IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Night Vision Surveillance: Object Detection using Thermal and Visible Images

K.P.SaiMahesh and Dr.J.Srinivasan
Department of Computer Applications
Madanapalle Institute of Technology & Science,
Madanapalle, India.

***Abstract*—Surveillance has become an important task in recent time mainly due to the increasing of crime rates. The existing research on surveillance for day time has achieved better performance by detecting and tracking objects using deep learning algorithms. However, it is difficult to achieve the same performance for night vision mainly due to low illumination and/or bad weather situation. One of the important tasks in surveillance is object detection which results in both class and location of the detected object with clear boundary of the objects from the image. We propose an efficient object detection module using fusion of thermal and visible images. Fusion module consists of encoder-decoder network in which encoder uses depthwise convolution to extracts efficient features from the given thermal and visible images. Then after, fused image is reconstructed using convolutional layers and final map is utilized in object detection algorithm (i.e., mask RCNN). The proposed method shows the effectiveness of utilization of pre-processing module i.e., fusion in object detection algorithm. Here, it is observed that the proposed method performs better for night vision when images are trained carefully with various features. Moreover, proposed method performs better on real time night vision images having no illumination condition.**

***Index Terms*—Depthwise convolution, encoder-decoder network, image fusion, night vision thermal images, object detection.**

I. INTRODUCTION

In the recent scenario of increased security and surveillance, more robust and sophisticated surveillance systems are demanded. Among many types of deep learning models, deep convolutional neural network (DCNN) is a powerful approach for low to high level feature learning. The main aim behind the use of DCNN is to extract features effectively from the data captured in low or no illumination situation during night time. Recently, thermal infrared camera is widely used to detect object in low illumination situation. The visible cameras have ability to capture images under natural/ artificial illumination conditions only. Hence, very limited visual information are captured in night vision and that makes difficult to perform surveillance in night time using visual sensors only. Moreover, thermal images contain higher information of objects which have high temperature. However, for the objects having low temperature, it provides poor information. On the other hand, visual imaging contains the high visual context of the particular object [1].

The idea that relies on the utilization of thermal and visible spectrum pairs that are situated around premises of objects is considered for night vision surveillance. To make the system automatic, development of efficient technique for robust object detection algorithm enables the utilization of images from two different sensors becomes necessary [2].

A. Image fusion:

In the last few decades, various methods have been developed in the literature to enhance the quality of the fused image of different imaging devices of same scene [3]. This includes the image fusion approaches based on multi-scale decomposition [4], sparse representation [5], deep neural network based methods [6], etc,. The other existing methods are based on fuzzy theory, gradient transfer and total variation, global entropy, saliency-based and hybrid methods [7–9].

Moreover, In [10], *DeepFuse* is presented with encoding and decoding networks and it performs better. However, it suffers from a drawback of inadequate extraction of salient features. Due to that, this approach fails to preserve meaningful information in the fused image. Therefore, authors in [6] proposed a CNN architecture with encoding and decoding networks called *DenseFuse* in which encoding network utilizes convolutional layers with dense block [11]. The main drawback of this method is the use of dense layers which increases the computational complexity to large extent and hence this method is not efficient for real time applications. To overcome this problem, a novel architecture of fusion is proposed in which the newly emerged depthwise convolution is utilized in encoder module to reduce the complexity of the network. As compared to other CNN based approaches, the proposed fusion module requires less number of parameters to tune.

B. Object detection:

Object detection from thermal pictures has recently shown a sharp rise in the performance of deep learning algorithms [2]. Even there have been significant advancements in recent years, the development of an effective strategy for efficient detection in real-world applications still presents a difficult problem. Due to training operations being performed only on visual input, it has been found that the majority of existing object detection algorithms are sensitive to variations inlight,

climate, and obstacles. Numerous research issues have been focused on the development of multi-spectral object identification methods for allowing robust target detection in order to overcome the limits for nighttime object detection previously discussed [12].

