IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 18, 2020, accepted June 30, 2020, date of publication July 6, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007481

Thermal Object Detection in Difficult Weather Conditions Using YOLO

MATE KRIŠTO, MARINA IVASIC-KOS, (Member, IEEE), AND MIRAN POBAR
Department of Informatics, University of Rijeka, 51000 Rijeka, Croatia

Corresponding author: Marina Ivasic-Kos (marinai@uniri.hr)

This work was supported by the Croatian Science Foundation through the Project Automatic Recognition of Actions and Activities in Multimedia Content from the Sports Domain (RAASS) under Grant IP-2016-06-8345.

ABSTRACT Global terrorist threats and illegal migration have intensified concerns for the security of citizens, and every effort is made to exploit all available technological advances to prevent adverse events and protect people and their property. Due to the ability to use at night and in weather conditions where RGB cameras do not perform well, thermal cameras have become an important component of sophisticated video surveillance systems. In this paper, we investigate the task of automatic person detection in thermal images using convolutional neural network models originally intended for detection in RGB images. We compare the performance of the standard state-of-the-art object detectors such as Faster R-CNN, SSD, Cascade R-CNN, and YOLOv3, that were retrained on a dataset of thermal images extracted from videos that simulate illegal movements around the border and in protected areas. Videos are recorded at night in clear weather, rain, and in the fog, at different ranges, and with different movement types. YOLOv3 was significantly faster than other detectors while achieving performance comparable with the best, so it was used in further experiments. We experimented with different training dataset settings in order to determine the minimum number of images needed to achieve good detection results on test datasets. We achieved excellent detection results with respect to average accuracy for all test scenarios although a modest set of thermal images was used for training. We test our trained model on different well known and widely used thermal imaging datasets as well. In addition, we present the results of the recognition of humans and animals in thermal images, which is particularly important in the case of sneaking around objects and illegal border crossings. Also, we present our original thermal dataset used for experimentation that contains surveillance videos recorded at different weather and shooting conditions.

INDEX TERMS Convolutional neural networks, object detector, person detection, surveillance, thermal imaging, YOLO.

I. INTRODUCTION

Because of global terrorist threats and illegal migration, concerns about the safety of citizens have been intensified. To prevent unwanted events and to protect people and their property, investment in security systems has reached record levels trying to utilize all available technological achievements to develop sophisticated systems.

Thermal cameras are now ubiquitous in video surveillance systems that take care of the safety of people and objects in urban areas, on state borders, and other monitored and guarded areas. Thermal cameras are important for surveillance and security because they can be used in such weather conditions when ordinary RGB cameras cannot be used or when they give poor results, such as in the night and darkness (Fig. 1.), or in the rain and fog.

In order to facilitate surveillance and increase security, it is important to detect unauthorized persons, suspicious movements in protected areas, and prevent illegal border crossings in a timely manner. The ability to automatically detect a person or object and alert a suspicious situation is very important for the security system.

So far, many successful machine learning algorithms have been developed for detecting and tracking objects such as the human face [1], [2] or the human figure [3] in RGB optical images. The purpose of object detection is to classify objects and mark their exact position in image or video.

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

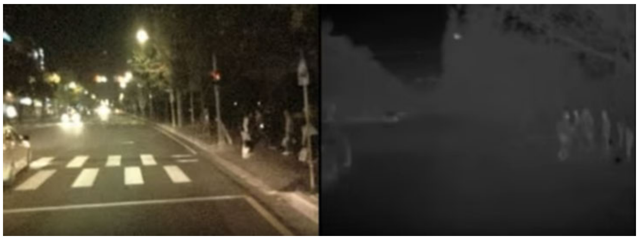


FIGURE 1. **Night vision (left) vs. thermal imaging (right) [20].**

Nowadays, the best object detection results are achieved in RGB images by models based on convolutional neural networks (CNN). The popularity of convolutional neural networks and deep learning began with the great success of AlexNet for the image recognition task in ImageNet Challenge in 2012 [4]. Since then, many successful CNN architectures for object detection have been developed, such as Fast R-CNN [5], SSD [6], Mask R-CNN [7], R-FCN [8] and YOLO [9]–[11] and adapted for different tasks [12]–[19].

Due to the differences in visual and thermal image features, the aim is to explore how deep learning methods that are successful for object detection in optical images will perform with thermal imaging.

To evaluate the detection performance, we introduce an original dataset of thermal videos and images that simulate illegal movements around the border and in protected areas and are designed for training machines and deep learning models. The videos are recorded in areas around the forest, in different weather conditions at night – in the clear weather, in the rain, and in the fog, and with people in different body positions (upright, hunched) and movement speeds (regular walking, running) at different ranges from the camera. In addition to using standard camera lenses, telephoto lenses were also used to test their impact on the quality of thermal images and person detection given different weather conditions and distance from the camera. The obtained dataset comprises 7412 manually labeled images extracted from video frames captured in the long-wave infrared (LWIR) segment of the electromagnetic (EM) spectrum.

For the detection task, the YOLOv3 network [11] was used, which achieves object detection results in RGB images at the state-of-the-art level. Models based on the YOLOv3 network were trained on subsets of our dataset and the results of human detection in thermal videos using the out-of-the-box YOLO neural network and the trained YOLO models were compared. The models were tested both on our own and seven different widely used and well-known thermal imaging datasets. The experimental results have shown the significantly improved performance of human detection in thermal imaging in terms of average precision for the trained YOLO model over the original model given the different weather and shooting conditions.

The main contributions of this paper are a) domain adaptation and use of convolutional neural network-based models that were originally intended for detection in RGB images in new settings of automatic object (person) detection in thermal images, b) analysis of the effect of using only training data acquired during clear weather, on model performance in difficult weather conditions where training data is harder to obtain (rain, fog) c) analysis of the effect of quantity of training data needed to achieve the detection successfully, d) original dataset of thermal images taken in different weather conditions that simulates realistic conditions of illegal movements around the border.

This paper is organized as follows: an overview of related work is presented in the next section. The basic information about thermal imagery is provided in Section 3, and the detection pipeline of the YOLO object detector is given in Section 4. Dataset and object detection experiments are described in Section 5. The results are presented and discussed in Section 6. The paper ends with the conclusion and direction for future research.

II. RELATED WORK

The use of convolutional neural networks (CNN, CONVnet) instead of standard classification algorithms is a trend in the research area of human detection, regardless of the task is nighttime detection using thermal imaging cameras or daytime detection using standard optical cameras. Thermal images are mainly used to detect the presence of people at night or in bad lighting conditions but can perform poorly in the daytime when there is insufficient thermal contrast between the people and their surroundings. Therefore, the authors in [21] proposed to augment thermal images with their saliency maps, to serve as an attention mechanism for the pedestrian detector. They trained the Faster R-CNN for pedestrian detection and report the added effect of saliency maps generated using static and deep methods (PiCA-Net and R3-Net). Their best performing model results in an absolute reduction of miss rate by 13.4% and 19.4% over the baseline in a day and night images respectively. The authors [22] also proposed the usage of deep learning and saliency maps for pedestrian detection at night. They integrate a hardwired adaptive Boolean-map-based saliency (ABMS) kernel with the YOLO detector, to generate a saliency feature map that boosts the pedestrian from the background based on the particular season. In [23], [24] YOLO detector was trained on a thermal image dataset for person detection. This paper greatly extends the scope of that work by analyzing different weather conditions separately, by testing on other datasets and including possibly confusing objects such as animals in the test.

In [25], the authors proposed the use of LWIR thermal images for counting people in public spaces such as classrooms. They developed a people counting algorithm on a custom dataset of 3000 thermal images recorded in student's workrooms, based on small CNNs that can be run on a limited-memory low-power platform such as Cortex M4, with reported error-free detection on 53.7% of the test images. Apart from the use of CNNs for human detection on thermal images and video, some authors proposed the use of CNNs for object tracking in thermal images as in [26]. In [27] the authors demonstrated enhanced target recognition and improved false alarm rates for a mid to long-range detection system, utilizing an LWIR sensor. They report an overall accuracy of over 95% for six object classes related to land defense using the CNN-based detector. A method for real-time human detection in thermal images based on background modeling and CNN is presented in [28]. For real-time implementation, the background modeling is done by modified running Gaussian average and the CNN-based human classification is performed only for the detected foreground objects.

Wang and Hosseinyalamdary in [29] applied deep convolutional neural networks for human detection on stacked frames from thermal video thus including some temporal information. The convolutional neural network that used stacked video frames had 21.4% higher accuracy than the neural network trained using single images on their test.

The infrared video-based automatic target detection/recognition (ATD/R) system presented in Zhang *et al.* [30] uses a Faster-RCNN detector trained on IR images combined with a super-resolution method to deal with the issue of a small number of pixels that targets at the long-range have. The system was tested using two datasets under different weather conditions, featuring pedestrians and six different types of ground vehicles as target types, and the tests show improved performance for long-range targets with the use of a super-resolution method.

The development of CNNs helped for moving surveillance systems to embedded devices like the Raspberry Pi. Khalifa *et al.* [31] presented a survey of different systems and techniques that have been deployed on embedded devices. They covered the characteristics of datasets, feature extraction techniques, and machine learning models. Also, they utilized a unified dataset to compare different systems concerning accuracy and performance time and suggested new enhancements, and future research directions.

Human detection and recognition on thermal images and videos and its applications are still growing and challenging research area, not only in the area of computer vision and deep learning but also in other areas like IR and technology. Researchers show interest in thermal imaging for human detection, as well in methods that combine thermal imaging with images recorded in other wavelengths like in [32]. The authors evaluate pixel-level image fusion of infrared and RGB images to improve the CNN-based pedestrian detectors so that they can work in a day and night conditions, which is crucial in advanced driver-assistance systems (ADAS), autonomous vehicles and video surveillance. Besides thermal imaging, some researchers used near IR in combination with CNNs [33] for pedestrian detection. They presented a method based on a 9-layer CNN model with self-learning soft-max for nighttime pedestrian detection and reported a 94.49% accuracy on a set of 15,000 testing samples. Imran *et al.* [34] presented a novel saliency-aware descriptor called Stacked Saliency Difference Image (SSDI) to model local and global spatio-temporal motion information for Human Action Recognition (HAR) in IR images. They use a four-stream deep framework built upon CNN and recurrent neural network (RNN) models and report results of 83.5% on InfAR dataset [35], and baseline result of 75.17% on their proposed IITR-IAR dataset [34]. The application of CNNs for human detection, recognition, and action classification are also presented in [36]–[40].

III. THERMAL IMAGERY

Infrared (IR) thermal cameras record the heat generated or reflected by objects being monitored and convert the detected energy into temperature values to form an image.

Cameras operating in the MWIR and LWIR bands (Fig. 2) do not require an additional source of light or heat, because the thermal radiation sensors in these ranges capture the emitted thermal energy of observed objects [41], [42].

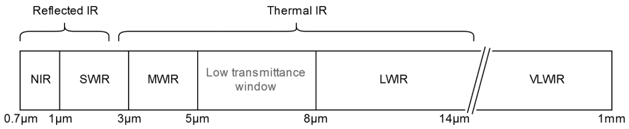


FIGURE 2. Electromagnetic spectrum with illustrated IR segments.

For this reason, unlike visible light cameras, they are invariant to illumination conditions, robust to a wide range of light variations [43], [44] and weather conditions, and can operate in total darkness. In this work, the focus is on using the LWIR subspectrum, which can also be a method of improving the visibility of objects in dark environments.

However, thermal imaging sensors provide much less detail than visible-light cameras, because instead of the information that color provides in the visible spectrum, they only provide the detected temperature ranges in the thermograms, Fig. 3, usually with much lower resolution.

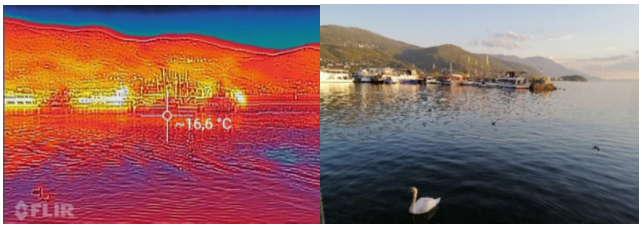


FIGURE 3. IR sensors provide much fewer details (left) than the optical sensor of visible light (right).

