Communication Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 29–35
DOI: 10.15439/2023F8409
ISSN 2300-5963 ACSIS, Vol. 37

Real-Time Detection of Small Objects in Automotive Thermal Images with Modern Deep Neural Architectures

Tomasz Balon*, Mateusz Knapik†‡ and Bogusław Cyganek†
*Department of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering,
Email: *tbalon@student.agh.edu.pl*
†Department of Computer Science, Electronics and Telecommunication
Email: *mknapik@agh.edu.pl, cyganek@agh.edu.pl*
AGH University of Science and Technology,
Al. Mickiewicza 30, 30–059 Kraków, Poland
‡MyLED Inc.
Email: *m.knapik@myled.pl*
Ul. W. Łokietka 14/2, 30–016 Kraków, Poland

***Abstract*—Thermal imaging has shown great potential for improving object detection in automotive settings, particularly in low light or adverse weather conditions. To help and further develop this industry, we extend our previously shared Thermal Automotive Dataset by more than 2000 new images and 2 novel object detecting models based on YOLOv5 and YOLOv7 architecture. We point how important is the size of the dataset. Additionally, we compare the performance of both models, to see which is more reliable and superior in terms of detecting small objects in thermal spectrum. Furthermore, we analysed how preprocessing affects thermal imaging dataset and models basing on it. The new dataset is available free from the Internet.**

I. Introduction

THE introduction of deep learning has revolutionized the computer vision field, bringing about remarkable advancements in object recognition and paving the way for significant progress in various domains. One particularly crucial area that greatly benefits from accurate and efficient object detection is the development of self-driving vehicles. With the ability to analyze the surrounding environment in real-time, these vehicles rely heavily on robust object detection algorithms to make informed and safe decisions [1], [2]. Although the prevailing source of information still constitute digital cameras, operating in visual spectrum, in the recent years the far infrared, so called thermovision cameras are gaining on importance [3]. In this paper we focus on this type of signals.

In a previous article [4], a thermal automotive dataset was introduced, specifically designed for object detection using the YOLOv5 model [5]. However, the presented dataset had certain limitations, as it contained only images captured during winter conditions. Nonetheless, even with this constraint, the dataset proved to be valuable for training object detection models and laying the foundation for further advancements in the field. Additionally, the previous article introduced the model based on YOLOv5 architecture.

In order to overcome mentioned limitations and push the boundaries of object detection in thermal automotive applications, we present an expanded thermal automotive dataset. This enhanced dataset incorporates over 2,000 new images, capturing a broader range of scenarios and weather conditions. By expanding the dataset, we aim to provide a more comprehensive and diverse collection of images, better reflecting the challenges faced in real-world automotive environments.

Furthermore, we introduce a novel object detection model, the YOLOv7, which builds upon the foundation of its predecessor, the YOLOv5. The YOLOv7 model incorporates improvements in architecture and training strategies, aiming to enhance object detection accuracy and speed [6]. By comparing the performance of the new YOLOv7 model with the previous YOLOv5 model, using the expanded dataset for evaluation, we can assess which model is superior in terms oh object detection in thermal imaging.

Moreover, we delve into the impact of dataset size on model training by conducting experiments with both the YOLOv5 and YOLOv7 models. We compare the performance of the models trained on the entire expanded dataset against those trained on only half of the dataset. This analysis allows us to examine the influence of dataset size on the training outcomes, shedding light on the relationship between dataset scale and object detection performance.

By undertaking this study, our objective is to contribute to the ongoing efforts aimed at enhancing object detection accuracy and speed in the automotive industry. Through the utilization of an expanded thermal automotive dataset and the introduction of the YOLOv7 model, we aspire to facilitate the development of safer, more efficient, and more reliable self-driving cars. Ultimately, our research aims to propel the advancement of autonomous driving systems, intelligent

transportation, and the broader field of computer vision in the automotive sector. Our new dataset is available free from the Internet [7].

II. Network architectures

A. You Only Look Once v5

The YOLOv5 deep convolutional neural network introduces novel advancements building upon breakthroughs in computer vision, particularly inspired by YOLOv4 [8] and other state-of-the-art approaches. Notably, YOLOv5 adopts the New CSP-Darknet53 structure as its backbone, an evolved version of the Darknet architecture used in previous iterations.

