Evaluation of Geometric and Photometric Data Augmentation for Pedestrian Detection with Thermal Cameras

Marco Fanfani, Matteo Marulli, and Paolo Nesi(✉)

DISIT Lab of Department of Information Engineering, University of Florence, Florence, Italy
{marco.fanfani,matteo.marulli,paolo.nesi}@unifi.it
https://www.disit.org, https://www.snap4city.org

Abstract. The capability to detect pedestrians using video surveillance systems is a fundamental task to assess the number of people in particular areas and to enable other applications like people tracking and trajectory estimation. Thanks to recent advancements of object detectors based on deep learning approaches, it is possible to localize pedestrians in recorded images with good accuracy. However, some limitations still exist. Most of the proposed solutions at the state-of-the-art deal with color RGB images, that could present privacy issues, hampering the deployment of such solutions. Moreover, changing the application scenario could worsen the performance, requiring time consuming fine tuning and transfer learning operation. In this paper, we respond to these issues by using thermal images analysis that can guarantee a higher privacy preservation instead of RGB data. In addition, a deep study on data augmentation techniques has been provided to generate more flexible trained models that could simplify their adoption in different scenarios without requiring additional acquisition, labeling work and fine tuning.

Keyword: Pedestrian Detection · Data Augmentation · Thermal Imaging

1 Introduction

Computer vision systems are increasingly deployed both in public and private sectors and video surveillance [1, 2], for example, to assess the presences and the flows of people, in several application scenarios of security [3], retails [4], mobility [5], etc. To detect the pedestrians by video cameras, many computer vision methods for object detection have been developed leveraging on machine learning solutions based on convolutional neural networks (CNN) [6]. Among the several proposed architectures, the YOLO approach [7] appeared to be one of the best solutions mostly thanks to efficient computational parameters guarantying higher detection speed with respect to other proposed region-detection methods [8–10], which is a fundamental characteristic for realizing real application scenarios, possibly requiring on-the-edge computations. Recently, different versions and different variants of the YOLO architecture emerged, in most cases increasing precision performance or optimizing computational aspects [11, 12].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14822, pp. 353–366, 2024.
https://doi.org/10.1007/978-3-031-65318-6_24

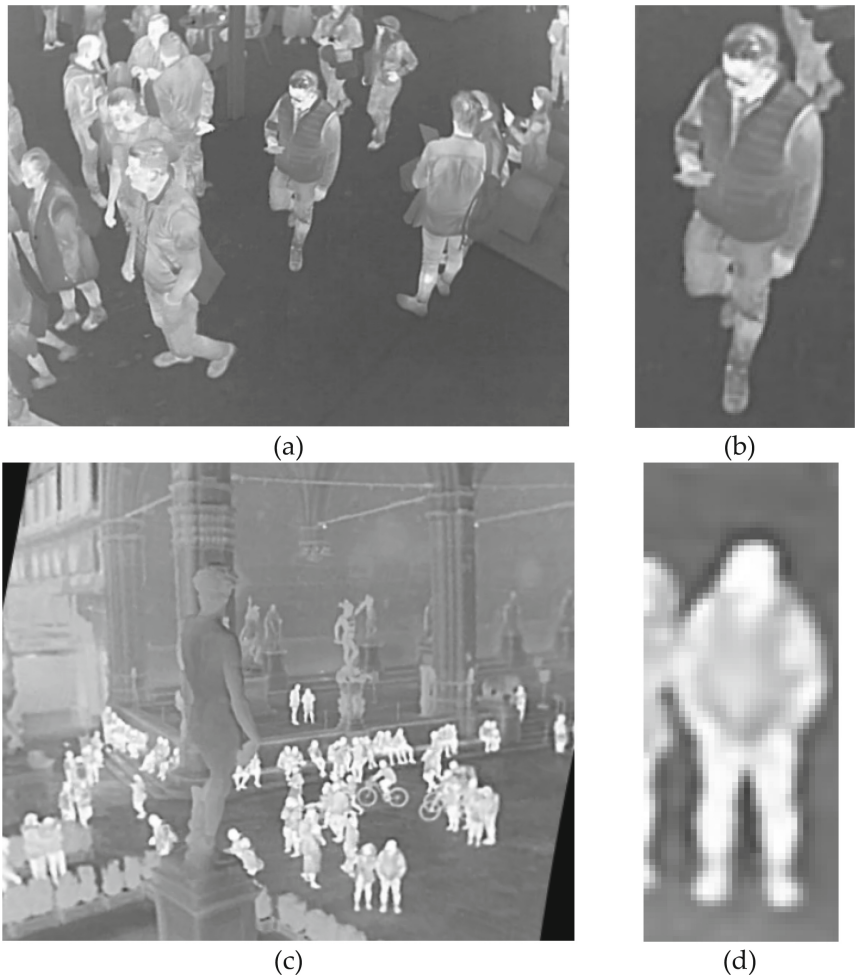
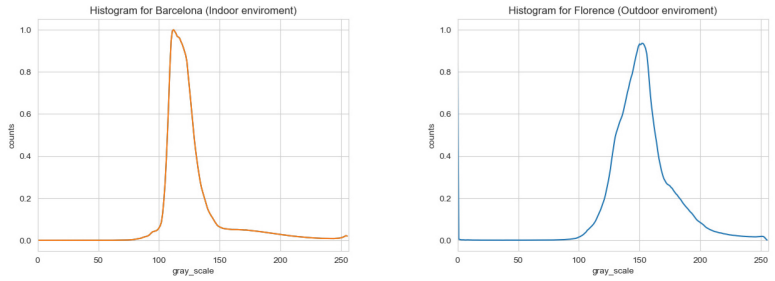


