Multiple Object Detection and Tracking in the Thermal Spectrum

Wassim A. El Ahmar
University of Ottawa
Ottawa, Ontario, Canada
welahmar@uottawa.ca

Dhanvin Kolhatkar
Sensor Cortek Inc
Ottawa, Ontario, Canada
dhanvin@sensorcortek.ai

Farzan Erlik Nowruzi
University of Ottawa
Ottawa, Ontario, Canada
fnowr010@uottawa.ca

Hamzah AlGhamdi
University of Ottawa
Ottawa, Ontario, Canada
halgh091@uottawa.ca

Jonathan Hou
Pleora Technologies
Kanata, Ontario, Canada
jonathan.hou@pleora.com

Robert Laganiere
University of Ottawa
Ottawa, Ontario, Canada
laganier@eecs.uottawa.ca

Abstract

Multiple Object Tracking (MOT) is an integral part of machine vision research. Most tracking-by-detection based MOT solutions utilize video streams from RGB cameras for their operation. However, for real-world applications, it is necessary to utilize sensors that operate in different spectrums to accommodate for varying lighting conditions. Since object detection is the first step of the tracking pipeline in tracking-by-detection approaches, we compare the performance of state-of-the-art object detectors when trained on color images to their performance when trained on thermal images. We introduce a new dataset for multiple object tracking with thermal images and corresponding RGB images and show that state-of-the-art trackers perform better on thermal images, especially in poor lighting conditions. Finally, we propose the use of a dynamic cut-off threshold for tracking-by-detection approaches that factors the size of a predicted box to enhance the tracker association. Our dataset and source code are publicly available at https://github.com/wassimea/thermalMOT

1. Introduction

Multiple object tracking is an important task in machine vision where the goal is to assign identities to different objects present in a video sequence, and effectively maintain the unique identity of objects across consecutive frames. With several applications including self-driving cars, human computer interaction, and virtual reality, MOT has been a popular area of research in the computer vision domain.

1.1. Multiple Object Tracking

Most state of the art MOT methods [1, 5, 21, 25] utilize the tracking-by-detection paradigm, which is a two-stage process. In the first stage, a standalone detector is run on the video sequence to predict the location and class of objects present in the frames. In the second stage, a tracker processes these detections to conduct association: assigning a unique ID to detections of the same object across consecutive frames. While the performance of an object detector is usually measured by how accurate it can localize and classify objects in a single frame, the performance of a multiple object tracker also factors how well it can correctly re-identify an object across consecutive frames of a video sequence.

While some approaches have been developed that perform one shot, end-to-end tracking [13, 23], such methods were still unable to surpass the performance of two-stage tracking methods. Zhang et al. [26] performed an empirical study that concluded that the tasks of object detection and object tracking often compete with each other during training, causing one-shot trackers to be less accurate than two-stage trackers.

1.2. Perception Systems and Thermal Sensors

Deep learning helped achieve significant breakthroughs in machine vision tasks (advanced perception modules, intelligent monitoring systems, and autonomous vehicles, to name just a few) [17]. Such systems rely on the fusion of information from multiple types of sensors (lidar, radar, RGB cameras, depth sensors, thermal sensors, etc.) to get a better perception of the environment. This information is processed by an artificial intelligence module to perform advanced analysis and make critical decisions. Accurate multiple object detection and tracking is an essential task as it allows the localization of objects of interest and the prediction of the trajectory of moving objects.

The use of thermal sensors in machine vision is becoming more popular [6, 12] as they offer a powerful perception of the thermal identities of objects in a scene. They are also suitable for outdoor applications as thermal sensors operate normally at night and are not significantly affected by poor weather conditions [20].

1.3. Contribution

In this paper, we study the feasibility of using thermal sensors to conduct accurate multiple object detection and tracking. The main contributions of our work are summarized below:

- We introduce a new dataset for multiple object tracking with images and ground truth annotations for RGB and corresponding thermal images.
- We compare the performance of two state-of-the-art object detectors (TOOD [9], VFNET [24]) when trained on thermal images to when they are trained on RGB images of the Teledyne FLIR Thermal Dataset for Advanced Driver-Assistance Systems [1].
- We study the efficacy of applying transfer learning of weights trained on a RGB dataset when training an object detector on thermal images.
- We develop a tracking-by-detection MOT method based on the current state-of-the-art approaches that operates on thermal images, and enhance the data association of the tracker by applying a dynamic cut off score for detections based on the predicted box area.