Furthermore, both YOLOv4 and YOLOv5 employ the CSP Bottleneck, originally proposed by WongKinYiu in the Cross Stage Partial Networks (CSP) paper [9], for feature formulation. The CSP architecture, built upon DenseNet [10], is designed to overcome challenges such as vanishing gradients in deep networks, facilitate feature propagation, encourage feature reuse, and reduce the number of network parameters. In CSPResNext50 and CSPDarknet53, the DenseNet structure has been tailored to separate the feature map of the base layer, thereby mitigating computational bottlenecks and enhancing learning by directly passing an unedited feature map to the subsequent stage.

YOLOv5 draws insights from YOLOv4's research inquiry to determine the optimal neck architecture. Both YOLOv4 and YOLOv5 feature the PA-NET neck for effective feature aggregation, where each "Pi" represents a feature layer in the CSP backbone. Other improvement is the auto-learning of YOLO anchor boxes when custom data is input, eliminating the need for manual anchor box tuning.

One of the main contributions of YOLOv5 repository is an introduction of a model scaling, first proposed in EfficientNet paper [11]. In contrast to conventional approach, that employ arbitrary changes in model architecture, proposed scaling method uniformly adjusts the network in depth, by changing the number of convolutional blocks repetitions, as well as in width, by changing number of filters in selected layers, using a set of fixed scaling coefficients. The rationale behind the compound scaling method is grounded in the intuitive understanding that to improve performance, the network requires additional layers to expand the receptive field and more channels to capture finer patterns in the larger image. YOLOv5 offers different pre-trained model sizes (e.g., YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x), which are variants of the same architecture, but with different scaling parameters, balancing computational costs and memory requirements.

B. You Only Look Once v7

The main focus of the introduced advancements in YOLOv7 was to achieve a superior balance between performance and efficiency in real-time object detection.

One of the key advancements in YOLOv7 is the adoption of the Efficient Layer Aggregation Networks (ELAN) architecture as its backbone. ELAN considers memory access cost and analyzes factors such as input/output channel ratio, number of branches, and element-wise operations. This careful analysis leads to reducing gradient propagation path, resulting in faster and more accurate network inference, significantly improving the overall efficiency of the model. Moreover, gradient flow propagation paths also aids the module level re-parameterization.

YOLOv7 also revisits the idea of auxiliary head proposed in the Inception paper [12], that aids the initial model training as well as reduces the vanishing gradient problem. Authors experiment with varying degree of supervision for aux head, settling on a coarse-to-fine definition where supervision is passed back from the lead head at different granularities.

The concept of model scaling is further refined by the authors, by compound scaling depth and width as well as layer concatenation. As shown by ablation studies, this technique keeps the model architecture optimal while scaling for different sizes. Based on this, YOLOv7 provides different models (e.g. YOLOv7-tiny, YOLOv7-X, YOLOv7-E6, YOLOv7-W6), that have various size and scaling parameters. Each version is tailored to different hardware configurations and requirements, allowing users to choose the one that best suits their specific needs and computing resources.

III. Data acquisition and description

A. Data acquisition

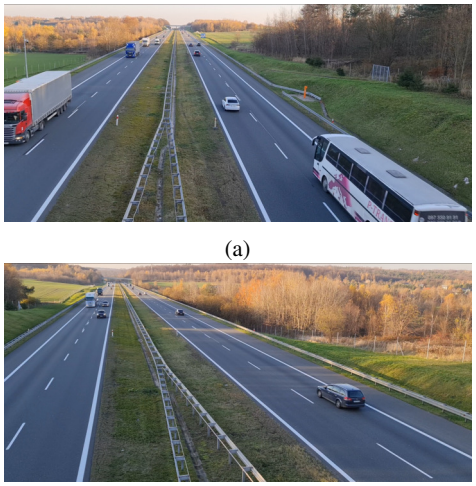
The video footage used in this study was captured using the FLIR® A35 thermal imaging camera. The data acquisition process was conducted during an autumn afternoon, specifically between 2:30 PM and 3:15 PM, when the ambient temperature ranged from 12°C to 14°C under clear weather conditions. The recording setup involved capturing real-life traffic scenes at high speeds. To achieve this, the camera was strategically positioned on an elevated bridge overlooking the road. In Figure 1, the provided images illustrate camera's field of view, showcasing the prevailing weather conditions and providing an approximate depiction of the time of day. These meticulous details ensure that the dataset encompasses realistic scenarios and accurately represents the thermal imaging perspective in a dynamic traffic environment.