Fig. 1. (a) example of a thermal image acquired by a thermal camera near to the subject sampled from the Barcelona dataset, and (b) detail of a pedestrian in the scene. (c) a different example captured by a thermal camera placed far from the scene (Florence dataset), and (d) a single pedestrian in the scene.

Usually, CNN image object detectors are designed to work on color RGB images. However, video surveillance systems have to work in challenging scenarios, often experimenting difficult light conditions [13] that could hinder the performance by altering the photometric characteristics of the image. Moreover, color images are not privacy preserving – in fact, in most cases people identities can be recognized and matched with some database [14, 15] – and their utilization in video surveillance can impose the use of anonymization solution (for example the blurring of faces) in order to respect privacy regulation, e.g., GDPR [16]. On the contrary, thermal cameras can provide an alternative solution to mitigate those problems. Since thermal imaging capture the heat signature



(a) (b)

Fig. 2. Cumulative histograms of the intensity distribution taken with the same Thermal camera (AXIS Q1952) of the indoor Barcelona (a) and outdoor Florence (b) datasets.

of the elements in the scene, it can help to build a robust detection method less affected by lighting conditions, and can reduce the critical privacy related issues of GDPR [17]. Nevertheless, the adoption of thermal cameras for people detection can still be challenging in particular when addressing domain shifts between data used for training and the actual application scenario, since the information collected is less. For example, changes in the environment, weather, and time of day, the camera position with respect to the scene, the number of persons in the scene, etc., can lead to a performance decrease [18]. For example, as it has been represented in Fig. 1 which depicts two thermal images, (a)/(b) are extracted from indoor case at an Expo (Barcelona dataset) and captured with a camera close to the scene, while the other (c)/(d) are samples from Florence dataset, obtained from a camera placed at greater distance. As, it can be noticed, they present strong visual differences due to the scale change. Pedestrians captured from close cameras (see Fig. 1b) show sharp edges and clear details while, when the camera is far from the scene, persons appear blurred with less characteristic features (see Fig. 1d). Moreover, with different camera placements that can produce images with different field of view, people can appear almost uniformly scattered over the image (see Fig. 1a) or be concentrated in specific part of the image, as in Fig. 1c where persons are found in the lower part. Still observing Figs. 1a and 1c, it can be noted that the two datasets exhibit also different intensities conditions, since images are captured in indoor and outdoor scenarios respectively. Indeed, we reported in Fig. 2 the cumulative histograms obtained by calculating the grayscale histogram for all images in both the Barcelona and Florence datasets and then summing up the values. The Barcelona dataset mainly includes pixels with values ranging from 100 to 150, see Fig. 2a, while the Florence dataset shows a higher concentration of pixel with values between 100 and 200, see Fig. 2b, demonstrating the photometric difference between the two datasets. To address such difficulties when using color images, data augmentation techniques [19] have been proposed to improve the generalization capabilities of the trained model creating additional examples by altering the original images in the training set using geometric or photometric transformations. While several works studied the data augmentation for color images [20–22], less attention has been devoted on assessing its importance for thermal images.

