



UNIVERSITÀ
DEGLI STUDI
FIRENZE

FLORE

Repository istituzionale dell'Università degli Studi di Firenze

Addressing Domain Shift in Pedestrian Detection from Thermal Cameras without Fine-Tuning or Transfer Learning

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

Original Citation:

Addressing Domain Shift in Pedestrian Detection from Thermal Cameras without Fine-Tuning or Transfer Learning / Fanfani, Marco; Marulli, Matteo; Nesi, Paolo. - STAMPA. - (2023), pp. 314-319. (2023 IEEE International Conference on Smart Computing (SMARTCOMP)) [10.1109/smartcomp58114.2023.00078].

Availability:

The webpage <https://hdl.handle.net/2158/1355513> of the repository was last updated on 2024-04-10T08:39:09Z

Publisher:

IEEE COMPUTER SOC

Published version:

DOI: 10.1109/smartcomp58114.2023.00078

Terms of use:

Open Access

La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (<https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf>)

Publisher copyright claim:

Conformità alle politiche dell'editore / Compliance to publisher's policies

Questa versione della pubblicazione è conforme a quanto richiesto dalle politiche dell'editore in materia di copyright.

This version of the publication conforms to the publisher's copyright policies.

La data sopra indicata si riferisce all'ultimo aggiornamento della scheda del Repository FloRe - The above-mentioned date refers to the last update of the record in the Institutional Repository FloRe

(Article begins on next page)

Addressing Domain Shift in Pedestrian Detection from Thermal Cameras without Fine-Tuning or Transfer Learning

Marco Fanfani, Matteo Marulli, Paolo Nesi

DISIT-Lab, DINFO Dept., University of Florence, Firenze, Italy

<https://www.disit.org>, <https://www.snap4city.org>,

marco.fanfani@unifi.it, matteo.marulli@unifi.it, paolo.nesi@unifi.it

Abstract— The use of thermal imaging to detect the presence of people in indoor and outdoor environments is gaining an increasing attention given its wide applicability in the tourism, security, and mobility domains. However, due to the particular characteristics of different contexts, it is necessary to train/finetuning specifically object detectors for each scenario in order to obtain accurate results. This is due to changes in appearance caused by camera position, scene size, environmental factors, etc. In this paper, we present a data augmentation method that can improve both versatility and robustness of pedestrian detection models based on thermal images. Thanks to our solution, the trained model can deal with unseen thermal data from both indoor and outdoor environments, reliably detecting pedestrians regardless of their apparent size and position in the image, without any fine-tuning or transfer learning, therefore avoiding time consuming labeling activities to fine-tune and deploy the system in different scenarios.

Keywords—YOLOV5, thermal imaging, data augmentation, pedestrian detection

I. INTRODUCTION

Video surveillance systems are increasingly used in public and private sectors [1, 2] given their relevance for several application fields like security, tourism and mobility management. Pedestrian detection, and more generally object detection, is typically addressed by Computer Vision techniques that exploit convolutional neural network and work on RGB images [3, 4]. However, color camera-based video surveillance solutions can reach their limits in certain situations [5], e.g., in low-light conditions or when people are obscured by objects or structures. Additionally, the use of color images, that can be exploited for recognizing people identity, requires special care in order to respect the GDPR (General Data Protection Regulation) [6] and privacy issues in general. Thermal imaging can provide a solution to these problems [7, 8], as it enables the detection of people based on their heat signature rather than visible light and limit the possibility to identify people, relaxing the privacy requirements. In addition, thermal imaging can be useful in the tourism sector, e.g., for monitoring crowded conditions ensuring people safety, and counting presences to address people flow estimation. The purpose of this study is to investigate the use of thermal imaging for the detection of people in indoor and outdoor environments with different camera-scene configurations. The development of people detection solutions based on thermal imaging presents some challenges. For



Figure 1: from left to right: (a) An example of image from Barcelona dataset, (b) An example of image from Florence dataset.