2. Literature Review

In this section, we provide an overview of the existing methods and approaches that we build upon in our work.

2.1. Object Detection

2.1.1 Task-aligned One-stage Object Detection (TOOD)

Feng et al. introduced the Task-aligned One-stage Object Detection (TOOD) [9] which strengthens the link between the localization and the classification tasks of object detection. This is accomplished by introducing ”Task-Alignment”: taking the network’s outputs for each task and passing them to a network head that modifies the score and the location predictions to align their optimal anchors. Their Task alignment learning (TAL) also pushes the network to predict better aligned bounding boxes. TOOD achieved an AP of 51.1 on the MS-COCO dataset.

2.1.2 VarifocalNet (VFNET)

Zhang et al. designed VFNet using Varifocal Loss [24]. This loss is meant to maximize the IoU-aware classification score (IACS) that takes into account both the classification and the location of a prediction. The Varifocal loss also modifies focal loss by weighing positive examples more heavily than negative ones. Additionally, VFNet uses a new nine-point deformable convolution representation for bounding boxes and a network head to refine the network’s box predictions by learning an additional offset to their locations.

2.2. Tracking By Detection

The tracking-by-detection paradigm is more suitable for real-world applications, where different detectors could be used as a first step in the tracking pipeline, and the training data does not necessarily need to contain tracking ground truth labels. In addition, the rapid breakthrough in deep learning has led to the emergence of faster and more accurate detectors [9, 14, 24]. This has led to more research in MOT utilizing state-of-the-art object detection models [3, 22].

Since object detectors are not perfect, and there will always be cases where the detector predicts false positive boxes or misses true detections (false negatives), state-of-the-art MOT approaches often eliminate predicted boxes with a low confidence by setting a cut-off threshold for detector confidence. However, this inevitably leads to some true detections being ignored, especially in cases where there is occlusion.

Of the systems utilizing the tracking-by-detection paradigm, ByteTrack [25] has achieved state-of-the-art performance on the test data of the benchmark MOT17 dataset [16] with an MOTA of 80.3%. ByteTrack utilizes YOLOX [10] to generate detections. Instead of ignoring detection boxes with low confidence, ByteTrack separates the predicted boxes into a set of high-score detection boxes, and a set of low-score detection boxes. The algorithm first predicts the locations of the tracklets in the next frame using a Kalman filter, then matches the tracklets with the high-score detection boxes by computing the IOU between the high-score detection boxes and the predicted tracklet location. Tracklets that remain unmatched through this first association are then matched with the low-score detection boxes through a second association step. At the end of this two-step association process, tracklets that remain unmatched are deemed to be lost, new tracklets are created from the unmatched high-score detection boxes, and the remaining low-score

[1] https://www.flir.ca/oem/adas/adas-dataset-form/

(a) Morning sample RGB image.

(b) Afternoon sample RGB image.

(c) Night sample RGB image.

(d) Morning sample Thermal image.

(e) Afternoon sample Thermal image.

(f) Night sample Thermal image.

Figure 1. Samples from the three testing sequences of our dataset taken during different times of the day.

detection boxes are ignored.

3. Proposed Approach and Experiments

In this section, we elaborate on the experiments we have conducted in the color and thermal domain on object detection, multiple object tracking, and our proposed dynamic confidence thresholding for MOT.

3.1. Datasets

3.1.1 City Scene RGB-Thermal MOT Dataset

For the purpose of comparing tracking methods on visible and infrared images, we collected and manually annotated a dataset using a FLIR infrared camera and a visible-light camera, both at a framerate of 10Hz. One in every 2 frames was annotated, resulting in an effective framerate of 5Hz for the dataset. During data collection, the cameras were static and aimed at a city intersection. Cars and pedestrians were annotated up to a distance of 300m and 100m respectively.

The dataset is composed of 15 sequences collected over the course of a day, for a total of 1,997 annotated frames. These sequences are divided into a training set of 12 sequences (1,591 frames) and a testing set of 3 sequences (406 frames). Samples from the three testing sequences are given in Figure 1.

The annotated frames contains 267 unique car instances, 25,985 car bounding boxes, 145 unique pedestrian instances, and 7,822 pedestrian bounding boxes.