In this paper, we propose a thorough experimental validation of the effect of data augmentation for people detection with thermal cameras. This work extends out previous

analysis [22] – where we investigated the effect of geometric augmentation (i.e., scaling and shifting) – by addressing photometric changes. The main contribution of this paper are:

- The definition of photometric augmentation procedures to create novel training examples with different lighting conditions;
- An in-deep assessment of results of training using the photometric augmented datasets with respect to two datasets not used to train the former neural network model;

This research was carried out in the context of CN MOST, the Italian National Center on Sustainable Mobility [23]. The work has been developed exploiting the Snap4City platform of DISIT lab for data collection and management [24, 25].

The paper is organized as follows: in Sect. 2, the geometric and the photometric augmentation algorithms are presented, while experimental results are reported in Sect. 3. Conclusions are drawn in Sect. 4.

2 Materials and Methods

In this section, the data augmentation and preprocessing algorithms used to produce the geometric and photometric transformed images are presented in detail.

2.1 Geometric Data Transformation

The geometric augmentation takes into account two main transformations: scale and shift. The former is used to train the CNN by using pedestrians appearing with different size in the image. The latter addresses the problem of finding people in specific part of the image. In order to produce such augmented examples, a set of original images are scaled and then the resized patch is put in different positions in the original image canvas padding with zeros the rest of the image. See Fig. 3 for an example.

More in detail, the procedure takes as input an annotated image dataset, and two vectors: the *resizeFactors* vector, that contains the scaling factors, and the *paddingPositions* vector that is a list of strings representing the positions where the resized image should be placed. At first an empty set called $\widehat{D}$ is created. For each resize factor and position strategy an image with its bounding box annotations is extracted from the original dataset D. Then, various functions are applied to manipulate the image and its bounding boxes. As a first step, the images are scaled according to the specified resize factor. Then, to remove possible aliasing and artifacts due to the subsampling, a Gaussian smoothing is applied using a 3×3 kernel. A horizontal flip effect in our experiments has been applied randomly with a probability of 0.5, to further augment the datasets. Finally, the subsampled image is placed in given position on the original image canvas and zero-padding is applied to the rest of the image to simulate a shift of the image content, and to avoid the trained detector focusing on empty areas in search of pedestrian. Note that, all these transformations work on the images and on the respective annotations that must remain consistent. Objects as their bounding boxes resulting after the transformation with an area inferior to 100 pixels are removed to avoid numerical instability and false negatives. Finally, images without bounding boxes are removed from the data set $\widehat{D}$. In Fig. 4, the pseudocode of the geometric augmentation procedure is reported.



Fig. 3. Example of images generated with the geometric data augmentation procedure. In (a) the original image, while (b), (c), and (d) presents three transformed examples by using scaling, shifting, and flipping.

2.2 Photometric Data Transformation

In order to obtain a trained model more robust with respect to photometric changes due to different illuminations in a real scenario, we implement and test two different approaches.

A first solution has been based on a brightness and contrast transformation. We used the augmentation method provided by the Albumentations library, which implements the following equation:

$$\hat{I} = (1 + \alpha) * I + \beta * I_{MAX} \tag{1}$$

where: I and $\hat{I}$ are the original and the modified image, respectively; α is the brightness parameter and β the contrast. I_{MAX} the is maximum admissible value that can be found in I according to its data type (i.e., $I_{MAX} = 255$ when I is encoded with unsigned integer at 8-bit). Intensity values falling out the admissible range (e.g., [0, 255]) after the transformation are clipped to the valid range.

The second solution has been an intensity normalization approach, where the images are normalized to have zero mean and unit standard deviation by using:

$$\hat{I} = \frac{I - \mu(I)}{\sigma(I)} \tag{2}$$

where: I and $\hat{I}$ are the original and the modified image respectively, while $\mu(I)$ is the mean value of I and $\sigma(I)$ the standard deviation. The idea is to create modified images with similar intensity distributions to reduce the impact of different illumination. Note that, in this case, the transformation is applied both on training and validation data.