B. Dataset description

Provided dataset extension consists of approximately 2000 annotated images with bounding boxes for 4 classes: car, motorcycle, bus and truck. In order to maintain consistency with previous version of dataset, all possible photos parameters were kept as they were - resolution of 320x256 pixels and 8-bit grayscale colors. For annotation we used DarkLabel [13] software and kept the same class IDs.

In total, the dataset contains over 8000 images and approximately 35000 annotations divided into 5 different classes, as shown on Figure 3.

New images introduce not only new weather conditions, but also other factors. The inclusion of highway traffic in our dataset brings forth several key factors that differentiate it from traditional city traffic datasets. Firstly, the higher speeds at which vehicles are traveling introduce motion blur, making



(a)

(b)

Figure 1: Camera's field of view

the detection task more demanding. Additionally, the bigger presence of larger vehicles, such as trucks compared to regular city traffic.

Dataset, together with object detecting models, is publicly available under the link: https://home.agh.edu.pl/~cyganek/AutomotiveThermo2_0.zip.

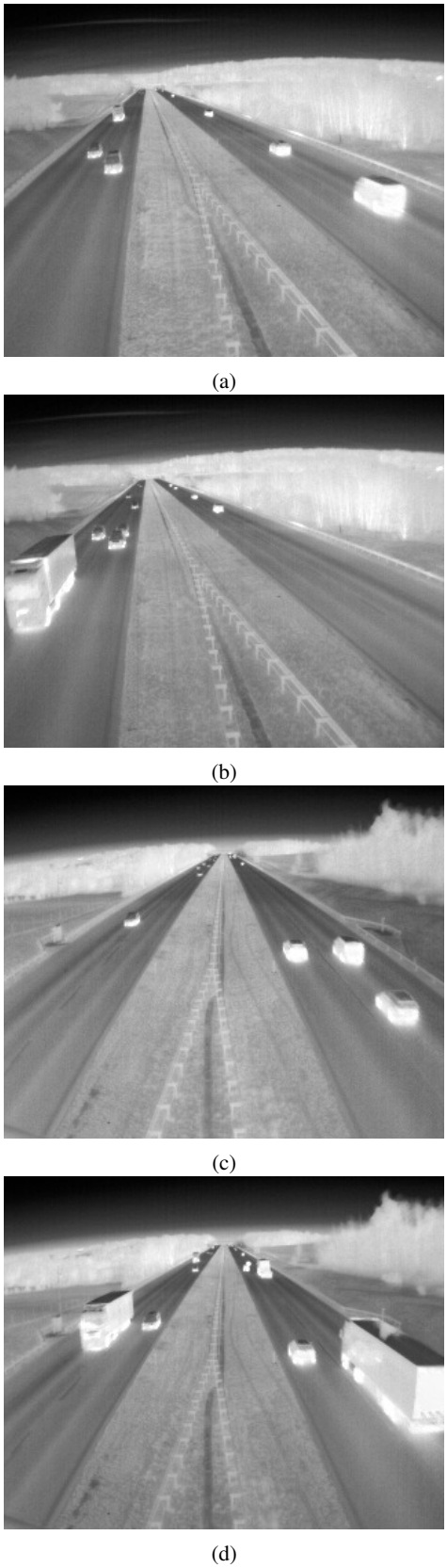
C. Data structure

Alongside with this paper, dataset and object detection models are provided. Dataset contains a total of over 8000 images divided into train, val and test subsets, each being separate folder. Additionally, trained YOLOv5 and YOLOv7 based models are published.

