DeepThermal Outdoor: A first-person thermal imaging dataset

Nguyen Hoang Bach[1*], Doan Quang Tu[1], Pham Duy Thai[2],
Pham Dang Quang[3], Nguyen Van Duy[4]

[1]Institute of Information Technology/ Academy of Military Science and Technology;
[2]Faculty of Control Engineering/Military Technical Academy;
[3]Faculty of Information Technology/Military Technical Academy;
[4]Faculty of Information Technology/Thuyloi University.
*Corresponding author: nhbach2203@gmail.com
Received 15 Sep 2022; Revised 28 Nov 2022; Accepted 15 Dec 2022; Published 30 Dec 2022.
DOI: https://doi.org/10.54939/1859-1043.j.mst.CSCE6.2022.92-104

ABSTRACT

Recently, thermal imaging modules equipped for infantry soldiers have been a trend to improve the combat ability of soldiers. Soldiers have to perform many different tasks at the same time, so it is necessary to equip them with the tools of automatic target detection, especially human objects detection, in practice. Hence, there is a need to intelligently optimize the effectiveness of thermal imaging equipment. New artificial intelligence and deep learning(DL) approaches are applicable methods that show superior accuracy compared to previous methods. However, state-of-the-art DL methods depend on the generality and diversity of the training data set. To address this issue, our paper presents the DeepThermal Outdoor thermal imaging data set, which is collected from equipment mounted on the body of infantry at various terrain locations. The labeled dataset focuses on human objects with different locomotion postures, and it contains 10,190 images and 22,464 labeled human-objects. Finally, the experiment is conducted with several DL methods using the proposed dataset, and the results show its contribution to the improvement of the performance of DL methods to detect humans on thermal images as well as to evaluate the practical applicability of a DL.

Keywords: Artificial intelligence; Thermal image; Human detection.

1. INTRODUCTION

There are two types of night vision devices, including thermal imaging cameras and low-light cameras; however, the former has more outstanding advantages in comparison with the latter. In fact, using thermal imaging cameras, it is possible to observe things without any sources of light. It is also possible to detect objects at a distance and observe them in conditions of light fog and thin smoke. Due to these advantages, thermal imaging cameras have been more and more widely applied in daily life, for example, they have been used as a security monitoring system or as supporting tools for autonomous control vehicles. In the military, thermal imaging vision devices are equipped for infantry popularly, contributing to the increase in the combat capabilities of forces in night conditions. Furthermore, with the need for infantry equipment, thermal imaging vision devices are chosen due to their compact sizes. With a detection distance of radiating objects of approximately 200 meters, the devices prove to be particularly suitable for combat and technical conditions for the Vietnamese art of war. This leads to the trend of attaching thermal imaging cameras to the integrated systems equipped for infantry soldiers.

The use of thermal imaging devices makes the integration of information-supporting features based on thermal image data for soldiers more convenient. Currently, infantries

are equipped with thermal imaging cameras to enhance the ability to fight at night, in which humans are the most focused objects. Due to the characteristics of combat conditions in modern warfare, infantry soldiers have to perform important and complex tasks, such as processing more information while manoeuvring and moving around. Therefore, it is necessary to have intelligent systems which are capable of automatically detecting human objects on thermal images. Thermal imaging detection is to maximize the effectiveness of the thermal imaging equipment and assist the infantry soldiers in avoiding eliminating important problems when performing their missions. If able to automatically detect objects continuously for a long time, the human detection feature on thermal images can be considered as an assistant to assist in tracking, detecting, and giving alarm signals to soldiers.

In practice, Image Processing (IP), Computer Vision (CV), and Deep Learning (DL) have been applied to recognize, detect and classify such objects and scenes. However, with the specific conditions and requirements of the devices attached to soldiers' helmets, the detection of human objects on the image requires the detection of the locations containing the human object, but not the accuracy in every single pixel. Therefore, in this context, DL models for object detection are preferred due to their advantages, including the ability to detect objects in a bounding box and quick process. Furthermore, DL methods are data-driven models, so the quality of the dataset determines the efficiency of the models in the performance. Hence, in this paper, we introduce the dataset DeepThermal which is collected to provide a dataset for DL models of human object detections. It helps to enhance the generality of the models, and it can be applied to evaluate the practical applicability of models and algorithms in real conditions due to the diversity of data types with different difficulty levels. The main contributions of this paper are as follows:

- The DeepThermal Outdoor dataset is proposed. The data set is collected to simulate the real movement conditions of infantry soldiers with a first-person view. Factors affecting the accuracy of deep learning models are taken into account, and they are implemented with different difficulty levels.

