

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/387348670>

Automatic Labeling for Thermal Imaging Datasets Generation

Conference Paper · November 2024

DOI: 10.1109/ROBOT61475.2024.10797435

CITATIONS

0

READS

7

2 authors, including:



[Maria Teresa Lazaro](#)

Instituto Tecnológico de Aragón

25 PUBLICATIONS 309 CITATIONS

SEE PROFILE

Automatic labeling for Thermal Imaging Datasets Generation

Daniel Cantón and María T. Lázaro
Instituto Tecnológico de Aragón (ITA), Zaragoza, Spain
Email: {dcanton, mtlazaro}@ita.es

Abstract—Low visibility scenarios present a significant challenge in the development of robust perception systems due to the difficulty in replicating real conditions and the lack of available data for training neural networks across a wide range of situations. This problem is even more pronounced when using less common sensors such as thermal (infrared) cameras, which operate in the non-visible spectrum and provide better performance in low light, dusty, or smoky conditions. This paper addresses the problem of dataset generation in such scenarios by proposing an automatic labeling method that leverages the capabilities of pre-trained networks on images captured in the visible spectrum using traditional cameras. We evaluate the proposed method by comparing the quality of automatically generated datasets with manually annotated datasets. Finally, we demonstrate the versatility of training a network in the non-visible spectrum and applying it to low visibility situations.

Index Terms—automatic labeling, neural networks, infrared camera, dataset generation

I. INTRODUCTION

With the advancement of robotic systems and their deployment in real-world scenarios, new challenges are emerging that have been minimally addressed in the literature. One significant challenge is operating in environments with low visibility conditions, such as those caused by dust, fog, and smoke, which are commonly found in construction sites, firefighting scenarios, and certain industrial settings. In these scenarios, the use of thermal imaging cameras is emerging as a viable technology to address the perception problems that robotic systems equipped with traditional cameras may not be able to solve.

Thermal imaging cameras, also known as infrared (IR) cameras, operate within the long-wave infrared spectrum, detecting the thermal or heat signature of objects. All everyday objects emit a certain level of invisible infrared radiation, which these cameras are designed to capture. This ability allows them to perform well in environments with dust, fog, or low-light conditions, where traditional (i.e., RGB) cameras often fail. By relying on infrared thermography, IR cameras can provide clear visibility and reliable images despite these common visual obstructions, making them essential for robotic applications in challenging conditions.

One of the key objectives of a perception system in these environments is to provide a comprehensive understanding of

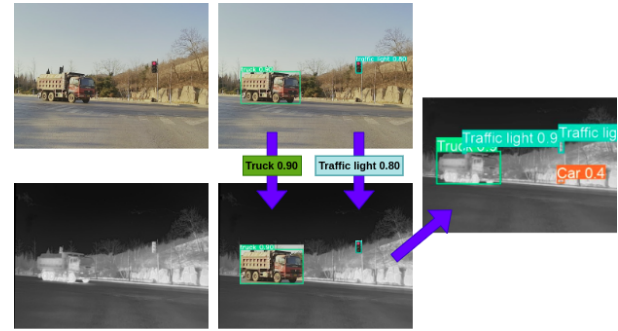


Fig. 1. Overview of our automatic labeling approach. Labels generated from visible spectrum images (top) are transferred to corresponding thermal images (bottom), which are then used to train new models.

the surroundings, which in a basic interpretation involves the detection and tracking of various entities, such as objects, other vehicles, and people. To achieve this, neural network-based techniques represent the state-of-the-art solutions.

Training neural networks for object detection tasks heavily relies on the availability of annotated datasets tailored to specific fields of interest. Creating such datasets is a labor-intensive and time-consuming process, but over the past decade, it resulted in an abundance of datasets using traditional images, which has significantly contributed to the development of robust and accurate object detection and tracking systems. Pre-trained networks on these datasets perform well in good visibility conditions, such as scenes without occlusions, dust, or fog. However, there still remains a notable scarcity of datasets for IR images in low visibility conditions and a variety of scenarios.

This paper addresses the significant challenge posed by the scarcity of datasets for IR cameras operating in low visibility environments. To tackle this issue, we propose an automatic labeling methodology that leverages the use of object detection networks pre-trained on RGB images to identify objects and transfer these labels to IR images. Specifically, we apply the YOLO [1] neural network to datasets containing paired RGB-IR images captured in visible RGB scenes, automating the labeling process for IR images. Figure 1 shows an example of this process. Subsequently, in our experiments, we analyze the quality and confidence of the generated dataset and demonstrate that retraining a network with these labeled

IR images allows us to effectively apply the trained model to IR images acquired in non-visible conditions.

