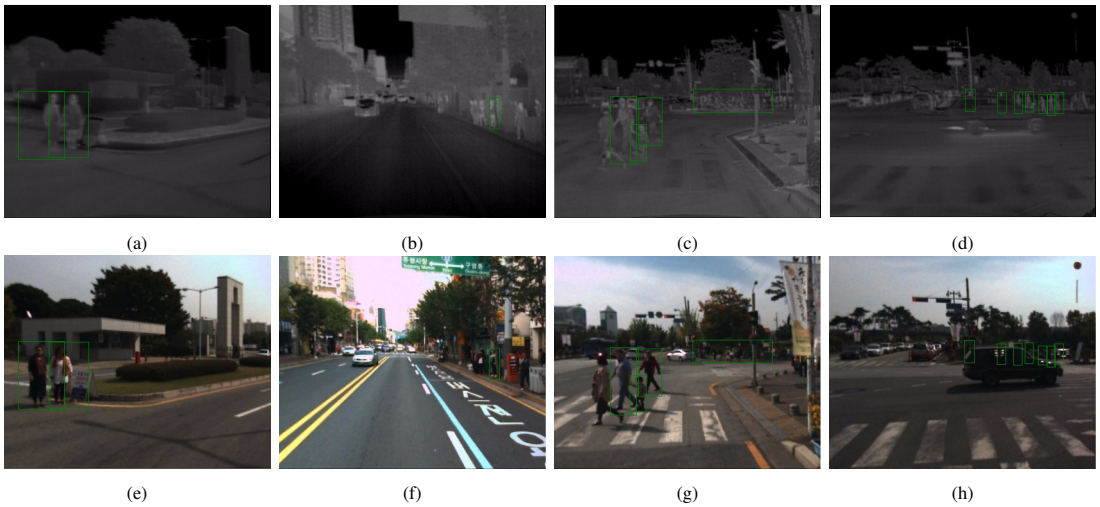


Figure 2. Examples of problematic annotations in thermal (top) and color (bottom) image pairs in KAIST multispectral pedestrian dataset: (2a,2e) improper bounding box, (2b,2f) missing annotations, (2c,2g) partial bounding box, (2d,2h) difference in acquisition rate

annotations of different salient object classes. These datasets can not be used here as the present work considers only pedestrian as salient object. In [3], a salient pedestrian dataset for thermal image is introduced by annotating a subset of the KAIST Multispectral pedestrian dataset. It contains 1702 training images (913 day and 789 night images) and 362 testing images (193 day and 169 night images). However, the masks for this dataset have two limitations and we propose to introduce an enhanced version of those as follows:

i) The masks do not fit properly with the pedestrian boundaries and hence their shapes do not resemble with pedestrian in many cases as shown in Figure 3a. In the enhanced version, the masks are created with more accurate shapes and smooth boundaries (Figure 3c).

ii) There are some missing annotations, particularly when the pedestrian size is small as shown in Figure 3b. The enhanced version also considers those pedestrians so that the saliency detector can detect the distant pedestrians as well (Figure 3d).

Again, in [3], only thermal images are used to generate binary masks as the saliency detection is done on thermal images. Due to misalignment in image pairs in the KAIST dataset as discussed in Section 4-A, the masks derived from the thermal images can not be used for the color channel. As the present work aims to consider color channels alongside thermal channels in saliency detection, we also create saliency masks separately for the color images corresponding to the 1702 training and 362 testing thermal images. Binary masks of the pedestrians are drawn by using the VGG image annotator [37]. This work uses the enhanced version of masks for saliency detection. The derived enhanced versions of the saliency masks contain a total of 5276 and 5302 instances in thermal and color training images respectively corresponding to 4170 instances in the dataset presented in [3]. Our versions of testing saliency masks contain 1118 and 1136 instances in thermal and color images respectively corresponding to 1029 instances in the saliency masks used in [3].

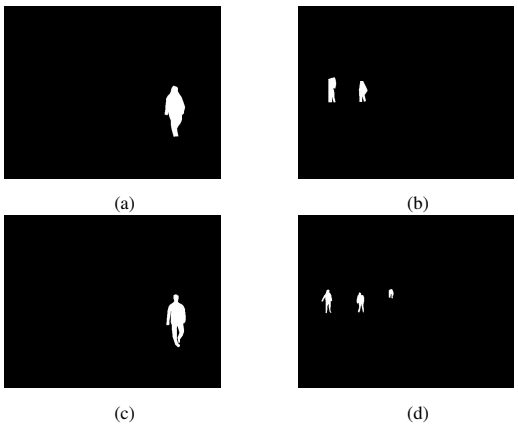


Figure 3. Examples of binary masks in salient pedestrian dataset (top) provided by [3] and (bottom) our modified versions: (3a) improper shape, (3b) missing annotations, (3c, 3d) modified annotations

5. Implementation Details

This section presents the parameters involved with the implementations of saliency networks and Faster R-CNN.

A. *Deep saliency networks*

The five deep saliency networks considered in this work for data augmentation are trained for both color and thermal images separately by using the saliency masks described in Section 3. The deep saliency networks are implemented by using the same architectures as per the works introduced in [4][5][6][7][8]. All these models are trained and tested on NVIDIA GeForce GTX 1070 GPU with 8 GB memory. All

parameters in the implementations are based on empirical analysis and the best combination parameters are considered as follows.

