Multimodal Human Detection Using YOLO and Representation Learning for Robot Perception

Kennedy O. S. Mota, Diogo S. de Oliveira, Luís Garrote, Cristiano Premebida
University of Coimbra, Institute of Systems and Robotics
Department of Electrical and Computer Engineering, Coimbra, Portugal
{kennedy.mota, garrote, cpremebida}@isr.uc.pt

***Abstract*—This work concentrates on the problem of multi-sensor people detection using YOLO trained on four distinct modalities: depth and intensity LiDAR-maps, RGB, and 'thermal' images. RGB cameras, ubiquitous in this application domain, offer great resolution but struggle with adverse lighting conditions resulting in overexposed or underexposed images which then impact negatively on the performance of the algorithms. Thermal (long-wave infrared) cameras are more resilient against varying light conditions and provide complementary textural features, although with lower resolution when compared to RGB cameras. LiDAR sensors, while having a significantly low resolution, contribute to a physically interpretable mapping of the environment providing precise information regarding size/dimension and location of the objects. The main goal of this work is to tackle people detection using deep-models trained on single and multi-modality representations. To support the experimental part this work introduces a new multimodal dataset (called MID-3K). MID-3K allows the development of data fusion strategies by leveraging four modalities (obtained from three distinct exteroceptive sensors mounted on a mobile robot). Leveraging on a single-modality YOLO framework, we propose a multimodal representation learning approach to improve the baseline performance and to capture more relevant features across all input modalities. The evaluation of the proposed detection pipeline is conducted on the MID-3K dataset, where the reported results are grounded on state-of-the-art performance measures. The new dataset is available in a GitHub repository[1].**

I. Introduction

In recent years the integration of advanced sensing technologies in robotics has seen significant advancements, particularly in the realm of object detection in indoor environments. Human detection, in particular, is a critical area of research with extensive applications in various fields such as security, healthcare, autonomous navigation, smart home systems, situational awareness, and more [1]. Robot's ability to accurately detect and interact with humans in real-world environments can greatly enhance autonomous navigation, emergency response, and assistive robotics, thereby improving safety and efficiency in these domains. Numerous deep learning (DL) and computer vision algorithms, embedded with filtering techniques and sensor fusion, have been developed to improve performance and achieve more reliable results in people detection, with cameras being widely used as the primary sensor for the perception process [2], [3].

Human detection in indoor settings presents unique challenges. Traditional vision-based systems offer high-resolution images allowing detailed information on people, objects, surfaces, textures, and surrounding environment. These images can be readily interpreted by both human observers and computer vision algorithms excelling in many computer vision tasks. However, despite being much more affordable and easily available they often struggle with varying lighting and occlusions.

Thermal imaging, on the other hand, can detect temperature differences between objects and their surroundings, visualizing them by their "heat" signatures. This capability makes thermal imaging suitable for low-light environments, adverse weather conditions and scenarios where the field of view may be obstructed by foliage, smoke or camouflage like caves, tunnels, and mines. However, this modality offers lower resolution compared to RGB cameras. Currently, thermal cameras have become more accessible, as until recently they were primarily used for defense applications and were difficult to acquire due to their high cost and limited use [4].

LiDAR also operates relatively independently of visible light. Although highly effective in mapping environments and detecting objects, it may not always accurately distinguish between humans and other objects due to its significantly lower resolution when compared to traditional RGB cameras or even thermal cameras.

Since these modalities have complementary characteristics, the use of spatially and temporarily aligned images of the three modalities becomes efficient in the realm of human detection particularly in adverse conditions, when combined with a proper sensor fusion technique.

In terms of **contributions**, this paper presents a new multi-modality dataset for indoor human detection using an RGB camera, a thermal camera and a LiDAR. Furthermore, a YOLO network is trained for each individual modality (serving as baseline), as well as for the numerous combinations of data-representations across early-fusion in order to evaluate the performance of single and multi-modalities. Finally, a new tailored multi-modal representation learning strategy is proposed to improve the single and early-fusion models' performance.