3.1.2 FLIR ADAS dataset

In our experiments, we study the performance of state-of-the-art object detectors when trained on color images compared to when trained on thermal images. To that end, we use the FLIR Thermal Dataset for Advanced Driver-Assistance Systems, which is one of the largest and most comprehensive thermal datasets that also provides color images corresponding to the thermal images. It is composed of city scenes captured using a thermal sensor and an RGB camera installed on top of a car. It is manually annotated for 15 classes (person, car, bike, etc.). There is a total of 11,886 training images and 3,749 testing images.

3.2. Object Detection Experiments

The object detector is an integral part of tracking-by-detection approaches. The performance of the tracker is significantly influenced by the performance of the object detector. In our experiments on object detection, we address the following two matters:

- We study the efficacy of applying transfer learning of weights trained on the Imagenet dataset [7] (color images) to train an object detector on thermal images.
- We compare the performance of object detectors when trained on thermal images against when they are trained on color images. As both the FLIR ADAS dataset and our City Scene RGB-Thermal

MOT dataset contain images taken at night, this experiment is important to examine the efficacy of thermal images under poor lighting conditions.

Table 1 shows a summary of the object detection experiments conducted. We use the MMDetection toolbox [4] for all object detection training experiments we perform.

We use Resnet50 [11] as the backbone of the detectors in all experiments. The models are trained for 6 epochs with a batch size of 8, with an initial learning rate of 0.01. We use focal loss [15] for box classification. We apply data augmentation (resizing and flipping) to enrich the dataset.

3.3. MOT Experiments

For our experiments on MOT, we use the state-of-the-art ByteTrack [25] approach. ByteTrack's simple design that conducts data association based on motion similarity make it ideal for our experiments, as the tracker does not require any domain-specific training. We fine tune the trained object detectors on the tracking dataset we collected as follows:

- Fine tune TOOD and VFNET on the thermal images of our City Scene RGB-Thermal MOT dataset with weights initialized from the trained detectors on the ADAS Thermal dataset.

- Fine tune TOOD and VFNET on the RGB images of our City Scene RGB-Thermal MOT dataset with weights initialized from the trained detectors on the ADAS RGB dataset.

3.4. Dynamic Confidence Cut-Off (DCC)

In the original implementation of ByteTrack, the authors set a minimum detection threshold $\mathcal{T}_{min}$ (0.1) and a threshold for high score detection boxes $\mathcal{T}_{high}$ (0.5). Bounding boxes that have a confidence below $\mathcal{T}_{min}$ are ignored, boxes that have a confidence between $\mathcal{T}_{min}$ and $\mathcal{T}_{high}$ are treated as low score detection boxes, and boxes that have a confidence higher than $\mathcal{T}_{high}$ are treated as high score detection boxes as explained in Section 2.2.

A drawback of this implementation is that it treats all objects of all sizes in the same manner strictly based on which range the confidence value falls in. This inevitably leads to false and missed detections as it does not take into account several factors like very small detections that are far from the camera, significant overlap and occlusion, especially as the size of the objects gets smaller as they move away from the camera (as is the case in some sequences in our City Scene RGB-Thermal MOT dataset, where the size of a tracked car for example becomes smaller as it moves further from the camera).

To address this issue, we propose the use of a dynamic confidence cut-off score for $\mathcal{T}_{high}$ in the implementation of ByteTrack inspired by the work conducted by Stalder et al. [19]. Our results show that the dynamic confidence cut-off significantly improves the performance of the trackers, especially for objects with a smaller area. We elaborate more on the findings in Section 4.2.

4. Results

In this section, we elaborate and analyze the results we achieved on object detection and MOT.

4.1. Precision Recall of Object Detectors

To study the performance of TOOD and VFNET on the thermal and RGB variants of the ADAS dataset, we plot the precision-recall curves of all the models from Table 1. The results are given in Figure 2 for TOOD and Figure 3 for VFNET.

From the analysis of the precision-recall curve of TOOD, we see that the detectors trained on thermal images perform significantly better than the ones trained on RGB images. This could be attributed to the fact that the ADAS dataset contains data captured at night, where the RGB images would contain little to no features of the objects of interest present in the frames. This explains why the maximum recall achieved by the detectors trained on RGB images is 79%, while the maximum recall achieved by the detectors trained on thermal images is 97%. This shows the superiority of detectors trained on thermal images, especially under poor lighting conditions. The ADAS dataset does not split the frames captured during the day from those captured during the night, so we were unable to conduct experiments to show the performance of the detectors exclusively on the frames taken at night. However, in the dataset that we collected, we split the sequences taken during the day from those taken at night, allowing us to compare the performance of the trackers in both settings.