R^3-Net: ResNeXt [38] is used here as the backbone of the feature extraction network. It is then optimized with stochastic gradient descent (SGD) and the momentum is set to 0.9. For training, the learning rate is kept as 0.001 and the weight decay as 0.0005 for 15000 iterations by using a batch size of 6.

PiCANet: The backbone architecture here is VGG-16 [39]. Here also, SGD optimizer is used with momentum of 0.9. We initially train the decoder with learning rate 0.01 and the encoder with 0.001 with a step size of 7000 for 16 epochs. They are then trained with learning rates decayed by a factor of 0.1 for another 16 epochs. The entire set-up of PiCANet is trained using batch size 2 and weight decay 0.0005. As this implementation generates saliency maps of size 224×224, resizing of the maps to the original size of the image is done by using bilinear interpolation [40].

BASNet: ResNet-34 [41] is the backbone architecture used here. It is optimized by using Adam optimizer with momentums (0.9, 0.999), learning rate 0.001 and zero weight decay. The network is trained with a batch size of 4 for 32,000 iterations. The saliency maps generated are of size 256×256, which are then resized to the input image size by using bilinear interpolation.

PFANet: Here, VGG-16 is used as backbone of the feature extraction network which is pretrained on ImageNet [42]. It uses SGD optimizer with a learning rate of 0.001 and a batch size of 10 for 150 epochs.

PoolNet: PoolNet also uses VGG-16 as backbone architecture. It is trained for 18 epochs with a batch size of 1. It uses Adam optimizer with weight decay 0.0005 and learning rate 0.00005 initially. After 15 epochs, the learning rate is reduced by a factor of 10.

The testing parameters for each networks are kept same as the training parameters. The output saliency map is post processed with dense CRF. The trained deep saliency network followed by dense CRF is used to augment the 7601 training and 2252 testing images (as mentioned in Section 3) for Faster R-CNN.

B. Faster RCNN

Faster R-CNN detector is used with VGG-16 as backbone architecture. The training of Faster R-CNN is done for a) only thermal images, b) augmented thermal images, c) only color images and d) augmented color images. In VGG-16, the first two convolutional layers are fixed and the rest of the model is tuned by using SGD optimizer with a momentum of 0.9. For training the network, the parameters that are taken into consideration are as follows: the anchor scales used are 0.05, 0.1, 0.25, 0.5, 0.75, 1, 2, 3, 4 and the anchor ratios are 0.5, 1, 2. The learning rate is set at 0.001. These values are chosen by empirical testing. The network is trained initially with the weights pretrained on ImageNet dataset [42] for 6 epochs. For testing, the parameters used are same as those used during training. We train the models in a hardware using NVIDIA GeForce GTX 1070 GPU with 8 GB VRAM.

6. Experimental results and analysis

To evaluate performance of the detectors, we use metrics called log-average miss-rate (LAMR) [43] and mean average precision (mAP) [16]. Miss-rate (MR) provides the proportion of false negatives among the ground truths and is defined as:

$$Miss\ rate = 1 - recall = \frac{FN}{TP + FN} = \frac{FN}{all\ ground\ truths} \tag{1}$$

LAMR is used to measure the detector performance, computed by averaging miss rate at 9 False Positive Per Image (FPPI) rates spaced evenly in logarithmic space in range 10^{-2} to 10^{0}. Average Precision (AP) is calculated individually for each class. These AP values are averaged altogether to obtain mAP. It is given as:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \quad for\ n\ classes \tag{2}$$

Performance of a detector is better when LAMR is lower and mAP is higher. We also plot MR versus FPPI [43] and precision versus recall (PR). If MR vs. FPPI plots are linear in range of 10^{-2} to 10^{0}, then LAMR is similar to the performance at 10^{-1} FPPI and gives a stable and informative assessment of evaluation performance. These studies are performed on day and night images separately for both the channels.

Two popular performance metrics used to analyse saliency networks are F-measure (F_β) and mean absolute error (MAE) [27]. F-measure [28] gives a measurement for the pixel-level performance of the model. Here, the non-negative weight β is set to 0.3 as suggested by previous works [27]. A smaller value of β is considered to emphasize more on precision over recall [27]. It is given as:

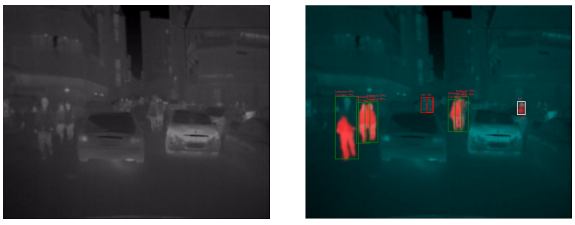
$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall} \tag{3}$$

MAE [28] compares the generated saliency map with ground truth by computing pixel-wise difference. The performance of a saliency network is considered to be good when F-measure is higher and MAE is lower. It is given as:

$$MAE = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\hat{S}(i,j) - \hat{Z}(i,j)\right| \tag{4}$$