Also, changes in the ambient temperature affect the quality of thermal images since images are formed by the radiated intensity difference between objects and their surroundings and thus a higher ambient temperature can decrease the contrast between the detected object and the background, Fig. 4.

The temperature scale in images changes depending on the ambient temperature and the temperature of objects. Thermal images are usually presented in pseudo-color where the lowest color (dark blue) corresponds to the coolest part of the

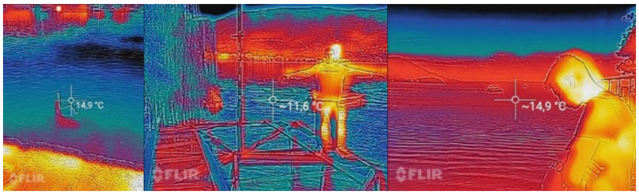


FIGURE 4. IR sensors are very sensitive to changes in ambient temperature.

image and the brightest (white) corresponds to the hottest part of the image. E.g., in Fig. 4. (left), light blue corresponds to 14.9 °C, and in Fig. 4. (right) the same temperature corresponds to red color. On the other hand, in Fig. 4. (middle), the blue color corresponds to a temperature of 11.6 °C.

The heat that objects themselves emit is also not constant but depends on the internal state of the object. For example, during running or intense exercise, the production of metabolic heat in the human body may increase 10 to 20 times compared to the heat production at rest which is reflected in the increase in body temperature [45].

A. THERMAL CAMERA CHARACTERISTIC FOR SURVEILLANCE APPLICATIONS

Recording distance and the type and quality of the thermal camera and its thermal imaging sensor are significant for human detection because they directly affect the image or video resolution and the size of the human figure in the recording. The impact of sensor resolution, recording distance, and image quality on object detection, recognition, and identification were studied in [46], where experiments based on TTP (Targeting Task Performance) were conducted to determine the effective distance for each of the tasks.

The probability of success for each of the tasks was found to be directly related to the distance at which images are captured and the quality or resolution of the camera. Considering the same rate of success, object detection is possible at longer distances than object identification, which is only viable at the smallest distances. Also, with higher camera resolution, larger distances can be achieved. For example, with a smaller resolution camera, object detection was found successful with 80% probability up to at about 1.4 km, recognition at up to 0.5 km and identification up to about 200 m, while with higher resolution camera, with the same rate of success, identification was possible up to 0.7 km, recognition up to 1.2 km and detection up to about 2.2 km [46].

Weather conditions are another major factor that affects the recording quality and thus the detection performance. With the deterioration of weather conditions, the distance at which it is possible to make a successful object or person detection is reduced [23]. For example, somebody parts that may be important for object recognition, such as human leg, are tiny in the case of long-distance shooting and are represented with only a few pixels in the image. In bad weather, the image quality may deteriorate so that these parts are heavily degraded or not visible at all. That loss of information can heavily affect the performance of the classifier.

In protected environments such as in border controls or airports for 24/7 all-weather surveillance, both mid-range and long-range thermal camera sensors can be used. Sensor operating in different bands have their strengths and limitations, so there is no perfect choice between LWIR and MWIR that covers all surveillance scenarios. Different environment, climatic, temperature, and weather conditions affect sensor performance in different ways, with typical operating conditions for surveillance applications summarized in Table 1. Both bands see negligible solar effects but are adversely affected by fog and rain, although the LWIR band has better performance than MWIR in foggy conditions. For most target ranges MWIR sensors are less affected by humidity than LWIR sensors and have higher atmospheric transmission than LWIR in most climates. MWIR sensors are favorable in warmer climates while LWIR is preferred for colder climates [43], [47].

TABLE 1. MWIR and LWIR band characteristic for surveillance applications.

Condition	MWIR band	LWIR band
Climate	warmer	colder
High humidity	> 2.5km	< 2.5km
Atmospheric constituents	rain	rain/smoke/aerosols/fog
Target temperature	high-temperature targets (airplanes or missiles)	more flux available at most scenes on ambient temperature (flux - thermal energy emitted by targets and environment background)
Atmospheric transmission	very long-range detection (>=10 km)	
Applications	coastal and vessel traffic surveillance, harbor protection	Firefighting, military

A convenient solution for surveillance systems is to use thermal cameras that are simultaneously active in both MWIR and LWIR bands, such as the widely used FLIR systems (Forward Looking Infrared) [48] cameras. A group of humans recorded at the distance of about 110 meters in nighttime and clear weather conditions with the FLIR Thermacam P10 is shown in Fig. 5 [49].

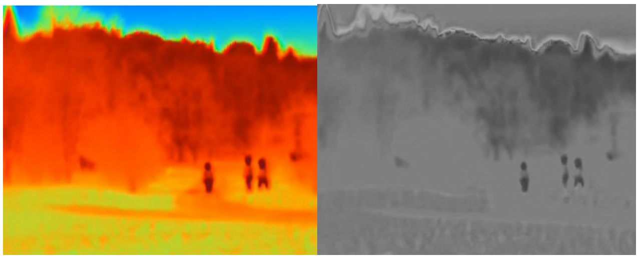


FIGURE 5. Comparison of the same scene in colorized, RGB (left) and Greyscale (right) showing group of humans at a distance of 110 m (night-time, clear weather), recorded using FLIR Thermacam P10.

For recordings at a larger distance or in fog conditions and heavy rain, telephoto lenses may be used instead of the

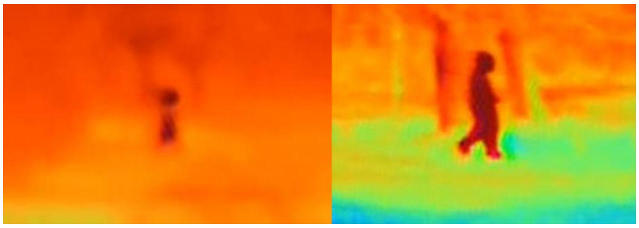


FIGURE 6. **Human silhouette recorded with standard lenses at 110 m (left) and with telephoto lenses at 165 m (right).**

standard lens. As seen in an example in Fig. 6, the visual properties of the object can vary significantly depending on the atmospheric conditions and distance. Some body parts that may be important for object recognition, such as legs, can be very tiny or blend into the background, especially when recording from a long distance.

IV. EXPERIMENT WORKFLOW

A. PROBLEM FORMULATION

The experiment aims to detect people in thermal images and videos taken in different weather (clear weather, heavy rain, and fog) and recording conditions during surveillance of a protected area, using a real-time object detector.

We examined some of the object detectors with state-of-the-art results on RGB images and choose a detector for further experiments that achieves the best results on thermal images in terms of average precision and inference speed. Namely, we considered the Faster R-CNN [50], SSD [6], Cascade R-CNN [51], YOLOv3 [11] and FCOS [52] detectors that achieve the best results today for person detection, and chose YOLOv3 as the one that best suits our goal and can be customized to thermal images.

Although thermal images significantly differ from RGB images both in color and detail, it is expected that individual layers of RGB images still sufficiently resemble the thermal image so that the shape features that the model learned to extract on RGB training data should still be useful for thermal images.

To build the model, we needed an appropriate thermal imaging database taken during the surveillance of supervised areas such as state borders. As there was no adequate public database, we created a thermal imagery database of surveillance scenes considering different shooting conditions and prepared data for supervised machine learning.

B. DATASET CREATION

Depending on the goals and tasks that should be solved, researchers can either use existing datasets or if a suitable dataset doesn't exist, create a specialized dataset that better fits the set goals [14], [53]. Since the existing thermal imaging datasets, e.g. OTCBVS Benchmark Database [54] and CASIA Infrared Night Gait Dataset [55] didn't fully match the intended goal of detecting persons in various security situations such as unauthorized movement in monitored areas, sneaking around protected objects and border crossing in the night at both favorable and unfavorable weather conditions, a custom dataset was recorded and prepared to simulate these real-life scenarios.

For the task of human detection in thermal images and videos, we use our IR thermal image dataset, named UNIRI-TID, that simulates realistic conditions for application in detection and recognition systems in difficult weather. All scenarios in the dataset are recorded in the night and during the winter period. Weather conditions in the recordings vary from clear weather, fog, and heavy rain [56].

The people in the recordings simulate intentional but unauthorized entry or walk through the area under surveillance, and thus move with different body positions and movement speeds at different distances from the camera such as crawling, hunched, or normal walking and running. In some cases, the persons were accompanied by dogs, so that the ability of the person detector to discriminate between a person and other objects with similar thermal characteristics can be tested. Since different weather conditions determine what is discernible in recorded images and videos, different scenarios were defined for each weather type, as follows.

The clear weather scenario serves for training the person detection model, as well as to estimate the maximum distance to the camera at which one can detect the person in the video with the naked eye. The reference distance for the normal lens was 110 m, while recordings at larger distances were done with a telephoto lens.

The dense fog scenario is used to prepare a test set for testing the robustness of the human detection model trained only on clear weather images (video) to changing recording conditions and to examine the use of a human detection model in as realistic as possible real-world surveillance conditions. As with the clear weather scenario, the reference ranges of the camera were determined in fog with low visibility, here the reference distances are 30 m.

The heavy rain scenario is also used for testing the robustness of the model trained on clear weather images, as well as increasing the coverage of different realistic conditions that can arise in monitored areas such as the state border. The reference distance at which the camera can record in rainy weather and weather with high humidity was examined from 30 to 215 m.

In all three scenarios the people were recorded while walking, running, and sneaking (walking while hunched).

To test the detector's ability to distinguish between human and non-human objects that may have similar thermal characteristics as humans, we created an additional dataset based on recordings that contained both humans and dogs.

1) DATA COLLECTION

The recording was done in several sessions during the wintertime using the FLIR ThermalCam P10 LWIR thermal imaging camera mounted on a tripod at a height of about 140 cm, with the standard 24° x 18° field of view (FOV) lens and with FLIR 7° FOV Telephoto Lens (P/B series) [57]. The camera sensor captures a thermal resolution of 320 × 240 pixels, which was upscaled to 1280 × 960 pixels using an external video recorder.

For the distance measurement, we used the ViewRanger application [58] installed on the CAT S60 [59] GPS-equipped smartphone.

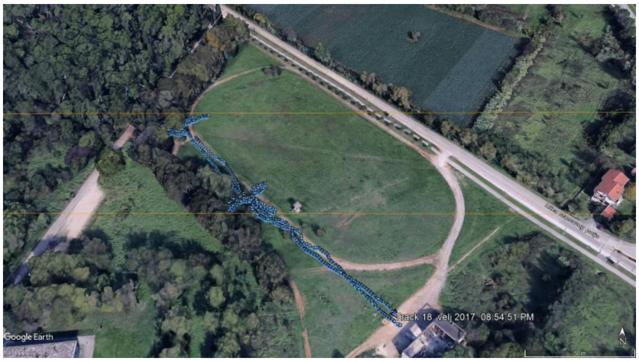
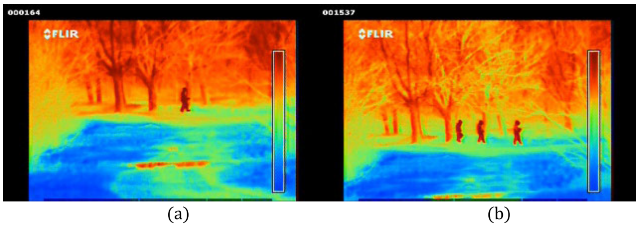


FIGURE 7. Aerial view of the recording field. Map data: Google, Europa Technologies.

a: CLEAR WEATHER

The clear weather recording was done out in a field bordering a small forest (Fig. 7), to include the situation when people are hiding behind trees and bushes. The air temperature was about 2 °C, in night conditions with good visibility and without affecting atmospheric conditions. A single person and a group of three persons were recorded as they walked away from the camera from 50 to 110 m, then turned and moved across the camera's field of view at a distance of 110 and 165 m.