II. Related Work

In this section a concise yet representative description of the related work is presented. YOLO [5], [6] (You Only Look Once) is a state-of-the-art real time object detection Convolutional Neural Network (CNN) that predicts bounding

[1]MID-3K dataset: https://kennedyk1.github.io/MID-3K/

boxes and class probabilities directly from full images in a single evaluation, making it extremely fast and capable of processing images in real-time at 45 frames-per-second. Different versions of YOLO have been developed, focusing on either faster or more precise results, with FastYOLO achieving double the mean Average-Precision (mAP) when compared to other real-time detectors, while processing images at an astonishing 155 frames-per-second.

Many works [3], [7], [8] showcase the benefits of sensor fusion for human detection applications. In this domain, relevant works combine data from RGB cameras, thermal cameras, and/or depth sensors, such as LiDARs, and use it to train CNNs (such as YOLO) or other vision-based detection algorithms. In order to merge the information from the available modalities, different fusion techniques are applied, from data level fusion (a.k.a. early fusion), feature level fusion to decision level fusion. Sensor fusion outperforming each modality individually, when trained with the same network, is the common result obtained by these papers for all types of sensor fusion. Other works [9]–[11] developed similar data fusion approaches and provided their own dataset comprising RGB and thermal images. A different approach for decision-level fusion was taken in [12] where RGB and depth (LIDAR-based) data were explored. The fusion weights in [12] are based on the lighting conditions of the RGB images thus, if the lighting conditions are optimal then the RGB detections have a higher weight in the final result; otherwise if the images are overexposed (too bright) or underexposed (too dark) the detections from the depth data have a bigger weight in the final result. This method presents a 'dynamic' approach as it changes the degree of confidence of each modality based on the current environment conditions.

III. Dataset, Modality Fusion and YOLO-models

A. Overview

In this section we present a new multi-modality dataset collected at the Polo II campus of the University of Coimbra, Portugal. The sensors were mounted on a robotic platform shown in Fig. 1. The dataset supporting this work, designated by MID-3K, includes four sensory-representations obtained from the three sensors (2×LiDAR, 1×RGB, 1×Thermal).

The dataset, the generation of the modalities by combining the (single) primary modalities, the methodology used to train models with four or more channels, as well as the results are described in the sequel.

B. MID-3K Dataset

To support the experiments the multi-modal dataset, MID-3K (Multimodal ISR Dataset with 3083 images) is used. The experimental setup for data collection was as follows:

- A mobile robot *Clearpath Jackal* model has been adapted to be equipped with a laptop and the sensors (see Fig.1).
- A laptop running *ROS Noetic Ninjemys* on *Ubuntu 20.04 LTS* with *32GB DDR5, Core i7-12700H 14C/20T and RTX3080Ti.*

