ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Object Detection on Thermal Images: Performance of YOLOv4 Trained on Small Datasets.

Maxence Chaverot[1,2], Maxime Carré[2], Michel Jourlin[3], Abdelaziz Bensrhair[1] and Richard Grisel[4]

1- INSA Rouen - LITIS
685 Avenue de l'Université 76800 Saint-Étienne-du-Rouvray - France
2- NT2I
10 Rue Jean Servanton 42000 Saint-Étienne - France
3- Hubert Curien Laboratory
18 Rue Professeur Benoît Lauras, 42000 Saint-Étienne - France
4- INSA Rouen
685 Avenue de l'Université 76800 Saint-Étienne-du-Rouvray - France

Abstract. Thermal sensors are underrepresented in the field of Advanced Driver-Assistance Systems whereas their capabilities to acquire images independently of weather or daytime can be very helpful to achieve optimal pedestrian and vehicle detection. This underrepresentation is due to the small amount of available public datasets. This lack of training samples and the difficulties of building such datasets are a real hurdle to the development of an object detector dedicated to thermal images. Thanks to YOLOv4 and its detection performance, we show in this paper that fine-tuning this neural network requires few samples to achieve satisfying performance, outperforming the results of state-of-the-art detectors.

1 Introduction

Pedestrian detection is a major topic in Advanced Driver-Assistance Systems, ADAS, or video surveillance. State-of-the-art algorithms using a Convolutional Neural Network, CNN, and multiple modalities fusion achieve high performance in this task, while not perfect. Infrared and Visible images fusion detectors are topics of high interest, these two modalities combining themselves decently. However, this implies costly and complex systems, especially because of the sensors alignment requirements. Moreover, in low-light scenes, visible images provide limited information for detection. At night, thermal objects detectors perform as well as visible and thermal fused objects detectors.

Thermal images present less details than visible ones, less texture information, poor spatial resolution, and are blurred due to various causes: atmosphere, misfocusing, high wavelength captured by the sensors and heat radiation. There is also a lack of available datasets. Some exist, like KAIST [1] and FLIR ADAS [2], but compared to visible imaging datasets, the number of labeled objects is much lower. Gathering, filtering, and labeling plenty of images becomes a tedious task. High resolution thermal sensors are highly expensive, and the recurrent low quality of thermal images makes difficult the annotation step.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

In this paper, it is shown that fine-tuning YOLOv4 only requires a small dataset to achieve satisfying results. This could lead to better database constitution and annotation, resulting in more certain and accurate detections.

2 Related Works

2.1 CNN for Object Detection

During the last years, significant progresses have been performed in the field of object detection, due to the emergence of CNN and their efficiency for object classification. In addition of classification, an objects detector needs localization.

Girshick et al. [3] used a region proposal module allowing a CNN to classify them, but to process a single image, around 2.000 classifications are required. Girshick improved this idea with Fast-R-CNN [4], computing the whole image in the CNN and then projecting the proposed region in the last layers of the network to determine the bounding boxes, improving the execution time. Later Ren et al. [5] proposed Faster R-CNN, a network using an object detection algorithm that permits the network to learn by itself the region proposal.

At the same time, Redmond et al. [6] proposed a one stage detector: YOLO, You Only Look Once. It suggests the likely regions and their classification based on a small set of candidate regions and features of the whole image. Single Shot Detector, SSD, is due to Liu et al. [7] combining YOLO and R-CNN to get a good compromise between the quality of detection and execution time. Now, YOLO evolved in new improved versions, achieving state-of-the-art compromise between accuracy and speed [8].

2.2 Object Detection in Thermal Images

While many recent works address multimodal detection using both RGB and thermal images, few works focus on thermal images only. John et al. [9] proposed a CNN for pedestrian detection, then Thermal Images augmented by Saliency maps have been used to train a Faster R-CNN network by Ghose et al. [10]. Domain adaptation is a topic of interest with the work of Hermann et al. [11] transforming thermal images into visible grayscale ones. Later Devaguptapu et al. [12] use Generative Adversarial Network, GAN, to generate synthetic thermal images from RGB, and in the same way, Self-Supervised Network has been explored to maximize information between thermal and visible spectra, increasing performance of objects detection in thermal images [13].

3 Methodology

3.1 YOLOv4

The authors of YOLOv4 have selected an architecture realizing the best compromise between the network resolution, the number of convolution layers and the parameters number at the learning step.