II. RELATED WORK

Thermal imaging technology is becoming widely popular in applications demanding robust perception and image analysis systems [2]. For instance, in firefighting, thermal cameras can enhance visibility for firefighters navigating in smoky scenarios [3], improving their situational awareness. In agriculture, these cameras can assist in crop health monitoring and irrigation scheduling [4]. In the medical field, thermal imaging aids in disease diagnosis by detecting abnormal temperature patterns [5] and, in the automotive sector, provides advanced driver assistance systems (ADAS) with enhanced security and vehicle safety [6].

Most of these applications require the availability of training datasets. While numerous datasets exist for traditional cameras, such as COCO [7], ImageNet [8], and others, there are very few annotated datasets containing thermal images. Some exceptions can be found in the FLIR dataset [9], which includes annotations in IR images, and the M³FD [10] and LLVIP [11], which also include annotated paired RGB-IR images. However, the annotated paired RGB-IR images are often restricted to annotations for a few classes, as is the case of the LLVIP, which is labeled exclusively for people detection. Additionally, there are other databases that include non-labeled paired RGB-IR images such as the OSU Color-Thermal [12].

The creation of annotated datasets is crucial for training neural networks to develop advanced perception systems. Traditional approaches often rely on manually labeled data, which requires significant effort and time. Tools like X-AnyLabelling [13] offer a semi-automated solution by utilizing an AI inference engine to assist in the labeling process; however, user guidance is still required, as the user must actively select the objects to label. As an alternative to human intervention, automatic or unsupervised labeling approaches have been proposed to annotate datasets efficiently. Several studies have demonstrated the effectiveness of using synthetic data generated from simulators to automate the labeling process. For instance, in [14], Blender was used to model 3D objects resembling real-world scenes with high fidelity, enabling the creation of diverse and accurately labeled datasets. In the context of robotics, Gazebo was used in [15] to create a comprehensive dataset for mobile robot navigation, including LiDAR and stereo images as annotated perception data. The autonomous driving simulator CARLA has also been used in [16], which proposes a tool for generating synthetic datasets by leveraging CARLA's capability to provide ground truth semantic segmentation. In [17], it is shown that relying on modern graphics engines to generate a training dataset can achieve comparable performance to manually annotated data.

A different set of studies aiming to reduce human intervention and the annotation efforts falls under the category of transfer learning [18]. For example, in [19], the authors discuss the use of supervisory networks to transfer mid-level CNN

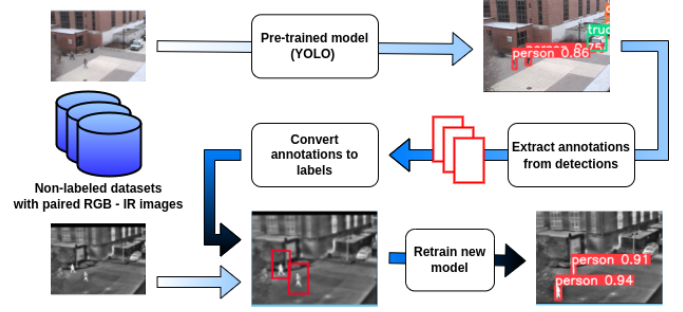


Fig. 2. Pipeline of the proposed automatic labeling approach. The process begins with object detection in RGB images using a pre-trained network. Detected labels are then transferred to corresponding IR images, followed by dataset creation and model training using the newly labeled thermal images.

layers between different modalities, specifically from the RGB to the depth domain. This approach leverages the extensive availability of datasets recorded with RGB-D cameras and demonstrates that faster training models can be achieved while enhancing performance across modalities.

In the realm of thermal imaging, automatic labeling techniques have also been explored to generate annotated thermal images datasets, essential for developing robust perception systems in low visibility conditions. For example, in [20], the authors propose a superimposition method between gray-scale and low resolution IR images to automatically annotate an IR dataset for people detection. In [21], an approach similar to ours was proposed, where RGB image labels are used for labeling thermal images. Additionally, their approach combines detections in the RGB images with bounding boxes generated from 3D LiDAR data, enhancing the accuracy of the transferred labels. We propose a similar methodology to train a network using different non-labeled datasets, which is later applied to images acquired in low visibility environments, where training in the visible spectrum would not have been feasible.