Due to the significant degree of variety in human appearance, including body size, enunciated motion, fractional obstruction, contradicting textile texture, excessively cluttered backgrounds, and low/no lighting conditions, the visual image-based object detector is ineffective. Additionally, multispectral photos of thermal-color spectrum pairs have proven to be more effective for object detection than using a single thermal-color spectrum, particularly under a variety of lighting conditions. For the purpose of detecting nighttime pedestrians, authors in [13] recently demonstrated the combination of deep features with quicker RCNN. Furthermore, they incorporate the information collected from deep convolutional networks in subsequent frames because pedestrians cannot be reliably spotted from a single night vision image. The correlation between light levels and the object detection confidence score from thermal or colour images is necessary. The authors of [14] therefore proposed illumination-aware deep neural networks (i.e., IATDNN). For improved performance, authors in that work using both visual and thermal images. When compared to other methods, the object detection algorithm in [13] without the use of thermal pictures has an extremely high missing rate (MR) in the case of night vision. Another slightly different method involves applying CNN layers before fusion to robustly extract features that achieve low MR at night [15]. Hence, It is verified that objectdetection algorithm for night time surveillance performs better,

due to utilization of feature enhancement module by fusing of features from thermal and visible images. The sequential training of those enhanced feature maps and object detection algorithm are performed better rather than the use of only object detection algorithm without preprocessing night time images.

Here, we provide a novel fusion module for effective feature extraction and reconstruction and demonstrate how to use it to an object identification algorithm (i.e., Mask RCNN [16]). When images are carefully trained with multiple features, a comprehensive pipeline for object detection using the fusion of thermal and visual images with MRCNN is developed, and it performs better for night vision.

II. PROPOSED METHODOLOGY

We propose an efficient object detection architecture for night vision images using fusion of thermal and visible images with MRCNN. First, We implemented MRCNN algorithm for object detection of night time images using thermal images and observed that MRCNN suffers from insufficient features extracted from it. Hence, it motivates us to implement image enhancement module as pre-processing which can enhance the features of the objects from night time images. Here, it is focused to implement efficient fusion module that can

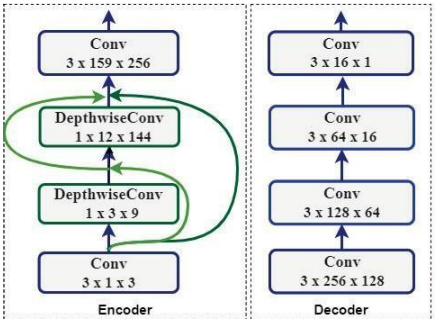


Fig. 2: The encoder and decoder network of fusion module

extracts more meaningful features from given pair of images, and utilized it to object detection algorithm which is shown in Fig. 1. Fig. 2 gives a description of the encoder-decoder architecture employed in the suggested fusion module. The fusion and MRCNN modules make up the suggested approach. Encoder and decoder are both part of the fusion module. In order to extract more significant features, encoder is used. Therefore, it is used to extract significant features from a pair of source photos. Simple convolution and depthwise convolution (DC) layers make up its structure. Due to the grayscale nature of the initial layer, depthwise convolution was not achievable. Thus, numerous channels are obtained using a straightforward CNN. The network, which is based on [17], is then given depthwise convolution. 3 1 3 stands for filter size (f), input channels (m), and output channels (n) in Fig. 2. (i.e. f m n for all layers). Each layer uses the Leaky ReLU (LReLU) activation function, which is defined as,

$$f(x = W[x] + b) = \begin{cases} x, & if \quad x > 0 \\ 0.01x, & otherwise. \end{cases} \quad (1)$$

Convolution is used in DC on a single channel at a time, making it computationally less expensive and requiring fewer parameter adjustments1. Each DC layer's channel multiplier is set to the number of input channels. The 1-1 filter is used in DC. Additionally, we make use of all features produced at earlier stages in the encoder module. As a result, at each layer, feature maps from all preceding layers are concatenated, as seen in the encoder module of Fig. 2. Therefore, the suggested encoder module aids in extracting more useful characteristics.