D. Object detection model training

To ensure a fair and comprehensive comparison between the YOLOv5 and YOLOv7 models, we adopted a systematic training approach. For the YOLOv5-based model, we retrained the previously published model, based on YOLOv5-M size architecture, on the complete thermal automotive dataset. This enabled us to evaluate the model's performance on the same dataset used for the YOLOv7 model.

Similarly, for the YOLOv7 model, we aimed to maintain consistency in the training process. As default YOLOv7 model size is comparable to YOLOv5-L, we have decided to scale it down, so that the final models have similiar number of parameters and FLOPS. Therefore, for all our tests, we set depth scaling parameter to 0.67 and width scaling parameter to 0.75. This results in a model that has computational requirements of 60.4 GFLOPS (49.2 GFLOPS for YOLOv5-M), 21.26 million parameters (21.19 million for YOLOv5-M) spread across 415 layers (291 layers in YOLOv5-M). We initially trained it using a subset of the dataset to establish a baseline performance. This step allowed us to gauge the model's initial capabilities before incorporating the expanded dataset. Subsequently, we retrained the YOLOv7 model, taking advantage of the new



(a)

(b)

(c)

(d)

Figure 2: New sample images from the dataset

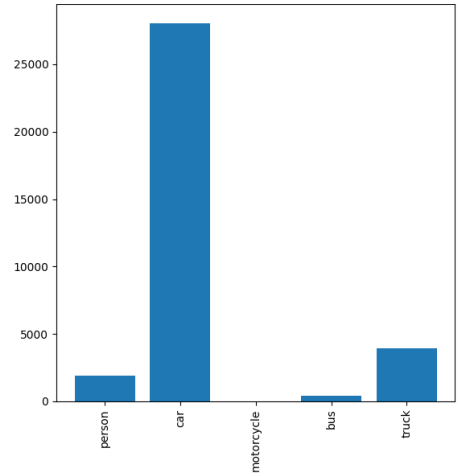


Figure 3: Number of instances in each class.

Model	Dataset	Precision	Recall	mAP	
				0.5	0.5:0.95
YOLOv5	Half	0.951	**0.971**	0.990	0.715
	Entire	**0.984**	0.965	**0.992**	**0.726**
YOLOv7	Half	0.834	0.899	0.933	0.587
	Entire	0.913	0.864	0.945	0.602

Table I: YOLOv5 and YOLOv7 models training results

Model	Dataset	Precision	Recall	mAP	
				0.5	0.5:0.95
YOLOv5	Half	0.976	0.985	0.994	0.722
	Entire	**0.989**	**0.995**	**0.995**	**0.749**
YOLOv7	Half	0.982	0.895	0.987	0.610
	Entire	0.903	0.970	0.984	0.626

Table II: YOLOv5 and YOLOv7 evaluation results on test subset

training examples in the dataset. This sequential training procedure facilitated a comprehensive analysis of the model's performance improvement with the addition of more data.

It is worth noting that both models underwent an initial pretraining phase on the COCO dataset, a widely used benchmark in computer vision. This pretraining step provided a foundation for the models to learn general object detection capabilities before being fine-tuned on the specific thermal automotive dataset. By leveraging the pretrained models, we harnessed the prior knowledge gained from the COCO dataset to enhance the object detection performance of both the YOLOv5 and YOLOv7 models on the thermal automotive dataset.

IV. Experimental part and models comparison

To evaluate the performance of the new dataset and compare the YOLOv5 and YOLOv7 models, we conducted several experiments using different training configurations.

A. Size of training dataset

Firstly, we tested how does dataset size affect training results. We took pretrained models on previous dataset and trained them using only half of the new dataset. Then we again took previously trained models and trained them on the entire new dataset. To avoid overfitting and yet to achieve best results in models training, all were trained for 50 epochs.

Results of training all four models are presented in Figure 4. These images present precision, recall, and mean average precision (mAP). As clearly visible, each model was increasing its' accuracy as with successive epochs. At first, advancements were made rather rapidly to then slow down while coming to the end of training, which was expected. Final numeric results are summarized in Tables I and II.