At 110 m, people were recorded while running in addition to normal walking, while for the one-person case, walking on all fours, crawling, and lying on the ground were recorded as well (Fig.9). Moving across the camera FOV at 165 m, people were recorded walking and running upright, and walking and running while hunched. At 165 m, the recordings were done only with the telephoto lens, since the images recorded using the standard lens were of too low quality to detect and recognize a person with the naked eye. At 100 m, both, standard and telephoto lenses were used (Fig. 8).

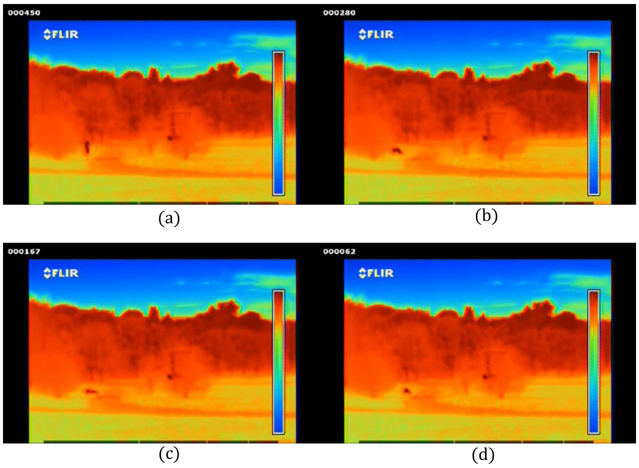


(a) (b)

FIGURE 8. Comparison of images taken in the clear weather and at the same distance with (a) standard thermal camera lens, (b) using telephoto lens [56].

b: FOGGY WEATHER

Recording in the foggy weather was the most demanding since the fog disperses LWIR radiation, due to the high density of waterborne particles in the air and significantly reduces the visibility for the thermal imaging camera compared to other atmospheric conditions [60]. For this reason, in very dense fog, people at distances of over 50 meters were not visible at all in the recordings and thus it was not possible to replicate the clear weather scenarios. After reviewing the preliminary recordings, distances from 0 to 30 m and at 50 m, were selected for the foggy weather scenario, using only the telephoto lens. Again, a single person and a group of three persons walked away from the camera from 0 to 30 m, then crossed the camera's field of view walking normally upright, walking hunched and running. The recording was done on the asphalt road in the forest. The air temperature was about 2 degrees Celsius, while the visibility was less than five meters due to dense fog.



(a) (b)

(c) (d)

FIGURE 9. Comparison of images taken at the clear weather using standard lenses – one person – changing positions: (a) normal walking, (b) four-leg walking; (c) lying on the ground - left side of the person; (d) lying on the ground - head in the direction of the camera [56].

c: RAINY WEATHER

The recording in the heavy rain was done at a site that provided the ability to record at distances of over 165 meters. An individual and two persons were recorded while walking normally, running, both upright and hunched at 30, 70, 110, 140, 170, 180 and 215 m from the camera. At 215 m, a recording was possible only with a telephoto lens, while for all other distances both standard and telephoto lenses were used.

d: HUMAN - NON-HUMAN DATASET

A part of the dataset was created to test the ability of the detector to distinguish human and non-human objects that may have similar thermal characteristics, such as animals, and in this case, a dog. The recordings where a dog accompanied the persons were made in dense fog only.

The human-non-human data subset consists of 5,420 images, 3,552 of them recorded in the clear weather, and 1,868 in the fog. The clear weather recordings correspond to the previously mentioned clear weather scenario (recording at distances of 110 m, walking from 110 to 165 and 165 m using standard and telephoto lenses with subjects changing movement speed (normal, running) and body positions.

In the dense fog, the recording was done at distances from 0 to 30 and 50 m, using standard and telephoto lenses, as previously for the fog scenario. The volunteers changed body positions and movement speed, and hid in the woods, while the dog behaved as normally – changing movement speed and body positions (walking, running, jumping. . .).

By class, the Human–Non-Human dataset contains 2,685 images annotated with only the Person class, 1,497 images are annotated with both human (person) and Non-human (dog) classes, and 1,238 images without either persons or animals and without annotations (negative examples).

2) DATA PREPROCESSING AND ANNOTATION

From all recordings, about 20 minutes of material from the clear weather scenario, 13 minutes from the fog scenario, and about 15 minutes from rainy weather were selected for further processing. The longer videos were cut into sequences according to the steps in previously defined scenarios and from these sequences individual frames were extracted, resulting in 11,900 images for the clear weather, 4,905 images for the fog, and 7,030 images for the rainy weather scenarios.

Of all the frames, 6,111 were selected for manual annotation so that they could be used to train the supervised model. When selecting the frames, it was taken into account that the selected frames include different weather conditions so that in the set there were 2,663 frames shot in clear weather conditions, 1,135 frames of fog and 2,313 frames of rain.

The annotations were made using the open-source Yolo BBox Annotation Tool [61]. It is a tool that runs within any web browser and it can simultaneously store annotations in the three most popular machine learning annotation formats YOLO [9], [62], VOC [63] and MSCOCO [64], saving time in later phases because no subsequent conversion is required in the needed format. The image annotation consists of a centroid position of the bounding box around each object of interest, size of the bounding box in terms of width and height, and corresponding class label (Human or Dog). After the dataset was annotated, the machine learning model was trained.

3) DATA ADAPTATION FOR TESTING ON BENCHMARK DATASETS

To test the trained model on benchmark datasets such as CVC FIR: Sequence Pedestrian Dataset [65], [66] VOT-TIR2015 [67], OTCBVS Benchmark Dataset Collection [54], Terravic Motion IR Database [68], that all contains exclusively grayscale thermal images (Fig. 10), we had to do a domain adaptation. We applied a simple method and transformed our basic pseudo-colored RGB images into grayscale. A similar domain adaptation was applied in [69], [70].

The grayscale conversion was done to better reflect the various representations of thermal images in different datasets and to better simulate the realistic conditions in which such a model would be implemented. Since people may be warmer or cooler than the background our basic pseudo-colored RGB images were converted into greyscale in two ways. In the first, the hotter areas are represented with bright shades (Fig, 11. middle), and in the second one, the hotter areas are represented with darker shades (Fig, 11. left). In this way, the 6,111 images in the dataset were supplemented with 12,222 grayscale images.

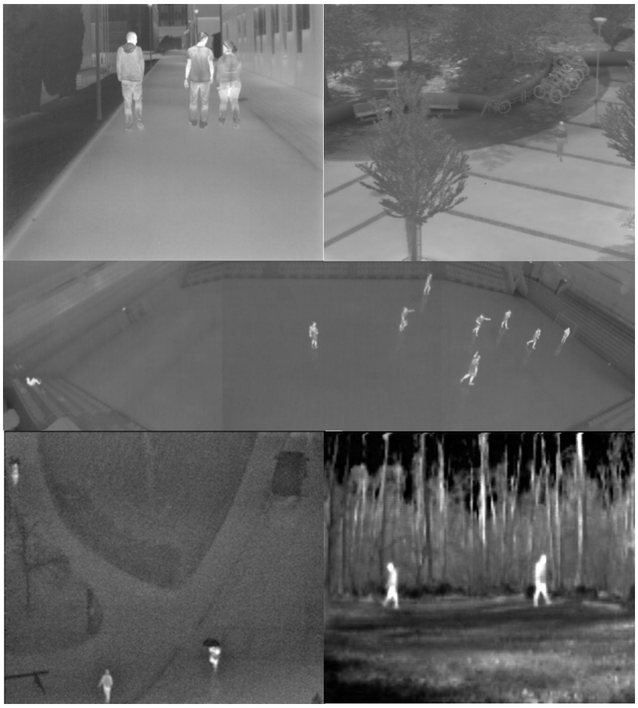


FIGURE 10. CVC IR 09 image example (up left); VOT-TIR2015 dataset image examples (upright, middle); OSU thermal dataset form OTCBVS Benchmark Dataset Collection (down left); Terravic motion dataset (down right).

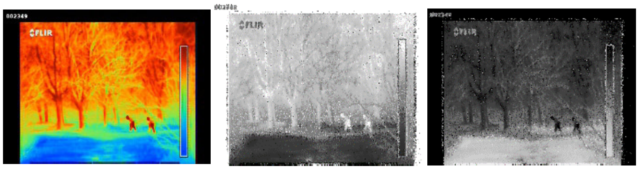


FIGURE 11. Original thermal image (left) is converted into greyscale in two ways: the hotter areas are represented with bright shades (middle), and the hotter areas are represented with darker shades (right).

Summary data about the collected and annotated subsets of UNIRI-TID dataset is shown in Table 2.

C. OBJECT DETECTOR

In order to select the neural network architecture to be used in further research, we conducted a preliminary experiment to test the performance of state of the art RGB CNN-based object detectors for person detection on our custom dataset. We have retrained the Faster R-CNN detector[50], SSD detector [6] with the Inception v2 [71] backbone, FCOS [52] with ResNet 50 backbone [72], Cascade R-CNN [51] with

TABLE 2. Subsets of UNIRI-TID dataset.

Dataset subset	Camera distance	Number of images	Object class
Clear	50-160 m	2663	Human
Rain	30-215m	2313	Human
Fog	0-30 m, 50m	1135	Human
All	0-215 m	6111	Human
Human-non-human	0-215 m	5419	Human and dog
Transform	30-215m	18333	Human

TABLE 3. Comparative results for person detection on thermal images.

Model	AP	Inference time	FPS
Faster RCNN	98,86%	0,141	7,097
SSD	94,02%	0,063	15,794
FCOS	97,05%	0,048	20,790
Cascade RCNN	98,80%	0,223	4,480
YOLOv3	97,93%	0,036	27,472

ResNet 101 [72] backbone, and YOLOv3 [11] without changing the original architecture on 4,270 thermal images from our dataset for 40000 iterations. Testing was performed on 1,841 images from the test set and comparative results are given in Table 3.

All examined models perform well with comparable results in terms of AP (Average Precision), and the top three models are Faster RCNN, Cascade RCNN, and YOLOv3. However, due to its architecture, YOLOv3 is significantly faster, has a processing speed of 27,5 consecutive frames per second (FPS), and has a shorter inference time of the top 3 models. Assuming the video has a frame rate of at least 24 (FPS) to have a smooth appearance, only YOLOv3 can process the video sequence "online" while other models don't have that capability because they are too slow. Therefore, we have used YOLOv3 further in our experimental work.

1) YOLO OBJECT DETECTOR

YOLO is an object detector that uses a single pass to detect the potential regions in the image where certain objects are present and to classify those regions into object classes. The authors [9] framed the object detection task as a regression problem from image pixels to coordinates of objects' bounding boxes and associated object class probabilities. Till now, the authors presented three versions of the YOLO detector and in this work, the YOLOv3 [11] network is used.

YOLOv3 [11] uses a network consisting of 53 convolutional layers of 3 × 3 and 1 × 1 filters, with some shortcut connections between layers (residual blocks) for feature extraction, Fig. 12. Instead of the max-pooling layers that are typically used in CNNs to decrease the dimension of the results from a convolutional layer, the convolutional layers with a stride of 2 are used to down sample the feature maps, to prevent the loss of low-level features often attributed to pooling [11].

To better handle the detection of objects of different sizes, YOLOv3 uses a structure similar to feature pyramid networks to predict boxes at two additional scales. To obtain the features for a finer scale than in the last layer, the features that are computed in a previous layer with finer feature maps are combined with up-sampled features computed further in the network at a coarser scale (Fig. 12). From these features, the box predictions for the finer scale are computed. This is repeated for the smallest scale, merging the features of a convolutional layer with a bigger feature map with previously computed features at a medium scale. At each scale, 3 sets of box predictions are generated for each location in the feature maps.

Final classification in YOLOv3 uses logistic classifiers instead of soft-max in the final layer, with binary cross-entropy loss during training, so in this case, multiple class labels may be assigned to a single detected object simultaneously, which can be useful in cases where partially overlapping classes exist in the data, e.g. vehicle and automobile [11].

V. EXPERIMENT SETUP

In the experiment, the application of the YOLOv3 detector in surveillance applications when using thermal imaging for human and non-human (animal) detection in different weather conditions is tested.