The results also show that using transfer learning when training the detectors improves the overall performance of the detector. However, the results show that the effect of transfer learning was more significant on the detectors trained on the RGB images of the ADAS dataset. This could be explained by the fact that pretrained weights are also a result of training RGB images (the Imagenet dataset [7]). While transfer learning did improve the overall performance of the detector trained on thermal images, the improvement was not as significant. This shows that transfer learning is most effective when the initial weights and the dataset on which the detector is being trained on both belong to the same spectrum.

Experiment #	Weight Initializatoin	Trained / Tested on	Object Detector
1	Imagenet	ADAS Thermal Train/Test Sets	TOOD
2	Random	ADAS Thermal Train/Test Sets	TOOD
3	Imagenet	ADAS RGB Train/Test Sets	TOOD
4	Random	ADAS RGB Train/Test Sets	TOOD
5	Imagenet	ADAS Thermal Train/Test Sets	VFNET
6	Random	ADAS Thermal Train/Test Sets	VFNET
7	Imagenet	ADAS RGB Train/Test Sets	VFNET
8	Random	ADAS RGB Train/Test Sets	VFNET

Table 1. Object detection experiments conducted

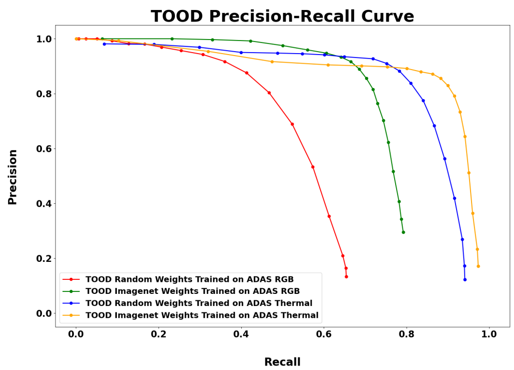

Figure 2. TOOD Precision-Recall curve. A detection is considered a true positive if it has at least a 0.5 IOU with a ground truth box.

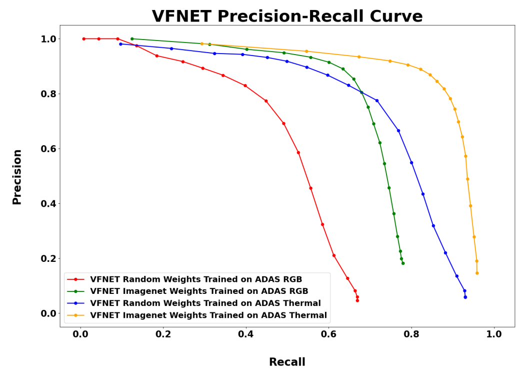

Figure 3. VFNET Precision-Recall curve. A detection is considered a true positive if it has at least a 0.5 IOU with a ground truth box.

The analysis of the precision-recall curve of VFNET confirms the above findings. The maximum recall achieved from the detectors trained on RGB images is 78%, while the maximum recall achieved by the detectors trained on thermal images is 96%. Applying transfer learning also improved the performance of detectors both in the color and thermal domains, but we notice that the improvement in the color domain is more significant.

4.2. MOT Metrics

To study the effectiveness of the trained detectors in the task of MOT, we calculate the standard MOT metrics [2, 18] of ByteTrack with DCC when using each trained detector on our City Scene RGB-Thermal MOT dataset. We retrain the detectors on our City Scene RGB-Thermal MOT dataset with the weights initialized from the weights of the detectors trained on the ADAS dataset. We study the performance of the tracker across three testing sequences, taken in the morning, afternoon, and at night to study the performance of the tracker under different lighting conditions. The results are given in Table 2.

The results on the sequence taken in the morning show that the trackers operating on RGB images are more effective than trackers operating on thermal images. This could be explained through analysis of the sunlight distribution in that sequence. From Figure 1, we notice that the field of view of the thermal camera covers an area that is partially in the shade. Since the dataset was collected in a hot day in the month of July, there is a considerable difference in temperature between the areas in the shade and the areas in direct sunlight. This could be noticed by comparing the appearance of the car present in the shade to the appearance of the car present in the sunlight. This difference makes it more difficult for the detector to operate on thermal images of this sequence.