They choose a backbone with high classification quality on MS COCO dataset [14], they use Spatial Pyramid Pooling, SPP, and Path Aggregation Network, PANet, to multiply the receiving fields and aggregate the classifier characteristics with those of

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

the detector. Finally, the head of this network has the same architecture than YOLOv3. Various methods of data augmentation like Flip, Random Scale, CutOut, MixUp, CutMix, Mosaic, and a modified Spatial Attention Module, SAM, permit to obtain better results on MS COCO.

Our objective is to apply object detection to real fields like ADAS, safety, video -surveillance. In these domains, the execution time constitutes an essential challenge. YOLOv4 quality detection appears very efficient compared to the state-of-the-art speed-to-cost ratio.

3.2 **FLIR ADAS Dataset**

FLIR ADAS proposes 10228 images with 80000 annotations. These images have been acquired during a large period, 60% by day and 40% by night in various scenes. The images of the FLIR dataset are available under different formats: the raw images directly issued from the sensor and the modified and enhanced ones thanks to specific FLIR algorithms.

To avoid effects of noise generated by FLIR enhancement algorithms, we use a dynamic expansion, to transform raw images into grayscale images, centered on a fixed mean of 0.5 and a fixed standard deviation of 0.25, as seen in Fig. 1. This transformation permits to obtain an 8-bits grayscale dataset closer to the MS COCO dataset used as a pre-training for YOLOv4 detector.

Figure 1: Comparison between different translations from raw to 8-bits grayscale (Image 1: Obtained from a Min-Max Linear Transformation, Image 2: Provided by FLIR, Image 3: Mean-Standard Deviation Linear Transformation)

Four different classes have been annotated in this dataset: Persons, bicycles, cars, and dogs. Dogs have very few annotations compared to the other classes, we choose not to use this label to train and evaluate the detector.

	Persons	Bicycles	Cars	Dogs	Total Images
Train	13725	3297	36642	178	7860
Validation	4955	441	5209	12	1360

Table 1: FLIR ADAS Classes Distribution on Train and Validation split

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

4 Experiments and Results

4.1 Training

The YOLOv4 detector was fine-tuned with images of size 640x512 pixels, with a batch size of 64. A stochastic gradient descent, SGD, was used with a learning rate of 0.001. Each training used pre-trained weight on MSCOCO and lasted for a minimum of 60 epochs. The score on the validation set was calculated for each epoch and the best one was kept at the end of each experiment.

4.2 Experimental Results

The YOLOv4 evaluation on the FLIR ADAS database is performed according to the train and validation split suggested by FLIR. We have also tested the YOLOv4 model with pre-trained weights on MSCOCO, without fine-tuning. To compare the detectors performance, we choose to use average precision, AP, defined in [15], and mean average precision, mAP. It has been observed that, in the validation set, the number of bicycles annotations is significantly smaller than the number of persons and cars: 4 %, compared to 46.7 % and 49.11 %, respectively. Such a distribution can result in an important bias on the results: in fact, an improvement of the bicycles detection has not the same weight than the same one on the persons or cars classes. That is why we define the weighted mean average precision, w-mAP, which considers the number of annotations in each class, given the following equation:

$$w\text{-}mAP = \sum_{i=0}^{N} AP_i * \frac{l_i}{L}$$

where *(i)* is the class index, *(AP_i)* is the Average Precision of the class, *(l_i)* is the number of labeled objects of the class (*i)* in the test set, and *(L)* the total number of labeled objects in the test set.

The Table 2 shows the comparison between our experiments and state-of-the-art results:

	Persons	Bicycles	Cars	mAP	w-mAP
Faster R-CNN (ResNet-101) [5]	54.8	42.76	67.99	55.18	60.77
SSD-512 (VGG-16) [7]	70.2	53.99	80.55	68.24	74.6
YOLOv4–MSCOCO [8]	81.19	46.27	79.93	69.13	79.11
YOLOv4–MSCOCO-FLIR **(Ours)**	**88.12**	**74.96**	**91.57**	**84.88**	**89.26**

Table 2: Average Precisions (%) between YOLOv4 and other detectors

Such results show the ability of YOLOv4 pre-trained on MSCOCO to recognize persons and cars, even when thermal images are not considered in the learning base. A fine-tuning step produces significantly improved results on the FLIR database.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

4.3 **Ablation Study**

For the considered dataset, various ablation levels have been realized. For each of them, the validation base remains unchanged. Each sub-base is represented by the number 1/N if only one image among N is processed. For example, 1/2 means that one in two images is selected, in order and without randomization. Different compositions are proposed in Table 3:

	Persons	Bicycles	Cars	Total Images
FLIR 1/2	6830	1654	18309	3930
FLIR 1/4	3377	835	9120	1965
FLIR 1/8	1704	409	4523	982
FLIR 1/16	873	206	2244	491

Table 3: FLIR Ablations Classes Distributions.