First, we recorded video materials and prepared an annotated database framework for supervised learning of a model for detecting a person when moving in real-life scenes at different distances and in different weather conditions.

As the baseline model, the original YOLOv3 network with input size 608 × 608, referred to as bY, is used. This model was pre-trained on the MS COCO RGB image dataset [73] to detect a large number of object classes.

The baseline performance was compared with the YOLOv3 network that was additionally trained on thermal image data for the class Person, here called tY. For the training and testing purposes, the YOLOv3 detector architecture within the Darknet framework [62] was used, (available in the GitHub repository [74]). The training was done on images from our dataset, described in detail before, where 4270 images were used for training the model, and 1841 images that were not part of the training set were used for testing. Both the custom trained model referred to as tY and the baseline model referred to as bY are tested using the same test set. Then, an experiment was conducted to determine if a smaller training set can achieve good detection results. A model that is trained on only 10% of the data of the Clear subset is referred to as tY_clear10, on 20% tY_clear20, and a model trained on 80% of the data tY_clear80.

Collecting data in the case of fog and rain is very demanding and it is much more difficult to collect sufficient data for training than in the case of clear weather, and therefore a model referred to as tY_clear was trained only on clear data

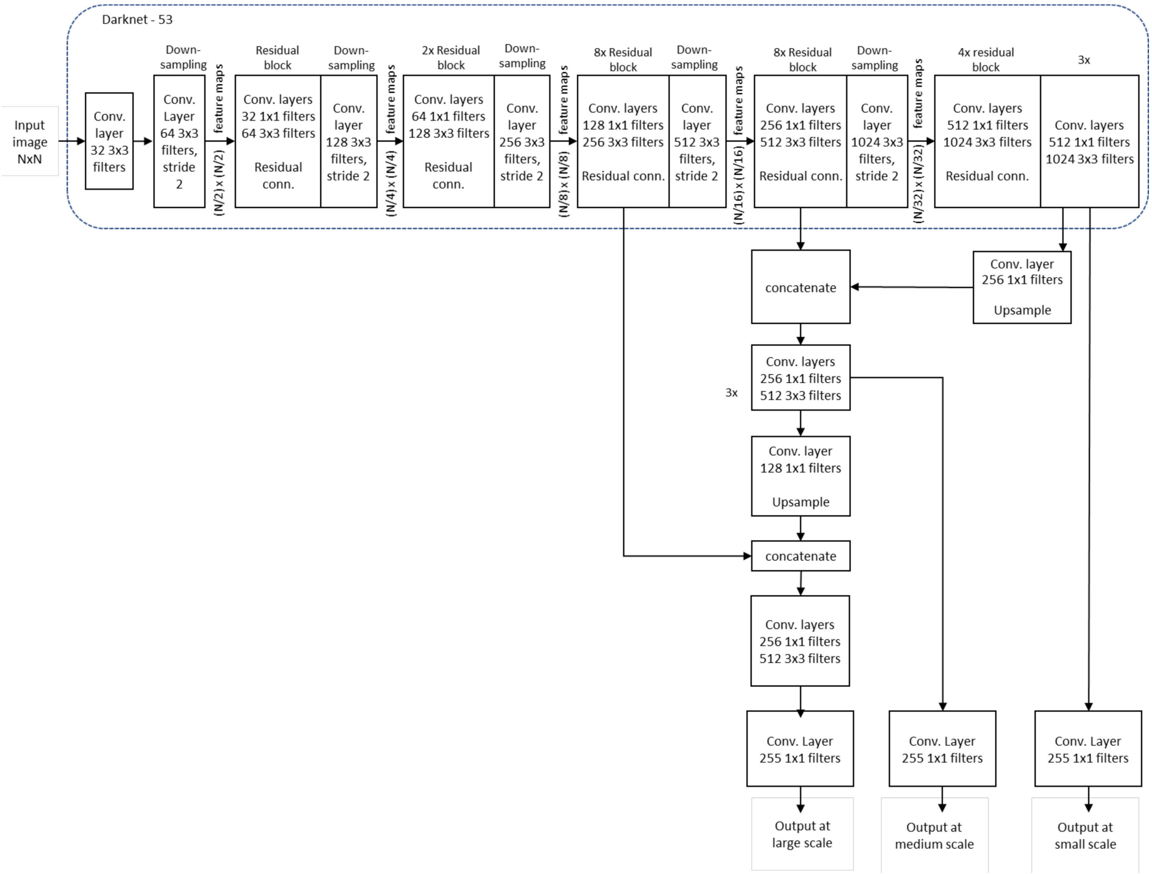

Darknet - 53
Input image NxN
Conv. layer 32 3x3 filters
Down-sampling
Conv. Layer 64 3x3 filters, stride 2
(N/2) x (N/2) feature maps
Residual block
Conv. layers 32 1x1 filters 64 3x3 filters
Residual conn.
Down-sampling
Conv. layer 128 3x3 filters, stride 2
(N/4) x (N/4) feature maps
2x Residual block
Conv. layers 64 1x1 filters 128 3x3 filters
Residual conn.
Down-sampling
Conv. layer 256 3x3 filters, stride 2
(N/8) x (N/8) feature maps
8x Residual block
Conv. layers 128 1x1 filters 256 3x3 filters
Residual conn.
Down-sampling
Conv. layer 512 3x3 filters, stride 2
(N/16) x (N/16) feature maps
8x Residual block
Conv. layers 256 1x1 filters 512 3x3 filters
Residual conn.
Down-sampling
Conv. layer 1024 3x3 filters, stride 2
(N/32) x (N/32) feature maps
4x residual block
Conv. layers 512 1x1 filters 1024 3x3 filters
Residual conn.
3x
Conv. layers 512 1x1 filters 1024 3x3 filters
concatenate
Conv. layer 256 1x1 filters
Upsample
3x
Conv. layers 256 1x1 filters 512 3x3 filters
Conv. layer 128 1x1 filters
Upsample
concatenate
Conv. layers 256 1x1 filters 512 3x3 filters
Conv. Layer 255 1x1 filters
Conv. Layer 255 1x1 filters
Conv. Layer 255 1x1 filters
Output at large scale
Output at medium scale
Output at small scale

FIGURE 12. YOLOv3 network architecture.

subset of the training set for tY and tested on fog and rain data to evaluate its generalization capabilities in bad weather conditions.

Furthermore, an experiment was performed on the Human–Non-Human dataset where the goal was to detect and distinguish two classes, Person and Animal (dog). The dataset used for training and testing consisted of 5,419 images and for this experiment, the dataset was divided 80:20 between training and test images.

Table 4 lists the models we trained and tested in this experimental work and the corresponding training and test set information.

The training was performed for 40000 iterations with a learning rate of 0.001, momentum 0.9, and decay 0.0005. During training, data augmentation was used in the form of random image saturation and exposure transformations, with maximum factors of 1.5, and with random hue shifts by at most 0.1. The input image size was kept at 608 × 608 pixels without multiscale training.

Finally, to test the generalization onto different datasets, a model referred to as tY_transform was tested on widely used and well-known thermal imaging datasets mentioned earlier. The tY_transform model was trained using a subset of the images from our dataset in all weather conditions, artificially augmented by grayscale images to better simulate conditions in which a model would be implemented. The total number of images used for the training tY_transform model was 11,437.

TABLE 4. List of trained models with salient data.

Model	**Additional training on** UNIRI-TID	**Train images**	**Testing part of** UNIRI-TID	**Test images**
bY	-		All	1841
tY	All subset	4270	All	1841
tY_clear10	10% of Clear subset	266	90% of Clear subset	2397
tY_clear20	20% of Clear subset	532	80% of Clear subset	2131
tY_clear80	80% of Clear subset	2131	20% of Clear subset	532
tY_clear	70% of clear subset	1862	All	1841
tY_hNh	Human-nonhuman subset	4335	Human-nonhuman subset	1084

EVALUATION MEASURE

Standard accuracy metrics for the tasks in the domain of computer vision are accuracy, precision, recall, and F-measure [75], [76]. For the object detection task, the mean average precision (mAP) measure is also used to evaluate the performance of the models [77]. The detection results are compared with the ground truth, and detection is considered true positive if the intersection-over-union (IoU) score of the detected bounding box and the corresponding ground truth bounding box is 50% or larger. An example of positive and negative object detection concerning intersection-over-union (IoU) score in case of person detection is shown in Fig. 13.

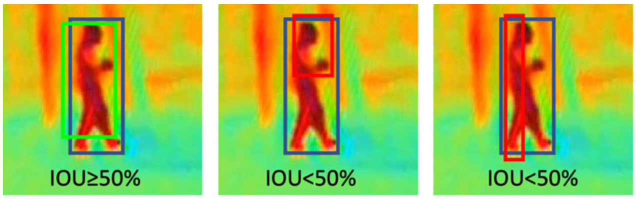


FIGURE 13. Visual negative (left) and positive (right) representation of IoU criteria.

Given an unseen image, the detector returns the bounding box coordinates of the detected objects, the corresponding class label, and a confidence value that indicates how certain the detector is about the detection. The precision-recall curve for a class, here Person, is obtained by varying the confidence threshold from 0 to 100% and calculating the precision and recall at each point. When the confidence threshold is 0, the recall is at its maximum value, while the precision is at its lowest, and the opposite is true for the threshold at 100%. The AP score is then the area underneath the obtained precision-recall curve.

VI. RESULTS AND DISCUSSION

Below are the person detection results for the average precision for all tested scenarios.

A. HUMAN DETECTION IN ALL WEATHER CONDITIONS USING YOLO ON THERMAL VIDEOS

Fig. 14 presents the precision-recall curve for the Person class of the baseline YOLO model bY, that was not trained on thermal images [24], and the same curve for the model tY that was additionally trained on the thermal images from our dataset for the class Person. The plots are computed on the whole test set. Additional training significantly improved the results over the baseline: the AP score achieved with the model tY is 97.93%, while the AP score of bY is 19.63%.

The model bY achieves the 100% precision with a recall of 15.5%, while the model tY achieves the same precision with a recall of approx. 50%, meaning that the tY model can detect a lot more people present in the images without false positives in comparison to the base model.

Looking separately at the performance in different weather conditions, the model achieves the AP score of 97.85% for clear and foggy weather, while in the rain the AP score is even better at 98.08%. With 100% precision, the tY model achieves a 35% recall in the clear weather, 75% in the rain, and 50% in the fog. Fig. 15 shows the complete precision-recall curves for the tY model computed on the parts of the test set corresponding to different weather conditions.

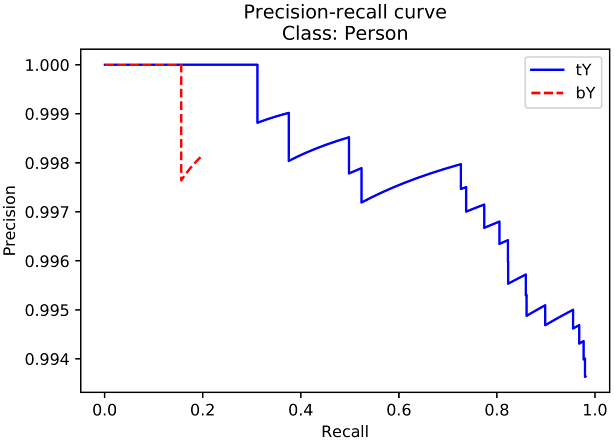


FIGURE 14. Precision/recall curve for baseline YOLO model, bY, and for custom trained YOLO model, tY.

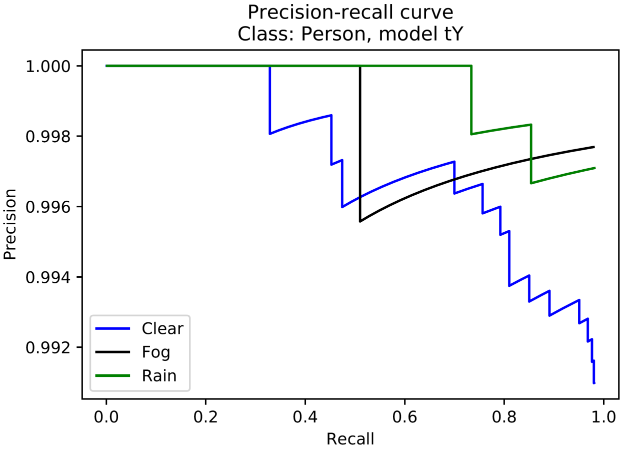


FIGURE 15. Precision/recall curve for model tY in different weather conditions.