In the second sequence taken in the afternoon, we can see from Figure 1 that the entire field of view of the thermal camera is almost entirely in the shade, so there is no significant variation in temperature values across different parts of the field of view. This results in the tracker performing better on thermal images, even having a higher MOTA than the tracker running on RGB images.

In the sequence taken at night, we notice that the trackers operating on thermal images perform significantly better than the trackers operating on RGB images

(34% higher MOTA when using TOOD, 48% higher MOTA when using VFNET). This is further proof that the detectors perform better on thermal images when there is not a significant variation in temperature across different parts of the field of view. In addition, the trackers operating on RGB images are expected to struggle in detecting objects at night as they are not clearly visible.

To study the effect of using a dynamic cut-off confidence, we compare the performance of the trackers with DCC against the performance of the trackers when using a fixed confidence for high-score detection boxes as in the original implementation of ByteTrack does. The results are given in Table 3. It can be seen that applying a DCC noticeably improves the MOTA of the trackers, especially since there are lots of objects with a small area in the dataset (far from the camera).

4.3. Speed

We benchmark the speed of the proposed trackers in two environments:

- A powerful machine with an NVIDIA RTX 3090 GPU, and an Intel Core i9 - 10900X 3.70 GHz Processor. (Referred to as AWP).
- A lower power edge device, NVIDIA Jetson Xavier, with a 512-core Volta GPU, and a 8-core ARM v8.2 64-bit Processor. (Referred to as Xavier).

The results are given in Table 4. Since the tracking process is independent from the detection process, and the tracking association runs on CPU while the detector inference runs on the GPU, the latency of the tracking process is constant.

4.4. Failure Cases

As elaborated in Section 4.2, the thermal trackers perform poorly under conditions where there are different intensities of sunlight in the thermal sensor's field of view. We also discuss the limitations of the RGB trackers at night. In addition, we notice that there is a noticeable number of incorrect annotations in the FLIR ADAS dataset, on which we heavily rely for the training of our detectors. There are several instances where pedestrians present in an image are not annotated. This would cause a problem during training when a large number of objects predicted as pedestrians by the model do not have corresponding ground truth annotations, causing the model to train on considering them negatives when they are actually true positives. Similarly, when evaluating the model, several false positives are actually true positives, which would negatively affect the precision and recall values. We randomly sample 100 testing images from the ADAS Thermal dataset, and found that 6 of them have annotation faults (2 of which are shown in Figure 4. This should also be taken into consideration when analyzing the precision-recall curves of the models.

5. Conclusion and Future Work

In this paper, we have conducted in-depth empirical studies to analyze the feasibility of using thermal sensors for multiple object detection and tracking. We train two state-of-the-art object detectors on the RGB and thermal variations of the FLIR ADAS dataset, and study the efficacy of transfer learning when applied to training a detector on a dataset from a different spectrum than the initial weights.

We show the superiority of detectors trained on thermal images compared to those trained on RGB images, especially under poor lighting conditions. We also show that transfer learning is more effective when used to train a detector on a dataset from the same spectrum as the initial weights. We introduce the use of a dynamic confidence cut-off, which factors the size of a predicted box, as an enhancement to the motion similarity association of tracking-by-detection based MOT, and show that it improves the accuracy of the tracker.

We have highlighted the limitations of trackers operating on RGB images under poor lighting conditions, and the limitations of trackers operating on thermal images when the thermal sensor field of view is covering areas of significantly different sunlight intensities.

Our experiments and analysis clearly highlight the importance of sensor fusion, especially in critical systems like ADAS. Each type of sensor is optimal under certain conditions and has limitations under other conditions. Being able to combine data sources from different spectrums significantly enhances the perception ability of an autonomous system.

In the future, our work can be expanded by examining further enhancements to the data association of the tracker that utilize the thermal information of an object. In addition, developing a tracker that utilizes information from both the visible and non-visible spectrum to enhance the tracking accuracy can be investigated.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 1

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5

[3] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the*

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	0.35	0.57	2	1	6	63	403
TOOD RGB	**0.86**	**0.92**	6	4	0	51	51
VFNET Thermal	0.33	0.55	2	1	6	56	417
VFNET RGB	0.66	0.84	6	3	1	185	61

(a) Results on the test sequence taken in the morning.

