8. Supplementary

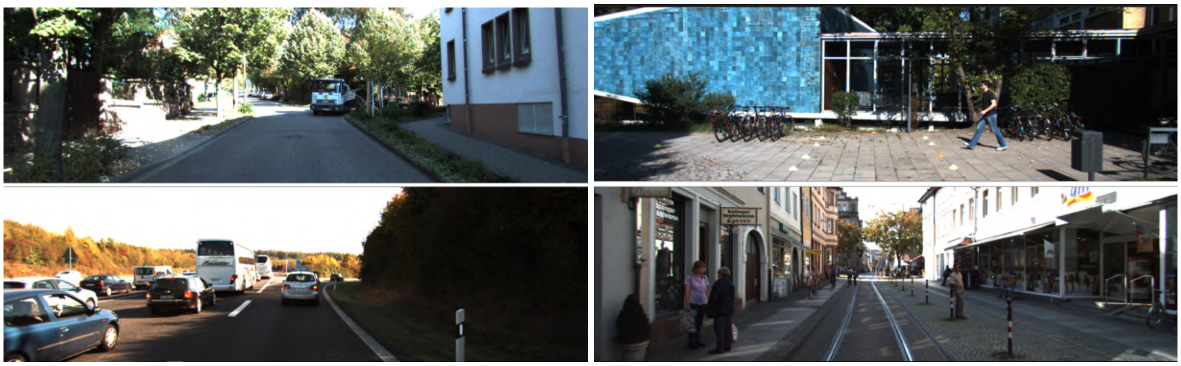
8.1. Extension of Situation Monitor on Unseen Novel Objects

In this section, we utilize the KITTI dataset to explore the Situation Monitor's (SM) capability in detecting novel, unseen objects within the same dataset. Specifically, to establish an optimal setup and ideal configuration, the training dataset intentionally omits any images containing instances of unseen objects, with their inclusion reserved exclusively for the validation and evaluation phases of both the model and the Situation Monitor. Consequently, we transition from $\mathcal{O}^{near}$ and $\mathcal{O}^{far}$ datasets problem statement to $\mathcal{O}^{near}$ and $\mathcal{O}^{far}$ objects. Our experimental approach encompasses two distinct sets. The first set involves the removal of van and truck classes during the training phase, while images containing these classes are introduced exclusively during the evaluation phase. In this set, van and truck classes can be considered as $\mathcal{O}^{near}$ objects, given the continued presence of the car class in the training data, which shares similarities with vans and trucks. Conversely, the second set of experiments focuses on removing person sitting, pedestrian, and cyclist classes during training. In this set, these two objects are categorized as $\mathcal{O}^{far}$ objects, as there is no resemblance between these objects and any others present in the training data. This meticulous experimental design allows for a nuanced assessment of the Situation Monitor's performance across various object classes and distances.

Furthermore, the final OOD value $\mathcal{U}_{SM}$ will now be monitored at the object level, departing from the earlier image-level approach as detailed in Section 4.2 and Eq. (6). In this object-level monitoring, there is no longer a need to compute the mean of object variances or centre the variances to maintain OOD values at the object level. This shift is attributed to the inclusion of raw variances, which encapsulate localization errors crucial for capturing novel objects. Despite this adjustment, the confidence score C of the detection remains instrumental in centring the variances. The refined $\mathcal{U}_{SM}$ monitoring values, as shown in Eq. (7) enhances the Situation Monitor's efficacy in identifying and characterizing novel objects within the given context. For evaluating the situation monitor on unseen novel objects, we categorise the detections that overlap with the unseen objects using the ground truth as a novel OOD object. The detections are the in-distribution objects; otherwise, they don't belong to these unseen novel object categories but are true positives. To assess the Situation Monitor's performance on unseen novel objects during training, we categorize detections that overlap with the ground truth of these unseen objects as novel Out-of-Distribution (OOD) objects. In contrast, in-distribution objects are detections that neither belong to the category of these unseen novel objects nor are considered false positives, indicating accurate identification.