- The dataset is labeled with a large number to serve in the training and evaluation tasks of DL models and to evaluate the applicability in practice. The dataset focuses on human object labels with sub-attributes to provide flexibility in assessing different levels of difficulty.

- Experiments are also conducted to evaluate the accuracy and operability of DL models, and it is found that the dataset is a huge challenge for all the tasks.

The rest of this paper is organized as follows. The related work is discussed in section 2. The dataset is introduced in section 3. Section 4 presents the experimental results and discussion. Finally, we conclude the paper in section 5 and provide some suggestions for several future research directions.

2. RELATED WORK

2.1. Thermal imaging cameras equipped for infantry soldiers

Thermal imaging cameras have been adopted by advanced countries such as Germany, Russia, The US, and The UK since the early years of the twentieth century.

Experiencing different stages of development, thermal imaging cameras have affirmed their great applicability in various fields of social life. Not stopping at the ability to support observations, Thermal imaging cameras are used in complete information integration systems as an important information channel and are widely equipped with intelligent features to help promote their advantages.

ARC4 system[1] (Augmented Reality Command Control Communicate and Coordinate) was designed and manufactured by Applied Research Associates of the US based on augmented reality glasses technology. The main components of the system include a specialized computer, communication equipment, a power supply battery, and night vision goggles. Specialized computers are installed with supporting software to synthesize information and data, displayed on smart glasses in the form of overlapping information layers such as maps, images, orientation, and command orders. For the night vision channel, the image is transmitted to a specialized computer, which analyzes and displays the identification information overlayed on the image received from the night vision goggles and displayed on the smart glasses.

The system of photo-electronic weapons integrated into the helmet of MOHOC[2] (The USA). MOHOC is equipment for infantry soldiers (NLBB), including a thermal imaging camera and smart glasses. Images obtained from thermal imaging cameras are transferred to a computer and displayed on smart glasses, combining map and terrain information to help soldiers increase observation, enrich information, and enhance combat fighting capabilities in night conditions.

2.2. Thermal datasets

There are now thermal and infrared datasets for images for a variety of visual tasks, such as TNO Image Fusion[3], OSU Color-Thermal Database[4], KAIST Multispectral Dataset[5], FLIR Thermal Dataset[6]. The TNO Image Fusion Dataset contains multispectral (enhanced visual, near-infrared, and long-wave infrared or thermal) nighttime imagery of different military scenes, and is recorded in different multi-band camera systems. It contains only 261 thermal images, including many sequences of consecutive similar images and human objects. KAIST and FLIR are datasets for autonomous driving, they capture various regular traffic scenes in day and night time. The dataset is more focused on analyzing urban traffic conditions than detecting people in various terrain conditions. The OSU Color-Thermal Database is a visible-infrared paired dataset for the fusion of color and thermal imagery and fusion-based object detection. The images were taken at a busy pathway intersection on the Ohio State University Campus, cameras mounted to each other on a tripod at two locations approximately 3 stories above ground. The images contain a large number of pedestrians. However, all images are collected in the daytime, so the pedestrians in visible images are already very clear. In such cases, the advantages of infrared images are not prominent.

The above datasets are all collected from the urban environment, under favourable environmental conditions and camera angles, so that they can be used for training and assessment purposes for the Vietnamese environment. Moreover, it is not really suitable in the military field.

2.3. Human detection in thermal images

There are two approaches for detecting humans in thermal images in the literature:

Based on classical computer vision and based on deep learning.

The first approach is an approach based on classical computer vision algorithms. There exist a fair number of approaches for detecting humans in thermal images in the literature. Davis and Keck presented a two-stage template-based method [7], which takes advantage of the invariance of edge information. In the first stage, human contours are obtained by creating a Contour Saliency Map (CSM) of thermal images. CSM represents the belief that each pixel belongs to an edge contour of a person. Then a screening template is produced by averaging the human samples cropped from the CSM images. Last, a multi-resolution screening procedure is applied to obtain candidates. In the second stage, four Sobel filters with different angles are applied to the human samples to get four projected edge images. An Adaboost classifier is trained with the projected images and is applied to new input images. This method proves that edge is a robust feature for object detection in thermal images. Arens and Jungling proposed a local-feature based pedestrian detector on thermal data[8]. In the training phase, they used a combination of multiple cues to find interesting points in the images and used SURF (SpeedUp Robust Features; Tuytelaars et al.[9]) as features to describe these points. Then a codebook is created by clustering these features and building the Implicit Shape Model (ISM) to describe the spatial configuration of features relative to the object center. In the detection stage, SURF features are first located in each image. Then the matching between the features and the codebook is conducted to locate the object center. The challenge of this detector is whether a high performance can be achieved when local features are not obvious, for example, in thermal images of poor quality. Wang et al. have presented a new method for detecting pedestrians in thermal images[10]. The method is based on the Shape Context Descriptor (SCD) with the Adaboost cascade classifier framework. In Qi et al., another advanced driver assistant system (ADAS) was developed, but this time sparse representation was proposed for pedestrian detection in thermal images[11]. Two types of dictionaries, i.e. a generic dictionary optimized by K-SVD and a naive dictionary with basis atoms being directly composed of training samples, were employed to represent image features. In the implementation, a boundary box shrinking scheme is applied to improve the accuracy of the detection by finding the proper size for the boundary box. The experimental results demonstrate a comparable performance of the proposed approach.

