Object Detection in Monocular Infrared Images Using Classification – Regresion Deep Learning Architectures

Raluca Brehar, Flaviu Vancea, Tiberiu Marita, Cristian Vancea, Sergiu Nedevschi
Technical University of Cluj-Napoca,
Computer Science Department

***Abstract*—The rapid development of deep learning architectures that have a good performance on object detection in visual monocular images has triggered and interest towards the application of these architectures on other image modalities such as stereovision or infrared images.**

We propose a framework for multi-class object detection in monocular infrared images that integrates and compares different classification-regression deep learning architectures [1] on a novel benchmark infrared dataset developed by FLIR.

The work described is evaluated using standard object detection metrics and an average precision of 82% for pedestrians, 86% for cars and 66% for bicycles is achieved while running at 40fps.

I. Introduction

The detection of objects in infrared scenes is of particular interest because infrared sensors can see where visible sensors don't: for example at night, in low visibility situations such as heavy rain, snow, fog, dust. The appearance of objects in infrared images is quite complex due to phenomena like heat diffusion that makes the borders of the object blurry or occlusions of cold / warm objects by others.

There are many solutions that perform object detection in the visible domain using either classical machine learning algorithms and, lately most of them explore deep learning architectures [2]. In order to encourage the scientific work in deep learning for infrared images, FLIR introduced a benchmark dataset that contains annotations for pedestrians, cars, bicycles and dogs [3]. Another interest dataset is [4] that has annotations for pedestrians in well aligned visible and infrared images. In recent years this dataset [4] has been largely explored for pedestrian detection in infrared images and for fusion of infrared and visible pedestrian detectors.

Our work is focused on the detection of multiple objects from monocular infrared images and as far as we know until now there are only a few relevant contributions in this field [3]. The original aspects of the proposed method reside in:

- The fine tuning of two deep convolutional neural network architectures to perform multi-class object detection in infrared images.
- The study and comparison of the performance of these networks on different objects (cars, bicycles, pedestrians).

978-1-7281-4914-1/19/$31.00 ©2019 IEEE

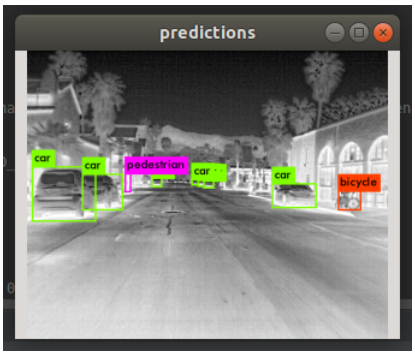


Fig. 1. Results of the proposed method

- A generic framework for multi-class object detection for infrared images.

In order to have a reliable solution and accurate results the environment temperature should be lower than the temperature of the human body. The proposed system has good results for night scenes and also for day scenes achieving a high accuracy for pedestrians and cars as shown in Figure 1.

The rest of the paper is structured as follows: in section II we present other existing approaches in the field. Section III describes the proposed framework including the network architectures with the fine tuning of parameters. The dataset used for evaluation, the parameters of the training procedure, the detailed accuracy results and the precision-recall curves are described in section IV. Section V concludes the paper and shows the main ideas and future development directions.

II. Related work

The rapid development in deep learning provides powerful learning architectures that are able to generate high-level deep features relevant for semantic segmentation or object detection. In what follows we revise the most important deep learning architectures with good results in the visible domain and we summarize the main approaches for pedestrian detection in infrared and visible images.

A. Deep learning object detection architectures for the visible domain

The problem of object detection has been studied extensively in the field of deep learning. One of the first

successful architectures for object detection is Region-based Convolutional Network (R-CNN) proposed by Girshick. et. al. [5] where the authors have used a region proposal module to generate bounding boxes proposals, a CNN that extracts features for each proposal and an SVM classifier that classifies each proposal using the computed features. An improved version of R-CNN called Fast R-CNN was proposed by R. Girshick [6] where features are initially computed for the whole image using a CNN and a Region of Interest Pooling (ROIPooling) layer is used to extract features for candidate regions.