$$\begin{aligned} xy_{var} &= \sqrt{\sum_{i\in(x,y)} (b_i^{\alpha} - b_i^{\beta})^2}, \\ wh_{var} &= \sum_{i\in(w,h)} (b_i^{\alpha} - b_i^{\beta})^2, \\ \mathcal{U}_{SM} &= \frac{\sqrt{xy_{var}} * wh_{var}}{\sqrt{C}} \end{aligned} \tag{7}$$

‘



(a) Vans and trucks classes as $\mathcal{O}^{near}$ objects

(b) Pedestrians, person sitting and Cyclists classes as $\mathcal{O}^{far}$ objects

Figure 11: Images showing the instances of unseen novel objects considered to study the extension of the situation monitor

8.1.1. Van and trucks classes as $\mathcal{O}^{near}$ objects

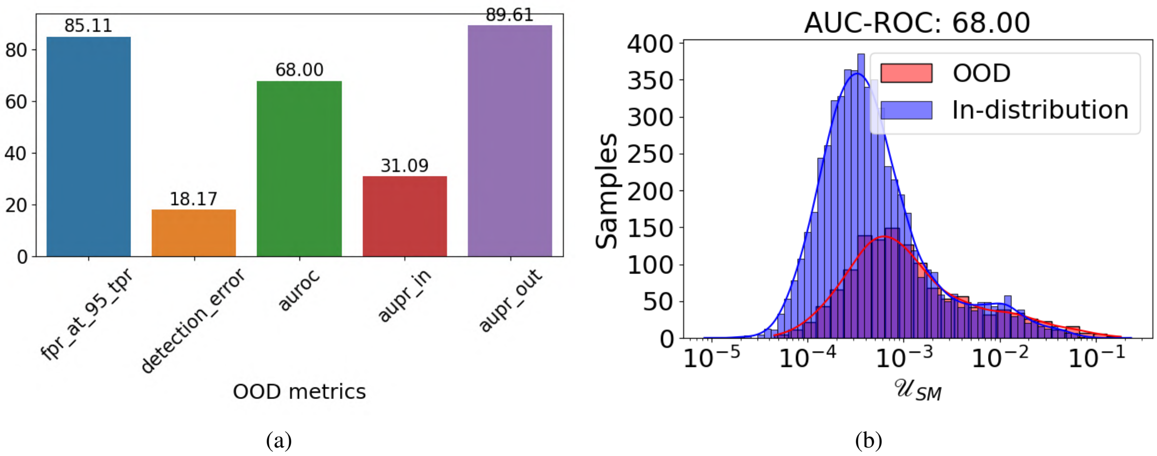
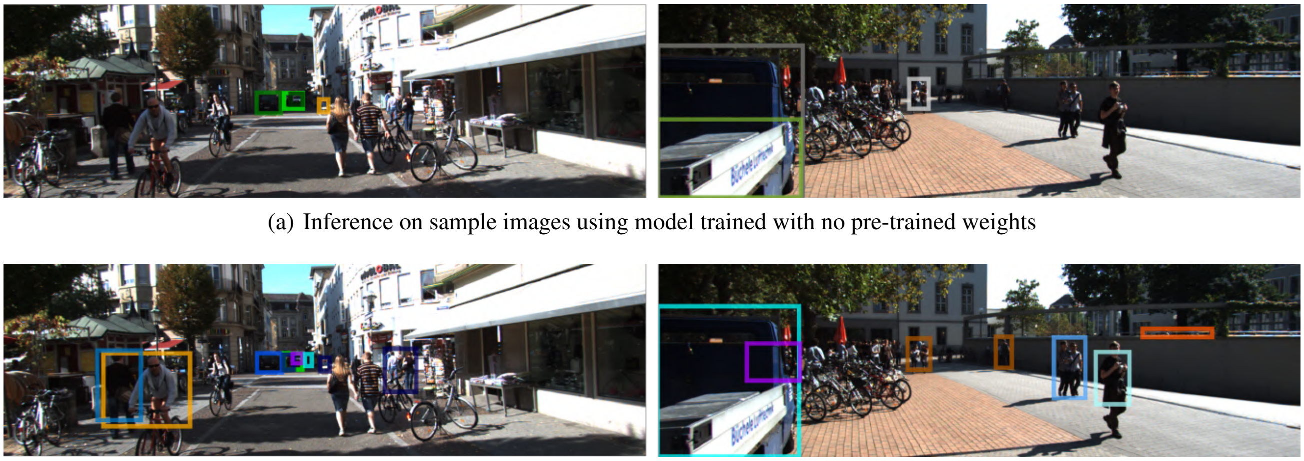