example, environmental factors such as temperature, lighting conditions [9], camera hardware [10], distance between the camera and the targets, and the number of people in the scene can affect detection accuracy. The effects of these factors can vary depending on the environment, weather, and time of day. In addition, using a detector trained for outdoor environments may not be effective indoors, and vice-versa, unless transfer learning is applied. Difference in the observed pedestrian dimensions, that can change due to the distance between the camera and the targets, is another effect that can dramatically affect detection performances. Indeed large-size pedestrians exhibit different visual characteristic wrt small-size ones: as can be seen in Figure 1 (a) and (b), while thermal images of large-size persons show sharp edges and clear details, small-size ones appear blurred with less characteristic features. Special care must be devoted in training to address such dramatic size changes. Additionally, due to the camera placement, people can be concentrated in particular image areas: for example, while in Figure 1 (a) people are almost uniformly scattered over the image, in Figure 1 (b) pedestrians appear only in the lower part of the image. Similar problems have been addressed for RGB images using data augmentation techniques [11] where the available labeled images are transformed in order to obtain a greater number of examples and increase their variance with the aim to obtain more general models. For example, scale invariance was tackled in [12] by using multi scaling, while resizing was used in [13]. Differently, in [14] Kisantal et al. propose to oversample images with small objects and copy and paste small objects to augment their representation in the dataset. However, oversampling requires at least few images with small objects, while copy and paste need accurate segmentation mask in order to avoid the introduction of artefacts during pasting. Therefore, simpler geometrical transformations as resizing and

scaling are more generally usable. Data augmentation can also address object concentrating in specific image locations by shifting and padding image patches. To the best of our knowledge, no particular effort has been devoted on assessing the relevance of data augmentation for detection in thermal images to tackle scale and position invariance.

In this paper, we aim to provide a solution for pedestrian detection in indoor and outdoor environments under strong changes in scale and position using thermal images. In particular, the paper proposes an easy and practical solution to increase flexibility in the range of applications to obtain relevant precision in different conditions without the need to fine-tune or retrain the model. The solution is based on a procedure for improving the learning process via data augmentation in order to address changes due to indoor and outdoor environments, size of the scene and apparent pedestrian size, different point of views/perspectives of the cameras. The research has been developed in the context of national Center on Sustainable Mobility, MOST, of Italy, with the aim of developing easy to use solution for people detection in mobility conditions: indoor and outdoor. The development exploited the facilities of DISIT lab and Snap4City platform (<https://www.snap4city.org>) [15], [16].

This paper is organized as follows. Section II presents the data and assessment method. Section III describes the proposed approach in terms of training procedure. In Section IV, the evaluation results are reported and discussed. Finally, in Section V conclusions are drawn.

II. DATA AND ASSESSMENT METHODOLOGY

According to the introduction, several infrared and thermal imaging datasets have been considered. The main datasets used for the training phase were (i) the benchmark datasets LLVIP (Low Light Vehicle Identification and Verification Program) [17], and (ii) Teledyne FLIR ADAS (Advanced Driver Assistance Systems) [18]. LLVIP is a dataset consisting of about 150000 images annotated for different object categories and acquired with infrared cameras under different lighting conditions. The second dataset, Teledyne FLIR ADAS, contains more than 10000 manually labeled thermal images taken under different lighting and visibility conditions. LLVIP and FLIR ADAS were used to develop and test video surveillance systems, advanced driver assistance systems, such as pedestrian, and cyclist detection. Both datasets were filtered to take the images with humans. For the evaluation phase we used two thermal imaging datasets acquired under different conditions in Barcelona and Florence. The Barcelona dataset has been obtained by randomly selecting 212 images from 392 videos. These videos were taken during the three days of the annual Smart City conference in Barcelona from Nov. 15, 2022, to Nov. 17, 2022. Each image was then manually annotated with a bounding box for the positions of the people. As can be seen in Figure 1 (a), some people may appear taller than others depending on the perspective on how far they are from the wide-angle camera, which is mounted on a tripod at a height of 3.5 meters. Since Barcelona is an indoor space, there are no natural light sources, and the temperature of the pavilion is controlled by ventilation systems. People can be numerous and appear anywhere in the picture. The view is a one of the cross points in

