Supplementary Material

The MONET dataset: Multimodal drone thermal dataset recorded in rural scenarios

Luigi Riz[1] Andrea Caraffa[1] Matteo Bortolon[1] Mohamed Lamine Mekhalfi[1] Davide Boscani[1]
André Moura[2] José Antunes[2] André Dias[2] Hugo Silva[2] Andreas Leonidou[3]
Christos Constantinides[3] Christos Keleshis[3] Dante Abate[3] Fabio Poiesi[1]

[1]Fondazione Bruno Kessler [2]INESC TEC [3]The Cyprus Institute

1. Introduction

We provide some additional material in support of the main paper. The content is organised as follows:

- In Sec. 2 we provide details about the training configuration of each detector and describe each data augmentation we used.
- In Sec. 3 we analyse the statistics of each dataset split for the dirt-road and runway scenarios.
- In Sec. 4 we provide examples of qualitative results taken from each detector that we evaluated.

2. Detector setups

In the main paper we have two sets of results that compare state-of-the-art detectors. For both sets we use non-maximum suppression with IoU threshold 0.5 and we evaluate all the bounding boxes with confidence above 10^{-8}. The reason behind this choice is that the models are calibrated differently, hence their output confidences can not be directly compared. Setting a low confidence threshold is to minimally filter detector predictions, resulting in a fairer comparison of their performance.

The first set compares nine detectors with training configurations that we set as similar as possible. Tab. 1 reports the details of the chosen configurations.

We train all the detectors with the same data augmentation strategy. Data augmentations are applied in the following order: i) RandomCrop: This crops a portion of the image with a size determined by randomly sampling two independent values within the interval [0.8, 1.0] and by multiplying them by the height and width of the original image; ii) Resize: This randomly resizes the eventually cropped image between (600, 800) and (300, 400) while keeping its original aspect ratio. iii) RandomHorizontalFlip: This randomly flips the image horizontally with a probability of 0.7. iv) Padding: This is applied to make all the images of the same size, i.e. (600, 800). v) Normalisation: This involves normalising the image pixels with a mean of 126.225 and a standard deviation of 73.338. Note that these normalisation factors differ from the standard ones computed on ImageNet and were explicitly calculated for the MONET dataset.

Table 1. Detectors setup. Keys: BS: Batch Size. LR: Learning Rate. CosAnn.: Cosine Annealing.

Detector	Epochs	BS	Backbone	LR	Schedule	Optimiser
F. R-CNN [6]	20	24	ResNet-50	1e-3	CosAnn.	AdamW
SSD [5]	20	24	VGG-16	1e-3	CosAnn.	AdamW
CornerNet [4]	20	9	HourglassNet-104	1e-4	CosAnn	AdamW
FCOS [7]	20	24	ResNet-50	1e-4	CosAnn	AdamW
DETR [1]	50	24	ResNet-50	5e-5	CosAnn	AdamW
Def. DETR [10]	20	9	ResNet-50	5e-5	CosAnn	AdamW
VarifocalNet [9]	20	24	ResNet-50	1e-4	CosAnn	AdamW
ObjectBox [8]	20	24	YOLOv5 v6.0	1e-2	Warmup CosAnn	SGD
YOLOv8 [3]	20	24	YOLOv8.0x	1e-2	Warmup CosAnn	SGD
ObjectBox† [8]	40	24	YOLOv5 v6.0	1e-2	Warmup CosAnn	SGD
YOLOv8† [3]	40	24	YOLOv8.0x	1e-2	Warmup CosAnn	SGD

The second set compares ObjectBox [8] and YOLOv8 [3] with their original data augmentation. We use the † in the main paper to represent these setups.

The data augmentations we use for ObjectBox are applied in the following order: i) Mosaic: This combines 4 images (600, 800) into a single image (1200, 1600). Padding is then applied to produce a squared image of size (1600, 1600); ii) RandomAffine: This applies translation and scale operations. The scale factor is randomly sampled from [0.5, 1.5] while the independent vertical and horizontals shifts are randomly sampled in the interval [-160, 160], i.e. using a maximum absolute fraction of 0.1. The image is then resized to (800, 800); iii) Blur: This blurs the image using a random kernel size sampled in the interval [3, 7] with a probability of 0.1; iv) MedianBlur: This blurs the image using a median filter with random aperture linear size

sampled in the interval [3, 7] with a probability of 0.1; v) RandomHSV: This firstly converts the image to HSV colorspace. Then, three scalars are sampled with the intervals [0.985, 1.015], [0.3, 1.7], and [0.6, 1.4], which are used to multiply the original values of Hue, Saturation and Value, respectively. Lastly, the image is converted back to RGB colorspace; vi) RandomHorizontalFlip: This randomly flips the image horizontally with a probability of 0.5.