In [7], an improved version of Fast R-CNN called Faster R-CNN was proposed. In Faster R-CNN, a Region Proposal Network (RPN) was introduced to eliminate the costly region proposal algorithm used by R-CNN and Fast R-CNN. The RPN is composed of a CNN that produces a feature map and a smaller network that is used in a sliding window fashion over the computed feature map. The smaller network produces an objectness score for a set of pre-defined bounding boxes, called anchors. The smaller network also produces a set of offsets for the anchors in order to refine the detections. A ROI pooling layer and a classifier is then used to classify the boxes proposed by the RPN and refine the boxes further.

Lin et. al. [8] introduces the Feature Pyramid Network (FPN) that is similar to Faster R-CNN but detects objects at multiple scales. The network is composed of two pathways, a bottom-up pathway and a top-down pathway. During the bottom-up pathway, feature maps at different scales are computed. The top-down pathway up-samples coarser features and combines them with features from the bottom-up pathway using lateral connections. Different branches of RoI Pooling followed by fully connected layers are applied on feature maps of different scales from the top-down pathway in order to classify and regress bounding boxes.

The introduction of YOLO [9] made real time object detection using deep learning possible, having a version running at 45 FPS and a faster version running at 150 FPS on a Titan X GPU. In YOLO, an image is first divided into a grid of a fixed size. For each cell in the grid, the network predicts multiple bounding boxes, a probability for an object to be in a bounding box, as well as class probabilities for each box. In YOLO, all predictions are done in the same time, as opposed to ther models like Faster R-CNN where bounding boxes are first proposed and then classified. YOLOv2 is an improvement over the original YOLO architecture, obtaining a higher mean average precision at a higher speed. Some of the improvements of YOLOv2 are the use of Batch Normalization [10] layers and the use of anchors, similar to Faster R-CNN. The main problem with YOLO and YOLOv2 is that they struggle with small objects. YOLOv3 [1] addresses this by predicting boxes at 3 different scales. YOLOv3 also replaces softmax with independent logistic classifier to deal with overlapping labels.

B. Deep learning for infrared and visible pedestrian detection

Infrared object detection has been explored from the perspective of pedestrian detection and fusion of visible and infrared data in order to enhance the accuracy of the results. A comparison of different convolutional network fusion architectures is employed by [11]. They discover that pedestrian detection confidences from color or thermal images are correlated with the illumination conditions and propose an illumination-aware Faster R-CNN (IAF R-CNN) that gives an illumination measure of the input image. They merge color and thermal sub-networks by a gate function that is defined over the illumination value.

Two stream deep convolutional neural networks are proposed by [12] that learn multi-spectral human-related features under different illumination conditions (daytime and nighttime). They use the illumination information with multi-spectral data in order to generate more accurate semantic segmentation that is used to boost the pedestrian detection accuracy. The proposed method is trained end to end and uses a multi-task loss function. The results outperform state of the art approaches on KAIST multi-spectral pedestrian dataset. An analysis of existing detection approaches from the perspective of their generalization ability in the combination of visual and thermal spectra for person detection is presented by [13]. The Yolov3 architecture has also been employed for pedestrian annotation enhancement in thermal images by [14] that aimed at the acceleration of pedestrian labeling in far-infrared image sequences. Multi-class object detection in infrared traffic scenes has not been largely explored due to the lack of multi-class annotations for infrared images. There are a few approaches that perform multiclass classification in video surveillance applications.

A transfer knowledge framework for object recognition of infrared image is proposed by [15]. The method extracts Hu moments based on which auxiliary feature data are generated. and The proposed transfer knowledge approach can transfer knowledge from the auxiliary data to help the tiny amount of training data to train a better classifier, which improve the performance of object recognition.