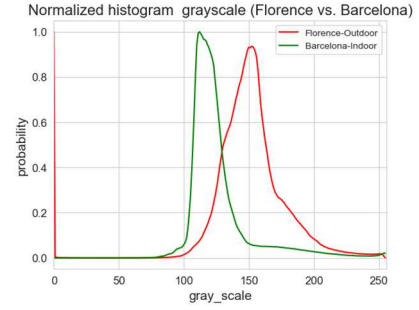


Figure 2: Histogram of gray scale for datasets.

which two large corridors intersect each other. The Florence dataset consists of images taken by a thermal imaging camera with a wide-angle lens in Piazza della Signoria in Florence on the days from 14/07/2022 to 19/07/2022. In this case, we applied the same sampling strategy as in Barcelona to obtain a sample of 219 images. As you can see in Figure 1 (b), because of the distance of the camera from the scene (the camera is 30 meters away from the scene), the people appear very small and are only in the lower part of the scene. Also, the scene is very crowded and there are usually between 80 and 120 people in the images. In this case, the images represent an outdoor environment, so the lighting of the scene is very dynamic, changing from daylight to night, illuminated by streetlamps. The temperature also changes, increasing during the day and decreasing at night. Manual labeling was performed using bounding boxes.

In both cases (Florence and Barcelona) the images were pre-processed to eliminate the wide-angle image effects, since most of the datasets used to train algorithms (such as LLVIP and FLIR ADAS) do not contain wide-angle images. It should be noted that such processing does not limit the applicability of the proposed detector, since in video surveillance applications the camera is usually accessible and offline calibration can be performed. These two datasets differ in terms of environmental conditions. It can be also noted by observing the cumulative histograms of the grey level distribution. These histograms were obtained by calculating the grayscale histogram for each image and then summing to obtain the cumulative histogram. The Florence dataset images show a higher concentration of pixels with values between 100 and 200, see Figure 2, while the Barcelona dataset mainly includes pixels with values ranging from 100 to 150. In the specific case, the Kullback-Leibler divergence [19] between the two datasets is 1.41. Thus, the above-described datasets represent relevant test benches for the experiments since they depict two environments with strongly different characteristics, in terms of scale, and perspective. In the following Sections, a particular care has been devoted to the training process to obtain a model network capable to deal with different environments and avoid a relevant loss in performance in passing from the usage of a model in the conditions similar to those in which has been trained to a new case in which the operative conditions are quite different as explained above. Table 1 reports the numbers of training and validation sets of images for the above-mentioned datasets.

A. Evaluation metrics

Typically, specific metrics are used to evaluate the performance of an object detection system [20] as reported in

Dataset	Train example	Validation example	TV cam Kind
LLVIP	12025	3463	Infrared
FLIR ADAS	10742	1144	Thermal
Florence	175	44	Thermal
Barcelona	169	43	Thermal

Table 1: Volume of data sets images for the used data.

the following. These metrics are influenced by a basic metric called the intersection over union, often abbreviated as IoU. Given the areas of the inferred bounding box A_1 and the area of the ground-truth bounding box A_2 , IoU measures the amount of area the two rectangles have in common as:

$$IoU = \frac{A_1 \cap A_2}{A_1 \cup A_2}$$

Higher IoU scores indicate better detection results. This metric directly affects the precision and recall metrics, as the IoU affects the detection of true positives, false positives, and false negatives, which determine the precision and recall rates. After selecting a threshold for the IoU, a detection that meets or exceeds the threshold is classified as a true positive, otherwise as a false positive. False negatives are detections that did not occur. Rather than using precision and recall separately, the mean average recision metric (mAP) is used. In order to evaluate the goodness of a detector, for all experiments mAP@50 is considered, that calculates the average precision (AP) for all classes with an intersection over union (IoU) greater than 0.5.