Another approach that has been used a lot in recent studies is to use deep learning for human detection of thermal images. Popular object detection deep learning networks such as YOLO [12], Faster-RCNN [13], and single-shot multi-box detector (SDD)[14] have been used for detecting people on thermal images. Jia, Xinyu, et al [15] use Yolov5 and Yolov3 for pedestrian detection. The model was first pre-trained on the COCO dataset and then fine-tuned on their LLVIP dataset which is a visible-infrared paired dataset for low-light vision. The results show that there are many missed detection phenomena in visible images. The infrared image highlights pedestrians and achieves a better effect in the detection task, which not only proves the necessity of infrared images but also indicates that the performance of the pedestrian detection algorithm is not good enough under low-light conditions. Akshatha, K. R., et al. [16] proposed using Faster - RCNN and SSD algorithms for human detection in aerial thermal Images. They carried out the performance evaluation of Faster-RCNN and SSD algorithms with different

backbone networks to detect human targets in aerial view thermal images. For this purpose, two standard aerial thermal datasets having human objects of varying scales are considered with different backbone networks, such as ResNet50, Inception-v2, and MobileNet-v1. The evaluation results demonstrate that the Faster R-CNN model trained with the ResNet50 network architecture out-performed in terms of detection accuracy, with a mean average precision (mAP at 0.5 IoU) of 100% and 55.7% for the test data of the OSU thermal dataset [2] and AAU PD T dataset [17], respectively. The SSD with MobileNet-v1 achieved the highest detection speed of 44 frames per second (FPS) on the NVIDIA GeForce GTX 1080 GPU. Fine-tuning the anchor parameters of the Faster-RCNN ResNet50 and SSD Inception-v2 algorithms caused remarkable improvement in mAP by 10% and 3.5%, respectively, for the challenging AAU PD T dataset. Devaguptapu, Chaitanya, et al. propose a 'pseudo-multimodal' object detector trained on natural image domain data to help improve the performance of object detection in thermal images [18]. The key idea is to borrow knowledge from data-rich domains such as visual (RGB) without the explicit need for a paired multimodal dataset. They achieve this objective by leveraging the success of recent image-to-image translation methods to automatically generate a pseudo-RGB image from a given thermal image, and then propose a multimodal Faster-RCNN architecture to achieve the objective of detecting humans on thermal imagery. They use the recently released FLIR ADAS [6] dataset and the KAIST Multispectral Pedestrian dataset [5] for experimental studies. They demonstrate that their framework achieves better performance than the baseline, even when trained only on a quarter of the thermal database.

3. DATASET

DeepThermal Outdoor dataset is collected for the purpose of detecting people from thermal images in a real combat environment. Therefore, data collection requires attentive selection with the collection strategies, device setting methods, and object attribute selection for labeling. This section contains three subsections, the first presents the data collection strategy, and the second describes the statistical data features of this dataset and compares it with other thermal datasets.

3.1. Data collection

To collect data, the first step is to select equipment and mount it on the soldier's body or equipment. In this study, the Boson 320 longwave infrared thermal camera module was chosen, with a 12 μm uncooled detector that comes in resolutions 320 × 256, FPS = 9 frames/second. This device meets the image quality, weight and low power consumption. To collect data, the first step contains selecting equipment, and in the second step the thermal camera was attached to the infantry helmet. Figure 1 illustrates the camera and attaching method in practice. Surface Pro 4 tablet, m3/4GB/128GB is worn on the back to collect data and power the camera via signal and power cables. The data collection process was carried out under night conditions, the temperature changed from 33 to 26 degrees Celsius, humidity of 62%.

DeepThermal Outdoor dataset is labeled from the frames contained in raw videos. Using the CVAT(Computer Vision Annotation Tool), a total of 10,190 frames were annotated with 22,464 human objects. Since the thermal imaging camera has a capturing

frequency of 9 images per second, and the camera is placed on the helmet of the soldier, the background scene is constantly changing. Therefore, all frames are used for labeling.