An experimental study on geometric and appearance features for outdoor video surveillance videos is proposed by [16]. They also analyse the classification performance under two dimensionality reduction techniques (i.e. PCA and Entropy-Based feature Selection) in the framework of an object classification system for infrared surveillance videos.

C. Pedestrian detection in the infrared domain

The detection of pedestrians in the infrared domain has been explored by the scientific community. Due to the lack of annotated benchmark datasets for the FIR field, only pedestrians have represented instances of interest for detection and recognition algorithms. An early approach was proposed by [17] that describe a multi sensor system consisting of a far infrared camera, a laser scanning device and ego motion sensors. To handle the combination of the information of the different sensors a Kalman filter based data fusion is used. They analyze the estimated optical flow and the shape parameters related to the human motion. A fast region of interest generator combined with fast feature pyramid object

detection is described by [18]. Using the appearance model of pedestrians in infrared images, edge and intensity based filters are used to generate regions for pedestrian hypotheses in order to speed up the detection process. On these regions the Aggregated Channel Features are computed and pedestrian detections are inferred. This work is extended by [19] that combine four channel features, infrared, histogram of gradient orientations, normalized gradient magnitude and local binary patterns (uniform, rotation invariant) in order to improve detection results. A temporally and spatially aligned multi-sensor system that combines monocular, stereo and infrared images is proposed by [20] that improve pedestrian detection the usage of aggregated channel features classifiers trained on images captured with two types of sensors: far infrared and stereo-vision sensors. Solutions for infrared pedestrian detection have also been ported to embedded systems [21] and use HOG descriptors and different classification methods. The results are promising and can be a baseline for future development in embedded solutions.

III. Proposed framework

We employ two architectures that have a similar backbone namely YOLO [1] and YOLO with spatial pyramid pooling that is partially described by [22]. These architectures have been chosen due to high accuracy results on visible data and fast detection time.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 2. Darknet-53 [1]

In Yolov3 [1] a single neural network is applied to the full image. The network splits the image into several regions and for each region predicts bounding boxes which are weighted by probability scores based on clusters and anchor boxes. Logistic regression is employed for predicting the objectness score for each bounding box. Binary cross entropy loss is used for predicting the classes that a bounding box may contain. The architecture of the network allows the prediction across three different scales. K-means clustering applied on the input training set is used for computing the bounding box priors. The network uses 53 convolutional layers as described in Figure 2 and detailed in [1].

Layer ID	Repeat	Type	Filters	Size	Output
		Input image			640x512
0 - 74		Darknet-53	1024		20 x 16
75		Conv	512	1x1	20 x 16
76		Conv	1024	3x3	20 x 16
77		Conv	512	1x1	20 x 16
				5x5	
78 - 83		**SPP Module**	**2048**	**9x9**	**20x16**
				13x13	
84 - 87	2x	Conv	512	1x1	20 x 16
		Conv	1024	3x3	20 x 16
88		Conv	27	1x1	20 x 16
89		YOLO			
90		Route [-4]			
91		Conv	256	1x1	20 x 16
92		Upsample	256	2x2	40 x 32
93		Route [-1, 61]			
94 - 99	3x	Conv	256	1x1	40 x 32
100		Conv	512	3x3	40 x 32
101		Conv	27	1x1	40 x 32
102		YOLO			
103		Route [-4]			
104		Conv	128	1x1	40 x 32
105		Upsample	128	2x2	80 x 64
106		Route [-1, 36]			
107 - 112	3x	Conv	128	1x1	80 x 64
		Conv	256	3x3	80 x 64
113		Conv	27	1x1	80 x 64
114		YOLO			

TABLE I

YOLOv3-SPP. THE YOLO LAYERS ARE RESPONSIBLE FOR DETECTING BOUNDING BOXES AND THE ROUTE LAYERS CONCATENATE THE OUTPUTS FROM LAYERS WITH THE SPECIFIED IDS. DARKNET-53 IS USED UNTIL THE AVERAGEPOOL LAYER.