III. METHODOLOGY

Our approach offers a solution to the challenge of creating annotated datasets for low visibility conditions. The overview of our methodology is shown in Fig. 2. In summary, we leverage pre-trained models on RGB images to identify objects in paired RGB-IR datasets and automatically transfer these labels to the IR images, allowing us to subsequently train new models on the IR data. The method is structured into several key steps, described in the rest of this section.

A. RGB and IR image collection

The first step in our methodology involves obtaining registered pairs of RGB and IR images. This registration ensures that there is a pixel-to-pixel correspondence between the two image types, which is crucial for accurate label transfer. Some commonly used sensors in robotics, such as the Orbbec Astra Pro, Intel RealSense, and FLIR thermal cameras, inherently provide registered RGB and infrared streams from the same sensor, simplifying the process.

In cases where registration is not readily available but the images have been captured simultaneously, it is possible to synchronize the images and calculate the homography between them. This homography can then be applied to align the IR images with the RGB images, ensuring proper registration.

B. Running YOLO for object detection

In low visibility conditions, such as those with low light, dust, or smoke, the direct use of paired RGB-IR images is not feasible due to both the lack of data and potential detection failures in the visible spectrum. Furthermore, as we will show in our experimental section, while IR images may resemble grayscale images, pre-trained models often perform inadequately in the IR spectrum. Therefore, our methodology relies on RGB-IR pairs captured under clear conditions to train our models effectively. In this process, we use a supervisory network to identify and transfer detections of objects of interest from RGB to IR images. This network can be a pre-trained YOLO model or any other network tailored to specific target domains, such as medical imaging or construction.

The first step to transfer detections from RGB images to IR images involves obtaining the bounding boxes of the detected objects from the RGB images using the supervisory network. Pre-trained YOLO models are capable of detecting a wide variety of objects, as they are trained on the COCO dataset. By leveraging these pre-trained models we can extract bounding boxes from the RGB images, which are subsequently used to generate labels for training a YOLO model on IR images. This methodology ensures that our approach is adaptable to various domains by using the appropriate pre-trained detection models.

C. From detection to labeling

Once all objects in the initial set of RGB images are detected by YOLO, the procedure described in this paper involves converting these detections into the format required by YOLO for training custom datasets. Specifically, this format requires the creation of YOLO .txt label files for each image containing detections, adhering to the following structure:

- One row per detection
- Each row follows the format:

```
class x_center y_center width height
```
- The bounding boxes must be in normalized xywh format (from 0 to 1).
- Class numbers are zero-indexed (starting from 0).

Furthermore, each bounding box has an associated object confidence value. In our methodology, we perform filtering to convert only those detections with a confidence value greater than a specified threshold into labels. This filtering step helps in obtaining a more robust dataset for training.

D. New model training

Once the IR images are annotated using the automatic labeling methodology described earlier, the next step involves training a YOLO model tailored for object detection in IR images. The dataset is split into training and validation sets,

TABLE I
GROUND-TRUTH LABELS IN M³FD [10] DATASET

Category	# of GT labels
People	11477
Car	18296
Bus	700
Motorcycle	521
Traffic-Light	2405
Truck	1008

typically with an 80-20 ratio for training and validation, respectively. This ensures the model is well-prepared for robust performance in real-world IR detection scenarios

IV. EXPERIMENTS

We conducted two sets of experiments to evaluate the effectiveness of our automatic labeling method, using YOLOv8 as the core detection model. These experiments allow us to validate both the quality of the automatically generated datasets and the applicability and robustness of the trained models in low visibility conditions.

A. Evaluation of the quality of the generated datasets

The first set of experiments is designed to assess the quality of the datasets generated by our method. Intuitively, we can consider the automatically generated dataset to be of high quality based on the number of detections a network trained with this dataset can achieve. Then, we can compare the results with respect to a network trained on a manually annotated dataset. To this end, we used the M³FD [10] dataset, which contains paired RGB-IR images with ground-truth labels provided by human annotators. It contains a total of 4200 synchronized image pairs, registered via homography matrix, with a total of 34407 labels along 6 categories: {*People*, *Car*, *Bus*, *Motorcycle*, *Traffic-Light*, *Truck*}. Table I summarizes the dataset information.