3 Experimental Results and Assessment

In this section, the carried out experimental analysis is described. At first, the datasets used for training and testing are presented. Then, the used CNN architecture and the implemented training procedure is described. Finally, quantitative results are reported and discussed.

Algorithm 1: Geometric augmentation procedure

Data: D: dataset, *resizeFactor*: list of floats, *paddingPositions*: list of strings
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);
      $\hat{B} \leftarrow$ *resizeBBoxs*(B, R);
      $\hat{I} \leftarrow$ *gaussianBlur*($\hat{I}$,$K \leftarrow 3$);
      $\hat{I} \leftarrow$ *horizontalFlip*($\hat{I}$, $prob \leftarrow 0.5$);
      $\hat{B} \leftarrow$ *horizontalFlipBBox*($\hat{B}$, $prob \leftarrow 0.5$);
      $\hat{I} \leftarrow$ *zeroPadding* ($\hat{I}$, P);
      $\hat{B} \leftarrow$ *zeroPadding* ($\hat{B}$, 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
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

Fig. 4. Pseudocode of the geometric augmentation algorithm.

3.1 Datasets

The main datasets used for the training phase are freely the available datasets: LLVIP (Low Light Vehicle Identification and Verification Program) [26], and Teledyne FLIR-ADAS (Advanced Driver Assistance Systems) [27]. LLVIP is a dataset counting about 15000 images with multiple object category annotations, acquired with infrared cameras with different lighting conditions. The second dataset, Teledyne FLIR-ADAS, includes more than 10000 thermal images taken under different lighting and visibility conditions with respective annotations. LLVIP and FLIR-ADAS were used to develop and test video surveillance systems and advanced driver assistance systems such as pedestrian and cyclist detection. Both datasets have been filtered to extract from them only the images with pedestrians.

Table 1. Description of the original datasets used.

Dataset	Number of Train examples	Number of Validation examples	Image kind
LLVIP	12025	3463	Infrared
FLIR ADAS	10742	1144	Thermal
Florence	175	44	Thermal
Barcelona	169	43	Thermal

For the evaluation phase we used two thermal imaging datasets acquired under different conditions: indoor in Barcelona and outdoor in Florence by the same thermal camera AXIS Q1952 with wide-angle lens, see respectively Figs. 1a and 1c for sample images. Barcelona dataset has been obtained by randomly selecting 212 images from 392 videos taken during the annual Smart City Conference in Barcelona from Nov. 15, 2022, to Nov. 17, 2022. The videos were recorded in an indoor scenario, with no natural light source and no controlled temperature, from a thermal camera mounted on a tripod at a height of 3.5 m over the intersection of two large corridors. Pedestrians appear uniformly scattered over the images. Florence dataset consists of images taken by a thermal imaging camera placed around 30 m above the ground in Piazza della Signoria in Florence on six days from 14/07/2022 to 19/07/2022. As for Barcelona, 219 images were randomly sampled from the recorded videos. In this case, people appear very small and only in the lower part of the scene, the scene is very crowded and, being an outdoor environment, 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 (the day videos are also taken when the sun is heating and illuminating the square and large portion of the images). Manual labeling was performed using bounding boxes for both the Barcelona and Florence datasets.

As anticipated in the Introduction, the above-described datasets strongly differ in terms of geometrical and photometrical characteristics (see the examples in Fig. 1 and histograms of Fig. 2). These datasets represent a relevant test benchmark for the experiments since they depict two different environments for characteristics and offer the possibility to test the generalization that can be achieved using data augmentation. In the following Sections, a particular care has been devoted to the training process using different augmented versions of the LLVIP and FLIR-ADAS dataset to test the proposed augmentations. Table 1 shows the number of training and validation examples for each dataset. Note that the augmentation procedures are only applied on the LLVIP and FLIR-ADAS training sets.