The second architecture we have employed is based on YOLO but adds the spatial pyramid pooling blocks and it is abbreviated Yolov3-spp. The layers of the architecture with spatial pyramid pooling are shown in Table III. This architecture provides slightly improved results with respect to YOLOv3.

In Yolov3-spp according to the scales that represent different layers of the feature pyramid, spatial pyramid pooling is responsible for dividing the input feature map into several bins. Then the maps of the features are pooled by the sliding windows of which the size is the same as that of the bins. experiments on several datasets [23] have shown that this network architecture provides slightly improved results.

Considering these two backbone architectures we develop a framework for object detection in infrared images. The pipeline of the proposed framework is described in Figure 3. In our pipeline we employ the FLIR-ADAS dataset [3].

We customize the annotations in order to respect YOLO format, that is each bounding box is described by

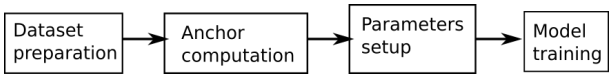


Fig. 3. Proposed processing pipeline

$class\_id, x_{center}, y_{center}, width, height$ where:

- class_id is the identifier of the class and it is in the range $0 \ldots$ number of classes minus 1.
- x_{center}, y_{center} are the coordinates of the center of the bounding box. They are normalized with respect to the image width and height.
- $width, height$ are the dimensions of the annotations normalized by the image width and height (width = bb_width / image_width, height = bb_height / image_height).

Based on the dimension of the labeled bounding boxes in the train set anchors are computed by using k-means clustering. Each cluster contains the representative bounding box dimensions (width, height). Nine clusters are generated. As described by [24] standard k-means clustering with Euclidean distance is employed.

IV. Experiments and results

All our experiments have been done on the FLIR-ADAS dataset [3] because to the best of our knowledge it is the only dataset to provide multi-class annotations for infrared images.

A. Dataset configuration

We have used the split into train and test sets as described by the authors of FLIR-ADAS in order to be able to compare the results with reported state of the art on this dataset.

Class	Train Samples	Test Samples
Person	21924	5779
Car	40711	5432
Bicycle	3581	471
Dog	226	14

TABLE II
Train and Test data distribution

The distribution of the data, as shown in Table II shows that the class dog is represented with only a few samples in the train and test set and this justifies the low detection rate that we have obtained. On the other hand the classes Person, Car and Bicycle are well represented and provide sufficient number of samples for a robust training procedure.

B. Network training parameters

The parameters of the network for each of the two architectures is shown in Table III.

Network	Batch	Subdivisions	learning rate
Yolov3	64	32	0.001
Yolov3-spp	16	8	0.001

TABLE III
Network parameters

The anchor values computed from the train dataset are pairs of width and height for each of the nine clusters. The following values have been obtained: (15.9367,21.7777), (39.2190,31.9081), (23.3501,57.3729), (67.5087,52.1448), (49.4228,115.3734), (107.6429,78.5150), (98.6828,202.4780), (166.9288,123.8407), (296.4098,169.3415).

C. Results

We have trained both YOLOv3 and YOLOv3-spp for 40000 iterations. Every 1000 iterations we compute the mean average precision and we pick the weights that provide the highest mAP. For YoloV3 the network reaches the best mAP after 12000 iterations while for Yolov3-spp the best results are obtained after 37000 iterations.

We compare our results with the ones reported as state of the art on the given dataset, namely with RefineDetect512 [25]. Table IV presents the average precision per class, the mean average precision computed for an IoU of the detection with respect to the annotation greater than 0.5. The precision-recall

Method	Person	Car	Bicycle	mAP
RefineDetect5120	79.4	85.6	58.0	0.587
YoloV3	78.68%	84.92%	66.27%	0.580
yolov3-spp	82.05%	85.78%	66.27%	0.586

TABLE IV
Accuracy of the proposed framework

curves have been computed for both architectures and for three classes: pedestrian, car and bicycle using VOC 2007 evaluation procedure. They are shown in Figures 4a and 4b.