B. YOLOV5

In this work, the YOLOV5 architecture, proposed by Ultralytics [21], was used. The first version of Yolo was introduced in [22]. YOLOV5 is made up of 3 main parts: the backbone CSPDarknet [23], the neck PANet [24], and the head Yolo layer. The backbone CSPDarknet is used as a feature extractor and is designed to provide a balance between accuracy and speed. The neck PANet is made up of a series of modules that connect the backbone to the head of the neural network that consists of a Spatial Pyramid Pooling [25] (SPP) module, which allows information to be captured at different spatial scales. Finally, the YOLOV5 head is responsible for predicting the bounding boxes and the classes of the detected objects. It uses a combination of convolutions, batch normalization and activation functions to produce the predictions.

C. Loss function of YOLOV5

The YOLOV5 loss is the sum of 3 losses:

$$Loss = l_{box} + l_{cls} + l_{obj}$$

which are the bounding box regression loss function, the classification loss function, and the confidence loss function, respectively defined as:

$$l_{box} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B I_{i,j}^{obj} (2 - \widehat{w}_i \times \widehat{h}_i) + \left[(x_i - \widehat{x}_i)^2 + (y_i - \widehat{y}_i)^2 + (w_i - \widehat{w}_i)^2 + (h_i - \widehat{h}_i)^2 \right]$$

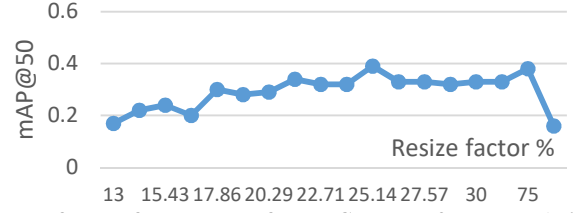


Figure 3: Performance of the Camera52 model (wide angle camera used in Florence and Barcelona) as the resize factor varies.

$$l_{cls} = \lambda_{class} \sum_{i=0}^{S^2} \sum_{j=0}^B I_{i,j}^{obj} \sum_{c \in \text{classes}} p_i \log(\widehat{p}_i)(c)$$

$$l_{obj} = \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^B I_{i,j}^{noobj} (c_i - \widehat{c}_i)^2 + \lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^B I_{i,j}^{obj} (c_i - \widehat{c}_i)^2$$

where λ_{coord} is the position loss coefficient, λ_{class} is the category loss coefficient, \widehat{x} and \widehat{y} are the true central coordinates of the target, \widehat{w} and \widehat{h} are the width and height of the target. $I_{i,j}^{obj}$ and $I_{i,j}^{noobj}$ are two binary variables that take values 1 or 0. Specifically, if there is an object in the anchor box at position (i,j), then $I_{i,j}^{obj}$ takes the value 1 and $I_{i,j}^{noobj}$ takes the value 0. Otherwise, $I_{i,j}^{obj}$ takes the value 0 and $I_{i,j}^{noobj}$ takes the value 1. $p_i(c)$ is the probability of the object belonging to class c, while \widehat{p}_i is the true probability of the object belonging to class c.

III. PROPOSED METHOD

This works started from a previous application where a domain adaptation strategy, the bottom-up layerwise [27], was applied on a YOLOV5 model (pretrained on COCO [26]) using a mixture of images taken by LLVIP and Florence, the model was named as Camera52. The results obtained by the usage of Camera52 model on Barcelona dataset was very poor, mAP@50 = 0.19. The main difference is the size of the people (e.g., Figures 1 (a) and 1 (b)). The retraining of the Camera52 model with a set of Barcelona images produced a relevant improvement, and took a relevant amount of time and resources, which is what we would like to avoid since the TV Cam should be typically installed in several different scenarios without the need of changing the model. Therefore, an initial study to identify the best image resize has been performed without changing the Camera52 model. The approach identified a resize factor of 25.19% obtaining a mAP@50=0.39. Figure 3 shows the mAP@50 trend of Camera52 model as a function of resize factor, confirming the fact that the resize is relevant for improving performance.

Other factors governing performance are the number of people to be detected and where they are placed in the scene, for example if no people appear on the left or bottom left of the scene then the detector will never look at those areas of the image for people.