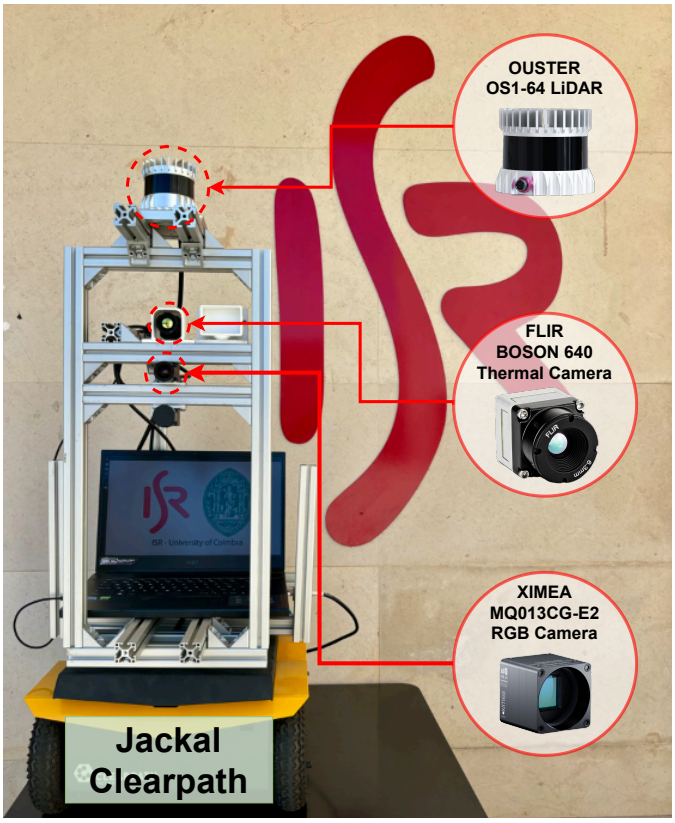

Fig. 1. The mobile robot Clearpath Jackal used to collect multi-modal data from three sensor technologies: (RGB) Ximea MQ013CG-E2, (Thermal) FLIR BOSON 640, Lens 50º 8.7mm, and (LiDAR) Ouster OS1-64.

- A *Ximea 1.3MP Colour Camera MQ013CG-E2*, which has a resolution of 1.3MP 1280x1024, recording at 40Hz and uses USB3.1 data interface.
- A *FLIR BOSON 640 LWIR (longwave infrared) Thermal Camera*, which has a resolution of 640x512, recording at 30Hz and uses USB-C interface.
- A mid-range digital LiDAR Sensor *Ouster OS1-64U*, with a vertical resolution of 64 channels and 2048 points horizontally at 10Hz, resulting in a total of 131072 points by 360º scan, connected with UDP over gigabit Ethernet.

Before collecting the data, the intrinsic and extrinsic parameters of the RGB and thermal cameras were computed, and the LiDAR was calibrated with the RGB camera using both a chessboard (see Fig.2) and *Computer Vision MATLAB toolbox*, so that the 3D-points of the point cloud can be projected onto an image plane with the same field of view as the cameras.

The dataset was collected over 5 days by the robot moving across floors 2, 3, and 4 of the Department of Electrical and Computer Engineering (DEEC) and floors 2 and 4 of the Department of Computer Engineering (DEI *i.e.*, a different building block) at the Polo II campus of the University of Coimbra, Portugal. The thermal and RGB images were rectified and aligned to match the field of view, resulting in a new smaller RGB image with the same dimensions as the thermal image. Therefore, the *RGB* and *thermal* modalities

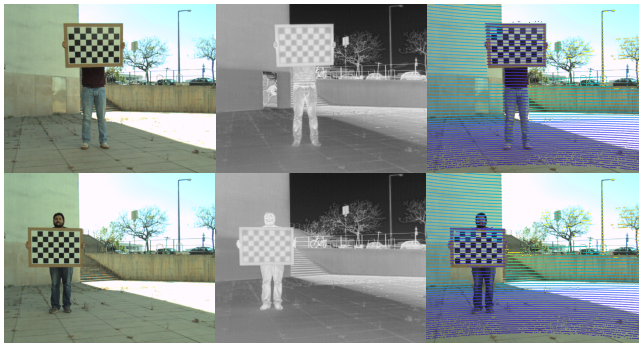

Fig. 2. Example images used for sensors calibration. The composite image shows the *RGB* (left) and *thermal* (center) images, and the projection of 3D-LiDAR points over the RGB image (right). The first two images are used for sensor calibrations, while the last one is used to verify if the LiDAR-camera calibration is properly aligned.

are ready to be directly used in a CNN however, regarding the LiDAR representation, it is necessary to perform appropriate processing steps to obtain *depth* and *intensity* representations based on the 3D-LiDAR point clouds.

To generate a dense (up-sampled) representation of *depth* and *intensity* from LiDAR data, a Bilateral Filter was employed with the Delaunay triangulation technique [13]. This technique has demonstrated significant effectiveness [14] in obtaining dense maps, nearing 100% density, crucial for applications requiring precise and detailed data acquisition.

Subsequently, the dataset was curated, temporally synchronized, aligned, and annotated (see Fig. 3). The available dataset contains 3,083 selected frames and includes *RGB*, *thermal*, *depth* and *intensity* map (both generated from LiDAR), totaling 12,332 image files. The *thermal* modality comprises 10,881 human/people annotations and the *RGB* contains 10,824 annotations. The *depth* and *intensity* modalities use the labels from the *RGB* modality, since the LiDAR is calibrated with the RGB camera.