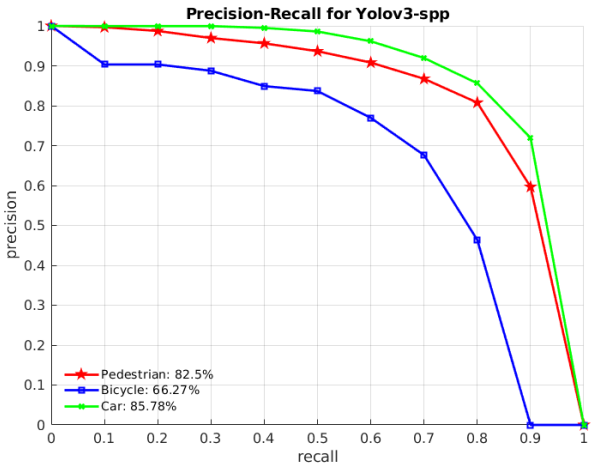
Another metric used for assessing the quality of an object detector is the log-average-miss rate [26]. We also plot the miss-rate against the number of false-positives per image (FPPI). This plot is obtained by varying the threshold on the detection confidence and compute for each threshold the miss rate = (false negatives + true positives)/false negatives and the number of false positives per image. The values are plotted on logarithmic axes. As described by [26] the log-average miss rate is computed by averaging the miss rate at nine FPPI rates that are evenly spaced in the log-space in the range $10^{-2} \ldots 10^{0}$.

These results show the per-class evaluation. In order to have a general view of the accuracy we also list the overall true positive (TP), false positives (FP) and false negatives(FN) for each network setup:

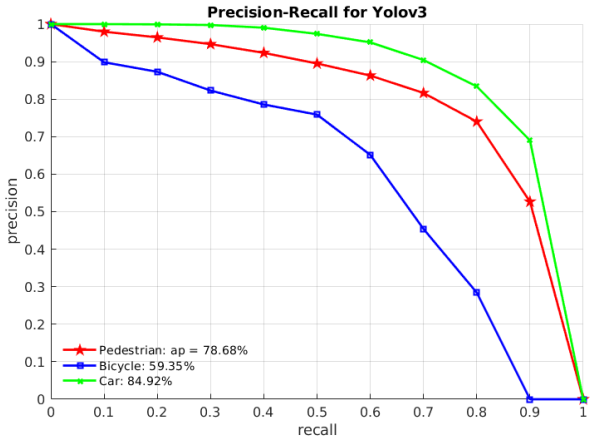
- Yolov3-spp:
  - TP = 8421, FP = 1351, FN = 3275
  - average IoU = 65.07 %
- Yolov3:
  - TP = 7853, FP = 1393, FN = 3843
  - average IoU = 64.08 %

The rate of false positives and false negatives generates the lower log-average miss rate than the reported average precision per class.

Discussion: The class Dog has few train / test samples and the detection score for it is pretty low (below 10%). Probably the results on this class could be improved if the number of labeled instances is larger. Due to this class imbalance issue