Region-level F-measure: The F-measure metric considers the quality of the saliency map at pixel-level and does not consider the count of pedestrians for which the masks are created. As saliency masks are used as an attention

mechanism for the Faster R-CNN, it is accordingly hypothesized in Section 3 that covering of the whole pedestrian by its mask is not always necessary. Hence, we introduce 'region-level F-measure' in terms of true positive (TP), false positive (FP) and false negative (FN) to study the performance of saliency network. It takes into account the quantity (number) of pedestrians covered by the saliency masks above an empirically selected threshold. We propose to consider a bounding box as TP when more than 50% of a pedestrian mask is included in the bounding box and the included part covers more than 30% of the bounding box. If the first condition is not met, the associated mask is considered as FP and bounding box is considered as FN. If the first condition is met and the second condition is not satisfied, then the mask is not considered as FP and is ignored. If no mask is created inside a bounding box, then the bounding box is considered as FN. If a mask is created outside the bounding box, it is treated as FP. Figure 4 demonstrates these cases with an example. The region-level F-measure value is evaluated by using the precision and recall values obtained with these TP, FP and FN counts. Figure 5 shows a flow diagram for evaluating TP, FP and FN which in turn is used for computing region-level F-measure. This implies that using the region-level F-measure does not require to create saliency ground truths for the test sets mentioned in Section 4-A.



(a) (b)

Figure 4. Example of TP, FP and FN for region-level F-measure calculation: (4a) original image, (4b) saliency output: TP(green BB), FP(white BB) and FN(red BB)

A. *Performance of the deep saliency networks*

Here, we first demonstrate the usefulness of a saliency network in the proposed pedestrian detection model presented in Section 3 by analyzing its performances in terms of F-measure and MAE at pixel-level. The performances of the saliency networks considered in this work are evaluated on the salient pedestrian datasets for color and thermal test images introduced in Section 4-B. Table II and Table III presents the observed F-measure values at pixel-level and MAE values respectively. It may be noted that PoolNet performs best in all the cases except that BASNet gives best result in case of thermal night images. The better performance of BASNet with thermal night images may be due to well defined pedestrians and minimal cluttered background in these images. It is observed that although the masks are of proper shape, there are lots of FP regions. R^3-Net yields fewer FP regions though the masks here are not as proper as BASNet. Again, the quality of masks

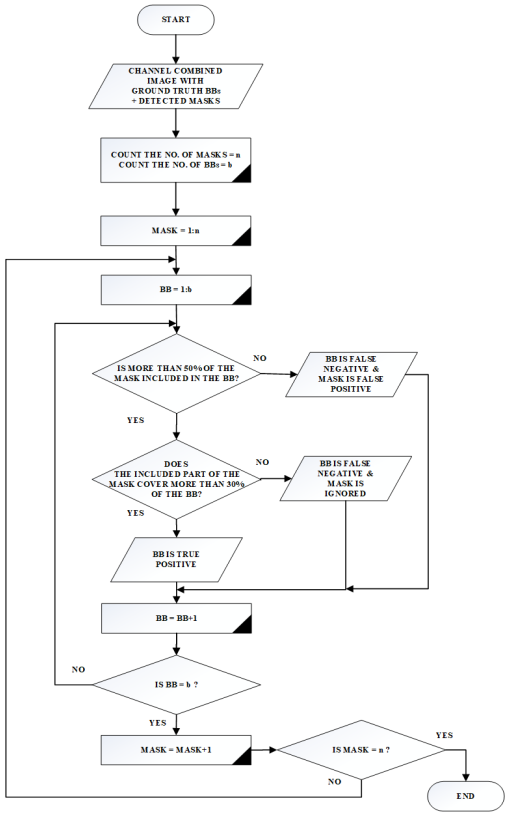


Figure 5. Flow diagram for evaluating TP, FP and FN for region-level F-measure (BB: Bounding box)

produced by PiCANet and PFANet are not proper and also these networks give more number of FPs. Table II does not provide insights to these cases. Therefore, we have proposed to use the region-level F-measure as discussed above. It may also be noted that PoolNet outperforms R^3-Net in all the cases. The best MAE values are also observed for PoolNet.

A saliency network will have better values of F-measure and MAE if saliency masks are created properly and number of misclassified regions is less. However, as hypothesized, it is more important in our case that the saliency network delivers masks which sufficiently cover the pedestrians than the quality of the mask to help Faster R-CNN to pay attention to the probable salient objects. Also, the number of FP regions is more critical than FN regions here as these regions make Faster R-CNN to pay attention to non-pedestrian regions. The observed region-level F-measure values for the saliency networks when tested on thermal and color images are presented in Table IV. From the table, it is evident that PoolNet has outperformed all the other saliency networks for both channels. It may be noted that R^3-Net has better precision values in all cases compared to PoolNet. This is due to the fact that R^3-Net yields fewer FPs than PoolNet. But, unlike recall, precision does not indicate the efficiency of the network in terms of TP count as it does not depend on FN count. The recall values for R^3-Net are

inferior to the corresponding values for PoolNet since the TP count is smaller than PoolNet by considerable margins. This in turn improves the F-measure of PoolNet over R^3-Net. It may be seen that BASNet does not perform good due to more number of FP regions produced by the network deriving a lower value of precision. PiCANet and PFANet perform marginally lower than R^3-Net.

B. Performance of detectors with thermal images

The Faster R-CNN network when trained with only thermal images in the training set acts as the baseline detector. To study the effectiveness of using saliency network in augmentation, the R-CNN network is also trained separately on the augmented thermal images obtained by using each of the five deep saliency networks. The performances of these six detection models are studied in terms of LAMR and mAP by using thermal test images with their enhanced annotations. Table V presents the observed LAMR and mAP values. The MR vs. FPPI and precision vs. recall plots for the detectors are also presented in Figure 6, 7 and Figure 8, 9 respectively for reference. Table VII shows MR values corresponding to some representative FPPI values.