In the repository, the metadata information is available, which includes all details for each scene, including the scene name (with timestamp in nanoseconds), the original *rosbag* filename, the capture date and time, the department where the data was collected, the floor, and the number of annotations for each modality. The metadata allows grouping the dataset by day, capture period (morning or afternoon), floor, or department (location) for future work, providing more flexibility in dividing the dataset into training and testing sets.

For experimentation, the dataset is split into 2 parts: training and test, as shown in Tab. I. The training set includes all scenes captured on days 2, 4, and 5, while the test set includes all scenes captured on days 1 and 3.

C. Early Fusion

Considering the main modalities available in the dataset - *thermal* (T), *RGB* (R), *intensity* (I) and *depth* (D) - new modalities/representations have been created through fusion (*i.e.*, combination) of them, generating combined modalities via stacking. The new modalities and their components are succinctly described in Table II.

TABLE I
MID-3K: DISTRIBUTION OF THE SETS BY DATE.

	Day	Date	Images	Thermal Annotations	RGB Annotations
MID-3K	#1	29-Apr	368 (11.9%)	1368 (12.6%)	1404 (13.0%)
	#2	07-May	332 (10.8%)	1075 (9.9%)	1072 (9.9%)
	#3	08-May	337 (10.9%)	1658 (15.2%)	1651 (15.3%)
	#4	09-May	1333 (43.2%)	4199 (38.6%)	4125 (38.1%)
	#5	16-May	713 (23.1%)	2581 (23.7%)	2572 (23.8%)
	-	**-**	**3083**	**10881**	**10824**
TEST	#1	29-Apr	368 (52.2%)	1368 (45.2%)	1404 (46.0%)
	#3	08-May	337 (47.8%)	1658 (54.8%)	1651 (54.0%)
	-	**-**	**705**	**3026**	**3055**
TRAIN	#2	07-May	332 (14.0%)	1075 (13.7%)	1072 (13.8%)
	#4	09-May	1333 (56.1%)	4199 (53.5%)	4125 (53.1%)
	#5	16-May	713 (30.0%)	2581 (32.9%)	2572 (33.1%)
	-	**-**	**2378**	**7855**	**7769**

TABLE II
COMBINED MODALITIES AND PRIMARY COMPONENTS.

Modality	Depth	Intensity	RGB	Thermal
D-I	✓	✓		
RGB-D	✓		✓	
RGB-I		✓	✓	
RGB-T			✓	✓
T-D	✓			✓
T-I		✓		✓
T-RD	✓		✓	✓
T-RID	✓	✓	✓	✓

D. YOLO-based models and baselines performance

For training each modality (single and combined), YOLOv5 [15] [16] small was employed as the baseline detector. Each model for each modality was trained using the same hyperparameters and on the same hardware for 50 epochs without pretrained weights. The hardware setup for the trainings phases included an *Intel Xeon E5-2680v4 14C/28T processor*, *32GB of DDR4 RAM*, and an *NVIDIA RTX3050 8GB GPU*.

The results for the single modalities are detailed in Table III, while the performance during training using *mAP50* as a metric is shown in Fig. 4. Among the tested single (primary) modalities, the *thermal* modality achieved the best performance across all the evaluated metrics as indicated in the table.

To train models with images containing up to 4 channels (the maximum supported by the dataloader), minor modifica-

TABLE III
SINGLE-MODALITIES RESULTS USING YOLOV5-SMALL. THE BEST RESULT IS HIGHLIGHTED IN BOLD.