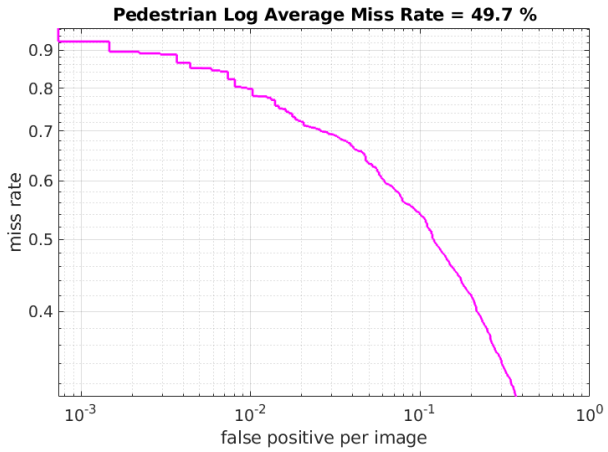


(a) Precision-recall Yolov3-spp

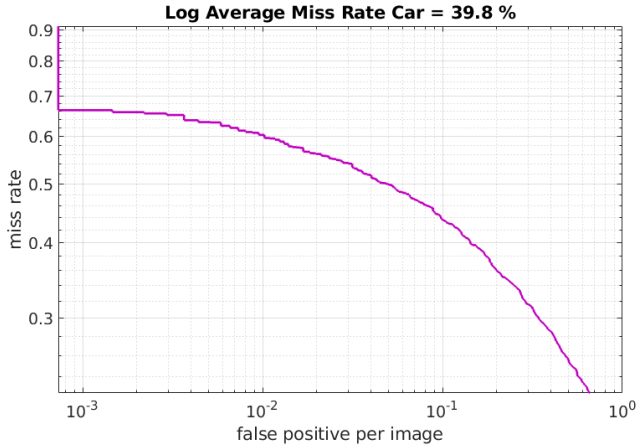


(b) Precision-recall Yolov3

Fig. 4. Precision-recall curves



(a) Log average miss rate for pedestrians



(b) Log average miss rate for cars

Fig. 5. Log average miss rates for pedestrians and cars

the total mean average precision of the proposed method is lower with 0.001 that the reported state of the art. But if the results are analysed per class an improvement with 3% can be noticed for pedestrians, and with 8% for bicycles while the accuracy for cars is quite similar with the state of the art results.

D. Hardware configuration

The described framework has been trained on the system having the following parameters:

- i7 Processor, 16GB memory, 2080Ti GPU.

Using this setup the inference time of the framework is about 40fps (an average of 25ms per frame).

The training process takes roughly 15 hours to complete about 40000 iterations for each of the employed network architecture.

E. Sample detection results

The proposed framework has been applied on each frame in the test set. Some detection results for reference are presented in Figure 6.

As future work we plan to evaluate the proposed framework based on the distance of the objects with respect to the ego-vehicle (that is objects that are smaller - far, medium objects and objects close to the car - large). A particular interest for future work can be given to occluded objects.

V. Conclusion

The paper presents a framework for detecting objects in infrared images. We study the behaviour of two different object detection architectures that have a common backbone, namely Yolo [1] and Yolov3-spp. We describe the actions taken for fine tuning these network architectures in order to work with infrared images.

For training and testing we have used a benchmark dataset [3] that contains night and day image sequences with representative objects of urban traffic. The evaluation based on performance and log average miss rate shows that the proposed topology has very good results, featuring an increased accuracy with respect to the reported state of the art.