Following feature extraction, a weighted sum fusion approach is used to combine the features from the source images. Here, weighted sum is taken into consideration to fuse photos based on various illumination conditions. Consequently, the weight of the thermal image is greater than the visible picture (i.e., 0.7 and 0.3, respectively). The calculation for image fusion using weighted sum is

$$f(x, y) = \alpha * I_1(x, y) + (1 - \alpha) * I_2(x, y), \quad (2)$$

where denotes the weight attached to each individual image (I). Finally, a decoder network with several convolutional layers is utilised to recreate the merged image.[1]

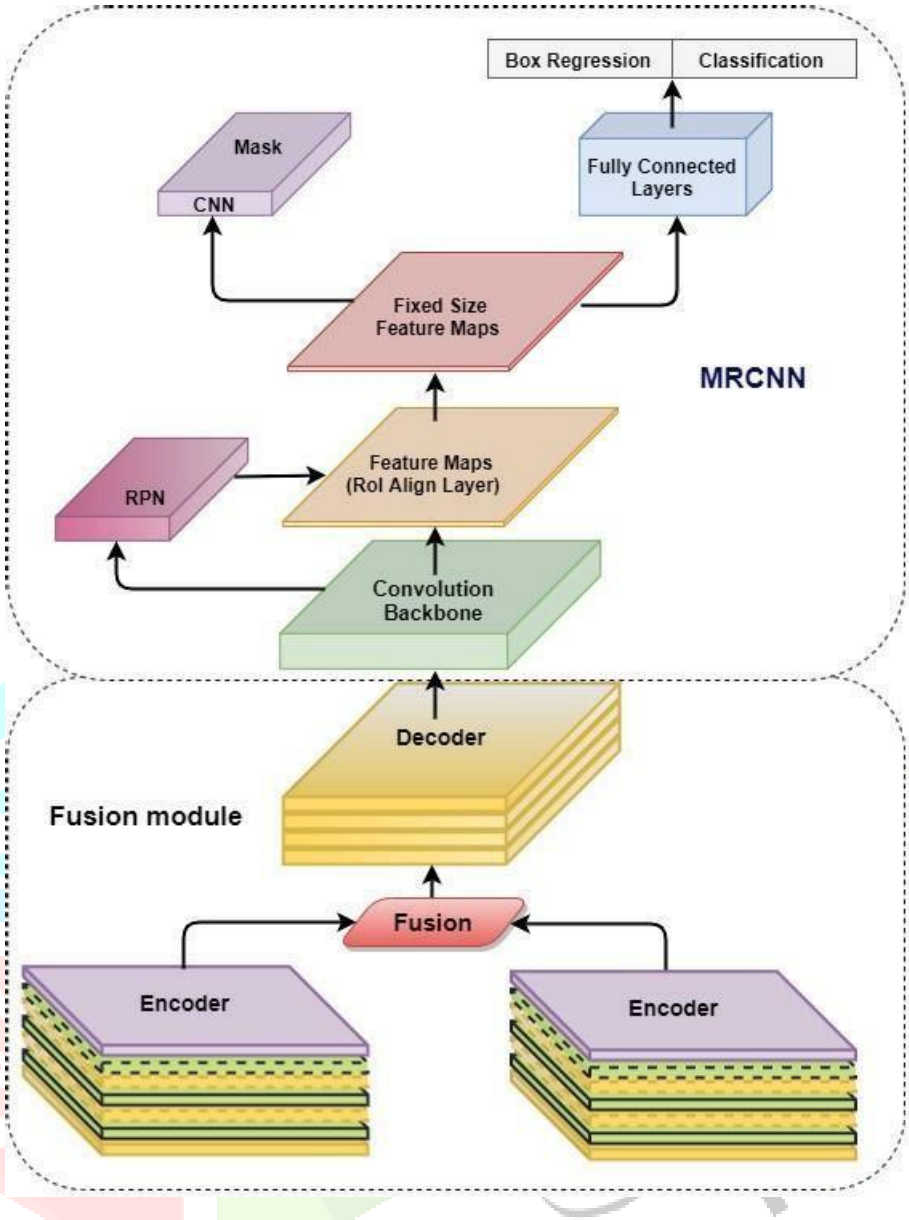


Fig. 1: The complete pipeline architecture of proposed approach of MRCNN with image fusion module

shown in decoder module of Fig. 2. Encoder and decoder modules of the fusion with MRCNN are used to train the network during the training phase without the inclusion of the fusion module.