The Table 4 presents the mAP and the w-mAP scores for FLIR 1/2, FLIR 1/4, FLIR 1/8 and FLIR 1/16:

	Persons	Bicycles	Cars	mAP	w-mAP
FLIR (Full Base)	88.12	74.96	91.57	84.88	89.26
FLIR 1/2	87.65	70.91	91.01	83.19	88.6
FLIR 1/4	86.72	71.87	90.59	84.84	88.0
FLIR 1/8	84.4	65.84	89.24	79.83	86.0
FLIR 1/16	80.52	59.76	88.04	76.10	83.35

Table 4: Average Precisions (%) of YOLOv4 fine-tuned on different ablation levels.

We have repeatedly trained the same dataset with the same parameters and seen a standard deviation of 1% in the results. It is shown that using only a fourth of the images of the FLIR Dataset leads to almost the same performance, within the range of the standard deviation, that with all the available data. We also remark that using one of sixteen images leads to poorer results than using a non-fine-tuned YOLOv4 detector on the Persons and Cars classes. Although fine-tuning on thermal images gives a good increase in results for all classes, the bicycle class gets the higher increase in AP. We believe that the FLIR Database has more variability of the representation of a bicycle than the MSCOCO dataset, explaining this large increase.

5 Conclusion

We have fine-tuned YOLOv4 on thermal images to improve pedestrian and vehicles detection on the FLIR ADAS dataset. This detector outperforms state-of-the-art methods without being trained on thermal images and fine-tuning leads to even better results. We have evaluated the performance between different dataset ablations. We have shown that YOLOv4 allows the use of smaller datasets, simplifying the setup of an object detector dedicated to thermal images. In the case of video-based dataset, we have observed that reducing training data to a fourth of the available images does not

significantly decrease the results. Moreover, we propose an alternative to the mAP score: the w-mAP, considering the distribution of labeled objects in the test set. In a near future, we will work on the possibility to use this detector in a multimodality framework, and more precisely to imagine novel methods able to reduce the training cost of a detector in the case of underrepresented multimodalities.

References

[1] S. Hwang, J. Park, N. Kim, Y. Choi and I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CPVR 2015), pages 1037-1045, 2015.

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

[3] R. Girshick, J. Donahue, T. Darrell and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR 2014), pages 580-587, 2014.

[4] R. Girshick, Fast R-CNN. In *IEEE International Conference on Computer Vision* (ICCV 2015), pages 1440-1448, 2015

[5] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pages 1137-1149, 1 June 2017, 2017.

[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR 2016), pages 779-788, 2016

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu and A.C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision* (ECCV 2016), pages 21-37, Springer, 2016.

[8] A. Bochkovskiy, C.Y. Wang and H.Y.M. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection. In *arXiv e-prints*, 2020

[9] V. John, S. Mita, Z. Liu and B. Qi, Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks. In *Proceedings of IAPR International Conference on Machine Vision Applications* (MVA 2015), pages 246-249, 2015.

[10] D. Ghose, S.M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau and T. Rahman, Pedestrian detection in thermal images using saliency maps. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops* (CVPR-W), 2019

[11] C. Herrmann, M. Ruf and J. Beyerer, CNN-based thermal infrared person detection by domain adaptation. In *Proceedings of Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, vol. 10643, International Society for Optics and Photonics, 2018.

[12] C. Devaguptapu, N. Akolekar, M. M Sharma and V. N Balasubramanian, Borrow from anywhere: Pseudo multi-modal object detection in thermal Imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (CVPR-W), 2019.

[13] F. Munir, S. Azam and M. Jeon, SSTN: Self-Supervised Domain Adaptation Thermal Object Detection for Autonomous Driving. In *arXiv preprint*, 2021

[14] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar and C.L. Zitnick, Microsoft coco: Common objects in context. In *European Conference on Computer Vision* (ECCV 2014), pages 740-755, Springer, 2014.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma and A.C. Berg, Imagenet large scale visual recognition challenge. In *International journal of computer vision*, vol. 115, no 3, pages 211-252, 2015.