The data augmentations we use for YOLOv8 are applied in the following order: i) Mosaic: This combines 4 images (600, 800) into a single image (1200, 1600). Padding is then applied to produce a squared image of size (1600, 1600); ii) MixUp: This averages two mosaic images with a probability of 0.15; iii) RandomAffine: This applies translation and scale operations. The scale factor is randomly sampled from [0.1, 1.9] while vertical and horizontals shifts are independently randomly sampled in the interval [-160, 160], i.e. using a maximum absolute fraction of 0.1. The image is then resized to (800, 800); iv) Blur: This blurs the image using a random kernel size sampled in the interval [3, 7] with a probability of 0.01; v) MedianBlur: This blurs the image using a median filter with a random aperture linear size sampled in the interval [3, 7] with a probability of 0.01; vi) CLAHE: This applies Contrast Limited Adaptive Histogram Equalisation with probability 0.01; vii) RandomHSV: This firstly converts the image to HSV colorspace. Then, three scalars are sampled in the intervals [0.985, 1.015], [0.3, 1.7], and [0.6, 1.4], which are used to multiply the original values of Hue, Saturation and Value, respectively. Lastly, the image is converted back to RGB colorspace; viii) RandomHorizontalFlip: This randomly flips the image horizontally with a probability of 0.5.

3. Additional dataset statistics

Figs. 1, 2, and 3, show the statistics of train, validation, and test splits of dirt-road, respectively, while Figs. 4, 5, and 6, show the statistics of train, validation, and test splits of runway, respectively. The statistics include i) the histogram of the bounding box instances, ii) examples of bounding boxes randomly sampled from the ground truth, iii) the distribution of the bounding box locations over the image plane, and iv) the distribution of the bounding box sizes as a function of the width and height. These figures are generated with the software provided with YOLOv5 [2] and ObjectBox [8]. It is interesting to observe the difference in bounding box sizes between dirt-road and runway splits.

4. Additional qualitative results

Because the detectors are calibrated differently, it is unfair to apply the same confidence threshold to visualise the results. Therefore, we choose a different threshold for each detector that corresponds to the maximum between γ and

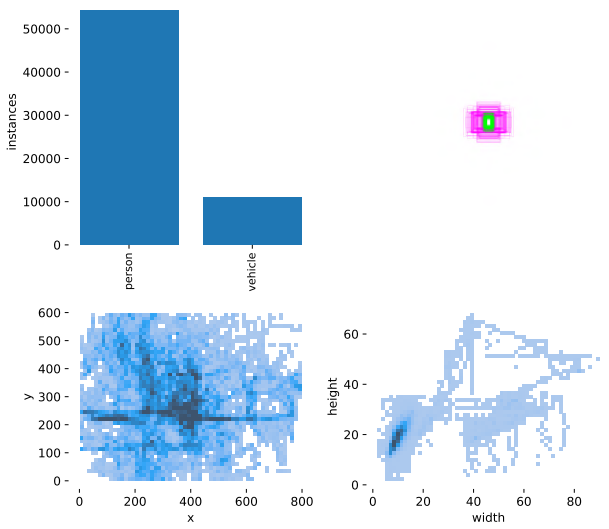

Figure 1. Bounding box ground-truth statistics of dirt-road train split. The top-left figure shows the histogram of bounding box instances. The top-right figure shows 1K examples of randomly sampled bounding boxes. The bottom-left figure shows the distribution of bounding box locations over the image plane. The bottom-right figure shows the distribution of bounding box sizes.

0.10, where γ is the confidence value of the true positive detection with lowest confidence value in a given frame. This approach allows us to visualise all detected targets, but it may result in more false alarms.

Fig. 7 shows the qualitative results of the different detectors when their models are trained on dirt-road and evaluated on dirt-road. We can observe that this scenario is very challenging because all the detectors fail to detect all the person targets. The vehicle is accurately detected by all the detectors except for CornerNet.

Fig. 8 shows the qualitative results of the different detectors when their models are trained on runway and evaluated on dirt-road. We can observe that SSD is the only detector that can detect some person targets. All the others either produce false alarms or do not detect any person targets. Like before, the vehicle is accurately detected by all the detectors except for CornerNet.