Therefore, as declared in the introduction, our proposed method aims to create a robust data augmentation that can help a detector model trained on a benchmark dataset to achieve reasonable performance on novel unseen conditions without

Algorithm 1: Thermal data augmentation procedure

Data: D : dataset, $resizeFactor$: list of floats, $paddingPositions$: list of string

Result: \hat{D} : augmented dataset

$\hat{D} \leftarrow \{\}$

for $R \in resizeFactors$ **do**

for $P \in paddingPositions$ **do**

for $I, B \in D$ **do**

$\hat{I} \leftarrow resizeImage(I, R)$;

$\hat{B} \leftarrow resizeBBBoxs(B, R)$;

$\hat{I} \leftarrow gaussianBlur(\hat{I}, kernel \leftarrow 3 \times 3)$;

$\hat{I} \leftarrow horizontalFlip(\hat{I}, prob \leftarrow 0.5)$;

$\hat{B} \leftarrow horizontalFlipBBBox(\hat{B}, prob \leftarrow 0.5)$;

$\hat{I} \leftarrow zeroPadding(\hat{I}, P)$;

$\hat{B} \leftarrow zeroPaddingBBBox(\hat{B}, P)$;

for $\hat{B}_i \in \hat{B}$ **do**

if $\hat{B}_i.area \leq 100$ **then**

$\hat{B}.remove(\hat{B}_i)$;

end

end

$\hat{D}.append(\hat{I}, \hat{B})$;

end

end

for $\hat{I}, \hat{B} \in \hat{D}$ **do**

if $\hat{B}.length = 0$ **then**

$\hat{D}.remove(\hat{I}, \hat{B})$;

end

end

using any labeling or fine-tuning activity. The approach is based on data augmentation with the aim of producing a training set which included different sizes and positioning in the images. The procedure has been applied to the LLVIP and FLIR ADAS datasets, resulting in a fourfold increase in the size of the LLVIP dataset, referred to as LLVIP4X dataset. Similarly, the FLIR ADAS dataset was increased by a factor of two, FLIR-ADAS-aug. Figure 4 shows some examples of augmented data generated by using the thermal data augmentation technique.

The algorithm of the proposed data augmentation method is reported in Algorithm 1. The procedure takes as input an annotated image dataset, along with two vectors for resize and padding. The $resizeFactors$ vector contains different scaling factors, while the $paddingPositions$ vector contains a list of strings representing different padding strategies, such as "padding_center", "padding_left", "padding_right", and so on. The procedure first creates an empty set called \hat{D} , and for each resize factor and padding strategy it extracts an image and its annotations from the D dataset. Then, various functions are applied to manipulate the image and its bounding boxes. The $ResizeImage$ function applies the resize to the image, and since the image is subsampled, the bounding boxes must be scaled according to the resize factor to be consistent with the new dimensions of the image. However, subsampling can cause aliasing and artifacts that degrade the quality of the resulting image. To remove these unwanted effects, Gaussian smoothing is applied using a 3×3 kernel matrix, and the library that implements Gaussian smoothing obtains the standard deviation sigma according to the dimensions of the kernel matrix. To get more images, a horizontal flip effect is applied with a certain probability, which is set to 0.5 by default. This effect helps to make the mesh independent of the positions that need to be viewed to identify people. The bounding boxes must be flipped

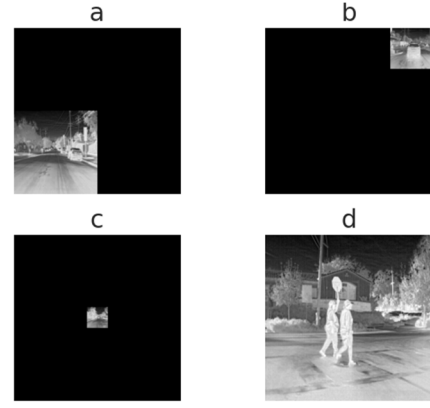


Figure 4: Some data images generated by the data augmentation procedure.

to match the flipped image. This is performed with the $horizontalFlipBBBoxs$ function. A padding is applied that executes the strategy specified by the P variable. The effect of padding returns the image to its original spatial dimensions since the original images were 640×640 pixels by default. In addition, the padding effect forces the detector to scan the entire image to find objects of interest. Again, the bounding boxes must be computed consistently by the $zeroPaddingBBBoxs$ function, which takes into account the padding strategy applied to the image. To avoid numerical instability and false negatives, objects present in image \hat{I} with an area less than 100 pixels have been removed. Finally, an additional check is performed to remove images without bounding boxes from \hat{D} .