Some examples of detection results with both bY and tY models are shown in Figs. 16 to 21 below, with different distances of subjects from the camera and in different weather conditions. In Fig. 16, the tY model has correctly detected the three persons in the image even if they were about 150 m away from the camera and only a few pixels tall, while the bY model missed two out of three persons present. This may still be an unexpectedly good result in this case because the silhouettes of persons are tiny and the relatively small temperature contrast between the persons and the surrounding vegetation.

Generally, images that were taken in the rain (Figs. 17-19), show a larger temperature difference between the cold environment (blue to green tones) and the warm person (shown in red), making it easier to detect persons, at least visually. In the example in Fig. 17, captured in the rain at about 70 m, the tY model correctly detects the person, however, the bY model falsely detects a TV monitor and misses the person, even though the temperature contrast is higher than in example in

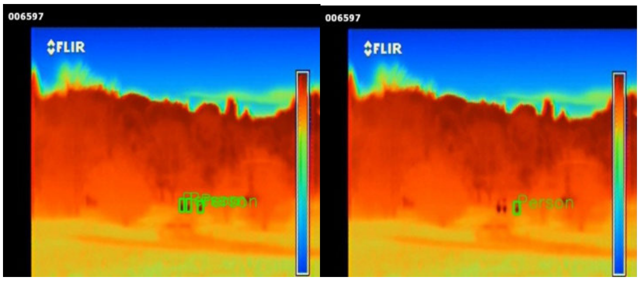


FIGURE 16. Example results of person detection using the bY (left) and tY model (right). Images recorded with a normal lens in clear weather, distance 110-160 m.

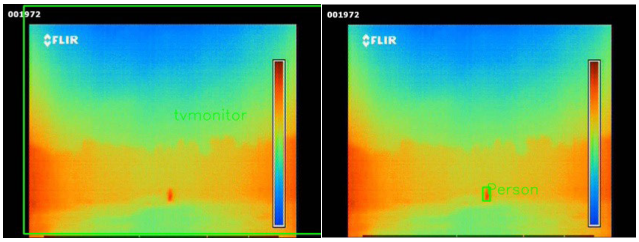


FIGURE 17. Results of person detection (hunched walk) using bY model (left) and tY model (right). Images recorded with a normal lens on rain condition, 70 m distance.

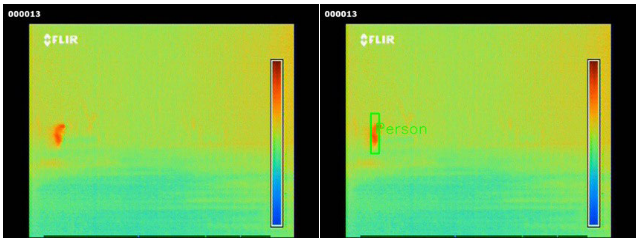


FIGURE 18. Results of person detection (hunched walk) using bY model (left) and tY model (right). Images recorded with a telephoto lens in the rain, 100 m distance.

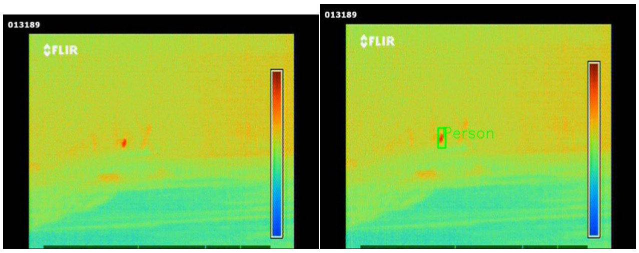


FIGURE 19. Results of person detection (running) using the bY model (left) and tY model (right). Images recorded with a telephoto lens in the rain, 215 m distance.

the Fig. 16. Similar results are commonly obtained in the case of a hunched walk, Fig. 18 or running, Fig. 19.

In other weather conditions, the temperature difference is often smaller, and the detection by the heat map is much harder, as in the fog example (Fig. 20. Here, the tY model has detected one of the two people present on the scene, Fig. 20 (right), while the bY model could not detect any person, Fig. 20 (left).

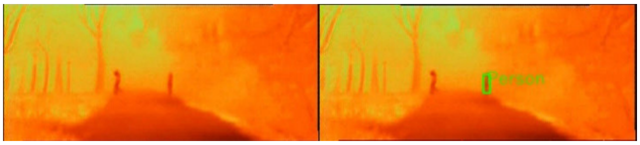


FIGURE 20. Results of detection using the bY model (left) and tY model (right). Images recorded with a telephoto lens in the fog, 50 m distance.

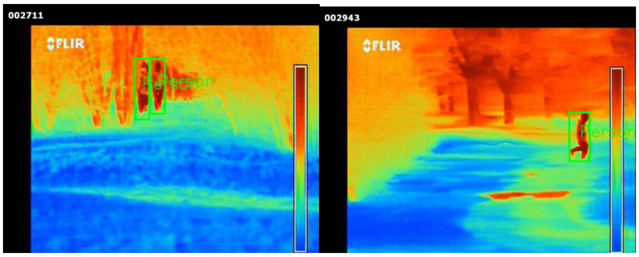


FIGURE 21. Results of person detection on images recorded with a telephoto lens in clear weather condition, 110 m distance: normal walk (left), running (right).

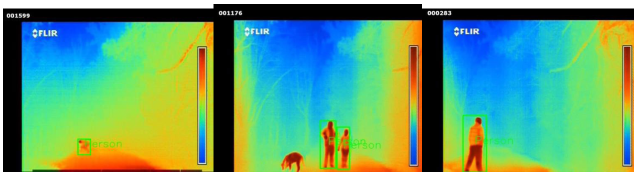


FIGURE 22. Results of person detection on images recorded with a normal lens in foggy weather: crouched walk, 50 m (left), standing still, <30 m (middle), normal walk, <30m (right).

Figs. 21 to 22 show additional examples of detection with the tY model in different scenarios. It can be noted that the model manages to detect people regardless of the mode of movement even when the thermal contrast between the person and the environment was low or at a large distance. The model also successfully distinguished between persons and other objects with similar contours or temperatures, such as tree trunks, detecting persons, and no false positives (Fig. 21, left).

Fig. 22 (left) is an example of a positive detection in the case of hunched movement in the fog. The presence of animals, as in this case the dog in Fig. 22. (right) did not confuse the detector which correctly detected both persons present. In this experiment, the model was not trained with the Non-Human class.

In the rain conditions (Fig. 23), there was a large temperature contrast between the person and the environment, especially when the telephoto lens was used (Fig. 23 (right)). The tY detector has successfully detected people regardless of body posture or camera distance. It is clear that the temperature contrast between person and environment varies greatly, even for the same weather conditions (Fig. 23. middle) and c)). Both Figs. 23. (middle) and (right) are recorded in the rain on the same day but in Fig. 23. (right) the environment is colder than the person and shown with blue and green while in Fig. 23 (middle) the immediate surroundings of the person seem to be warmer and are represented with tones as the person.

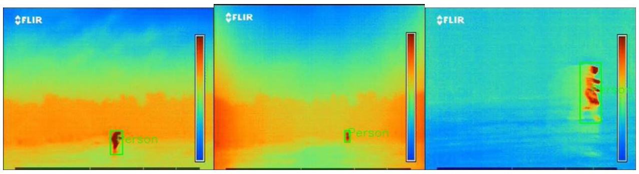


FIGURE 23. Results of person detection on images recorded in the rainy weather: 30 m normal lens (left), 70 m normal lens (middle), 30 m telephoto lens (right).

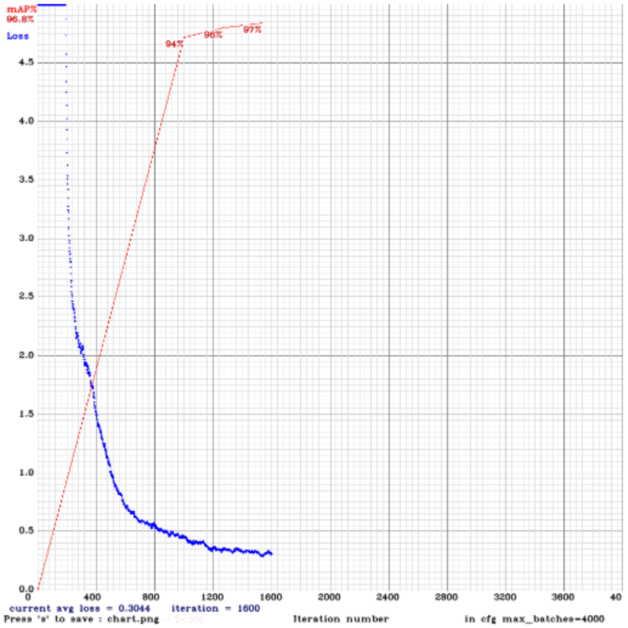


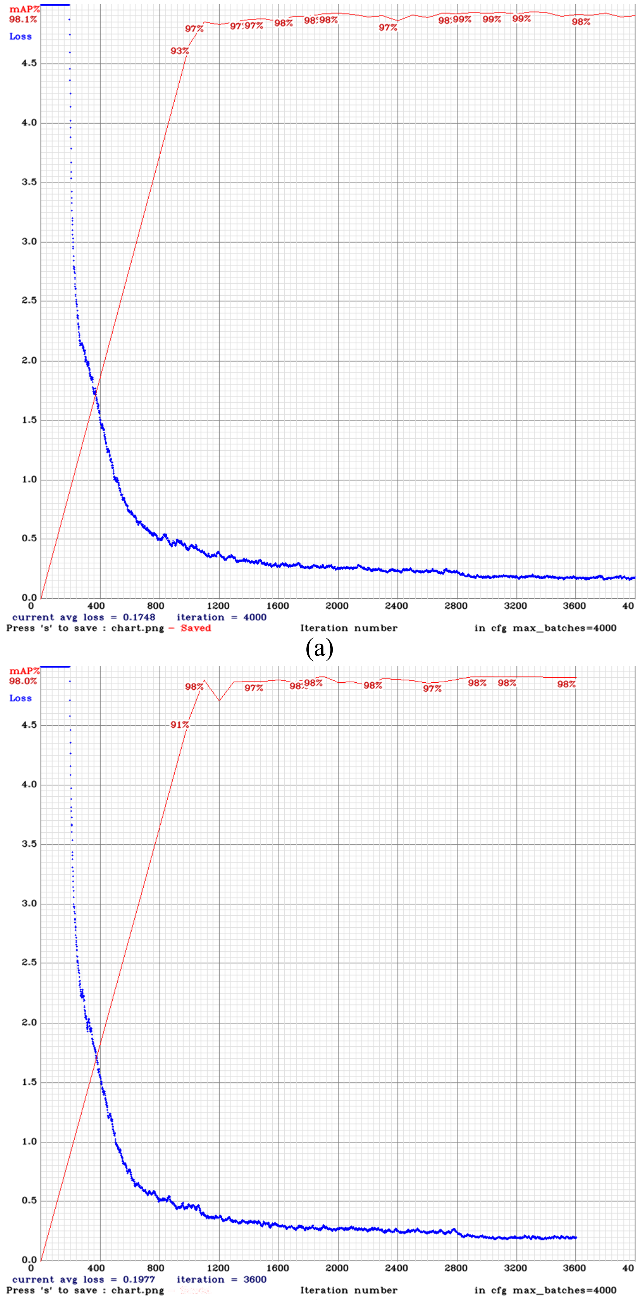
FIGURE 24. mAP and Loss function chart for Clear dataset with training: testing ratio 80:20.