For thermal images, it may be noted from Table V that all the detectors perform better with night images. It is also observed that only the detectors using R^3-Net and PoolNet perform better than the baseline detector. For the detector using R^3-Net, LAMR improves by 1.6% on day and 1.3% on night images. The corresponding improvements in mAP are 0.9% and 0.1%. For the detector using PoolNet, the LAMR is better by 2.8% on day and 4.3% on night images in comparison to baseline detector. The corresponding improvements are 1.2% and 3.2% over the detector using R^3-Net. Likewise, the mAP values are also better for the detector using PoolNet. Hence, PoolNet is the best performing saliency network here.

However, some detectors using augmented images show degraded performance than the baseline detector. These detectors are the ones using BASNet, PiCANet and PFANet. The detector using BASNet performs the worst and shows a LAMR degradation of 10% and 6.4% for day and night images respectively. The corresponding reduction in mAP values are 7% and 3.6%. The detectors using PFANet and PiCANet saliency networks have also degraded performances, but they perform very closely to the baseline detector.

In a nutshell, it may be noted that proper selection of saliency network is important for improving the performance of the Faster R-CNN. From Figure 6, 7 and Table VII, it can be observed that PoolNet based detector has performed better than all the detectors for most of the FPPI values which is also observed in Table V in terms of LAMR. However, for lower values of FPPI close to 0.01, the performance of PoolNet based detector is inferior to some of the detectors which can also be observed in Figure 6 and Figure 7. Since LAMR gives a stable and informative evaluation of performance [43], it is considered as the primary metric for assessment of the detectors in this work.

C. Performance of detectors with color images

Similar kind of evaluation is performed for color images as done for thermal images in Section 6-B. Similar to the previous case, here also separate Faster R-CNN based detectors are trained on the augmented set of images using different saliency networks. The detector trained with the original color images acts as the baseline detector. We provide the results of detection using our proposed systems in Table VI. The MR vs. FPPI and precision vs. recall plots are also shown in Figure 10, 11 and Figure 12, 13 respectively.

From the Table VI, it can be seen that in case of color images the performance of all the detectors are better in day light conditions than night conditions due to low contrast night images where pedestrians are hardly visible. The baseline detector has LAMR of 51.2% for day images and 67% for night images. It can also be observed that the detectors using R^3-Net and PoolNet saliency networks perform better than the baseline detector for color images as well. The performance improvement for the detector using R^3-Net is 0.7% for day images and 3% for night images in terms of LAMR. The detector using PoolNet performs the best among all the detectors. Here, the percentage improvement from the baseline detector in terms of LAMR are 1.8% and 4.2% for day and night images respectively. Similarly, improvements in performance are also observed in terms of mAP. These improvements for R^3-Net based detector are 2.6% and 3.3% and PoolNet based detector are 6% and 7% for day and night images respectively. It clearly indicates that there is a reduction in FP count using R^3-Net and PoolNet based detectors.

However, there are certain detectors using augmented color images also that perform inferior to the baseline detector. The detector using BASNet for augmentation performs the worst in case of color images similar to the case of thermal images. It can be seen that this detector performs 8.6% and 11.5% inferior to the baseline detector in terms of LAMR for day and night images respectively. The mAP values also degrade largely from the corresponding values for the baseline detector which indicates increased number of FPs. Similarly, PiCANet and PFANet based detectors also perform poorly than the baseline detector. These two detectors yield performances close to the baseline detector for night images, but suffer for day images. The observed LAMR degradations are 7.4% and 4.2% for PiCANet and PFANet based detectors respectively.

To summarize, we can say that a similar trend of LAMR and mAP performances of the detectors has been observed using color images as observed for thermal images. The PoolNet based detector performs the best throughout all the FPPI values for day images which can be observed from Figure 10 and Table VII. For night images, it performs slightly inferior to R^3-Net based detector for lower values of

TABLE II. Performance (pixel-level) of saliency networks on proposed salient pedestrian datasets in terms of Precision (P), Recall (R) and F-measure (F)

Saliency Network	Thermal						Color					
	Day			Night			Day			Night		
	P	R	F	P	R	F	P	R	F	P	R	F
BASNet	0.666	**0.475**	0.610	**0.788**	**0.563**	**0.722**	0.562	0369	0.502	0.365	0.168	0.287
PiCANet	0.645	0.335	0.532	0.773	0.429	0.652	0.576	0.315	0.484	0.516	0.256	0.418
PFANet	0.554	0.241	0.426	0.540	0.382	0.493	0.605	0.219	0.430	0.592	0.116	0.304
R^3-Net	0.697	0.418	0.604	0.771	0.500	0.685	0.617	0.364	0.531	**0.664**	0.207	0.441
PoolNet	**0.768**	0.470	**0.670**	0.776	0.513	0.694	**0.631**	**0.436**	**0.572**	0.593	**0.286**	**0.476**

TABLE III. MAE values of saliency networks

Saliency Network	Thermal		Color	
	Day	Night	Day	Night
BASNet	0.0046	**0.0038**	0.0069	0.0085
PiCANet	0.0063	0.0061	0.0073	0.0078
PFANet	0.0119	0.0153	0.0121	0.0132
R^3-Net	0.0050	0.0049	0.0060	0.0068
PoolNet	**0.0044**	**0.0038**	**0.0059**	**0.0065**

FPPI close to 0.01 FPPI. From Figure 10, 11 and Figure 12, 13, the BASNet based detector performs poorly for night images which is due to large count of FNs and FPs.