IV. EXPERIMENTAL RESULTS ASSESSMENT

The default recommended hyperparameters of YOLOV5 were initially used. The YOLOV5s architecture was used and initialized with the pre-trained weights from COCO. The process of tuning hyperparameters led us to change the batch size to 64, and the number of epochs to 1000. All experiments were run on a computer with an Intel(R) Xeon(R) W-2235 @ 3.80GHz processor, 32GB DDR5 RAM, and an Nvidia RTX3900 GPU.

A. Experimental Results

A series of experiments were conducted to evaluate the effects of the data augmentation procedure. Initially, it was created a YOLOV5s COCO model retrained on the LLVIP (called Model 1) and another YOLOV5s trained with the augmented data set produced with the data augmentation above-described LLVIP4X (called Model 2). Both models have been evaluated wrt to the LLVIP validation set, LLVIP-VAL as reported in Table 2. Both models achieved excellent results in terms of $mAP@50$, with values of 0.967 and 0.966, respectively. The difference in performance between the two models was minimal, only 0.001, indicating that the proposed data augmentation had no negative impact since the model trained on the augmented data was able to cover all the cases in the dataset, the results were reported in Table 2.

A similar analysis has been repeated for FLIR ADAS dataset. Two more models were created, one by using

Model	training			validation	
	Coco	LLVIP	LLVIP4X	mAP@50 LLVIP- Val	mAP@50 difference
1	X	X		0.967	—
2	X		X	0.966	0.001

Table 2: Comparison between YOLOV5s-COCO trained on LLVIP and on LLVIP4X.

Model	Training recipes			validation	
	Coco	FLIR- ADAS	FLIR- ADAS- aug	mAP@50 FLIR- ADAS- VAL	mAP@50 difference
3	X	X		0.831	
4	X		X	0.817	0.014

Table 3: Comparison between YOLOV5s models trained on FLIR ADAS and on FLIR ADAS augmented.

YOLOV5s COCO model tuned with FLIR ADAS (Model 3) and one by using FLIR-ADAS-aug (Model 4), with both models evaluated on the FLIR ADAS validation set. Once again, both models achieved excellent results in terms of mAP@50, with values of 0.831 and 0.817, respectively, and the differences in performance between the two models was 0.014. Again, these results proof that the data augmentation had no negative effects for FLIR ADAS, see Table 3.

This type of analysis was performed for other training combinations as reported in Table 4. Model 1 and Model 2 were retrained on FLIR ADAS and FLIR-ADAS-aug, respectively. Four new models emerged from these trainings and were evaluated using the LLVIP validation set. The results of the validation set show that all Models (identified as Model 5, 6, 7, 8, respectively) are above 0.7 in terms of mAP@50, in particular model 8 reaches a value of 0.766.

The models have been compared with former Model 1 to measure how much they omitted from the previous data set. This provides information on whether the data augmentation procedure improves the generalization capacity of the models. The Models 5, 6, 7, and 8 had a 20% to 25% loss of information, especially the models that were trained with at least one augmented dataset such as Model 5 and Model 7, were found to be less forgetful than Model 6, which was trained without ever seeing augmented data. The model with the best recall of all was Model 8, which saw only augmented data from both data sets, allowing it to forget only 20% of the information learned during initial training. Thus, the approach with direct training using the augmented training set resulted to produce better results.