The results given in figs. 21-23 and the presented examples show that additional training of the original YOLO model (bY) for the human detection (class Person) in thermal imaging results in the new tY model that achieves excellent results of detection in different weather conditions. The detection has proved successful even when persons were in the distance or tried to avoid detection by sneaking or walking in a hunched position.

B. PERFORMANCE AGAINST DIFFERENT TRAINING SET SIZE

To see how the size of the training set affects the detection performance, additional experiments with a varying number of training images were performed. We intended to examine the results by taking 80%, then 70% and so successively down to 10% of the data in the Clear subset for training. The rest of the data in the Clear subset is used for testing.

We randomly took 80% of the data from the Clear subset for training set in the first case, 20% in the second case and 10% in the third case but we did not examine other divisions of the sets as it soon became obvious that already with 10% of the data in the training set comparative results were obtained as with 80%. Fig. 24 shows the values of the mAP and loss functions for model with 80:20 training/test split and Fig 25. shows the same functionality for models with 10:90 and 20:80 splits. The models achieved no significant performance gains after about 1600 iterations, showing that in this case, training of the YOLO model is possible on a small set of training images with a small number of iterations, without losing much of the performance of the model.



(a)

(b)

FIGURE 25. mAP and loss function chart for Clear dataset with train:test ratio 10:90 (a), 20:80 (b).

C. HUMAN DETECTION IN ALL WEATHER CONDITIONS USING YOLO MODEL TRAINED ON CLEAR WEATHER

Model tY_clear is trained only on a clear weather subset of the training set of tY model that contains 1862 images.

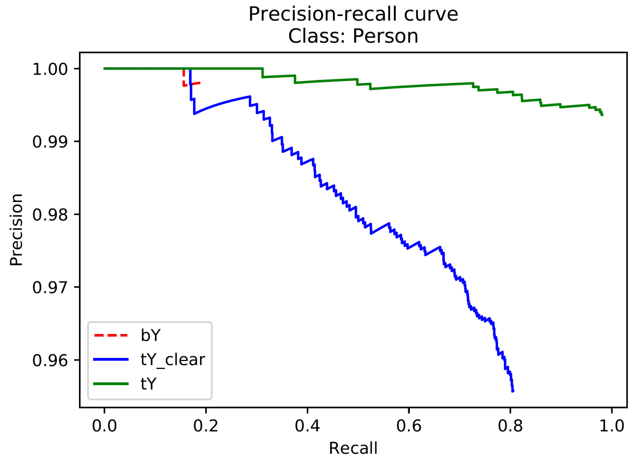


FIGURE 26. **Precision/recall curve for the baseline YOLO model bY, custom trained YOLO model, tY and model trained only on clear weather images, tY_clear.**

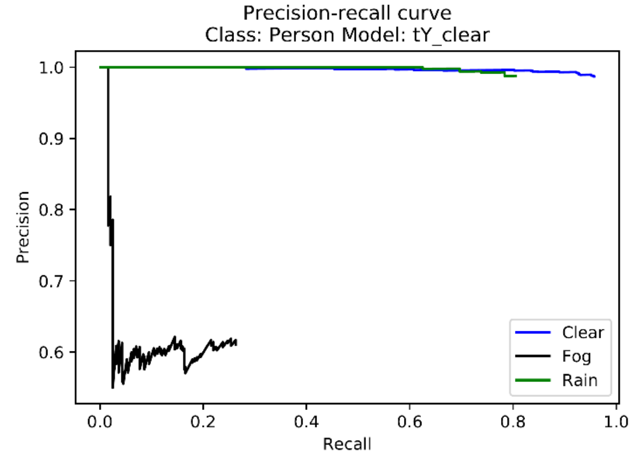


FIGURE 27. **Precision/recall curve for model tY_clear in different weather conditions.**

The generalization capability of the tY_clear is presented in Fig. 26 and compared to performances of models bY and tY in terms of a precision-recall curve for the Person class on the whole test set.

The basic YOLO model bY was not trained on thermal images, so the model trained only on clear weather tY_clear achieved significantly better results than the bY model with an AP score of 79.39% for the class Person compared to AP score 19.63% of the model bY. On the other hand, since less data was used for training the model tY_clear and the data of rain and fog were excluded from the training set, the model tY_clear does not perform as well as the model tY with AP score 97.93%, trained on all weather conditions.

The experiment has shown that additional model learning on a set of thermal images significantly improves performance results, but also that the data collected during clear weather can be successfully used in rainy weather conditions, Fig. 27. The results show that fog has a much greater impact than rain when it comes to detecting people in thermal imaging.

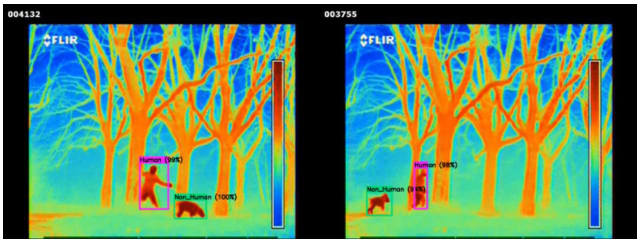
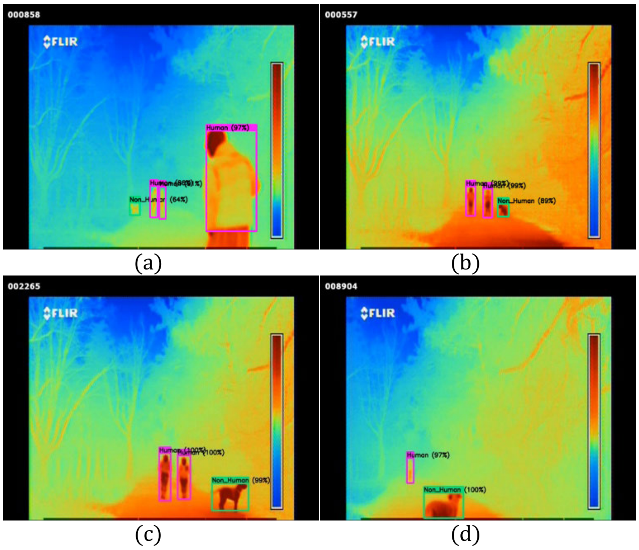


FIGURE 28. **Detection and recognition of humans and an animal in thermal images in dense fog, at the distance of 30 m, with changing body positions: normal walk (left), running (right).**



(a) (b)

(c) (d)

FIGURE 29. **An example of the detection of humans and animals in dense fog at different distances of persons and animals.**

D. HUMAN - NON-HUMAN RECOGNITION IN THERMAL IMAGES AND VIDEOS

For the task of detection of two classes, person and dog, the achieved mAP after training was 97.98%. For the Person class, the average precision was 97.86%, while for the Non-Human class the AP was 98.10%. In this case, Recall was 98%, and F1 score 97%. The obtained results show that YOLOv3 with additional training fully satisfies the basic requirements of the experiment, which is detection and distinction between humans and animals on thermal images and video. Some example results of detection and recognition are shown in Figs. 28 and 29. On the presented examples it can be seen that different distances, different visibility, and body positions did not affect the detection and recognition of humans and animals in thermal images. On the test set, there was a small number of false-positive recognition, that does not affect the applicability of the detector, as a security system based on the thermal imaging could track a detected object for at least a few seconds before sounding alarm or an increased confidence threshold could be used if false positives are to be avoided.

E. BENCHMARK TEST OF TRAINED YOLO MODEL

The detection performance of the tY_transform model trained on our dataset was tested on different publicly available

TABLE 5. Number of images in benchmark datasets for person detection.

Dataset	Number of images	Camera type/ Resolution
ASL ETH FLIR [78]	4381	FLIR Tau 320/324×256 pixels
LITIV2012 Dataset [79]	6325	320x240 pixels
KAIST Multispectral Pedestrian Detection Benchmark [20]	3500	FLIR-A35/640x480 pixels
OSU Thermal Pedestrian Database from OTCBVS Benchmark Dataset Collection [54]	6799	Raytheon 300D/360 x 240 pixels
Terravic Motion IR Database [68]	20255	320x240 pixels
CVC-09: FIR Sequence Pedestrian Dataset [65, 66]	10006	FLIR Tau 2/ 640 × 512 pixels; IDS UI-3240CP/ 1280 × 1024 pixels
VOT-TIR2015 Dataset [67]	7279	320×240 to 1920×480 pixels

TABLE 6. Detection metrics for tY_transform model.

Dataset	mAP	Avg. IOU	Recall	F1 score
ASL ETH FLIR [78]	0.36	0.38	0.27	0.35
LITIV2012 Dataset [79]	0.71	0.44	0.75	0.64
KAIST [20]	0.35	0.19	0.69	0.36
OSU Thermal Pedestrian [54]	0.84	0.50	0.87	0.72
Terravic Motion IR Database [68]	0.97	0.66	0.98	0.92
CVC-09: FIR Pedestrian [65, 66]	0.49	0.23	0.66	0.42
VOT-TIR2015 Dataset [67]	0.83	0.54	0.86	0.77

datasets, namely, ASL_ETH_FLIR dataset [78], LITIV2012 Dataset [79], KAIST Multispectral Pedestrian Detection Benchmark [20], OSU Thermal Pedestrian Database from OTCBVS Benchmark Dataset Collection [54], Terravic Motion IR Database [68], CVC-09: FIR Sequence Pedestrian Dataset [65], [66], and VOT-TIR2015 Dataset [67].

The basic information about the datasets is shown in Table 5. and the achieved detection metrics are shown in Table 6.

It can be noted that the best results of 97% mAP and 92% F1 score tY_transform model has achieved on the Terravic motion dataset since the thermal silhouettes from that set are most similar to the images from our original set. Slightly worse, but still very good detection results of about 83% mAP and 77% F1 score, the model achieves on VOIT-TIR2015 and OSU Thermal datasets, which indicates the fact that the model generalizes well, especially when taking into account the specificity of thermal images, i.e. significant difference in silhouettes of persons depending on distance and shooting conditions. The worst results of about 35% mAP were achieved on the ADT ETH and KAIST datasets which have significantly different recording positions, shooting distances, and shooting conditions.

TABLE 7. Detection metrics for tY models trained on benchmark datasets.

Dataset	mAP	Avg. IOU	Recall	F1 score
ASL_ETH_FLIR	0.75	0.76	0.79	0.58
LITIV_2012	0.83	0.84	0.87	0.69
KAIST	0.63	0.69	0.70	0.49
OSU Thermal Pedestrian	0.86	0.67	0.89	0.89
Terravic	0.96	0.95	0.95	0.75
CVC_IR	0.62	0.66	0.69	0.52
VOT-TIR2015	0.67	0.64	0.75	0.65

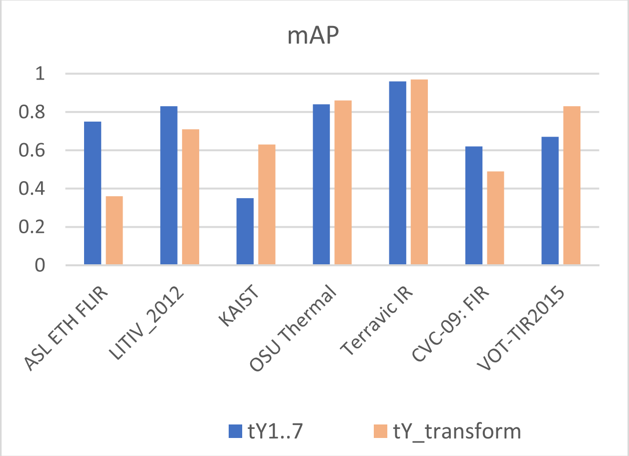


FIGURE 30. mAP results on different datasets for the model trained on our dataset (tY_transform) and each dataset (tY1. . .tY7).