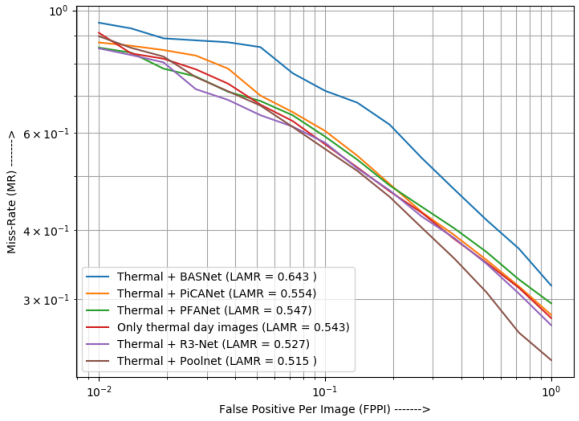


Figure 6. Miss-rate vs. FPPI plots for thermal images (Day)

D. Discussion and comparison

From the results shown in Table V and Table VI and also as discussed above, when saliency masks are fused with thermal and color images by channel replacement method, we have observed that the PoolNet based detector derives the best performance. Again, the performances of the five saliency networks have been analysed for day and night images and found that PoolNet gives the highest region-level F-measure value as presented in Table VII. The region-level F-measure is used to verify the hypothesis

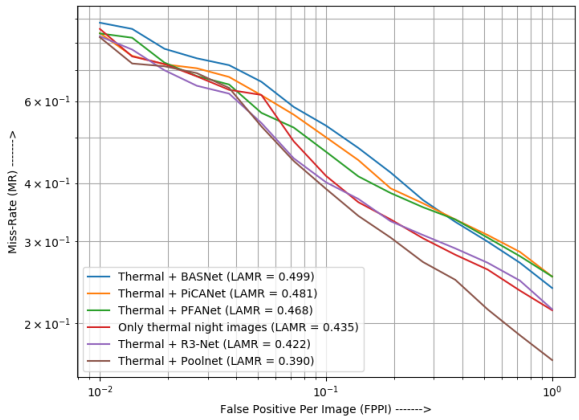


Figure 7. Miss-rate vs. FPPI plots for thermal images (Night)

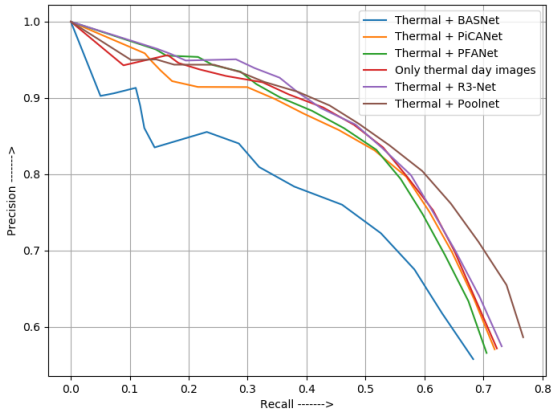


Figure 8. Precision vs. Recall plots for thermal images (Day)

mentioned in Section 3 that it is more important for a saliency network to provide less misclassified information rather than providing quality of the masks. The PoolNet saliency maps do not only highlight the correct salient objects but also maintain their sharp boundaries in almost

TABLE IV. Performance (region-level) of saliency networks on augmented images using enhanced annotations in terms of Precision (P), Recall (R) and region-level F-measure (R-F)

Saliency Network	Thermal						Color					
	Day			Night			Day			Night		
	P	R	R-F	P	R	R-F	P	R	R-F	P	R	R-F
BASNet	0.688	0.654	0.669	0.794	0.857	0.823	0.671	0.807	0.732	0.345	0.473	0.398
PiCANet	0.863	0.585	0.696	0.934	0.739	0.824	0.823	0.685	0.747	0.718	0.497	0.586
PFANet	0.823	0.611	0.701	0.809	0.844	0.825	0.812	0.709	0.756	0.753	0.484	0.588
R^3-Net	**0.928**	0.566	0.702	**0.972**	0.765	0.856	**0.960**	0.701	0.809	**0.911**	0.455	0.606
PoolNet	0.920	**0.774**	**0.840**	0.951	**0.899**	**0.923**	0.937	**0.896**	**0.915**	0.822	**0.645**	**0.722**

TABLE V. Performance comparison of Faster R-CNN networks using thermal images

Detector	Day		Night	
	LAMR	mAP	LAMR	mAP
BASNet saliency	0.643	0.569	0.499	0.667
PiCANet saliency	0.554	0.627	0.481	0.668
PFANet saliency	0.547	0.632	0.468	0.668
Baseline	0.543	0.639	0.435	0.703
R^3-Net saliency	0.527	0.648	0.422	0.704
PoolNet saliency	**0.515**	**0.692**	**0.390**	**0.749**

TABLE VI. Performance comparison of Faster R-CNN networks using color images

Detector	Day		Night	
	LAMR	mAP	LAMR	mAP
BASNet saliency	0.598	0.551	0.785	0.260
PiCANet saliency	0.586	0.555	0.675	0.381
PFANet saliency	0.554	0.588	0.674	0.391
Baseline	0.512	0.595	0.670	0.393
R^3-Net saliency	0.505	0.621	0.640	0.426
PoolNet saliency	**0.494**	**0.655**	**0.628**	**0.463**

all cases. Due to the pooling modules plugged into the FPN architecture, PoolNet could outperform the other saliency networks. Hence, it can be concluded that PoolNet saliency masks are more suitable for augmentation in both thermal and color images with Faster R-CNN.