Fig. 9 shows the qualitative results of the different detectors when their models are trained on runway and evaluated on runway. Unlike before, all the targets are correctly detected in this setting. Moreover, we can observe that the confidence value of each detector is rather different from each other.

Fig. 10 shows the qualitative results of the different detectors when their models are trained on dirt-road and evaluated on runway. We can observe that Deformable DETR is the best performing one, followed by VarifocalNet. The most noisy one resulted to be CornerNet.

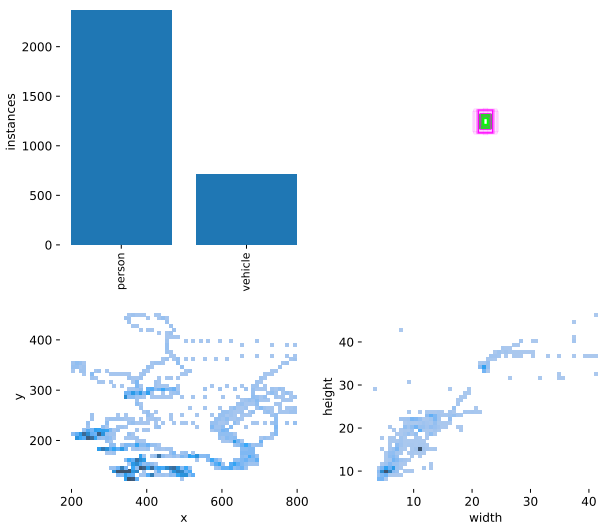

Figure 2. Bounding box ground-truth statistics of dirt-road validation split. The top-left figure shows the histogram of bounding box instances. The top-right figure shows 1K examples of randomly sampled bounding boxes. The bottom-left figure shows the distribution of bounding box locations over the image plane. The bottom-right figure shows the distribution of bounding box sizes.

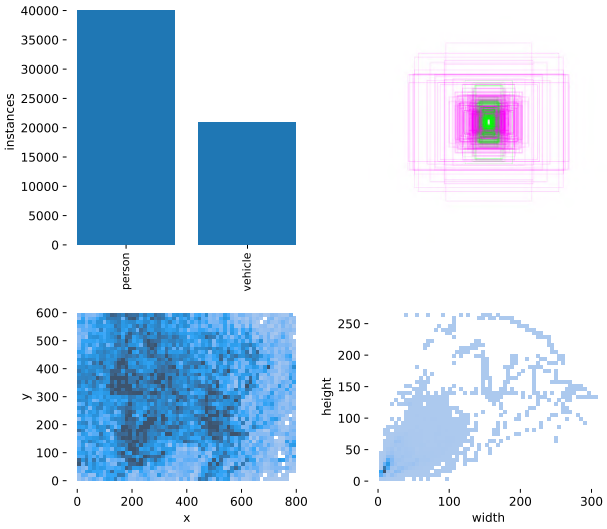

Figure 4. Bounding box ground-truth statistics of runway train split. The top-left figure shows the histogram of bounding box instances. The top-right figure shows 1K examples of randomly sampled bounding boxes. The bottom-left figure shows the distribution of bounding box locations over the image plane. The bottom-right figure shows the distribution of bounding box sizes.

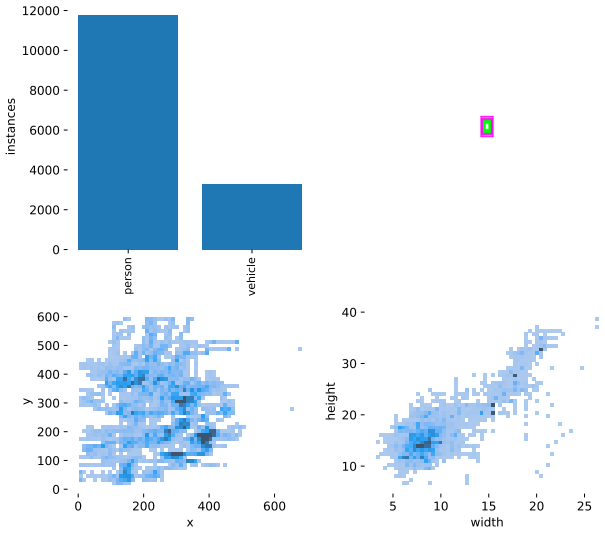

Figure 3. Bounding box ground-truth statistics of dirt-road test split. The top-left figure shows the histogram of bounding box instances. The top-right figure shows 1K examples of randomly sampled bounding boxes. The bottom-left figure shows the distribution of bounding box locations over the image plane. The bottom-right figure shows the distribution of bounding box sizes.