Although the differences between using only half or entire dataset are not very substantial, they display the overall trend – the more data available, the more accurate the model is. These numbers also show that using only part of the dataset, provides us with acceptable results which might be enough for object detection. However, we aim higher than that. The main goal is for the model to be as accurate and as robust as possible.

B. YOLOv5 vs YOLOV7

After examination of what impact does dataset size have on training results, we compared the two mentioned architectures. Head-to-head numeric results are stored in Tables I and II.

Advancements made to YOLO architecture between v5 and v7, would suggest newer version to be more accurate and have better results than it's predecessor. However it is not reflected in our results. According to outcome received after training both models, YOLOv5 outperforms YOLOv7. Particularly in mAP_0.5:0.95 – 0.726 for YOLOv5 in contrary to 0.602 for YOLOv7. The remaining results, although also in favor of YOLOv5, are not as substantial as mean average precision. These lead to a conclusion that in case of small 8-bit grey scale images, YOLOv5 would be more reasonable to use, rather than the newer YOLOv7.

During the training process of the YOLOv7 model on our thermal automotive dataset, we observed a sudden drop in precision, recall, and mAP scores. This unexpected decline in performance raised the need for investigation to identify the potential reasons behind this phenomenon. We search through known issues with the YOLOv7 implementation (and YOLOv5, as v7 codebase is heavily based on a code released by Ultralytics) code repository. We discovered that similiar problem was present in v5 code [14] and was possibly a code error triggered by very small objects present in our dataset. It was subsequently fixed in later releases, but it seems that it was transfered to the v7 repository when forked [15].

We tried to mitigate it, firstly by changing different hyperparameters such as different losses, learning rate ComputeLossOTA, as those might have lead to miscalculating loss function. Unfortunately though, changing values of these hyperparameters did not result in great improvement. It only shifted the sudden drop in epochs (e.g. drop happening in 3rd epoch, not 23rd).

C. Virtual High Dynamic Range

In our pursuit of further enhancing the quality and information content of our thermal automotive dataset, we explored the