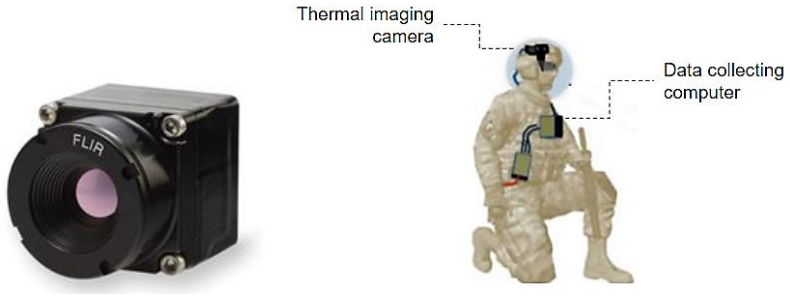

Figure 1. *Camera Boson 320 and infantry soldier with data collection equipment.*

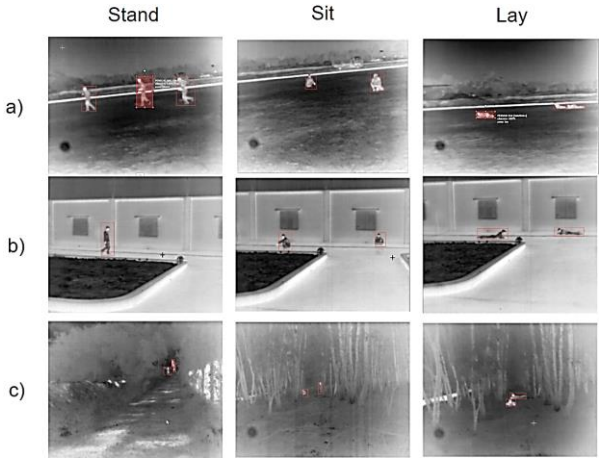

Figure 2. *Illustration of human annotations in different poses and terrain categories. a) Flat terrain; b) Urban area; c) Mountainous terrain.*

Based on different human objects occurring in the image, 3 attributes of each label were defined for the object detection task:

- **Bounding box:** the rectangle surrounding the human region in the frame.

- **Human pose:** there are 3 pose categories. The stand represents standing and running postures. Sit represents sitting and kneeling. Lay represents lying and crawling on the ground.

- **Obscure rate:** There are 4 levels of occlusion based on the proportion of the human body obscured by an obstacle namely 0%, 25%, 50%, and 75%. Depending on the specific task or application, this attribute is used to filter the data corresponding to different difficulty levels corresponding to the minimum occlusion rate of the human object.

Statistics are conducted to show the correlation ratio of the number of images and labels between factors such as posture, terrain, and occlusion ratio.

Statistics are performed to show the correlation ratio of the number of images and labels between factors such as posture, terrain, and occlusion ratio.

Figure 3 shows the correlation ratio of the number of images to the collected topographic factors and the number of labeled objects on each image, respectively. It can be seen that the number of instants on the image is mainly from 1 to 3 since the dataset is aimed at cases

where it is difficult for infantry to observe a small number of objects. The distribution of the number of frames with the different numbers of people on the image from 0 to 5 is 0.13%, 24.07%, 38.00%, 33.68%, 1.27%, and 2.86%, respectively. Additionally, the distribution of the number of frames according to the terrain types is more uniform, respectively 26.23% for flat, 36.29% for urban, and 37.46% for mountainous.

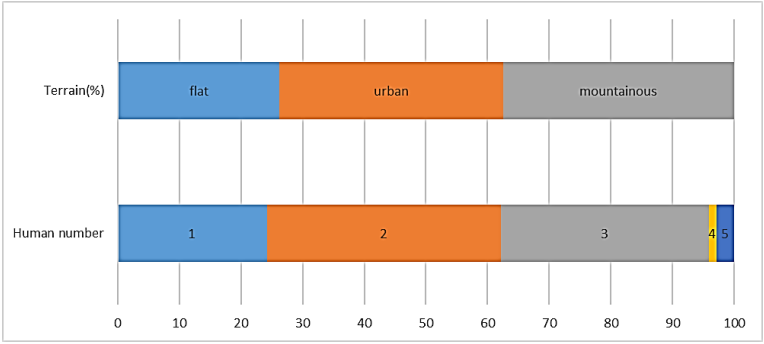

Figure 3. *Distribution of annotated frames. a) Terrain: terrain category; b) Human number: the number of human labels in the frame.*

Figure 4 shows the correlation ratio of the number of human object labels on the dataset, with the factors being the area of the bounding box, and the two attributes of the label, the occlusion ratio and the posture of the human object. The size distribution of the instances is 7.94% for large, 43.38% for medium, and 48.68% for small. The obscured rate distribution of the instances is 87.76% for 0, 3.79% for 25, 4.25% for 50, and 4.10% for 75. The human pose distribution of the instances is 11.01% for lay, 17.99% for sit, and 71.00% for stand, respectively.