BASNet is another network whose pixel-level performance is better than all other networks except PoolNet. However, in terms of region-level performance, it is observed that BASNet has achieved a better recall except PoolNet. But its precision is poor by a large margin than other networks. It implies that although BASNet derives more TPs and good quality masks, more count of FPs makes the performance weaker than other networks. As a result, the overall performance of the detector becomes significantly lower compared to detectors using other saliency networks. PiCANet and PFANet have similar region-level F-measure. However, PiCANet exhibits better precision

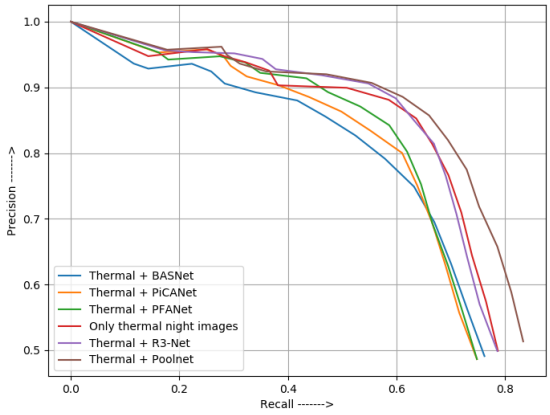


Figure 9. Precision vs. Recall plots for thermal images (Night)

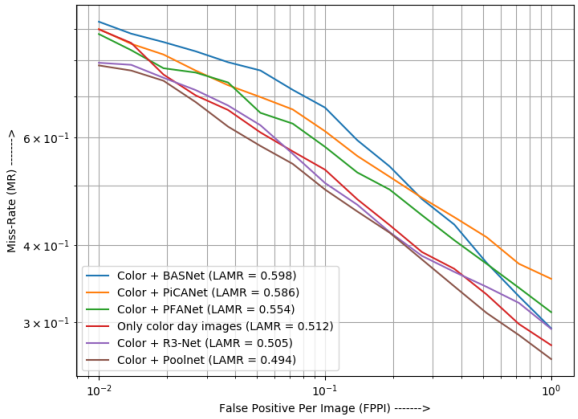


Figure 10. Miss-rate vs. FPPI plots for color images (Day)

than PFANet in case of thermal (day and night) images and color day images. Low region-level recall and precision of these two networks make them less recommended for augmentation. Moreover, the saliency masks produced by these two networks are not proper which is reflected by

TABLE VII. MR (%) of the pedestrian detectors at different FPPI values

Detector	Thermal						Color					
	Day			Night			Day			Night		
	FPPI values											
	0.01	**0.1**	**1**	**0.01**	**0.1**	**1**	**0.01**	**0.1**	**1**	**0.01**	**0.1**	**1**
BASNet saliency	94.9	71.5	31.7	88.4	53.1	23.7	92.5	67.0	29.3	97.8	83.2	54.3
PiCANet saliency	87.4	60.4	28.0	83.9	50.1	25.1	90.0	61.3	35.2	85.9	69.7	47.7
PFANet saliency	85.6	59.0	29.4	83.8	46.6	25.1	88.3	57.9	31.1	84.9	68.7	49.2
Baseline	91.0	57.2	27.7	85.7	41.4	21.3	89.9	53.1	27.5	84.0	68.3	49.1
R^3-Net saliency	**85.3**	57.5	26.8	82.6	40.1	21.4	79.3	50.5	29.2	**77.8**	64.9	47.9
PoolNet saliency	89.7	**56.1**	**23.2**	**82.2**	**38.8**	**16.6**	**78.5**	**49.3**	**26.1**	79.2	**63.9**	**45.0**

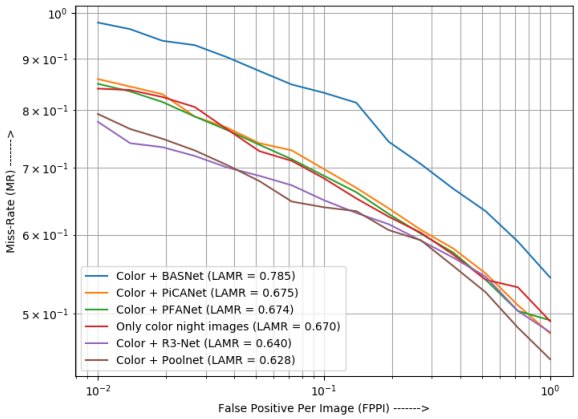


Figure 11. Miss-rate vs. FPPI plots for color images (Night)

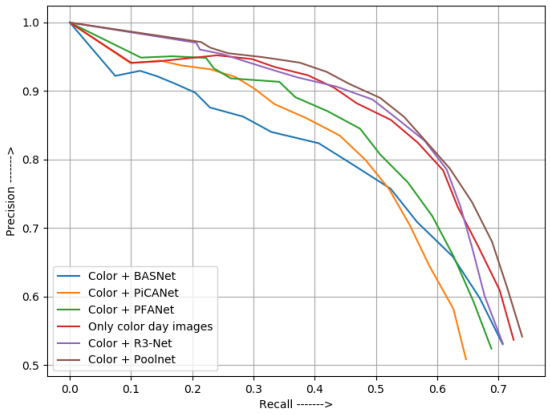


Figure 12. Precision vs. Recall plots for color images (Day)

the pixel-level F-measure and MAE values. Again, R^3-Net shows the best region-level precision among all the saliency networks despite that its pixel-level performance is lower than BASNet and PoolNet in case of thermal images.