The region proposal network, which is the first step of the object detection module, proposes the bounding boxes of objects (RPN). The following stage involves region of interes t pooling (RoIPool) to extract features from each feature map, followed by classification and bounding-box regression. The improvement of the ROIPool layer is the main contribution of MRCNN. Using a technique called RoIAlign, this layer is modified to be more precisely aligned. Each RoI receives the application of the tiny fully connected network (FCN), also known as the mask branch. With pixel-to-pixel segmentation, it predicts the mask. As a result, it produces a binary mask for each.

RoI . This method aims to cut the pixels of the same category of various objects starting from FCN outputs.

(i.e., per-pixel classification results). The spatial arrangement of an input object is encoded using a mask. Convolutions' pixel-to -pixel correspondence so readily addresses the extracted spatial structure of masks. Once these masks are created, segmentations are created by combining them with the classifications and bound boxes. Here, it is seen that when images are carefully trained with a variety of attributes, the suggested technique works better for night vision.

III. EXPERIMENTAL RESULTS

In this section, the implementation of proposed model i.e., fusion module of thermal and visible images followed by

(a) Thermal Image (b) Visible Image (c) Fused image

(d) MRCNN (e) fusion+FRCNN [15] (f) Proposed

Fig. 3: The testing results of proposed method with effect of fusion module utilizes as pre-processing task on FLIR dataset

(a) Thermal Image (b) Visible Image (c) Fused Image

(d) MRCNN (e) fusion+FRCNN [15] (f) Proposed

Fig. 4: The testing results of proposed method with effect of fusion module utilizes as pre-processing task on TN image fusion dataset

The object detection module is discussed, and the nighttime object detection test results are analysed. Without using any pre-learned weights, the proposed architecture is initially trained using FLIR dataset 2 with 1000 epochs. The number of steps in a single epoch is set to 100, and the batch size is set to 2. An Adam optimizer is employed, and the learning rate is set to 0.001.

[2]https://www.flir.in/oem/adas/adas-dataset-form/

help enhance the learning experience. It has been found that training networks on multiple datasets is essential for performing detection. Because thermal pictures only preserve edge features, it has also been found that network training using just those features is insufficient. Consequently, suggested architecture is once more trained on a pretrained COCO dataset [18] module, which contains thousands of TABLE I: Comparison of the MR acquired using the suggested methodology on different datasets. MR 1, MR 2, and MR 3 denote the MR on the FLIR, TN image fusion, and our own dataset, respectively.

	MR 1(%)	*MR* 2(%)	*MR* 3(%)
MRCNN	31.56	28.14	41.13
Fusion + FRCNN	33.65	30.89	43.56
Fusion + MRCNN	30.54	27.14	38.42

In order for objects to get a sufficient number of different features from both thermal and visual photographs, visible images must come first, followed by thermal images.