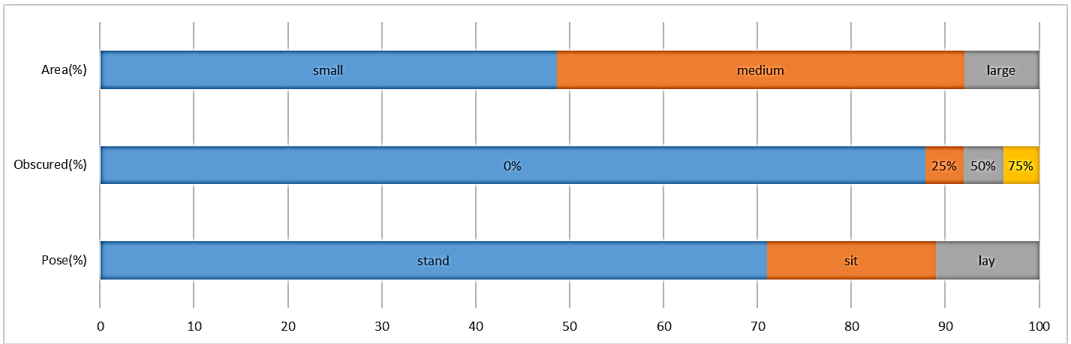

Figure 4. *Distribution of labels by attributes. a) Area: size category of label bounding box area; b) Obscured: obscure percentage; c) Pose: human pose of instance.*

Table 1. *Details of datasets used for training, validation and testing.*

Dataset	Frames	Resolution	Camera angle	Human
TNO	261	768 x 576	On the ground	Few
OSU	285	320 x 240	Surveillance	✓
KAIST	4750	640 x 480	Driving	✓
FLIR	5258	640 x 512	Driving	✓
LLVIP	15488	1080 x 720	Surveillance	✓
DeepThermal Outdoor	10408	320x256	Egocentric	✓

4. EXPERIMENTS AND DISCUSSION

In this section, we describe in detail the experiments of various object detection models on DeepThermal Outdoor dataset and evaluate the results. The experiments are conducted on HPC with GeForce GTX 1080 Ti GPU 16GB.

4.1. Dataset

The entire labeled DeepThermal dataset is used for training and evaluation purposes. The dataset is divided in the ratio of 8:1:1 for the training, validation, and test sets, respectively. The above ratio is calculated on the total number of images of each data set. Each data set is extracted from different videos to increase the accuracy of the evaluation process. During this experiment, we utilize all images and labels, regardless of the pose, occlusion ratio, and size of each instance on the image, for the purpose of evaluating results across a variety of conditions.

4.2. Models and Implementation

The experiment is conducted with three models namely YOLOv3 with backbone DarkNet-53, Faster-RCNN with backbone R-50-FPN and SSD with backbone VGG16. The mentioned above models are implemented using MMDetection platform and Pytorch backend. The detection network is updated via the Stochastic Gradient Descent (SGD) optimizer, and the learning rate, momentum, and weight decay are set to 0.02, 0.9 and 0.0001, respectively. All the training data is divided into mini-batches for network training, and the mini-batch size is set as four during the training stage. Data augmentation was performed, including vertical flipping, horizontal flipping, random rotation, and random scaling. The models are trained in 200 epochs and the models generated at the epoch with a minimum validation value on the validation set are the final models.

4.3. Evaluation metric

The accuracy of the algorithm is evaluated based on the Accuracy Precision (AP) and Accuracy Recall metrics, which are calculated based on the overlap ratio (IoU-Intersection over Union) between the bounding box of the human area on the detected thermal image and the bounding box of a labeled object.

$$IOU = \frac{area\ of\ overlap}{area\ of\ union} \tag{1}$$

Where "area of intersection" is the area of intersection between the area containing the detected human and the area surrounding the human of the label data, "area of union" is the area of the union of the two regions mentioned above.

True Positive (TP): A correct detection. Detection with IOU ≥ threshold;

False Positive (FP): A wrong detection. Detection with IOU < threshold;

False Negative (FN): A ground truth not detected.

Accuracy Precision

Precision is the ability of a model to identify only the relevant objects. It is the percentage of correct positive predictions and is given by:

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{all\ detection} \tag{2}$$

The AP metric is calculated based on the accuracy at each IoU metric threshold. For example:

AP50 means the AP at IoU threshold of 0.5, AP75 means the AP at IoU threshold of 0.75, and AP means the average of AP at IoU threshold of 0.5 to 0.95, with an interval of 0.05.