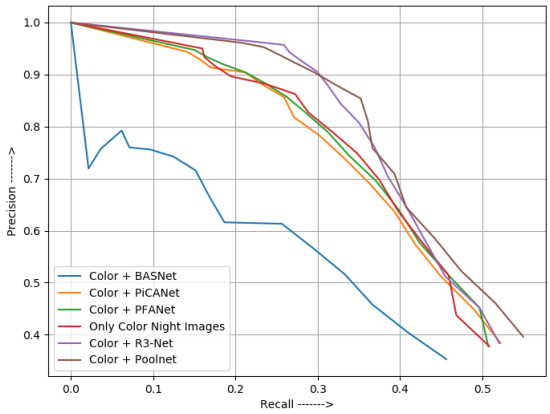


Figure 13. Precision vs. Recall plots for color images (Night)

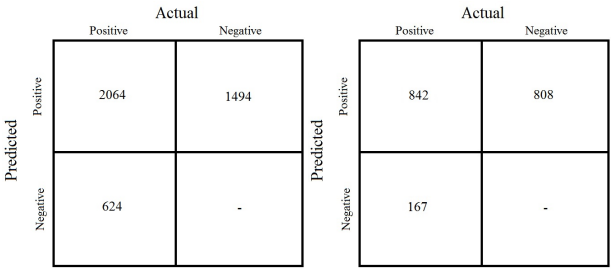
However, its recall at region-level is significantly lower than PoolNet which in turn reduces the F-measure. The overall performance of the R^3-Net based detector is better than other detectors except PoolNet based detector.

The baseline detectors have shown performances better than saliency network based detectors other than R^3-Net and PoolNet based detectors. The reason behind it is that some of the saliency networks are providing the detectors with data which either have large number of FPs (i.e. low precision as in BASNet) or low count of TPs (like PiCANet and PFANet).

From the analysis of region-level F-measure, it is found that R^3-Net and PoolNet based detectors are the two best performing detectors. It is also found that these detectors maintain good trade-off between precision and recall. Again, the performance analysis of the final detector in terms of LAMR and mAP also shows that these detectors produce good overall performance which validates the hypothesis mentioned in Section 3.

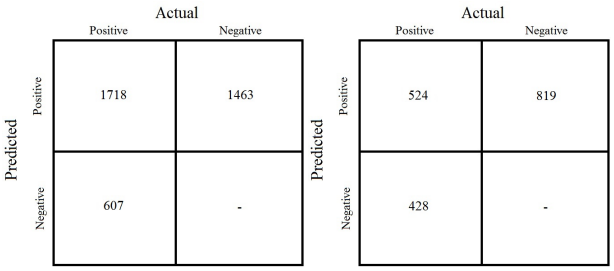
In [3], the authors have used R^3-Net and PiCANet with Faster R-CNN and reported improvement over the baseline

detector with thermal images. We have also used these two saliency networks for augmentation. These augmented images are used to train the Faster R-CNN network. Same parameters as in [3] are used to train the saliency networks and Faster R-CNN. From Table V and Table VI, it can be observed that R^3-Net based detectors have performed better than the baseline detectors. However, using the enhanced annotations, the PiCANet based detectors have shown inferior performances than the baseline detectors. The reason behind the inferior performance using PiCANet here might be due to consideration of more challenging conditions in our enhanced test set annotations. The PoolNet based detectors have performed better than both R^3-Net and PiCANet based detectors. This better performance is due to the ability of PoolNet to detect more small sized and partially occluded pedestrians in the enhanced annotations. PoolNet also produces good quality saliency masks due to improvement in the training saliency dataset. In Figure 14 and Figure 15, we present the confusion matrices for the PoolNet based detectors for thermal and color images for different illumination conditions are shown respectively.



(a) Day (b) Night

Figure 14. Confusion matrix (PoolNet) for thermal images



(a) Day (b) Night

Figure 15. Confusion matrix (PoolNet) for color images

The computation time to process a frame by using each saliency network with Faster R-CNN is also studied. Table VIII shows the average computation time for the detectors presented in this work. From the table, it can be observed that BASNet and PFANet based detectors have smaller computation time compared to the other detectors. However, R^3-Net based detector is the slowest among all with around 1.74 FPS. The computation time for the PoolNet based detectors is found to be around 5.1 FPS which may be considered for real-time applications [16]. In a nutshell, the PoolNet based detectors maintain a good trade-off between speed and detection performance as the detection performance of the BASNet, PiCANet and PFANet based detectors are inferior to the baseline detectors and R^3-Net based detector is the slowest one.