As explained in the previous section, the first step involves extracting the labels that will be used to annotate the IR images by running YOLO on the RGB images. At this step, we aim to evaluate the quality and precision of the YOLO pre-trained model in detecting objects belonging to the selected categories. This allows us to generalize the potential label transfer rate in cases where annotated datasets are not available. Note that when performing automatic labeling, both true and false positives will be transferred from the RGB to the IR images. With the availability of ground-truth data, we can evaluate the rate of true positive transfers over the total possible detections (i.e., recall). We consider true positive detections to be those with an Intersection over Union (IoU) value of at least 0.5. We then perform two types of evaluations: one where no confidence filtering is applied, thus transferring all labels, and another where only detections with a confidence threshold of 0.5 or higher are transferred. Table II(a) summarizes YOLO performance metrics on the RGB images of the M³FD dataset, establishing a baseline for the rest of the evaluation in this section. Furthermore, we observed that YOLO is able to

TABLE II
YOLO EVALUATION IN M³FD [10] DATASET

Category	No confidence threshold				Confidence threshold = 0.5			
	# Detections	# TP	Precision	Recall	# Detections	# TP	Precision	Recall
People	4553	3608	0.7924	0.3144	2837	2655	0.9358	0.2313
Car	14194	10604	0.7471	0.5796	9174	8589	0.9362	0.4694
Bus	719	398	0.5535	0.5686	536	367	0.6847	0.5243
Motorcycle	158	129	0.8165	0.2476	82	76	0.9268	0.1459
Traffic-Light	1209	653	0.5401	0.2715	535	390	0.7290	0.1622
Truck	1380	370	0.2681	0.3671	758	264	0.3483	0.2619

(a) Performance metrics on RGB images

Category	No confidence threshold				Confidence threshold = 0.5			
	# Detections	# TP	Precision	Recall	# Detections	# TP	Precision	Recall
People	5511	4578	0.8307	0.3989	3777	3597	0.9523	0.3134
Car	10520	8353	0.7940	0.4565	6524	6169	0.9456	0.3372
Bus	474	193	0.4072	0.2757	236	144	0.6102	0.2057
Motorcycle	33	26	0.7879	0.0499	6	5	0.8333	0.0096
Traffic-Light	177	84	0.4746	0.0349	42	24	0.5714	0.0099
Truck	750	147	0.1960	0.1458	261	57	0.2184	0.0565

(b) Performance metrics on IR images

provide detections on the IR images as well, which are also 3-channel images. Therefore, we also evaluate performance metrics on IR images. These values are summarized in Table II(b). Interestingly, YOLO detects more people in the IR images, which we believe is due to people being highly salient features in this type of imagery. However, it performs worse for other classes. This observation justifies and motivates the use of RGB images for the label transfer method proposed in this paper.

Based on the previous criteria regarding the different confidence thresholds used for label transfer, we created two datasets for the IR images, with which we retrained YOLO to obtain two neural networks. Furthermore, we retrained YOLO using the ground-truth annotations for comparison. Figure 3 shows the results of these training sessions. A maximum of 300 epochs was considered, with training stability achieved after approximately 150 epochs.

According to the metrics obtained from the training sessions, as expected, when no confidence threshold is applied to transfer labels to the dataset, more labels—both good and bad—are considered, resulting in better recall but lower precision. Conversely, by being more restrictive in label transfer, greater precision is achieved. Overall, observing the precision-recall curves, we obtain a slightly better mAP@0.5 when applying a 0.5 confidence threshold on the transferred labels, being closer to the results obtained with training using ground-truth labels, which, as expected, achieve the best performance. In general terms, we can say that the proposed automatic labeling method achieves a performance similar to the manual method.

B. Application in low visibility environments

In the following set of experiments we want to demonstrate the applicability and the utility of our method to generate

large-scale datasets from unlabeled data. To this end, we trained a YOLO network on IR images using datasets containing paired RGB-IR images, employing our method to transfer detections from RGB to annotations in the non-visible spectrum. Specifically, we used the following datasets from the OTCBVS Benchmark Dataset Collection: OSU Color-Thermal Database [12], LITIV-VAP [22], Bilodeau et al. dataset [23] and M³FD [10]. Each of those datasets are obtained with different sensor setups and environments and allows us to train a network over more than 31k images. Following our approach, we automatically labeled the dataset using a confidence threshold of 0.5 for label transfer. Figure 5 shows examples of automatically labeled IR images. As explained, we used this set of annotated images to retrain YOLO, considering the previous set of categories. Figure 4 shows the training results of this network.