To further test the generalization capabilities of the different models obtained, they were also tested on two additional validation sets of the Florence and Barcelona datasets, which have been described in Section II. These datasets have never been used to train or tuning the different models, and thus represent an additional evaluation step for the data augmentation procedure. In this experiment, see Table 5, all the models obtained in the previous experiments were used, taking Model 1 as the baseline. The best performance for Barcelona

Model	training					validation	
	Coco	LLVIP	LLVIP4X	FLIR ADAS	FLIR- ADAS- aug	mAP@50 LLVIP-val	mAP@50 difference
1	X	X				0.967	
2	X		X			0.966	0.001
5	X	X			X	0.715	0.252
6	X	X		X		0.703	0.264
7	X		X	X		0.750	0.217
8	X		X		X	0.766	0.201

Table 4: Summary of models obtained in the different experiments forgetting the information learned from the LLVIP dataset.

Model	Training recipes					Validations	
	Coco	LLVIP	LLVIP4X	FLIR ADAS	FLIR- ADAS- aug	mAP@50 Florence- val	mAP@50 Barcelona- val
1	X	X				0.305	0.719
2	X		X			0.450	0.738
3	X			X		0.576	0.719
4	X				X	0.546	0.773
5	X	X			X	0.530	0.761
6	X	X		X		0.554	0.735
7	X		X	X		0.580	0.728
8	X		X		X	0.549	0.738

Table 5: Comparison between YOLOV5s on unseen data taken from Florence and Barcelona datasets.

was obtained by Model 4, which obtained a mAP@50 of 0.773. For Florence, the best performance was obtained with the Model 7, which obtained a mAP@50 of 0.580. Note that a direct comparison of the results of Table 5 and those reported in Section III (i.e., Camera52 Model) is not straightforward. In Section III, the YOLO model was trained on COCO, LLVIP and Florence data sets: this led to excellent scores on Florence and poor results on Barcelona (even after resizing) obtaining a mAP@50 of 0.19 (0.39 after resizing). The fine-tuning of the YOLO-COCO-LLVIP with Florence dataset let the detector focus on small-size pedestrians, losing the capability to detect big-size ones. Differently, using the proposed data augmentation (as for Models 4 and 7) the detector can work sufficiently well with both Florence and Barcelona datasets, that were not used in model training, confirming the validity of our solution.

The results presented in Table 5 shown that the data augmentation procedure improved the generalization capabilities of YOLOV5, even for data that were not part of the original training set, which was the goal of the study. On the other hand, training a model directly on new dataset (e.g., obtained with YOLO-COCO-LLVIP fine-tuned with Florence or Barcelona) yield significantly better results, as shown in Table 6. Nevertheless, such an approach requires a training and not a direct usage of the model on the TV Cam as one expect when a new TV camera is installed in a new and different context. The training also implies a manual labeling of the new

Model	mAP@50	mAP@50 (exp)	mAP@50 (diff)
Florence	0.778	0.580	0.198
Barcelona	0.914	0.773	0.141

Table 6: Results of the models trained on the previous datasets and their comparison with the best models that used data augmentation.

dataset, a tedious activity that should be avoided when deploying the model in new scenarios.

V. CONCLUSIONS

This paper addressed the problem of pedestrian detection in indoor and outdoor environments under flexibility in terms of large-scale and position changes using thermal images. To solve this problem, a data augmentation approach has been proposed, which strengthen the generalization capabilities of a YOLOV5 model, making it stronger to produce better results with respect new conditions and data having very different characteristics. The validation has been demonstrated by using a set of benchmarks. Several tests have been performed to verify that the procedure did not degrade the performance of the models, and to quantify how much the models trained on the augmented data forgot the information from the previous training. Finally, a number of tests to measure the generalization capabilities towards new data not used during training were carried out. Results show that the procedure improve the performance of the obtained models, preserves most of the information from the previous training, and allows obtaining models with good generalization capabilities on unseen data even if they have different characteristics.

ACKNOWLEDGMENT

The authors would like to thank the MIUR, the University of Florence and the companies involved for co-founding the national Center on Sustainable Mobility, MOST. A thanks to the many developers on Snap4City platforms. Snap4City (<https://www.snap4city.org>) and platform is an open technology of DISIT Lab.