3.2 YOLOv5

In this work, the YOLOv5 neural network [28] was used, which is made up of 3 main parts: CSPDarknet [29], PANet [30], and the final Yolo layer. CSPDarknet, designed to provide a balance between accuracy and speed, is used as a backbone for the feature extraction. The PANet includes a series of modules that connect the backbone to the Yolo layer that is responsible for producing the predictions and assess the object classes by using a combination of convolutions, batch normalization and activation functions. The Yolov5 loss is composed by three loss functions, as shown in Eq. 3:

$$Loss = l_{box} + l_{cls} + l_{obj} \tag{3}$$

where: l_{box} is the bounding box regression loss, l_{cls} is the classification loss, and l_{obj} is the confidence loss.

In our experiment, a YOLOv5 pretrained on COCO [31] dataset has been used. All the training carried out in this work are the finetuning processes of the network weights (models) by using different versions of the LLVIP and FLIR-ADAS datasets, with and

without data augmentations described in the previous section. For the training, the default hyperparameters of the YOLOv5 has been used, except for the batch size which has been set at 64, and the number of epochs set at 1000. Early stopping with a patience of 50 was used to stop the training when no improvements in the loss function were detected.

3.3 Evaluation Metric

In order to assess the results coming from the experiments, the average precision (AP) metric has been adopted. It is defined as the area under the precision/recall curve $p(r)$:

$$AP = \int_{r=0}^{1} p(r)dr \tag{4}$$

Precision is obtained as the ratio between the true positive (TP) detections over the sum of true positive and false positive (FP); Recall is the ratio of true positive results over the sum of true positive and false negative (FN). To evaluate the TP, FP, and FN the Intersection over Union (IoU) between the ground truth bounding boxes and the predicted ones has been used to measure the area that two bounding boxes have in common. Higher IoU scores indicate better detection results. After selecting a threshold for the IoU, a detection having an IoU equal or greater than the threshold is classified as a TP, otherwise as a FP. FN are missed detections. For all experiments, AP@50 has been considered, i.e., a threshold of 0.5 was used to determine the TP and FP considering the IoU scores. Higher AP@50 scores indicate better performance.

Table 2. Results in terms of AP@50 of the YOLO models trained with LLVIP and FLIR-ADAS, when they are used for people detection in different domains as Florence and Barcellona data sets.

Models created with train dataset	Validation on same dataset (AP@50)	Validation on Barcelona (AP@50)	Validation on Florence (AP@50)
Model LLVIP	0.967	0.732	0.305
Model FLIR-ADAS	0.831	0.719	0.576

3.4 Experimental Analysis

As a baseline reference, in Table 2, the results obtained using the YOLOv5 trained on the original LLVIP and FLIR-ADAS are reported. Results are shown for the models obtained by using the datasets used for training and for both Barcelona and Florence datasets (not used during training). As it can be seen, the performance on the model produced by using training sets LLVIP and FLIR-ADAS with the corresponding validations data set are very good. On the other hand, the results obtained by the trained models by using the new above-mentioned datasets coming from different domains produces a relevant drop in the AP, in particular for the Florence dataset.

Assessment of Geometric Augmentation Impact

A new set of models has been produced by using the generation of the augmented datasets ($LLVIP_{Geom}$ and $FLIR\text{-}ADAS_{Geom}$, respectively) with the procedure described in Sect. 2.1. They have been named as Model $LLVIP_{Geom}$ and Model $FLIR\text{-}ADAS_{Geom}$. As a first step, we assessed if the use of augmented examples would degrade the performance on the corresponding validation datasets since examples with a greater variance are used in training. To this end, in Table 4, the obtained results are reported. As can be seen, there are no evident loss in the performance: a decrease of 0.001, and 0.014 have been measured respectively using Model $LLVIP_{Geom}$ and Model $FLIR\text{-}ADAS_{Geom}$ with respect to the results obtained with the corresponding original dataset for validation, the same used to produce column "validation on same dataset" in Table 2. We can conclude that the geometric augmentation does not have provoked negative impact on the detection results.

Moreover, it is also interesting to verify if the results obtained using the new models trained by using the augmented datasets improve with respect to the different domain of data sets named Barcelona and Florence. The results are reported in Table 3, where for convenience we included the results of Barcelona and Florence on the net trained without augmentation, provided in former Table 2.

Table 3. Impact of training with geometrically augmented examples. Results provided in terms of AP@50.