Finally, we aim to validate the use of the previously trained network on IR images obtained in low visibility conditions, where the pre-trained YOLO model may fail to perform detections on their corresponding RGB images. Figure 6 shows examples of these detections, demonstrating the applicability and potential use of our proposed method.

V. CONCLUSION

In this work, we have proposed an approach to automate the thermal imagery labeling process, leveraging pre-trained RGB models and the abundance of existing datasets with paired RGB-IR images. Throughout the experiments, we have analyzed the quality of the generated datasets and demonstrated the practicality of the method for generating large-scale datasets in thermal imagery.

Our future work will focus on applying this methodology to train a neural network capable of detecting entities within

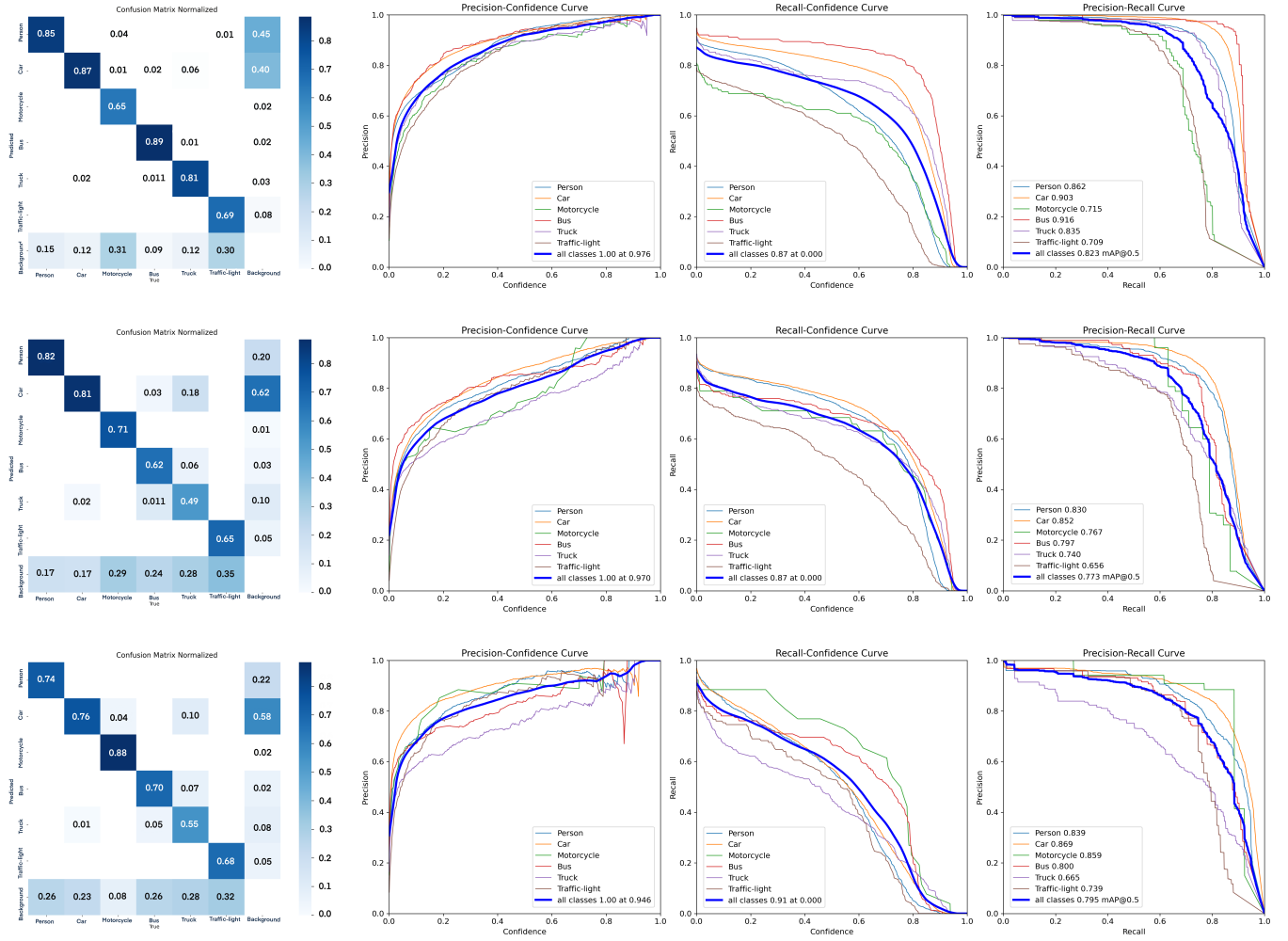


Fig. 3. Results from re-trained YOLO models using the IR images: Normalized confusion matrix, precision, recall and precision-recall curves. Top row: Training results using ground-truth labels. Middle row: Training results with no confidence threshold applied. Bottom row: Training results with a confidence threshold of 0.5 applied.