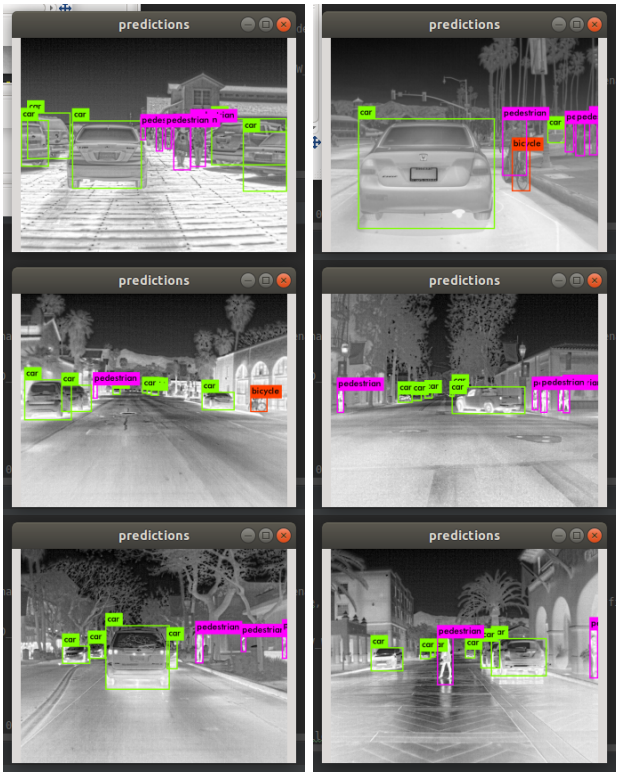


Fig. 6. Sample detection results

As future work we plan to combine the infrared with the visible field in order to check if the two modalities complement each other and provide increased detection results.

ACKNOWLEDGMENT

The results presented in this paper were partially supported in the framework of the GNaC 2018 ARUT grant "Detectia obiectelor in imagini monoculare termale FIR pentru viziune pe timp de noapte", research Contract no. 3091/05.02.2019, with the financial support of the Technical University of Cluj-Napoca, and partially supported in the framework of "Multispectral environment perception by fusion of 2D and 3D sensorial data from the visible and infrared spectrum MULTISPECT", CNCS-UEFISCDI, PN-III-P4-ID-PCE-2016-0727, grant no. 60/2017.

REFERENCES

[1] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[2] Z. Zhao, P. Zheng, S. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2019.

[3] FLIR. Flir thermal datasets for algorithm training. https://www.flir.com/oem/adas/dataset/.

[4] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, Jan 2016.

[6] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Dec 2015.

[7] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[8] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, July 2017.

[9] 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)*, pages 779–788, June 2016.

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[11] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161 – 171, 2019.

[12] 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.

[13] Kevin Fritz, Daniel König, Ulrich Klauck, and Michael Teutsch. Generalization ability of region proposal networks for multispectral person detection. *CoRR*, abs/1905.02758, 2019.

[14] P. Tumas and A. Serackis. Automated image annotation based on yolov3. In *2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, pages 1–3, Nov 2018.

[15] Zhiping Dan, Nong Sang, Jing Hu, and Shuifa Sun. A transfer knowledge framework for object recognition of infrared image. In Tieniu Tan, Qiuqi Ruan, Xilin Chen, Huimin Ma, and Liang Wang, editors, *Advances in Image and Graphics Technologies*, pages 209–214, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[16] Mohamed Elhoseiny, Amr Bakry, and Ahmed M. Elgammal. Multiclass object classification in video surveillance systems - experimental study. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 788–793, 2013.

[17] B. Fardi, U. Schuenert, and G. Wanielik. Shape and motion-based pedestrian detection in infrared images: a multi sensor approach. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pages 18–23, June 2005.

[18] R. Brehar, C. Vancea, and S. Nedevschi. Pedestrian detection in infrared images using aggregated channel features. In *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 127–132, Sep. 2014.

[19] R. Brehar and S. 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, Oct 2014.

[20] R. Brehar, C. Vancea, T. Maria, I. Giosan, and S. Nedevschi. Pedestrian detection in the context of multiple-sensor data alignment for far-infrared and stereo vision sensors. In *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 385–392, Sep. 2015.

[21] M. P. Muresan, R. Brehar, and S. Nedevschi. Vision algorithms and embedded solution for pedestrian detection with far infrared camera. In *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 133–136, Sep. 2014.

[22] Zhanchao Huang and Jianlin Wang. DC-SPP-YOLO: dense connection and spatial pyramid pooling based YOLO for object detection. *CoRR*, abs/1903.08589, 2019.

[23] Redmon J. Darknet: Open source neural networks in c. https://pjreddie.com/darknet/yolo/.

[24] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, July 2017.

[25] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018.

[26] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, April 2012.