REFERENCES

- [1] Elharrouss, Omar, Noor Almaadeed, and Somaya Al-Maadeed. "A review of video surveillance systems." *Journal of Visual Communication and Image Representation* 77 (2021): 103116.
- [2] Ingle, Palash Yuvraj, and Young-Gab Kim. "Real-Time Abnormal Object Detection for Video Surveillance in Smart Cities." *Sensors* 22.10 (2022): 3862.
- [3] L. R. Barba Guamán, J. Naranjo, A. Ortiz, and J. Pinzon Gonzalez, "Object Detection in Rural Roads Through SSD and YOLO Framework," 2021, pp. 176–185. doi: 10.1007/978-3-030-72657-7_17.
- [4] A. Khalfaoui, A. Badri and I. E. Mourabit, "Comparative study of YOLOv3 and YOLOv5's performances for real-time person detection," 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), 2022, pp. 1-5, doi: 10.1109/IRASET52964.2022.9737924.
- [5] Indhuja, U. S., and R. Amutha. "Pedestrian Detection in Extreme Weather." *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2021.
- [6] GDPR: General Data Protection Regulation, <https://gdpr.eu/>
- [7] Hwang, Soonmin, et al. "Multispectral pedestrian detection: Benchmark dataset and baseline." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
- [8] Li, Yi-Hao, et al. "Weighted HOG for Thermal Pedestrian Detection." *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 2018.
- [9] Kim, JongBae. "Pedestrian detection and distance estimation using thermal camera in night time." *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*. IEEE, 2019.
- [10] Rujikietgumjorn, Sitapa, and Nattachai Watcharapinchai. "Real-time hog-based pedestrian detection in thermal images for an embedded system." *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017.
- [11] P. Kaur, B. S. Khehra and E. B. S. Mavi, "Data Augmentation for Object Detection: A Review," 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), Lansing, MI, USA, 2021, pp. 537-543, doi: 10.1109/MWSCAS47672.2021.9531849.
- [12] R. Girshick, "Fast R-CNN", *Proc. Int. Conf. Comput. Vis.*, pp. 1440-1448, 2015.
- [13] Montserrat, D. M., Lin, Q., Allebach, J., & Delp, E. J. (2017). Training object detection and recognition CNN models using data augmentation. *Electronic Imaging*, 2017(10), 27-36.
- [14] Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., & Cho, K. (2019). Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*.
- [15] A. Arman, C. Badii, P. Bellini, S. Bilotta, P. Nesi, M. Paolucci, Analyzing demand with respect to offer of mobility, *Applied Science*, MDPI, 2022. <https://www.mdpi.com/2076-3417/12/18/8982>
- [16] F. Alberti, A. Alessandrini, A. Masiero, M. Meocci, A. Paliotto, D. Bubboloni, C. Catalano, M. Fanfani, P. Nesi, M. Loda, A. Marino, "Mobile Mapping to Support and Integrated Transport-Territory Modeling Approach", 12th International Symposium on Mobile Mapping Technology (MMT 2023).
- [17] Jia, Xinyu, et al. "LLVIP: A visible-infrared paired dataset for low-light vision." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
- [18] "Teledyne flir dataset" Teledyne flir. <https://www.flir.it/oem/adas/adas-dataset-form/> (accessed April 04, 2023).
- [19] Shlens, Jonathon. "Notes on kullback-leibler divergence and likelihood." *arXiv preprint arXiv:1404.2000* (2014).
- [20] "Evaluating-Object-Detection-guide" <https://manalelaidouni.github.io/> , <https://manalelaidouni.github.io/Evaluating-Object-Detection-Models-Guide-to-Performance-Metrics.html> (accessed April 04,2023)
- [21] "YoloV5" YoloV5 by Ultralytics. <https://ultralytics.com/yolov5> (accessed April 04, 2023).
- [22] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [23] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934* (2020).
- [24] Liu, Shu, et al. "Path aggregation network for instance segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [25] He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015): 1904-1916.
- [26] Xu, Renjie, et al. "A forest fire detection system based on ensemble learning." *Forests* 12.2 (2021): 217.
- [27] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer International Publishing, 2014.