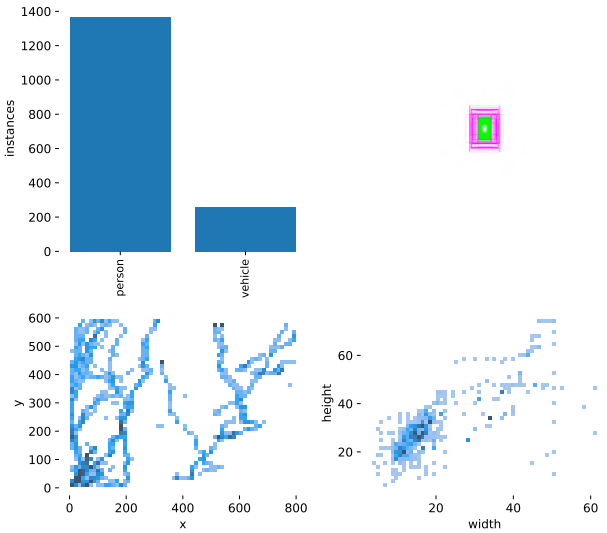

Figure 5. Bounding box ground-truth statistics of runway validation split. The top-left figure shows the histogram of bounding box instances. The top-right figure shows 1K examples of randomly sampled bounding boxes. The bottom-left figure shows the distribution of bounding box locations over the image plane. The bottom-right figure shows the distribution of bounding box sizes.

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[2] G. Jocher. YOLOv5 by Ultralytics, 2020. 2

[3] G. Jocher, A. Chaurasia, and J. Qiu. YOLO by Ultralytics,

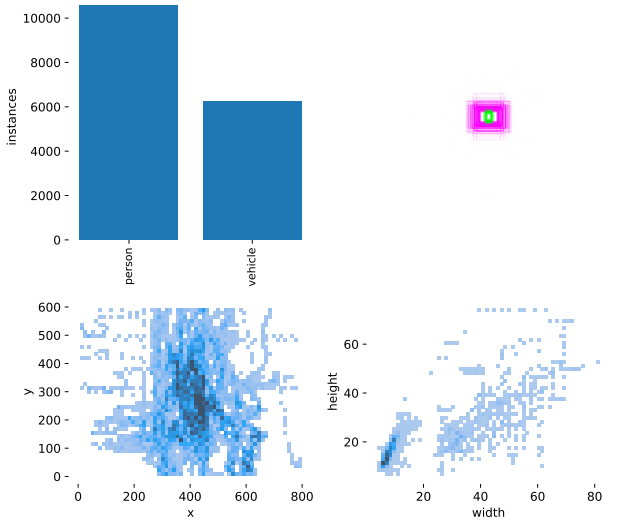

Figure 6. Bounding box ground-truth statistics of runway test split. The top-left figure shows the histogram of bounding box instances. The top-right figure shows 1K examples of randomly sampled bounding boxes. The bottom-left figure shows the distribution of bounding box locations over the image plane. The bottom-right figure shows the distribution of bounding box sizes.

2023. 1

[4] H. Law and J. Deng. CornerNet: Detecting objects as paired keypoints. In *ECCV*, 2018. 1

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 1

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[7] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 1

[8] M. Zand, A. Etemad, and M. Greenspan. Objectbox: From centers to boxes for anchor-free object detection. 2022. 1, 2

[9] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf. VarifocalNet: An IoU-aware dense object detector. *CVPR*, 2021. 1

[10] X. Zhu, W. Su, L. Lu, B Li, X Wang, and J Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 1