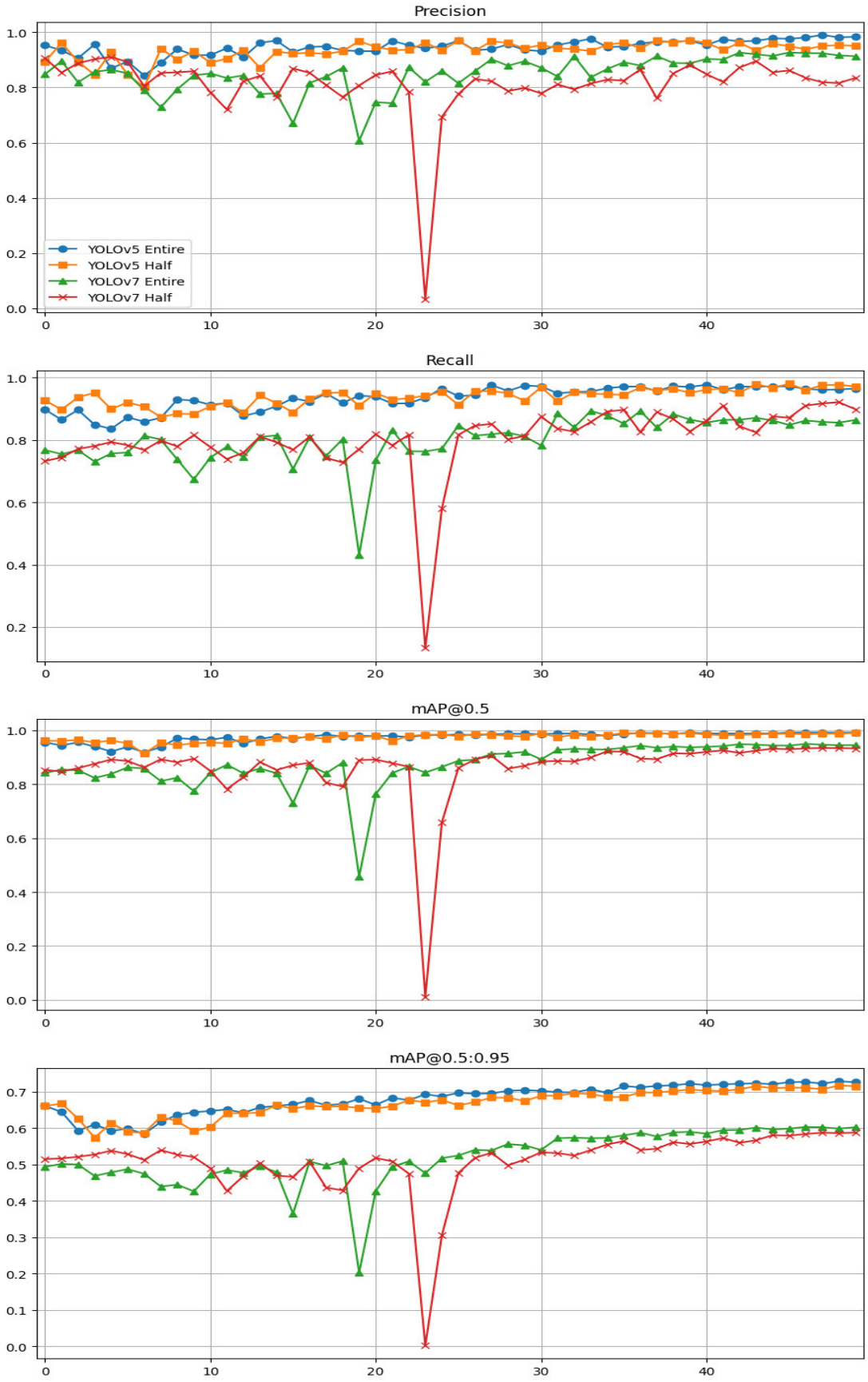


Figure 4: YOLOv5 and YOLOv7 models train results

implementation of the Virtual High Dynamic Range (VHDR) technique. Cyganek et al. [16] proposed another approach towards the VHDR method for images enhancement. To receive an VHDR image, an LDR image is taken as an input and is being processed by a set of tone adjustment curves to potentially reveal hidden details. Then it is fused to HDR image. Lastly image range conversion and contrast enhancement is done.

Based on our previous positive results with this kind of image preprocessing [1], [2], we applied the VHDR technique to our dataset and trained both the YOLOv5 and YOLOv7 models on this augmented dataset. However, in this case the results did not demonstrate a significant improvement in object detection performance. The mAP scores for both models remained relatively unchanged when compared to the models trained on the original dataset.

V. Discussion

Our results show that the YOLOv5 model outperforms YOLOv7 in terms of object detection accuracy on our thermal automotive dataset. This is an interesting observation, because v7 is both faster and achieves better results on regular datasets [6]. However, this improvements were generated by means of training procedure optimization and techniques like model re-parametrization and dynamic label assignments [6]. These can lead to increase in performance, but it also needs sufficiently big dataset to achieve that. When other modalities are used, such as long wave infrared, obtaining large scale training datasets is often unfeasible or even impossible. Older methods, such as YOLOv5, are less prone to such problems as their architecture is less data-specific. Additionally, the spatial resolution of thermal images is relatively low, posing a significant challenge for object detection, similar to the task of detecting small objects in RGB images. Furthermore, the limited input channel in thermal images further decreases the availability of extracted features during the initial stages of the network. However, the YOLOv5 algorithm addresses this issue by incorporating a unique first layer known as the Focus layer [17]. The primary purpose of this layer is to mitigate the impact of the small number of input channels compared to the significantly larger number of feature maps in deeper layers of the network. This is achieved by dividing the input layers into odd columns and rows, which are then redistributed as additional channels, enhancing the representation of features, similarly conceptually to dilated convolutions. Interestingly, we also found that YOLOv5 achieved good performance when trained on half the dataset, suggesting that it could be a more practical choice for those with limited computational resources. Furthermore, when the entire dataset is used, YOLOv5 also performs better, indicating that it is the better choice for smaller datasets and less demanding applications.