Model created with train dataset	Validation on the original dataset (AP@50)
Model LLVIP	0.967
Model $LLVIP_{Geom}$	0.966
Model FLIR-ADAS	0.831
Model $FLIR\text{-}ADAS_{Geom}$	0.817

Table 4. Impact of geometric augmentation on Barcelona and Florence. Results are provided in terms of AP@50.

Model created with train dataset	Validation on Barcellona (AP@50)	Validation on Florence (AP@50)
Model LLVIP	0.732	0.305
Model $LLVIP_{Geom}$	0.738	0.450
Model FLIR-ADAS	0.719	0.576
Model $FLIR\text{-}ADAS_{Geom}$	0.773	0.546

In Table 3, for the Barcellona dataset, the geometric augmentation improves the results for both Model LLVIP and Model FLIR-ADAS with AP@50 respectively passing from 0.732 to 0.738 (with an increase of 0.006) and from 0.719 to 0.773 (with an increase of 0.054). For the Florence dataset instead, there is a strong improvement when using Model LLVIP with the AP@50 rising from 0.305 to 0.450 (0.145 higher), while a decrease of 0.030 was measured using Model FLIR-ADAS. This decrement was probably due to the fact Model FLIR-ADAS and its datasets includes bounding boxes (of pedestrian) of smaller size with respect to those which are present in the LLVIP dataset (see Fig. 5). Therefore, performing the downscaling in the geometric augmentation can lead to very small pedestrians appearing in the FLIR-ADAS images, so small that some of them disappear during the augmentation procedure and the remaining information/shapes are not sufficiently representative to help in detecting people in the Florence dataset.

Assessment of Photometric Augmentation Impact

For the assessment of photometric augmentation impacts, we tested different combinations of range of values for parameters of brightness (α) and contrast (β):

$$(\alpha, \beta) \in [(0.2, 0.0), (0.0, 0.2), (0.2, 0.2), (0.5, 0.2), (0.2, 0.5)]$$

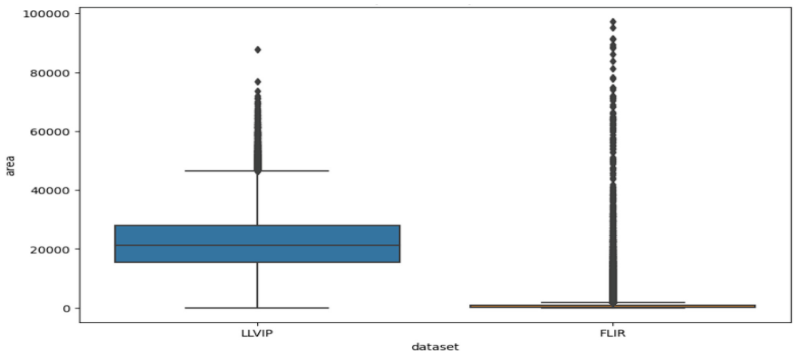


Fig. 5. Ranges of the bounding box dimension (area) comparison between LLVIP and FLIR-ADAS original datasets.

The applied $\left(\widehat{\alpha}, \widehat{\beta}\right)$, in the data sets, have been randomly extracted from intervals defined by the set parameters: $\widehat{\alpha} \in (-\alpha, +\alpha)$ and $\widehat{\beta} \in (-\beta, +\beta)$. These combinations generated specific $\text{LLVIP}_{\text{Photo}}$ and $\text{FLIR-ADAS}_{\text{Photo}}$ datasets from the corresponding originals and they have been used to generate a set of corresponding YOLO Models.

The assessment of the resulting Models $\text{LLVIP}_{\text{Photo}}$ (α, β) and Models $\text{FLIR-ADAS}_{\text{Photo}}$ (α, β) are reported in Table 5. As it can be seen, comparing results presented in Table 4 with respect to those of Table 5, the results provided by using the original validation data set have been improved for the LLVIP and get worst for the FLIR-ADAS.

Moreover, comparing the results presented in Table 3 with respect to those of Table 5, the improvement of precision with respect to the different domains of Barcellona has been registered only for most of the FLIR-ADAS$_{Photo}$ (α, β), namely those with (0.2, 0.0), (0.0, 2.0), (0.2, 0.2), (0.2, 0.5), where AP@50 of 0.722, 0.729, 0.737, and 0.751 were achieved (with an improvement of 0.003, 0.010, 0.018, and 0.032 respectively). All the other cases led to inferior performance with respect to those of Table 3.