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	**0.77**	**0.85**	6	1	3	83	14
TOOD RGB	0.7	0.82	5	2	3	46	14
VFNET Thermal	0.58	0.79	6	0	4	153	0
VFNET RGB	0.57	0.75	5	1	4	175	0

(b) Results on the test sequence taken in the afternoon.

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	**0.85**	**0.92**	6	4	1	40	81
TOOD RGB	0.51	0.67	3	4	3	126	268
VFNET Thermal	0.82	0.91	8	2	1	43	106
VFNET RGB	0.34	0.63	5	3	2	337	189

(c) Results on the test sequence taken at night.

Detector	MOTA	IDF1	MT	PT	ML	FP	FN
TOOD Thermal	0.67	**0.81**	14	6	10	186	608
TOOD RGB	**0.69**	**0.81**	14	10	6	223	585
VFNET Thermal	0.59	0.77	16	3	11	252	748
VFNET RGB	0.52	0.74	16	7	7	697	533

(d) Overall results across the three testing sequences.

Table 2. Overall ByteTrack with dynamic cut-off results on the 3 sequences of our City Scene RGB-Thermal MOT dataset. a) First test sequence captured during daytime. b) Second test sequence captured during daytime. c) Third test sequence taken at night. d) Overall performance across all 3 testing sequences.

Detector	Overall MOTA without DCC	Overall MOTA with DCC	Improvement
TOOD Thermal	0.58	0.67	9%
TOOD RGB	0.62	0.69	7%
VFNET Thermal	0.51	0.59	8%
VFNET RGB	0.46	0.52	6%

Table 3. Comparison of the performance of the trackers before and after the use of a DCC on our City Scene RGB-Thermal MOT dataset.

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6247–6257, 2020. 2

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 4

[5] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6172–6181, 2019. 1

[6] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51(3):1244–1261, 2021. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 4

[8] Farzan Erlik Nowruzi, Wassim A El Ahmar, Robert Laganiere, and Amir H Ghods. In-vehicle occupancy detection with convolutional networks on thermal images. In

	AWP		Xavier	
	Tood + DCC	**VFNET + DCC**	**Tood + DCC**	**VFNET + DCC**
Detection Time (ms)	0.05783	0.06197	0.54908	0.58723
Tracking Time (ms)	0.00085	0.00085	0.00998	0.00998
Total (ms)	0.05868	0.06282	0.55906	0.59721
Total Frames Per Second	17.04	15.92	1.79	1.67

Table 4. Benchmarking tracker speed on the AWP machine and the Jetson Xavier module.

Figure 4. Samples showing missing annotations from the FLIR ADAS Dataset. Green rectangles represent ground truth boxes, red rectangles represent detected boxes, orange arrows point to instances where a pedestrian is detected by a model but no ground truth annotation exists.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019.

[9] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021. 2

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[12] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. Thermal object detection in difficult weather conditions using yolo. *IEEE access*, 8:125459–125476, 2020. 2

[13] Chao Liang, Zhipeng Zhang, Yi Lu, Xue Zhou, Bing Li, Xiyong Ye, and Jianxiao Zou. Rethinking the competition between detection and reid in multi-object tracking. *arXiv preprint arXiv:2010.12138*, 2020. 1

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[16] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 2

[17] Erkan Oğuz, Ayhan Küçükmanisa, Ramazan Duvar, and Oğuzhan Urhan. A deep learning based fast lane detection approach. *Chaos, Solitons & Fractals*, 155:111722, 2022. 1

[18] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 5

[19] Severin Stalder, Helmut Grabner, and Luc Van Gool. Cascaded confidence filtering for improved tracking-by-detection. In *European Conference on Computer Vision*, pages 369–382. Springer, 2010. 4

[20] Michael Teutsch, Angel D Sappa, and Riad I Hammoud. Computer vision in the infrared spectrum: challenges and approaches. *Synthesis Lectures on Computer Vision*, 10(2):1–138, 2021. 2

[21] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1

[22] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3988–3998, 2019. 2

[23] Kwangjin Yoon, Jeonghwan Gwak, Young-Min Song, Young-Chul Yoon, and Moon-Gu Jeon. Oneshotda: Online multi-object tracker with one-shot-learning-based data association. *IEEE Access*, 8:38060–38072, 2020. 1

[24] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8514–8523, 2021. 2

[25] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 1, 2, 4

[26] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. 1