Three types of testing datasets have been used to evaluate the performance of the proposed method: FLIR [19], TN image fusion dataset [20], and our own dataset that was created at night using a FLIR E8-XT camera. Both visual and thermal photos of the same scene are included in these collections. Each dataset has unique properties. The FLIR dataset includes pictures of typical daytime and nighttime traffic situations to account for changes in lighting. Images from the TN image fusion datasets are in a variety of weather conditions, including foggy, sunny, cloudy, wet, etc. Our own dataset, which contains photographs of cars and people in both stationary and moving situations under various lighting and weather circumstances, has been prepared to operate with real-time night vision scenarios. Figs. 3- 5 display the outcomes of the suggested architecture and the performance of the architecture with/without the use of the fusion module. For the purpose of confirming the effectiveness of object detection on night vision images, the suggested technique is contrasted with an existing state-of-the-art method [15]. The outcomes from the FLIR and TN image fusion datasets are shown in Figures 3 and 4, respectively. The performance of the object detection technique with/without employing the fusion module is displayed in the second and fourth rows of that. Here, one can see that these datasets include photographs taken at night under low lighting conditions, where objects are situated close to a light source, such as a street or a moving car. Additionally, Fig. 5 displays the performance on a dataset that we created specifically for use with real-time night vision monitoring. The viewable visuals in this case are virtually black and aren't providing any information. Therefore, the suggested module works well to obtain enough features that are useful for object detection. Additionally, Fig. 5 shows that the suggested method outperforms the current method [15] on fused features. The missing rate discovered using different datasets is also contrasted and displayed in Table

I. CONCLUSION

In this research, we offer a technique for object recognition in night vision surveillance utilising thermal and visual images. The suggested network incorporates fusion and MR- CNN modules, where fusion employs an encoder and decoder module with a depthwise convolution to extract important features from the supplied input images. Then, to effectively detect objects, a fused image is used. Various datasets have been used for the studies, and the missing rate is also evaluated to assess how well the suggested strategy performs.

based on live night vision footage. It demonstrates that the suggested object detection method beats other cutting-edge existing technologies.

REFERENCES

[1] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[2] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: a survey," *Information Fusion*, vol. 45, pp. 153–178, 2019.

[3] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Physics & Technology*, vol. 76, pp. 52–64, 2016.

[4] R. Gao, S. A. Vorobyov, and H. Zhao, "Image fusion with cosparse analysis operator," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 943–947, 2017.

[5] H. Li and X. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, pp. 1–10, 2019.

[6] S. Rajkumar, Mouli, and Chandra, "Infrared and visible image fusion using entropy and neuro-fuzzy concepts," in *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I*. Springer, 2014, pp. 93–100.

[7] J. Zhao, Y. Chen, H. Feng, Z. Xu, and Q. Li, "Infrared image enhancement through saliency feature analysis based on multi-scale decomposition," *Infrared Physics & Technology*, vol. 62, pp. 86–93, 2014.

[8] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information Fusion*, vol. 24, pp. 147–164, 2015.

[9] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *2017 IEEE International Conference on Computer Vision (ICCV)* . IEEE, 2017, pp. 4724–4732.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2261–2269.