Table 5. Impact of photometric augmentation of the original LLVIP and FLIR-ADAS data sets and their corresponding YOLO Models in person detection with respect on Barcelona and Florence data sets. In bold the cases in which the baseline has been improved. Results are provided in terms of AP@50.

Model created with			Validation on the same dataset (AP@50)	Validation on Barcellona (AP@50)	Validation on Florence (AP@50)
Train dataset	α	β			
Model LLVIP$_{Photo}$	0.2	0.0	0.973	0.634	0.175
Model LLVIP$_{Photo}$	0.0	0.2	0.970	0.617	0.173
Model LLVIP$_{Photo}$	0.2	0.2	0.973	0.652	0.183
Model LLVIP$_{Photo}$	0.5	0.2	0.967	0.556	0.138
Model LLVIP$_{Photo}$	0.2	0.5	0.975	0.620	0.179
Model FLIR-ADAS$_{Photo}$	0.2	0.0	0.808	**0.722**	0.504
Model FLIR-ADAS$_{Photo}$	0.0	0.2	0.810	**0.729**	0.515
Model FLIR-ADAS$_{Photo}$	0.2	0.2	0.808	**0.737**	0.502
Model FLIR-ADAS$_{Photo}$	0.5	0.2	0.808	0.699	0.529
Model FLIR-ADAS$_{Photo}$	0.2	0.5	0.805	**0.751**	0.573

Table 6. Impact of image normalization by mean and standard deviation on Barcelona and Florence. Results are provided in terms of AP@50.

Model created with train dataset	Validation on the same dataset (AP@50)	Validation on Barcellona (AP@50)	Validation on Florence (AP@50)
Model LLVIP μ, σ	0.942	0.533	0.067
Model FLIR-ADAS μ, σ	0.759	0.609	0.451

Finally, in Table 6 results using the image normalization by mean and standard deviation according to Eq. 2, are reported. In this case, each image of the training datasets (LLVIP and FLIR-ADAS) and of the validation datasets (Barcelona and Florence) have been normalized using the mean and the standard deviation of the intensity values. Note that each image is normalized independently using its mean and standard deviation. As

can be seen, results do not improve performance with respect to the baselines results presented in Table 2: after training by using the normalized LLVIP dataset the obtained AP@50 resulted to be 0.533 and 0.067 by validating the model with the normalized version of the data sets Barcellona and Florence, respectively. Using the normalized FLIR-ADAS to train the YOLO network, we achieved an AP@50 score of 0.609 for Barcelona, 0.451 for Florence.

From these results we can argue that the generative approaches for altering the photometric characteristics of the images of LLVIP and FLIR-ADAS dataset do not provide consistent generalization in the trained models that do not improve with respect to the baselines. This can be due to particular characteristics of thermal images where the intensities are not related to lighting effects but instead are descriptions of the temperatures of the elements in the scene. Therefore, applying transformation valid for color RGB images on thermal data could spoil sensitive information and cause inferior detection performance.

4 Conclusions

In this paper, a study on augmentation techniques for generating extended data sets for pedestrian detection using thermal images was presented. We investigated the effects of geometric and photometric transformations of freely available datasets (LLVIP and FLIR-ADAS) used to train the detector, evaluating the results with respect to two different datasets (Barcelona and Florence) with different characteristics and not seen during training. Results have shown that the geometric augmentations can provide some benefits by creating new examples with different pedestrian sizes appearing in different image regions, improving the performance on data from different domains, and providing more flexibility. On the other hand, photometric alterations that can work on color RGB images were found inadequate to deal with changes in intensities in thermal data, probably due to the different photometric characteristic between color images, where intensity are related to lighting effects, and thermal images, where the intensities depend on the temperature of the scene elements.