Modality	Precision	Recall	mAP50	mAP50:95
Depth	0.592	0.404	0.465	0.206
Intensity	0.373	0.308	0.290	0.117
RGB	0.617	0.491	0.541	0.250
Thermal	**0.641**	**0.560**	**0.606**	**0.302**

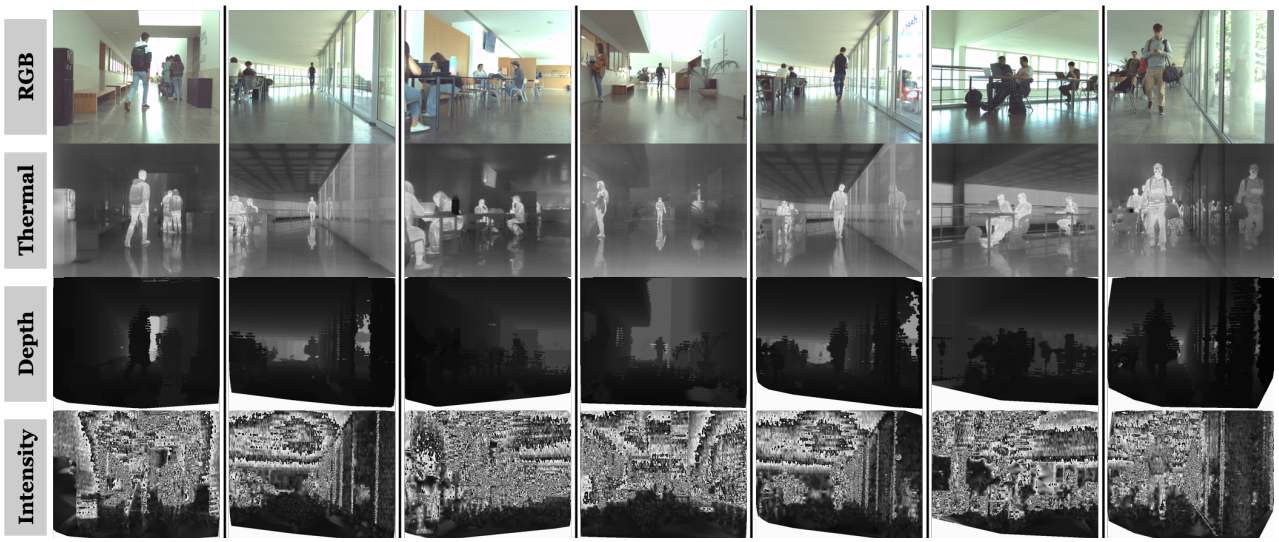

Fig. 3. Examples of the dataset frames showing RGB images, thermal images, depth maps, and intensity maps.

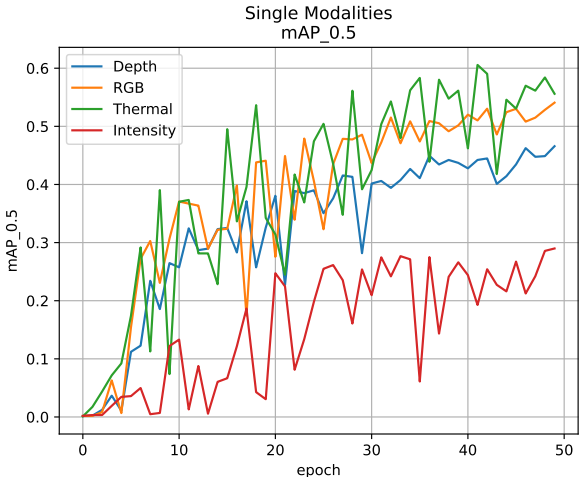

Fig. 4. The curves show that among all modalities the *thermal* and *RGB* had the best performance (after 30 epochs) followed by *depth*. The *LiDAR-intensity* modality achieved inferior performance.

TABLE IV
COMBINED MODALITIES RESULTS USING YOLOV5-SMALL. THE BEST MODEL RESULTS IS HIGHLIGHTED IN BOLD.

Modality	Precision	Recall	mAP50	mAP50:95
D-I	-	-	-	-
RGB-D	0.633	0.524	0.572	0.274
RGB-I	0.636	0.508	0.564	0.267
RGB-T	0.668	0.560	0.617	0.296
T-D	0.670	0.598	0.640	0.337
T-I	0.653	0.601	0.636	0.325
T-RD	0.670	0.591	0.637	0.332
T-RID	**0.686**	**0.580**	**0.643**	**0.338**

TABLE V
COMPARISON OF SINGLE, COMBINED AND COMBINED-RL MODALITIES USING YOLOV5-SMALL.