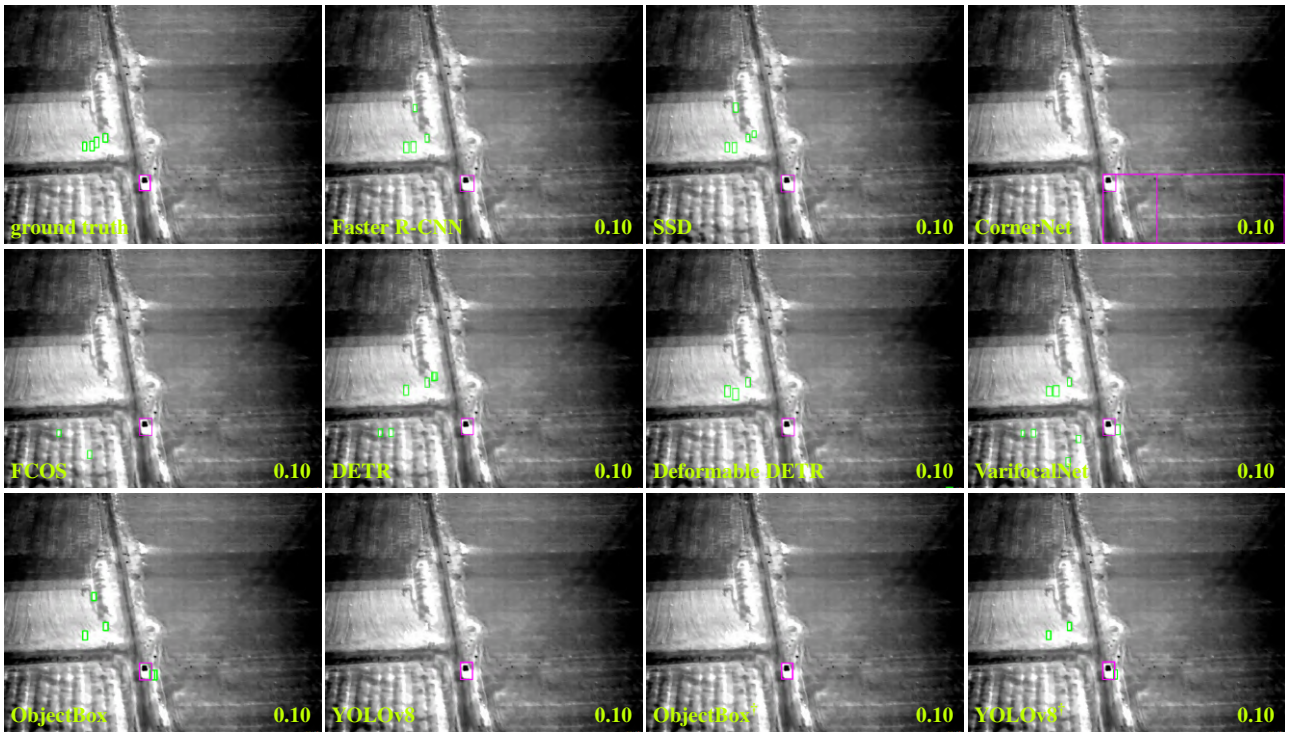

Figure 7. Qualitative results of the different detectors trained on dirt-road and evaluated on dirt-road. Bottom right value in each image represents the conference threshold adopted at inference time. Bounding boxes: green for `person`, magenta for `vehicle`. Recording altitude: 80m.