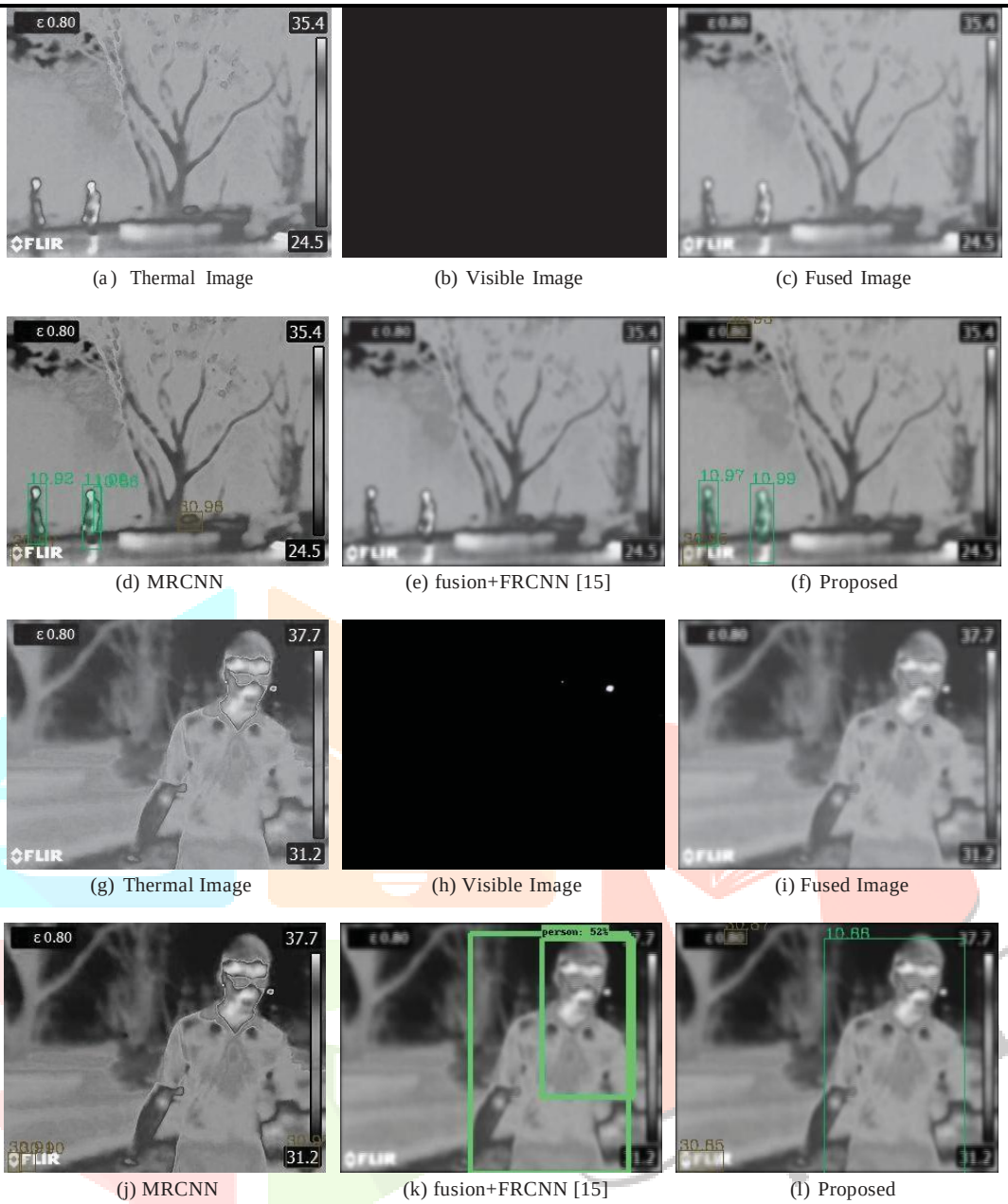
[11] T. T. Zin, H. Takahashi, T. Toriu, and H. Hama, "Fusion of infrared and visible images for robust person detection," *Image fusion*, pp. 239–264, 2011.

[12] J. H. Kim, G. Batchuluun, and K. R. Park, "Pedestrian detection based on faster r-cnn in nighttime by fusing deep convolutional features of successive images," *Expert Systems with Applications*, vol. 114, pp. 15–33, 2018.

[13] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, 2019.

[14] R. Gade and T. B. Moeslund, "Thermal cameras and applications: a survey," *Machine vision and applications*, vol. 25, no. 1, pp. 245–262, 2014.

[15] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, pp. 161–171, 2019.



(a) Thermal Image (b) Visible Image (c) Fused Image

(d) MRCNN (e) fusion+FRCNN [15] (f) Proposed

(g) Thermal Image (h) Visible Image (i) Fused Image

(j) MRCNN (k) fusion+FRCNN [15] (l) Proposed

Fig. 5: The testing results of proposed method with effect of fusion module utilizes as pre-processing task on our own dataset.

[2] X. Wu, S. Wen, and Y.-a. Xie, "Improvement of mask-rcnn object segmentation algorithm," in *International Conference on Intelligent Robotics and Applications*. Springer, 2019, pp. 582–591.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[5] *https:// www.flir.in/ oem/ adas/ adas-dataset -form/* .

[6] *https:// figshare.co m/ articles/ TNImageF u sion Data set/ 1008029*.