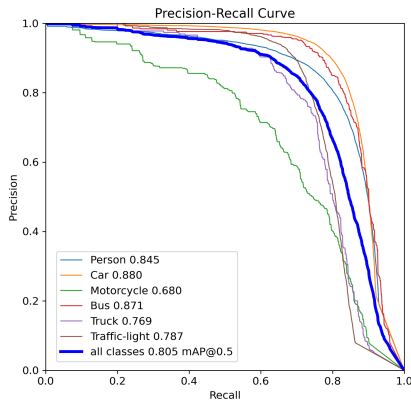


Fig. 4. Performance Metrics: Precision-Recall (PR) curve for the newly trained model.

the construction domain. This will involve training a specific supervisory network to detect workers, trucks, machinery,

cranes, and signs in construction environments using RGB images. We will then employ the proposed label transfer method to annotate images acquired in the same domain using a thermal camera. We will apply the trained network to develop a robust perception system capable to work in a post-explosion tunneling scenario, where smoke and dust make RGB cameras an impractical option.

In the long term, we plan to explore the use of multi-modal datasets, specifically RGB and LiDAR, to extend this approach to low visibility domains. We aim to apply the method to IR and RADAR data, particularly in dusty environments where LiDAR may fail. This extension will allow us to develop robust perception systems capable of operating in challenging low visibility conditions, further enhancing the utility and adaptability of our approach.

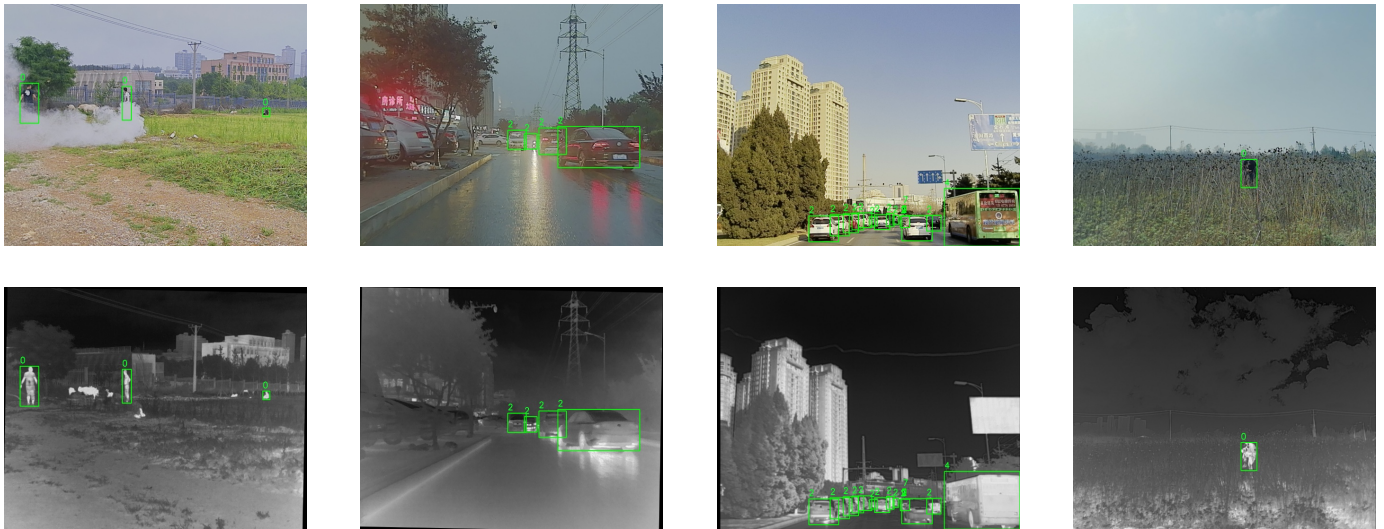


Fig. 5. Automatically labeled IR images from RGB detections.