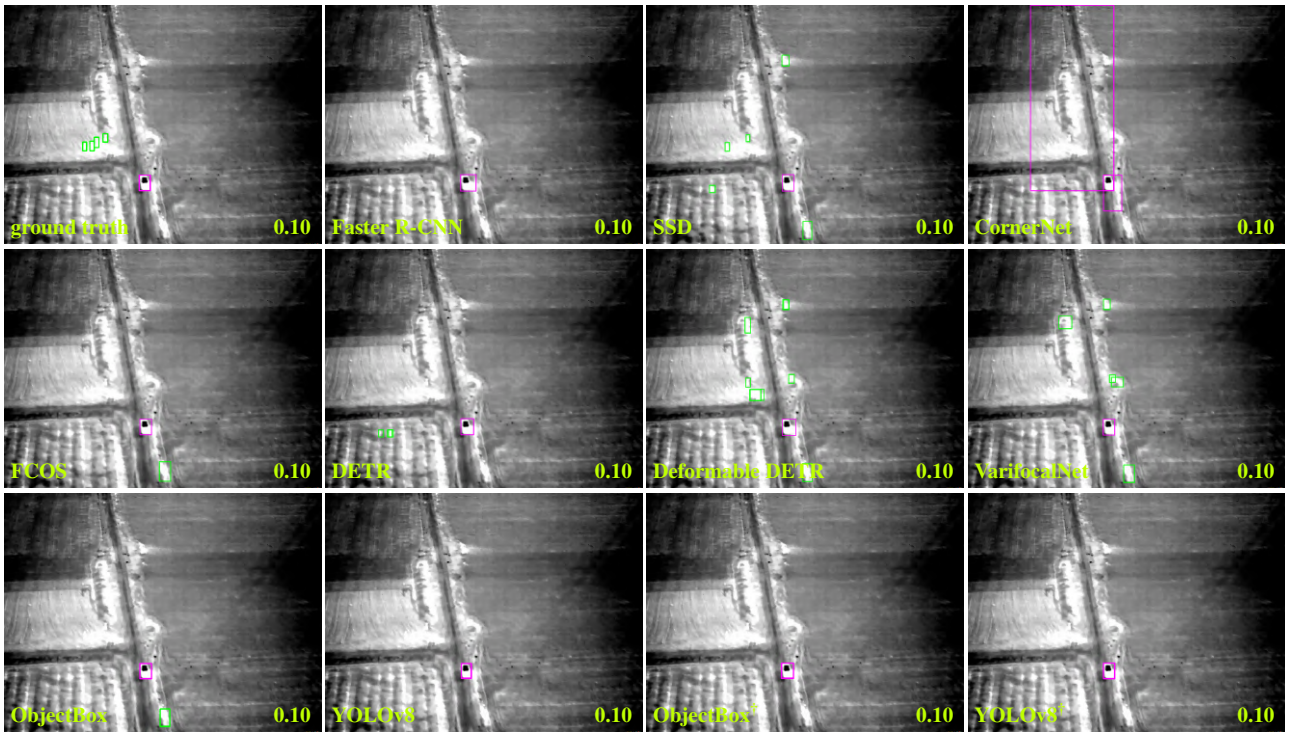

Figure 8. Qualitative results of the different detectors trained on runway and evaluated on dirt-road. Bottom right value in each image represents the conference threshold adopted at inference time. Bounding boxes: green for `person`, magenta for `vehicle`. Recording altitude: 80m.