TABLE VIII. Comparison of computation time

Detectors based on	Computation time (sec)		Frames/second (FPS)
	Saliency network	Faster R-CNN	
PFANet	0.05281	0.09953	6.56
BASNet	0.05689	0.09953	6.39
PoolNet	0.09634	0.09953	5.10
PiCANet	0.10513	0.09953	4.88
R^3-Net	0.47202	0.09953	1.74

7. Conclusion

This paper introduced Faster R-CNN based pedestrian detectors for thermal and color images augmented with saliency maps derived using deep salient networks. An enhanced version of annotations for thermal and color images in KAIST dataset was proposed. Pixel-level annotations for thermal and color images were also proposed for a subset of KAIST dataset. Different saliency networks were used for augmentation and their performances were analysed in terms of a newly introduced "region-level F-measure" metric to determine the best suitable one for use with Faster R-CNN. It was also observed that detectors using some saliency networks yield degraded performances than the baseline detectors for both the channels. The saliency networks were also studied in terms of F-Measure and MAE. The performances of the overall detectors were studied in terms of LAMR, mAP, MR vs. FPPI and precision vs. recall. The PoolNet based detectors outperformed the other saliency network based detectors. It is also observed that PoolNet based detectors maintain a good trade-off between detection performance and computation speed. As a future work, this saliency based augmentation technique may be explored using different object detectors. The performance of the presented approach may also be tested for multi-class detection problems.

Acknowledgements

This work is supported by the Department of Science and Technology, Govt. of India, under the INSPIRE fellowship program and also partly by TEQIP-3, Govt. of India.

References

[1] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1037–1045.

[2] J. Wagner, V. Fischer, S. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *ESANN*, 2016.

[3] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *CVPR Workshops*, 2019.

[4] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] Z. Deng, X. Hu, L. Zhu, X. xu, J. Qin, G. Han, and P.-A. Heng, "r^3-net: Recurrent residual refinement network for saliency detection," 07 2018.

[7] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3089–3098.

[8] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE CVPR*, 2019.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," vol. 1, 02 2001, pp. I–511.

[11] D. Huang, C. Shan, M. Ardabilian, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 41, pp. 765–781, 11 2011.

[12] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 01 2009.

[13] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[14] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," vol. 2049, 01 2001, pp. 249–257.

[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[17] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788.

[19] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," 06 2015, pp. 5079–5087.

[20] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Pedestrian detection: Domain generalization, cnns, transformers and beyond," 2022. [Online]. Available: https://arxiv.org/abs/2201.03176

[21] A. K. M. F. Rahman, M. Raihan, and S. Islam, "Pedestrian detection in thermal images using deep saliency map and instance segmentation," *International Journal of Image, Graphics and Signal Processing*, vol. 13, pp. 40–49, 02 2021.

[22] M. Kieu, L. Berlincioni, L. Galteri, M. Bertini, A. D. Bagdanov, and A. Del Bimbo, "Robust pedestrian detection in thermal imagery using synthesized images," 2021. [Online]. Available: https://arxiv.org/abs/2102.02005

[23] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3166–3173.

[24] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[25] M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, March 2015.

[26] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2214–2219.

[27] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," *arXiv preprint arXiv:1904.09146*, 2019.

[28] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 815–828, April 2019.

[29] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 236–252.

[30] T. Kim, K. Kim, J. Lee, D. Cha, J. Lee, and D. Kim, "Revisiting image pyramid structure for high resolution salient object detection," 2022. [Online]. Available: https://arxiv.org/abs/2209.09475

[31] M. S. Lee, W. Shin, and S. W. Han, "Tracer: Extreme attention guided salient object tracing network," 2021. [Online]. Available: https://arxiv.org/abs/2112.07380

[32] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *ECCV*, 2018.

[33] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 717–729, April 2016.

[34] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NIPS*, 2011.

[35] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian

detection via simultaneous detection and segmentation," in *BMVC*, 2018.

[36] Tzutalin, "Labelimg," Free Software: MIT License, 2015.

[37] A. Dutta, A. Gupta, and A. Zissermann, "VGG image annotator (VIA)," http://www.robots.ox.ac.uk/ vgg/software/via/, 2016.

[38] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 07 2017, pp. 5987–5995.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[40] P. Hurtik and N. Madrid, "Bilinear interpolation over fuzzified images: Enlargement," 08 2015.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F. F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, 09 2014.

[43] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, April 2012.

Amlan Jyoti Das has completed his Ph.D. from Electronics and Communication Engineering, Gauhati University. He is currently working as a Data Scientist at Obaforta India Pvt. Ltd. His areas of interest are computer vision, machine learning and deep learning.

Simantika Choudhury has completed her Bachelor of Engineering in Electronics and Telecommunication Engineering from Gauhati University. She has also completed her Master of Technology in Signal Processing and Communication from Gauhati University. She is currently pursuing her Ph.D. in Electronics and Communication Engineering, Gauhati University. Her research areas include image processing, computer vision, machine learning and deep learning.

Navajit Saikia is presently working as an Associate Professor in the Department of Electronics and Telecommunication Engineering of Assam Engineering College. Signal Processing and Communications are two areas of his teaching interests. His research interests include image processing, speech processing, information security and reversible logic. He has published several research papers in journals and conference proceedings.

Subhash Chandra Rajbongshi completed his M.Tech. and Ph.D. degree from Gauhati University, Assam, India. He is currently working as Scientific Officer at Gauhati University. His area of interest includes computer vision, image processing, signal processing etc. He has also published research papers in journals and conference proceedings.