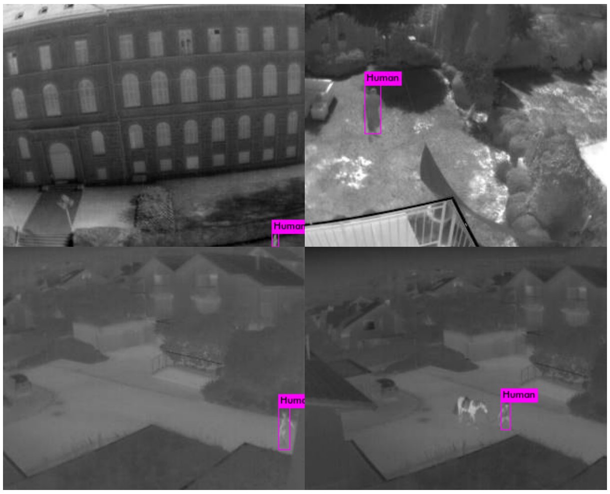


FIGURE 31. ASL ETH FLIR detections.

Additionally, to be able to more credibly validate the performance of the tY_transform model on benchmark thermal

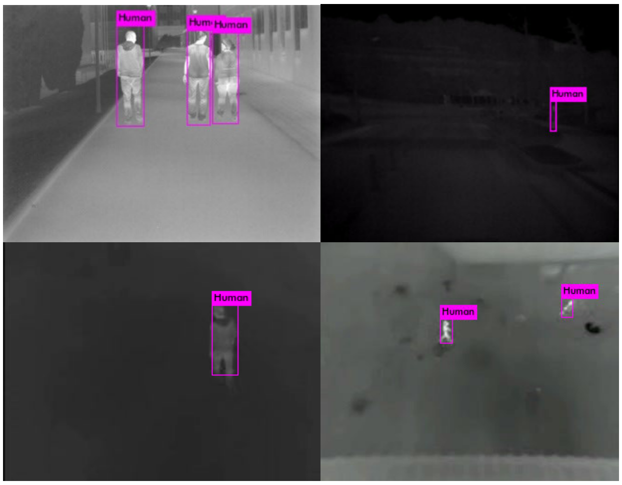


FIGURE 32. CVC IR 09 (up left); KAIST dataset (up right); LITIV 2012 (down) detection examples.

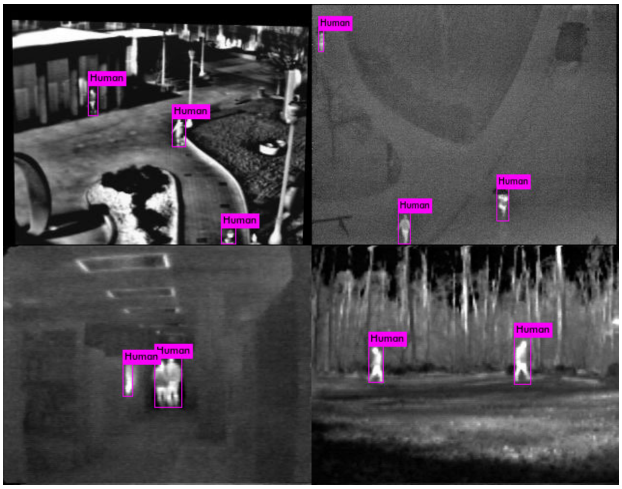


FIGURE 33. OSU thermal (up); Terravic motion dataset (down) detection examples.

image datasets, we performed an additional comparative experiment. We trained and tested the basic Yolo model tY1. . .tY7 on each of the thermal sets from Table 5 in a train: test ratio of 10:90 for 20000 iterations. The achieved results when the model was trained at the same set at which it was tested are given in Table 7.

It is expected that the models would achieve better results when trained on images from the original set from which the images for testing were extracted than when trained from a completely different set as in the case of the tY_transform model. This proved to be true for those sets that are very different from our set of images such as ASL ETH and KAIST and where the tY_transform model did not achieve good results. However, it is interesting that on the sets where the tY_transform model achieves good results, it achieved similar or even better results as the model learned on the original set.

The comparative results of models tY1. . .7 trained at the original set and tY_transform model trained at our set for mAP metrics are shown in Fig. 30.

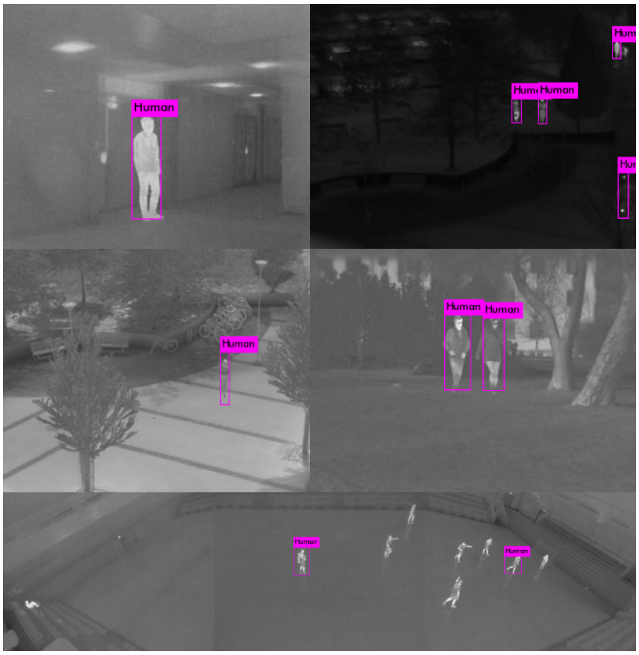


FIGURE 34. VOT-TIR2015 dataset detection examples.

Some detection examples on images from considered datasets are shown in Figures 31.-34.

VII. CONCLUSION

In this paper, the performance of common deep learning methods that are successful for object detection and recognition in RGB visual images was tested on thermal surveillance scenarios. The experiment was conducted on a custom dataset captured during the winter in different weather conditions (clear weather, rain, fog), during the night, and with different distance from the camera, ranging from 30 m to 215 m. The movement of persons varied from normal walking, running to trying to stay out of sight by sneaking or walking hunched to simulate illegal movements around the border and in protected areas.

To select a proper detector for the detection of humans in thermal images, we made a preliminary examination of the selected state of the art object detectors such as Faster R-CNN, SSD, FCOS, Cascade R-CNN, and YOLOv3 that achieve excellent detection results in RGB images. All models were additionally trained for the person class on a subset of thermal images from our dataset without any change od original architecture. The R-CNN, Cascade R-CNN, and YOLOv3 detectors achieved similar detection results in thermal images but YOLOv3 was significantly faster and was used further in the experiment.

The performance of the original YOLOv3 network trained on the COCO RGB dataset was used as the baseline model (labeled bY) and was compared to a model additionally trained for the person class on a subset of our thermal image dataset.

Despite thermal images being very different in appearance from the images recorded in the visible spectrum, it was assumed that the features that YOLO has learned on the COCO dataset of visual images for the class Person will still capture enough shape features that are similar in thermal images and thus provide a reasonable baseline for detection in thermal images as well. However, the original YOLO model (bY) achieved the average precision (AP) for the Person class of only 19.63% in thermal images with a recall of 15.5% at 100% precision, a significantly lower result than the reported AP of about 90% for the Person class in the RGB images. This model could still recognize persons in a number of thermal images, so it served as a good starting point for training a model specifically for thermal imaging.

A model (labeled tY) was trained on a set of about 3000 thermal images from our thermal image dataset, and achieved significantly better results on the test set, with an AP score of 97.93% for all weather conditions. The modestly sized training set proved to be sufficient for achieving excellent results of detection with all tested scenarios, with different weather conditions, pose, and camera distance variations. It was also showed that data collected during clear weather can be successfully used for training the model (labeled tY_clear) that perform well, with AP almost 100% in clear and rain weather condition.

Also, a model was trained to detect both Human and Non-Human objects (here dogs) in thermal images and achieved the mAP score of 97.98%, indicating the possibility of using this or similar CNN models for the development of a standalone system for the automatic monitoring of protected objects and areas.

Additional training of the YOLOv3 model has shown that it is possible to obtain a reliable model by using a relatively small number of images and with a small number of iterations, which greatly shortens the required training time. In addition, the trained YOLOv3 model shows good generalization properties with respect to the results achieved by testing on external image sets. An even better model can be obtained by combining all the sets used in this paper, which will further expand the set with different thermal silhouette representations, which in the case of training models for detecting persons in thermal images is crucial to achieving extremely reliable results for model implementation in real-world conditions.

In the future work, we plan to further examine the performance of person detection in other weather conditions such as exceptionally hot weather and extend the test of non-human objects with other potentially confusing examples such as wild animals.

Finally, we plan to examine the possibility of using a similar detector for the task of human action recognition (running, walking, hunched walking, four-leg walking, etc.), as well for the task of gait recognition.

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, pp. 511–518.

[2] C. K. Eveland, D. A. Socolinsky, and L. B. Wolff, "Tracking human faces in infrared video," *Image Vis. Comput.*, vol. 21, no. 7, pp. 579–590, Jul. 2003.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA: IEEE Computer Society, vol. 1, Jun. 2005, pp. 886–893.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multi-box detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[7] K. He, G. Gkioxari, and R. Dollár, "Mask r-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[8] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[12] M. Buric, M. Pobar, and M. Ivasic-Kos, "Adapting YOLO network for ball and player detection," in *Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods*, 2019, pp. 845–851.

[13] M. Buric, M. Pobar, and M. Ivasic-Kos, "Ball detection using YOLO and mask R-CNN," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2018, pp. 319–323.

[14] M. Ivasic-Kos and M. Pobar, "Building a labeled dataset for recognition of handball actions using mask R-CNN and STIPS," in *Proc. 7th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Nov. 2018, pp. 1–6.

[15] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Goncalves, W. R. Schwartz, and D. Menotti, "A robust real-time automatic license plate recognition based on the YOLO detector," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–10.

[16] D. Shen, X. Chen, M. Nguyen, and W. Q. Yan, "Flame detection using deep learning," in *Proc. 4th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2018, pp. 416–420.

[17] X. Zhang, W. Yang, X. Tang, and J. Liu, "A fast learning method for accurate and robust lane detection using two-stage feature extraction with YOLO V3," *Sensors*, vol. 18, no. 12, p. 4308, Dec. 2018.

[18] J. George, S. Skaria, and V. V. Varun, "Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans," *Proc. SPIE*, vol. 10575, Feb. 2018, Art. no. 105751I.

[19] V. Kharchenko and I. Chyrka, "Detection of airplanes on the ground using YOLO neural network," in *Proc. IEEE 17th Int. Conf. Math. Methods Electromagn. Theory (MMET)*, Jul. 2018, pp. 294–297.

[20] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 07–12.

[21] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[22] D. Heo, E. Lee, and B. C. Ko, "Pedestrian detection at night using deep neural networks and saliency maps," *Electron. Imag.*, vol. 17, pp. 060403-1–060403-9, Jan. 2018.

[23] M. Ivasic-Kos, M. Kristo, and M. Pobar, "Human detection in thermal imaging using YOLO," in *Proc. 5th Int. Conf. Comput. Technol. Appl. (ICCTA)*, New York, NY, USA, 2019, pp. 20–24.

[24] M. Ivasic-Kos, M. Kristo, and M. Pobar, "Person detection in thermal videos using YOLO," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2019, pp. 254–267.

[25] A. Gomez, F. Conti, and L. Benini, "Thermal image-based CNN's for ultra-low power people recognition," in *Proc. 15th ACM Int. Conf. Comput. Frontiers*. New York, NY, USA: ACM, May 2018, pp. 326–331.

[26] Q. Liu, X. Lu, Z. He, C. Zhang, and W.-S. Chen, ''Deep convolutional neural networks for thermal infrared object tracking,'' *Knowl.-Based Syst.*, vol. 134, pp. 189–198, Oct. 2017.

[27] I. Rodger, B. Connor, and N. M. Robertson, ''Classifying objects in LWIR imagery via CNNs,'' *Proc. SPIE*, vol. 9987, Oct. 2016, Art. no. 99870H.

[28] N. Shahid, G.-H. Yu, T. D. Trinh, D.-S. Sin, and J.-Y. Kim, ''Real-time implementation of human detection in thermal imagery based on CNN,'' *J. Korean Inst. Inf. Technol.*, vol. 17, no. 1, pp. 107–121, Jan. 2019.