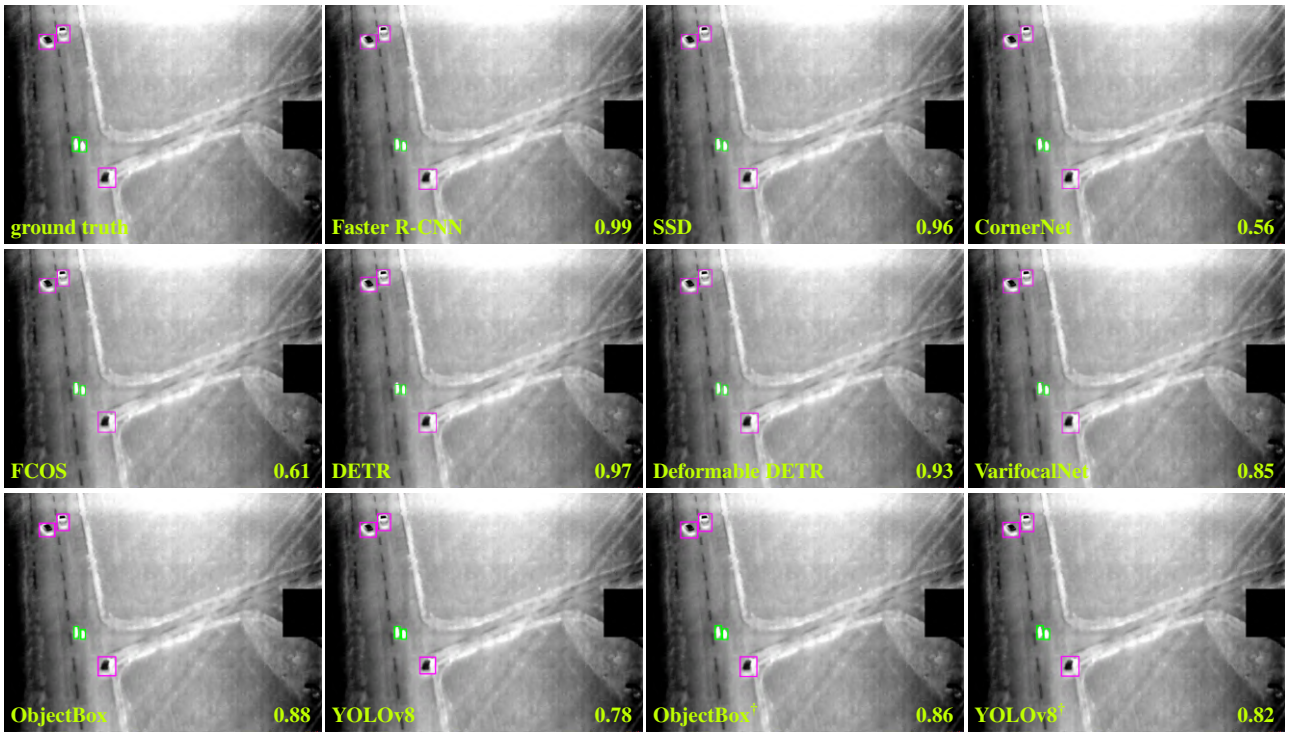

Figure 9. Qualitative results of the different detectors trained on runway and evaluated on runway. Bottom right value in each image represents the conference threshold adopted at inference time. Bounding boxes: green for `person`, magenta for `vehicle`. Recording altitude: 82m.

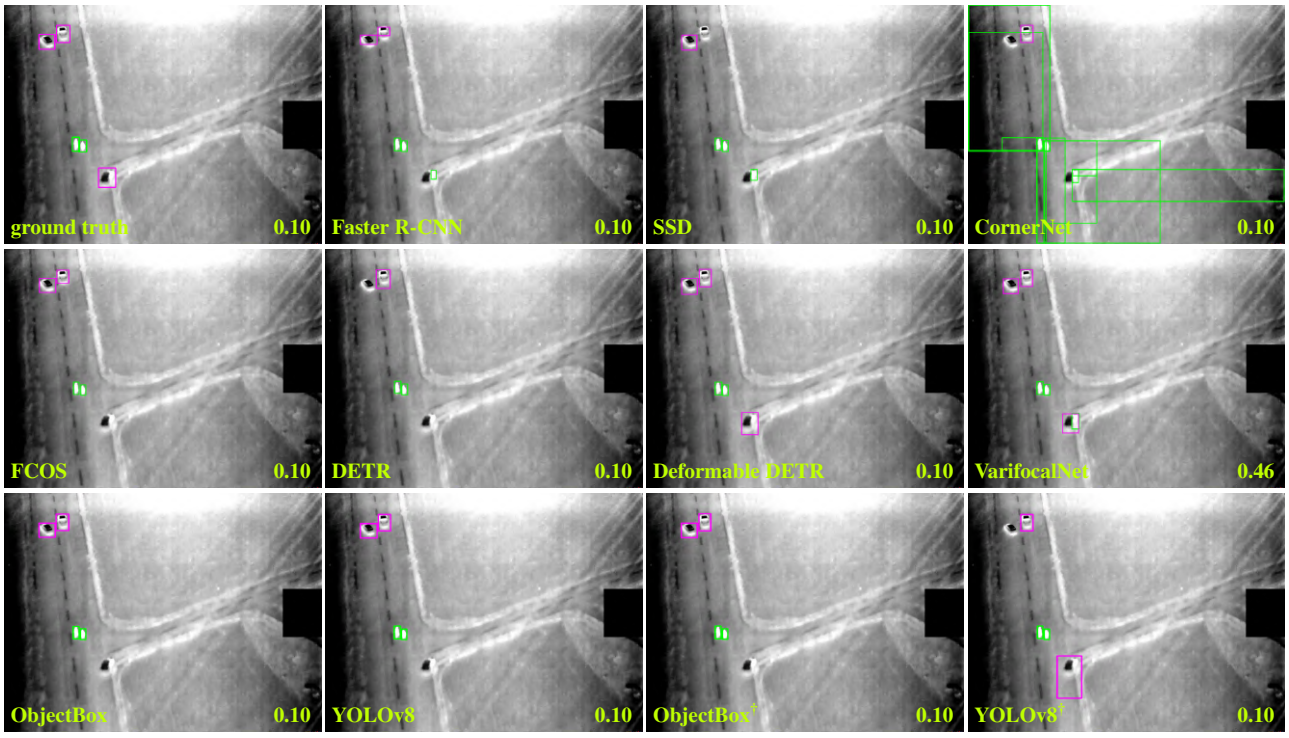

Figure 10. Qualitative results of the different detectors trained on dirt-road and evaluated on runway. Bottom right value in each image represents the conference threshold adopted at inference time. Bounding boxes: green for `person`, magenta for `vehicle`. Recording altitude: 82m.