Modality	Loaded Weights	Precision	Recall	mAP50	mAP50:95
Thermal	No	0.641	0.560	0.606	0.302
T-RID	No	0.686	0.580	0.643	0.338
T-RID-RL	**Yes**	**0.760**	**0.679**	**0.746**	**0.443**

tions were made to the YOLOv5 framework to accommodate the additional input channels. The "scratch-low" hyperparameter profile has been used with a slight adjustment in the data augmentation section where HSV (Hue, Saturation and Value) adjustments were disabled since they only work in the RGB color-space. For all combined modalities that include the *thermal* component, the models are trained with the *thermal* labels because this modality achieved the best results among the individual modalities. In other cases *RGB* labels are used since this strategy resulted in better experimental performance compared to using *RGB* labels for all combined modalities. For the combined modalities, the results and training performance are presented in Table IV and Fig. 6. The *D-I* modality, which is obtained from LiDAR-sensor alone, showed very poor performance compared to the individual primary components. On the other hand, the *T-RID* modality which uses all the primary components achieved slightly better performance. This performance can be explained by the fusion of multiple sensors and their representations, providing the model with a more detailed perception of the environment. The training process for single and combined models took between 50 and 58 minutes.

E. Representation Learning pipeline

For this experiment, a UNET-like [17] architecture was integrated into the YOLOv5 architecture to capture the most relevant features from the input data (images with up to four channels) and provide the framework with an enhanced

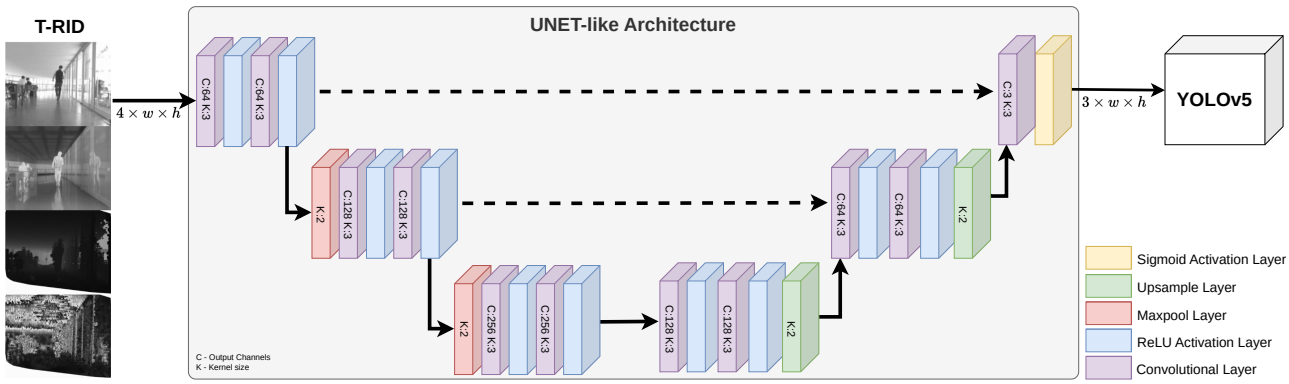

Fig. 5. The proposed YOLOv5-RL pipeline where a 4-channel image (*T-RID*) is fed into a UNET-like architecture. The architecture begins with an encoder that reduces the spatial resolution of the input images while increasing the number of channels, extracting important features from the image. Then, with the features extracted by the encoder, the decoder reconstructs the spatial resolution of the image, creating a 3-channel representation for YOLO.