Acknowledgments. Financed by the European Union—NextGenerationEU (National Sustainable Mobility Center CN00000023, Italian Ministry of University and Research Decree n. 1033— 17/06/2022, Spoke 9).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Elharrouss, O., Almaadeed, N., Al-Maadeed, S.: A review of video surveillance systems. J. Vis. Commun. Image Represent. **77**, 103116 (2021)
2. Ingle, P.Y., Kim, Y.-G.: Real-time abnormal object detection for video surveillance in smart cities. Sensors **22**(10), 3862 (2022)

3. Socha, R., Kogut, B.: Urban video surveillance as a tool to improve security in public spaces. Sustainability **12**(15), 6210 (2020). https://doi.org/10.3390/su12156210
4. Ansari, M.A., Singh, D.K.: An expert video surveillance system to identify and mitigate shoplifting in megastores. Multimed. Tools Appl. **81**, 22497–22525 (2022)
5. Bellini, P., Bilotta, S., Collini, E., Fanfani, M., Nesi, P.: Data sources and models for integrated mobility and transport solutions. Sensors **24**(2), 441 (2024). https://doi.org/10.3390/s24020441
6. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey. Proc. IEEE **111**(3), 257–276 (2023). https://doi.org/10.1109/JPROC.2023.3238524
7. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 580–587 (2014). https://doi.org/10.1109/CVPR.2014.81
9. Girshick, R.: Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
11. Khalfaoui, A., Badri, A., Mourabit, I.E.: Comparative study of YOLOv3 and YOLOv5's performances for real-time person detection. In: 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1–5 (2022). https://doi.org/10.1109/IRASET52964.2022.9737924
12. Hussain, M.:. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines **11**(7), 677 (2023). https://doi.org/10.3390/machines11070677
13. Indhuja, U.S., Amutha, R.: Pedestrian detection in extreme weather. In: 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE (2021)
14. Jin, K., Xie, X., Wang, F., Han, X., Shi, G.: Human identification recognition in surveillance videos. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, pp. 162–167 (2019). https://doi.org/10.1109/ICMEW.2019.00-93
15. Wang, M., Deng, W.: Deep face recognition: a survey. Neurocomputing **429**, 215–244 (2021). ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2020.10.081
16. GDPR: General Data Protection Regulation. https://gdpr.eu/. Accessed 19 Jan 2024
17. Collini, E., Ipsaro Palesi, L.A., Nesi, P., et al.: Flexible thermal camera solution for Smart city people detection and counting. Multimed. Tools Appl. (2023)
18. Kim, J.: Pedestrian detection and distance estimation using thermal camera in night time. In: 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE (2019)
19. Kaur, P., Khehra, B.S., Mavi, E.B.S.: Data augmentation for object detection: a review. In: 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), Lansing, MI, USA, pp. 537–543 (2021). https://doi.org/10.1109/MWSCAS47672.2021.9531849
20. Montserrat, D.M., Lin, Q., Allebach, J., Delp, E.J.: Training object detection and recognition CNN models using data augmentation. Electron. Imaging **2017**(10), 27–36 (2017)
21. Kisantal, M., Wojna, Z, Murawski, J., Naruniec, J., Cho, K.: Augmentation for small object detection. In: CS & IT Conference Proceedings, vol. 9, no. 17. CS & IT Conference Proceedings (2019)

22. Fanfani, M., Marulli, M., Nesi, P.: Addressing domain shift in pedestrian detection from thermal cameras without fine-tuning or transfer learning. In: 2023 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 314–319. IEEE (2023)
23. Alberti, F., et al.: Mobile mapping to support an integrated transport-territory modelling approach. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **48**, 1–7 (2023)
24. Han, Q., Nesi, P., Pantaleo, G., Paoli, I.: Smart city dashboards: design, development, and evaluation. In: 2020 IEEE International Conference on Human-Machine Systems (ICHMS), pp. 1–4. IEEE, September 2020
25. Badii, C., et al.: Industry 4.0 synoptics controlled by IoT applications in node-red. In: 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), pp. 54–61 (2020)
26. Jia, X., et al.: LLVIP: a visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
27. Teledyne FLIR dataset. https://www.flir.it/oem/adas/adas-dataset-form/. Accessed 19 Jan 2024
28. YOLOv5 by Ultralytics. https://github.com/ultralytics/yolov5. Accessed 19 Jan 2024
29. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
30. Liu, S., et al.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
31. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, Part V, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48