Fig. 6. Detections on low visibility conditions. First and second rows: RGB and corresponding IR images. Third row: YOLO detections on the RGB images. Notice missed and wrong detections in the visible range. Forth row: Detections on the IR images using the trained network over the automatically labeled dataset.

REFERENCES

- [1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [2] A. N. Wilson, K. A. Gupta, B. H. Koduru, A. Kumar, A. Jha, and L. R. Cenkeramaddi, "Recent advances in thermal imaging and its applications using machine learning: A review," *IEEE Sensors Journal*, vol. 23, no. 4, pp. 3395–3407, 2023.
- [3] Y. Abdelrahman, P. Knierim, P. W. Wozniak, N. Henze, and A. Schmidt, "See through the fire: evaluating the augmentation of visual perception of firefighters using depth and thermal cameras," in *Proc. of the 2017 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing and Proc. of the 2017 ACM Int. Symposium on Wearable Computers*. New York, NY, USA: Association for Computing Machinery, 2017, p. 693–696.
- [4] R. Ishimwe, K. Abutaleb, and F. Ahmed, "Applications of thermal imaging in agriculture - a review," *Advances in Remote Sensing*, vol. 3, pp. 128–140, 2014.
- [5] S. Shaikh, N. Akhter, and R. Manza, "Current trends in the application of thermal imaging in medical condition analysis," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, pp. 2708–2712, 2019.
- [6] M. A. Farooq, W. Shariff, D. O'callaghan, A. Merla, and P. Corcoran, "On the role of thermal imaging in automotive applications: A critical review," *IEEE Access*, vol. 11, pp. 25 152–25 173, 2023.
- [7] 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 *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.
- [8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [9] "Teledyne FLIR ADAS dataset," <https://www.flir.eu/oem/adas/adas-dataset-form/>.
- [10] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811.
- [11] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvp: A visible-infrared paired dataset for low-light vision," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3496–3504, 2021.
- [12] J. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, pp. 162–182, 2007.
- [13] W. Wang, "Advanced auto labeling solution with added features," <https://github.com/CVHub520/X-AnyLabeling>, CVHub, 2023.
- [14] A. Y. Barrera-Animas and J. M. Davila Delgado, "Generating real-world-like labelled synthetic datasets for construction site applications," *Automation in Construction*, vol. 151, p. 104850, 2023.
- [15] M. Sánchez, J. Morales, J. L. Martínez, J. J. Fernández-Lozano, and A. García-Cerezo, "Automatically annotated dataset of a ground mobile robot in natural environments via gazebo simulations," *Sensors*, vol. 22, no. 15, 2022.
- [16] Z. Benčević, R. Grbić, B. Jelić, and M. Vranješ, "Tool for automatic labeling of objects in images obtained from carla autonomous driving simulator," in *2023 Zooming Innovation in Consumer Technologies Conference (ZINC)*, 2023, pp. 119–124.
- [17] M. D. Cicco, C. Potena, G. Grisetti, and A. Pretto, "Automatic model based dataset generation for fast and accurate crop and weeds detection," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017.
- [18] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [19] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2827–2836.
- [20] A. Avi, M. Zuccatti, M. Nardello, N. Conci, and D. Brunelli, "Infrared dataset generation for people detection through superimposition of different camera sensors," in *2021 IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW)*, 2021, pp. 307–316.
- [21] A. Ligocki, A. Jelinek, L. Zalud, and E. Rahtu, "Fully automated dcnn-based thermal images annotation using neural network pretrained on rgb data," *Sensors*, vol. 21, no. 4, 2021.
- [22] S.-C. P.-L., G.-A. Bilodeau, and R. Bergevin, "Mutual foreground segmentation with multispectral stereo pairs," in *Int. Conf. on Computer Vision Workshops (ICCV Workshops)*, Venice, Italy, 2017.
- [23] G.-A. Bilodeau, A. Torabi, P.-L. St-Charles, and D. Riahi, "Thermal-visible registration of human silhouettes: a similarity measure performance evaluation," *Infrared Physics & Technology*, vol. 64, pp. 79–86, 2014.