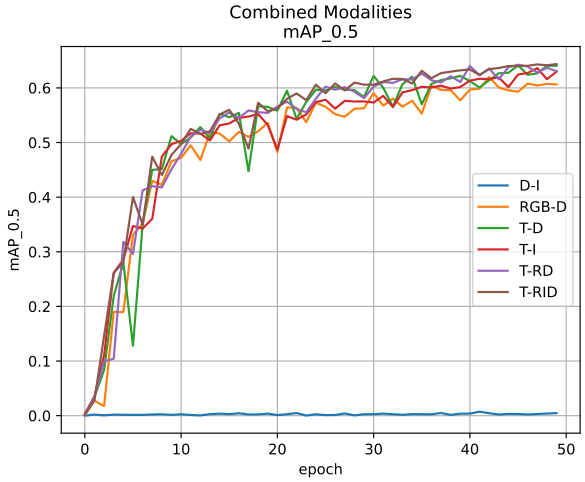

Fig. 6. Performance comparison in terms of mAP50, during training, using YOLOv5-small on combined modalities. In this plot, the modalities with the primary component provided by the thermal or RGB camera showed performances above 0.5. Only the *D-I* modality, which is single-sensor, had its performance significantly reduced compared to its primary components.

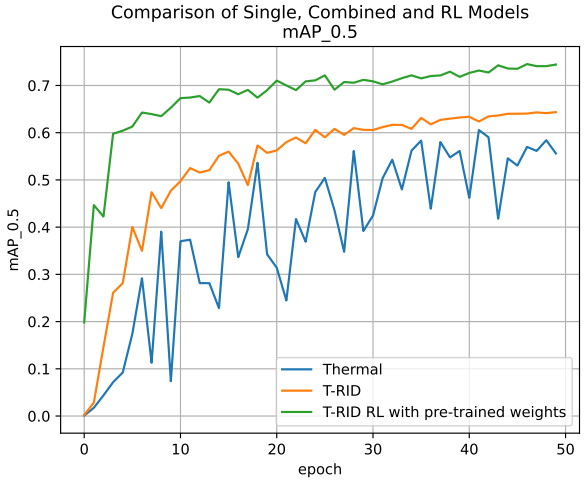

Fig. 7. Performance comparison in terms of mAP50, during training, using YOLOv5-small on Single and Combined modalities. It is possible to observe that the thermal modality showed greater performance fluctuations during training compared to the others. After the fusion of all modalities, creating the new *T-RID* modality, there was a significant improvement in performance using a slightly modified version of YOLO. The training of *T-RID* the multimodal representation-learning model, using pre-trained weights, achieved the best performance (curve in green).

representation of the most important features in a three-channel format.

The UNET-like model (see Fig. 5) is composed of three modules in the encoder and three modules in the decoder, with two shortcut connections after the first and second modules of the encoder connecting to the last two modules of the decoder. During the training of the main network (YOLO), the UNET is also trained, with the final loss being propagated through the entire network including both the UNET and YOLO. The weights of both architectures are adjusted via backpropagation and updated at each training iteration. This model is flexible because it enables the YOLOv5-RL to use pre-trained weights, contrary to the baseline that has been modified to accept inputs with more than three channels.

IV. Experiments and Results

In this section, the results of the baselines and techniques applied to the models implemented in this work are presented and discussed. Initially, the dataset was divided into training and test sets by grouping the sets accordingly to the dates/days they have been collected. This resulted in approximately 77.1% of the images in the training set and 22.9% in the test set, with a distribution of approximately 71.9% and 28.1% of the annotations, respectively.

From the images of the primary modalities (*depth*, *intensity*, *RGB*, and *thermal*), other modalities were created by a stacking method. For training the models, the YOLOv5 version from the official repository was used and some modifications were made to create two new YOLO-based frameworks: one

to allow the framework to accept images with up to four channels, and another to implement multimodal representation learning (designated by YOLOv5-RL).

For the training of the models the same hardware and the same hyperparameters are used. Among the models of the primary modalities, the *thermal* modality achieved the best performance. For the combined modalities, it was observed that data fusion (early fusion) improved the performance of the models. Except for the *D-I* modality, which had its performance significantly reduced, all other combined models outperformed the best model trained on primary component modality (analyzing the *mAP50*). Among the combined modality models the *T-RID* achieved the best performance in the *precision*, *mAP50*, and *mAP50*:95 metrics, making it the best candidate for testing in the framework with the implemented multimodal representation learning.