APs: AP measurement only on small instances, area < 32x32

APm: AP measurement only on medium instances, 32x32 < area < 96x96

APl: AP measurement only on large instances, area > 96x96

Accuracy Recall

Recall is the ability of a model to find all the relevant cases (all ground truth bounding boxes). It is the percentage of true positives detected among all relevant ground truths and is given by:

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{all\ ground\ truths} \quad (3)$$

In this experiment, AR is calculated with mean AR. In addition, with object detection models, the confidence threshold is set at 0.5(50%).

4.4. Results and discussion

The results of the assessment of the accuracy of the models are calculated on two metrics, namely AP and AR.

Table 2. *Evaluation of model in AP metrics.*

Models	AP	AP50	AP75	APs	APm	APl
YOLO	0.495	0.934	0.477	0.472	0.558	0.320
SSD	0.423	0.750	0.445	0.403	0.549	0.172
Faster-RCNN	0.600	0.952	0.641	0.582	0.625	0.811

It can be seen that the FasterRCNN model gives the highest accuracy of 0.600, generally, the accuracy increases as the IoU ratio decreases (from 75% to 50%), and decreases as the size of the object gets smaller.

Table 3. *Evaluation of model in AP metrics.*

Models	AR	ARs	ARm	ARl
YOLO	0.297	0.312	0.292	0.000
SSD	0.173	0.215	0.127	0.100
Faster-RCNN	0.649	0.642	0.654	0.825

The AR metric represents the ability to avoid missing objects in the image. Similar to the results on the AP measure, FasterRCNN showed the highest accuracy on the AR measure. For practical application, we can consider adjusting the minimum size of the object so that the model can meet the required accuracy.

Table 4. *Evaluation of model in AP metrics in flat terrain test set.*

Models	AR	ARs	ARm	ARl
YOLO	0.382	0.310	0.4502	0.000
SSD	0.423	0.355	0.217	0.100
Faster-RCNN	0.789	0.722	0.654	0.945

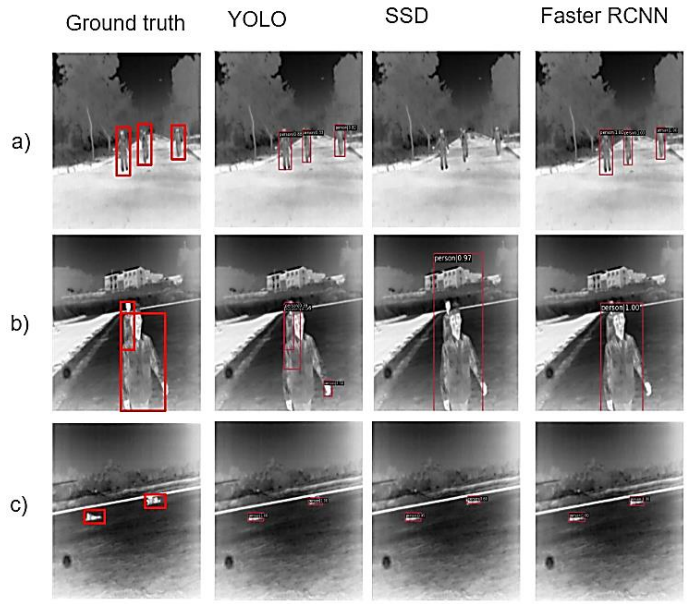

Figure 5. *Detection results in flat terrain. a) Humans walk within far distance; b) humans walk within near distance; c) humans lay on the ground.*

Table 5. *Evaluation of model in AP metrics in urban terrain test set.*

Models	**AR**	**ARs**	**ARm**	**ARl**
YOLO	0.117	0.242	0.192	0.000
SSD	0.111	0.090	0.057	0.130
Faster-RCNN	0.321	0.421	0.325	0.642