[29] X. Wang and S. Hosseinyalamdary, ''Human detection based on a sequence of thermal images using deep learning,'' *Int. Arch. Photogramm., Remote Sens., Spatial Inf. Sci.*, vol. 42.2/W13, pp. 1–6, Jun. 2019.

[30] H. Zhang, C. Luo, Q. Wang, M. Kitchin, A. Parmley, J. Monge-Alvarez, and P. Casaseca-de-la-Higuera, ''A novel infrared video surveillance system using deep learning based techniques,'' *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26657–26676, Oct. 2018.

[31] A. Farouk Khalifa, E. Badr, and H. N. Elmahdy, ''A survey on human detection surveillance systems for raspberry pi,'' *Image Vis. Comput.*, vol. 85, pp. 1–13, May 2019.

[32] Y.-L. Hou, Y. Song, X. Hao, Y. Shen, M. Qian, and H. Chen, ''Multispectral pedestrian detection based on deep convolutional neural networks,'' *Infr. Phys. Technol.*, vol. 94, pp. 69–77, Nov. 2018.

[33] X. Dai, Y. Duan, J. Hu, S. Liu, C. Hu, Y. He, D. Chen, C. Luo, and J. Meng, ''Near infrared nighttime road pedestrians recognition based on convolutional neural network,'' *Infr. Phys. Technol.*, vol. 97, pp. 25–32, Mar. 2019.

[34] J. Imran and B. Raman, ''Deep residual infrared action recognition by integrating local and global spatio-temporal cues,'' *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103014.

[35] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, ''InfAR dataset: Infrared action recognition at different times,'' *Neurocomputing*, vol. 212, pp. 36–47, Nov. 2016.

[36] E. J. Lee, B. C. Ko, and J.-Y. Nam, ''Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night,'' *Infr. Phys. Technol.*, vol. 76, pp. 261–270, May 2016.

[37] A. Lakshmi, A. G. J. Faheema, and D. Deodhare, ''Pedestrian detection in thermal images: An automated scale based region extraction with curvelet space validation,'' *Infr. Phys. Technol.*, vol. 76, pp. 421–438, May 2016.

[38] W. Qi, J. Han, Y. Zhang, and L.-F. Bai, ''Infrared object detection using global and local cues based on LARK,'' *Infr. Phys. Technol.*, vol. 76, pp. 206–216, May 2016.

[39] T. Yu, B. Mo, F. Liu, H. Qi, and Y. Liu, ''Robust thermal infrared object tracking with continuous correlation filters and adaptive feature fusion,'' *Infr. Phys. Technol.*, vol. 98, pp. 69–81, May 2019.

[40] X. Zhang, Q. Ding, H. Luo, B. Hui, Z. Chang, and J. Zhang, ''Infrared small target detection based on an image-patch tensor model,'' *Infr. Phys. Technol.*, vol. 99, pp. 55–63, Jun. 2019.

[41] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay, ''Bag of soft biometrics for person identification: New trends and challenges,'' *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 739–777, Jan. 2011.

[42] M. Kristo and M. Ivasic-Kos, ''An overview of thermal face recognition methods,'' in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 1098–1103.

[43] M. K. Bhowmik *et al.*, ''Thermal infrared face recognition—A biometric identification technique for robust security system,'' in *Reviews, Refinements and New Ideas in Face Recognition*, P. M. Corcoran, Ed. Rijeka, Croatia: IntechOpen, 2011.

[44] T. Bourlai and B. Cukic, ''Multi-spectral face recognition: Identification of people in difficult environments,'' in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, Jun. 2012, pp. 196–201.

[45] G. Tanda, ''The use of infrared thermography to detect the skin temperature response to physical activity,'' *J. Phys., Conf. Ser.*, vol. 655, no. 1, 2015, Art. no. 012062.

[46] J. Bareła, M. Kastek, K. Firmanty, P. Trzaskawka, and R. Dulski, ''Determining detection, recognition, and identification ranges of thermal cameras on the basis of laboratory measurements and TTP model,'' *Proc. SPIE*, vol. 8355, May 2012, Art. no. 83551E.

[47] Z. Wu, N. Fuller, D. Theriault, and M. Betke, ''A thermal infrared video benchmark for visual analysis,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 201–208.

[48] B. Blecha, ''Forward looking infrared (FLIR),'' in *Encyclopedia of Optical and Photonic Engineering*. New York, NY, USA: Marcel Dekker, Inc., 2003, pp. 560–568.

[49] FLIR Systems. (2002). *ThermaCAM P10 ThermaCAM P10 Datasheet*. Accessed: May 15, 2018. [Online]. Available: https://www.termogram.com/pdf/therma_cam_p_series/therma_cam_p10/P10_datasheet.pdf

[50] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[51] Z. Cai and N. Vasconcelos, ''Cascade R-CNN: Delving into high quality object detection,'' 2017, *arXiv:1712.00726*. Accessed: Jun. 8, 2020. [Online]. Available: http://arxiv.org/abs/1712.00726

[52] Z. Tian, C. Shen, H. Chen, and T. He, ''FCOS: Fully convolutional one-stage object detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Nov. 2019, pp. 9626–9635, doi: 10.1109/ICCV.2019.00972.

[53] M. Buric, M. Pobar, and M. I. Kos, ''An overview of action recognition in videos,'' in *Proc. 40th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*. Rijeka, Croatia: IEEE, May 2017, pp. 1310–1315.

[54] J. W. Davis and M. A. Keck, ''A two-stage template approach to person detection in thermal imagery,'' in *Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION)*, vol. 1, Jan. 2005, pp. 364–369.

[55] D. Tan, K. Huang, S. Yu, and T. Tan, ''Efficient night gait recognition based on template matching,'' in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, Aug. 2006, pp. 1000–1003.

[56] M. Kristo and M. Ivasic-Kos, ''Thermal imaging dataset for person detection,'' in *Proc. 42nd Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2019, pp. 1126–1131.

[57] *The Thermal Camera Lens*. Accessed: Dec. 1, 2019. [Online]. Available: https://www.pass-thermal.co.uk/flir-131-mm-7-degree-telephoto-p-b-series-lens

[58] Augmentra Ltd. (2018). *ViewRanger: Trail Maps for Hiking, Biking, Skiing*. Accessed: May 10, 2018. [Online]. Available: https://play.google.com/store/apps/details?id=com.augmentra.viewranger.android

[59] Gsmarena.com. *CAT S60—Full Phone Specifications*. Accessed: May 10, 2018. [Online]. Available: https://www.gsmarena.com/cat_s60-7928.php

[60] L. Cheow, E. K. Tiong, and H. Y. Elizabeth, ''Performance challenges for high resolution imaging sensors for surveillance in tropical environment,'' in *DSTA HORIZONS* (Common Sense Approach to Thermal Imaging), G. C. Holst, Ed. Washington, DC, USA: SPIE, 2015, pp. 80–88.

[61] *Annotation Tool*. Accessed: Dec. 11, 2018. [Online]. Available: https://github.com/drainingsun/boobs

[62] J. Redmon. *YOLO: Real-Time Object Detection*. Accessed: Dec. 1, 2019. [Online]. Available: https://pjreddie.com/darknet/yolo/

[63] *VOC Dataset*. Accessed: Dec. 1, 2019. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/

[64] *COCO Dataset*. Accessed: Dec. 1, 2019. [Online]. Available: http://cocodataset.org/#home

[65] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. López, ''Pedestrian detection at Day/Night time with visible and FIR cameras: A comparison,'' *Sensors*, vol. 16, no. 6, p. 820, Jun. 2016. [Online]. Available: http://adas.cvc.uab.es/elektra/enigma-portfolio/item-1/

[66] Y. Socarrás, S. Ramos, D. Vázquez, A. M. López, and T. Gevers, ''Adapting pedestrian detection from synthetic to far infrared images,'' in *Proc. ICCV-Workshop Vis. Domain Adaptation Dataset Bias.*, Dec. 2013, pp. 1–3. [Online]. Available: http://adas.cvc.uab.es/elektra/enigma-portfolio/item-1/

[67] M. Felsberg, A. Berg, G. Hager, J. Ahlberg, M. Kristan, J. Matas, A. Leonardis, L. Cehovin, G. Fernandez, T. Vojir, and G. Nebehay, ''The thermal infrared visual object tracking VOT-TIR2015 challenge results,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 639–651. [Online]. Available: https://www.votchallenge.net/vot2015/dataset.html.

[68] R. Miezianko. (2005). Terravic Research Infrared Database. IEEE OTCBVS WS Series Bench. [Online]. Available: http://vcipl-okstate.org/pbvs/bench/

[69] M. Kieu, A. D. Bagdanov, M. Bertini, and A. D. Bimbo, ''Domain adaptation for privacy-preserving pedestrian detection in thermal imagery,'' in *Image Analysis and Processing—ICIAP*, vol. 11752, E. Ricci, S. R. Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham, Switzerland: Springer, 2019, pp. 203–213.

[70] C. Herrmann, M. Ruf, J. Beyerer, ''CNN-based thermal infrared person detection by domain adaptation,'' *Proc. SPIE*, vol. 10643, May 2018, Art. no. 1064308, doi: 10.1117/12.2304400.

[71] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ''Rethinking the inception architecture for computer vision,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[72] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' 2015, *arXiv:1512.03385*. Accessed: Jun. 8, 2020. [Online]. Available: http://arxiv.org/abs/1512.03385

[73] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ''Microsoft COCO: Common objects in context,'' in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[74] J. Redmon. *Pjreddie/Darknet: Convolutional Neural Networks*. GitHub. Accessed: Dec. 12, 2019. [Online]. Available: https://github.com/pjreddie/darknet

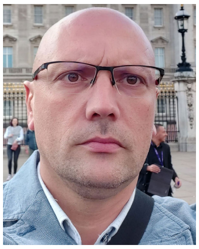
[75] E. Fernandez-Moral, R. Martins, D. Wolf, and P. Rives, ''A new metric for evaluating semantic segmentation: Leveraging global and contour accuracy,'' in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1051–1056.

[76] M. Ivasic-Kos, M. Pobar, and S. Ribaric, ''Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme,'' *Pattern Recognit.*, vol. 52, pp. 287–305, Apr. 2016.

[77] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ''The Pascal visual object classes (VOC) challenge,'' *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[78] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, ''People detection and tracking from aerial thermal views,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jun. 2014, pp. 1794–1800, doi: 10.1109/ICRA.2014.6907094.

[79] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi, ''Thermal–visible registration of human silhouettes: A similarity measure performance evaluation,'' *Infr. Phys. Technol.*, vol. 64, pp. 79–86, May 2014. [Online]. Available: https://www.polymtl.ca/litiv/en/codes-and-datasets



MATE KRIŠTO graduated in management from the Faculty of Economics, University of Rijeka, in 2006. In 2012, he started his doctoral studies in computer science at the Department of Informatics, University of Rijeka. He works in the field of public safety. His research interests include artificial intelligence and computer vision. The aim of his doctoral thesis is to develop algorithms that could be used in the surveillance of borders and protected buildings to increase security for citizens and property.

MARINA IVASIC-KOS (Member, IEEE) received the Ph.D. degree in computer science from the Faculty of Electrical Engineering and Computing, Zagreb. She is currently an Associate Professor with the Department of Informatics and the Head of the Laboratory for Computer Vision, Virtual and Augmented Reality, Centre for Artificial Intelligence, University of Rijeka. She is also the Leader with the National Research Project dealing with automatic recognition of actions in sports and a Researcher at a project dealing with crowd analysis in surveillance. She was involved in numerous business and research projects. Her research interests include AI, computer vision, knowledge representation, and soft computing. She is a Technical Committee Member and a Reviewer for numerous scientific conferences and high-cited journals like *PR* and *ESWA* (Elsevier), the IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE ACCESS, and *Signal Processing*.

MIRAN POBAR received the M.S. degree in electrical engineering from the Faculty of Engineering, University of Rijeka, Rijeka, Croatia, in 2007, and the Ph.D. degree in computer science from the Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia, in 2014. He is currently an Assistant Professor with the Department of Informatics, University of Rijeka. His current research interests include computer vision and action recognition.