Since the *T-RID* modality achieved the best performance among the previous modalities using the modified YOLOv5 to support more channels, this modality was used to train a new model with YOLOv5 considering an initial stage of multimodal representation learning (YOLOv5-RL) under the same training conditions. The proposed approach was trained with the default pre-trained weights confirming that the new *T-RID i.e.*, incorporating a multimodal representation learning model, is capable of loading these weights, learning to interpret modalities with 4 or more channels, and representing them in the 3-channel format. As shown in Table V and Fig. 7, this framework achieved superior performance by reusing the weights trained on large datasets. The downside aspect of the training process using combined models with multimodal representation learning is that it took between 6 to 7 times longer compared to single modalities training.

Finally, the the multimodal representation-learning framework (when trained with the *T-RID* modality using pre-trained weights) showed improved performance in all metrics: with gains of 18%, 21%, 23%, and 46% in the *precision*, *recall*, *mAP50*, and *mAP50*:95 respectively, compared to the *thermal* modality. Compared to the *T-RID* modality (that only uses a common fusion method on the modalities) the gains were 10%, 17%, 16%, and 31%, respectively.

V. Conclusion

This work tackles the problem of humans/people detection using RGB and thermal cameras and LiDAR, as well as on the fusion of these sensors to achieve better performance. To support this, a new multi-sensor dataset is made available, consisting of *depth*, *intensity*, *RGB*, and *thermal* modalities. From the primary modalities, new modalities are created by combining them. In addition to exploring the YOLO detector to obtain baselines for single modalities, new models are trained using the combined modalities and as result the image-fusion approach achieved superior performance compared to single modalities. Finally, a multimodal representation learning approach is proposed to create a 3-channel representation from a multi-channel image, which enabled training with pre-trained weights and further improved model performance thus, obtaining better results than the combined modalities.

Acknowledgment

This work has been supported by the project GreenBotics (PTDC/EEI-ROB/2459/2021), funded by the Portuguese Foundation for Science and Technology (FCT). This work was partially supported by ISR-UC FCT grant UIDB/00048/2020 (DOI: 10.54499/UIDB/00048/2020).

References

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2004.

[2] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing LIDAR and images for pedestrian detection using convolutional neural networks," in *2016 IEEE ICRA*.

[3] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining RGB and dense LIDAR data," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[4] M. Vollmer, "Infrared thermal imaging," in *Computer Vision: A Reference Guide*. Springer, 2021, pp. 666–670.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE CVPR*, 2016.

[6] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proceedings of the IEEE CVPR*, July 2017.

[7] Y. Yue, C. Yang, J. Zhang, M. Wen, Z. Wu, H. Zhang, and D. Wang, "Day and night collaborative dynamic mapping in unstructured environment based on multimodal sensors," in *2020 IEEE ICRA*, 2020.

[8] J. Kim, J. Kim, and J. Cho, "An advanced object classification strategy using YOLO through camera and LiDAR sensor fusion," in *2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2019.

[9] V. Knyaz, "Multimodal data fusion for object recognition," in *Multimodal Sensing: Technologies and Applications*, E. Stella, Ed., vol. 11059, International Society for Optics and Photonics. SPIE, 2019.

[10] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-Thermal calibration, dataset and segmentation network," in *2020 IEEE ICRA*.

[11] E. Sousa, K. O. S. Mota, I. P. Gomes, L. Garrote, D. F. Wolf, and C. Premebida, "Late-fusion multimodal human detection based on RGB and thermal images for robotic perception," in *2023 European Conference on Mobile Robots (ECMR)*.

[12] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," in *2016 IEEE/RSJ IROS*, 2016.

[13] L. Garrote, J. Perdiz, L. A. da Silva Cruz, and U. J. Nunes, "Point cloud compression: Impact on object detection in outdoor contexts," *Sensors*, vol. 22, no. 15, 2022.

[14] C. Premebida, L. Garrote, A. Asvadi, A. P. Ribeiro, and U. Nunes, "High-resolution lidar-based depth mapping using bilateral filter," in *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)*.

[15] G. Jocher, "YOLOv5 by Ultralytics," 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[16] Ultralytics, "YOLOv5 - Ultralytics," https://docs.ultralytics.com/models/yolov5/.

[17] O. Ronneberger, P. Fischer, and T. Brox, "n advanced object classification," in *18th international Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.