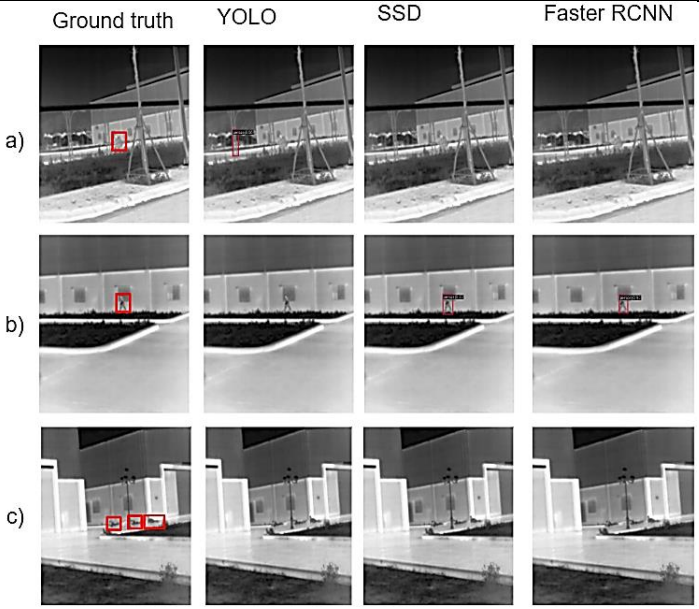

Figure 6. *Detection results in the urban area. a) Human sits behind the obstacles; b) Human walks along a building; c) Human lays on the facade of a building.*

On flat terrain, the object detection models show better performance, with different poses and distances. This can be explained because the contour of the object is more

clear than the background due to the difference in temperature. Additionally, in b) only YOLO can detect 2 objects when these 2 labels obscure each other in the frame, while Faster RCNN detects only 1 object and YOLO also results in a bounding box beyond the notation bounding box.

There are many missed detection phenomena in the urban area, when in conditions of many obstacles. Specifically, the background in the generated image is messy, the contours of the human are not clear and the details are wrong, and there are many artifacts in the image. There are SSD and FasterRCNN has the ability to detect humans in b), In addition, the models missed or detected incorrectly in the remaining cases.

Table 6. *Evaluation of model in AP metrics in urban terrain test set.*

Models	**AR**	**ARs**	**ARm**	**ARl**
YOLO	0.458	0.524	0.290	0.000
SSD	0.326	0.421	0.342	0.100
Faster-RCNN	0.724	0.635	0.701	0.856

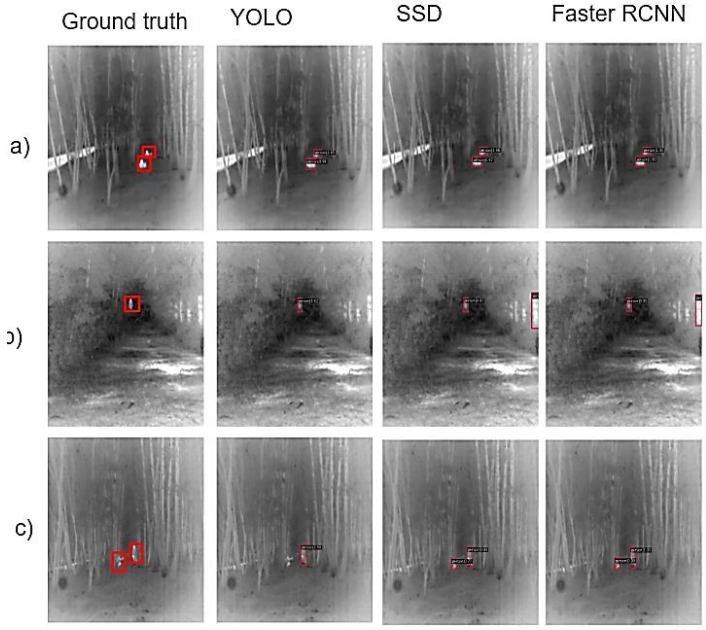

Figure 7. *Detection results in mountainous terrain. a) Two people lay on the ground in a forest; b) One person walks on a ridgeline; c) Two people stand and sit behind a bush.*

On mountainous terrain, despite many obstacles, the contours of the object are still clear compared to the background, the above conditions result in a correction in a) and c) cases. In b) only YOLO gave correct detection results, the rest SSD and FasterRCNN both detected the wrong object when the object appeared with grayscale and similar shape to the human object in the image.

5. CONCLUSIONS

In this paper, the DeepThermal Outdoor dataset is proposed by collecting data from human-attached thermal cameras at nighttime, and it contains large labeled images in various terrain conditions and human poses. The environment and movement conditions of people in the dataset are simulated according to the situations of infantry soldiers in

the terrain of Vietnam. The dataset is suitable with DL models to solve the problem of detecting people on thermal images with images collected from the observation device mounted on the helmet of a soldier. Additionally, the labels of the dataset are annotated with attributes for the purpose of filtering objects on the image with different levels of complexity. Dataset is also used to train and evaluate object detection models for human detection of thermal images. The experiments on the dataset indicate that it is applicable for the state of art DL methods. However, in the future, the dataset will be improved to adapt to the performance of the models.