In a parallel investigation, Yang [18] conducted a comprehensive analysis comparing the performance of YOLOv5, YOLOv6, and YOLOv7 models. Interestingly, Yang's findings align closely with our own research, as he observed that the YOLOv6 model exhibited superior performance compared to its counterparts. This convergence in results reinforces the efficacy of the YOLOv6 model and underscores its potential for advancing object detection capabilities in various domains.

Olorunshola et al. [19] conducted comparable investigations in the field, focusing on the performance of the YOLOv5 and YOLOv7 models. Their study employed the Google Open Images Dataset, incorporating specific classes such as Person, Handgun, Rifle, and Knife. Although their dataset comprised slightly larger color images in contrast to our thermal dataset, their findings echoed our own observations: YOLOv5 exhibited superior performance across various metrics, with the exception of Recall. These parallel outcomes indicate a consistent trend in the comparative analysis of YOLOv5 and YOLOv7, further affirming the potential advantages of YOLOv5 in object detection tasks.

VI. Conclusion

In this article, we introduced an expanded thermal automotive dataset with approximately 2,000 new images and classes, and a new object detection model based on YOLOv7. We compared the performance of the new model with the previous YOLOv5 model using the expanded dataset, and provided insights into the acquisition process of the dataset. We made our newest dataset available free from the Internet [7].

The results showed that the YOLOv5 model outperformed the YOLOv7 model in terms of accuracy. This study contributes to the development of safer and more efficient self-driving cars by providing a better tool for object detection.

Future work will involve expanding the dataset further, including adding images taken in different weather conditions such as summer, which presents a harsher environment for thermal imaging. These additions will help to improve the robustness and versatility of our dataset and enable the development of more accurate and reliable object detection models for thermal automotive images.

In summary, this study provides valuable insights into the use of advanced object detection models and thermal imaging for object detection in the automotive industry. The expanded thermal automotive dataset and both YOLOv5 and YOLOv7 models introduced in this article can be used as a benchmarks for future research in this field.

Acknowledgments

This work has been supported by the AGH University of Science and Technology under subvention funds no. 16.16.230.434.

References

[1] M. Knapik and B. Cyganek, "Fast eyes detection in thermal images," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3601–3621, 2021.

[2] ——, "Driver's fatigue recognition based on yawn detection in thermal images," *Neurocomputing*, vol. 338, pp. 274–292, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219302280

[3] 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.

[4] T. Balon, M. Knapik, and B. Cyganek, "New thermal automotive dataset for object detection," *Annals of Computer Science and Information Systems*, vol. 31, pp. 43–48, 2022.
[5] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, K. Michael, Lorna, A. V, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M. T. Minh, "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," Feb. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6222936
[6] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022.
[7] "Automotivethermo2_0," 2023. [Online]. Available: https://home.agh.edu.pl/~cyganek/AutomotiveThermo2_0.zip
[8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.
[9] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "Cspnet: A new backbone that can enhance learning capability of cnn," 2019.
[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018.
[11] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.
[13] "Darklabel annotation software," https://github.com/darkpgmr/DarkLabel, accessed: 2022-06-07.
[14] "Sudden performance decrease in training · Issue #5721 · ultralytics/yolov5 — github.com," https://github.com/ultralytics/yolov5/issues/5721, 2021, [Accessed 31-07-2023].
[15] "Unstable training · Issue #974 · WongKinYiu/yolov7 — github.com," https://github.com/WongKinYiu/yolov7/issues/974, 2022, [Accessed 31-07-2023].
[16] B. Cyganek and M. Woźniak, "Virtual high dynamic range imaging for underwater drone navigation," *Proceedings of the 6th IIAE International Conference on Industrial Application Engineering*, pp. 393–398, 2018.
[17] A. Song, Z. Zhao, Q. Xiong, and J. Guo, "Lightweight the focus module in yolov5 by dilated convolution," in *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, 2022, pp. 111–114.
[18] L. Yang, "Investigation of you only look once networks for vision-based small object detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0140410
[19] O. E. Olorunshola, M. E. Irhebhude, and A. E. Evwiekpaefe, "A comparative study of yolov5 and yolov7 object detection algorithms," *Journal of Computing and Social Informatics*, vol. 2, no. 1, pp. 43–48, 2023.