Acknowledgements: *This research is funded by Academy of Military Science and Technology (AMST) under the young researcher's orientation foundation in 2022.*

REFERENCES

[1]. ARC4: Augmented Reality Command Control Communicate and Coordinate. https://www.ara.com/arc4/

[2]. MOHOC production. https://www.mohoc.com/product/

[3]. A. Toet et al. Tno image fusion dataset. https://doi.org/10.6084/m9.figshare.1008029.v1.

[4]. J. W. Davis and V. Sharma. *"Otcbvs benchmark dataset collection".* http://vcipl-okstate.org/pbvs/bench/, (2007).

[5]. S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. *"Multispectral pedestrian detection: Benchmark dataset and baselines"* in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015).

[6]. O. Ronneberger, P. Fischer, and T. Brox, *"U-Net: Convolutional networks for biomedical image segmentation"* in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, (2015).

[7]. M. A. Keck, J.W. Davis, *"A two-stage template approach to person detection in thermal imagery".* In: Proc. Wkshp. Applications of Comp. Vision, (2005).

[8]. M. Arens, K. Jungling, *"Feature based person detection beyond the visible spectrum"* in IEEE CVPR Workshops, (2009).

[9]. T. Tuytelaars, H. Bay, L. V. Gool, *"Surf: Speeded up robust features"* in: Proc. 9th European Conference on Computer Vision, Graz, Austria, (2006).

[10].W. Wang, J. Zhang, C. Shen, *"Improved human detection and classification in thermal images"* in: IEEE 17th International Conference on Image Processing, (2010).

[11].B. Qi, V. John, Z. Liu, S. Mita, *"Use of sparse representation for pedestrian detection in thermal images"* in: CVPR workshop, IEEE, (2014).

[12]. Redmon, Joseph, et al. *"You only look once: Unified, real-time object detection."* in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016).

[13].Ren, Shaoqing, et al. *"Faster r-cnn: Towards real-time object detection with region proposal networks."*, Advances in neural information processing systems 28 (2015).

[14].Liu, Wei, et al. *"SSD: Single shot multibox detector."*, in European conference on computer vision. Springer, Cham, (2016).

[15].Jia, Xinyu and Zhu, Chuang and Li, Minzhen and Tang, Wenqi and Zhou, Wenli, *"LLVIP: A Visible-infrared Paired Dataset for Low-light Vision",* in: Proceedings of the IEEE/CVF International Conference on Computer Vision, (2021).

[16].K. R. Akshatha, et al. *"Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms."* in Electronics, (2022).

[17].N. U. Huda, B. D. Hansen, R. Gade, T. B. Moeslund, *"The effect of a diverse dataset for transfer learning in thermal person detection",* in Sensors, (2020).

[18].Devaguptapu, Chaitanya, et al., *"Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery."*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2019).

TÓM TẮT

DeepThermal Outdoor: Bộ dữ liệu ảnh nhiệt thu thập từ góc nhìn thứ nhất

Ngày nay thiết bị quan sát ảnh nhiệt được trang bị cho người lính bộ binh là xu thế nhằm nâng cao khả năng tác chiến của người lính. Người lính cùng một thời điểm phải thực hiện nhiều nhiệm vụ khác nhau, do đó việc trang bị tính năng tự động phát hiện mục tiêu đặc biệt là đối tượng người, và cảnh báo là cần thiết cho việc thông minh hóa và phát huy tối đa hiệu quả của khí tài quan sát ảnh nhiệt. Trí tuệ nhân tạo và học sâu là các phương pháp thể hiện được độ chính xác vượt trội so với các phương pháp trước đây. Tuy nhiên, các phương pháp học sâu tiên tiến đều phụ thuộc vào mức độ tổng quát và sự đa dạng của tập dữ liệu huấn luyện. Bài báo này giới thiệu tập dữ liệu ảnh nhiệt DeepThermal Outdoor, được thu thập từ thiết bị gắn trên cơ thể người lính bộ binh tại nhiều địa điểm địa hình khác nhau. Bộ dữ liệu được gán nhãn tập trung vào đối tượng người với các tư thế vận động khác nhau. Bộ dữ liệu bao gồm 10,190 ảnh và 22,464 nhãn đối tượng. Một số phương pháp học sâu phát hiện đối tượng được huấn luyện và kiểm thử trên bộ dữ liệu này và kết quả chỉ ra rằng còn nhiều thách thức cần giải quyết đối với phát hiện người trên ảnh nhiệt đối với điều kiện tác chiến đặc thù của người lính bộ binh. Bộ dữ liệu sẽ góp phần tăng độ chính xác các phương pháp học sâu phát hiện người trên ảnh nhiệt cũng như đánh giá khả năng áp dụng trên thực tế của một phương pháp học sâu.

Từ khóa: Trí tuệ nhân tạo; Ảnh nhiệt; Bài toán phát hiện người.