Figure 12: Metrics evaluating the Situation Monitor's detection of $\mathcal{O}^{near}$ objects (Vans and trucks classes)

The findings in Figure 12 indicate that the Situation Monitor (SM) faces challenges in identifying the Van and Truck classes as novel objects. This observation aligns seamlessly with the conceptual framework of SM refraining from categorizing the Near-OOD ($\mathcal{O}^{near}$) as Out-of-Distribution (OOD). This difficulty distinguishing Van and Truck classes as novel objects stems from the inherent resemblance between their features and those of the familiar Car class.



(a) Inference on sample images using model trained with no pre-trained weights

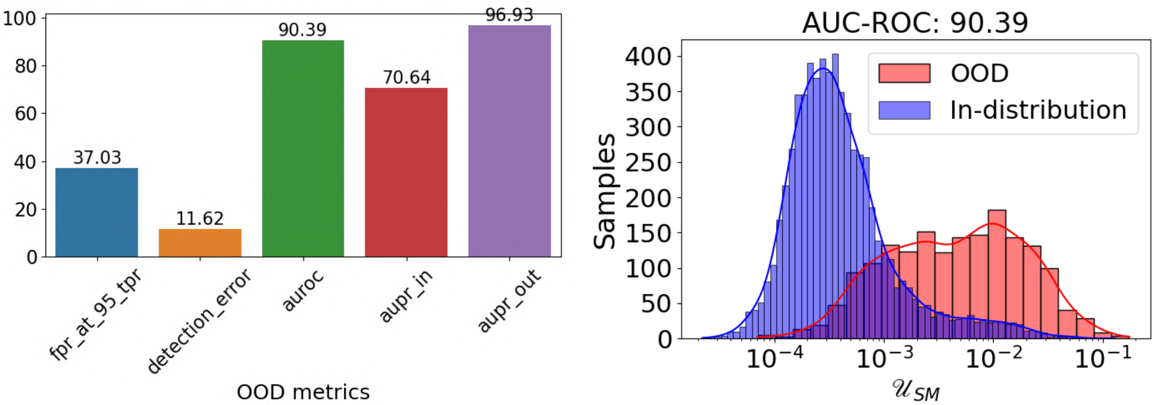
(b) Inference on sample images using model trained with pre-trained weights

Figure 13: Visualizing the impact of pre-trained weights on previously unseen and novel objects that are not part of the training dataset.

8.1.2. Pedestrians, person sitting and Cyclists classes as $\mathcal{O}^{near}$ objects

In many training processes, the common practice is to initiate training using pre-trained weights. These weights are often derived from well-established datasets such as CoCo or ImageNet, containing classes that may align with $\mathcal{O}^{far}$. Consequently, this initialization can impact the model's ability to detect unseen objects during real-time inference, even if these objects were not part of the training dataset. In such scenarios, the significance of an OOD detector becomes pronounced. This detector plays a crucial role in classifying novel or unseen objects, especially in safety-critical situations where previously not encountered instances require accurate identification for effective decision-making and response. The capability to distinguish and appropriately handle unforeseen objects enhances the overall robustness and reliability of the model in real-world applications.

Conversely, maintaining an ideal configuration for the setup involves training without the utilization of pre-trained weights. Instead, the model is trained from scratch, initializing the weights randomly. However, this approach does not guarantee



(a) (b)

Figure 14: Metrics evaluating the Situation Monitor's detection of $\mathcal{O}^{far}$ objects (Pedestrians, person sitting and Cyclists classes)

that the model will refrain from detecting unseen objects, as illustrated in Figure 13, which are not part of the ground truth. Ground truth typically refers to the annotated or labelled objects the model has been trained on. If an object is present in an image but is not part of the training data, the DNN may still attempt to identify and classify it based on learned features.

In the majority of model training procedures, pre-trained weights are commonly used. Hence, we will assess the SM's capability to detect $\mathcal{O}^{far}$ objects on a model trained with these pre-trained weights.

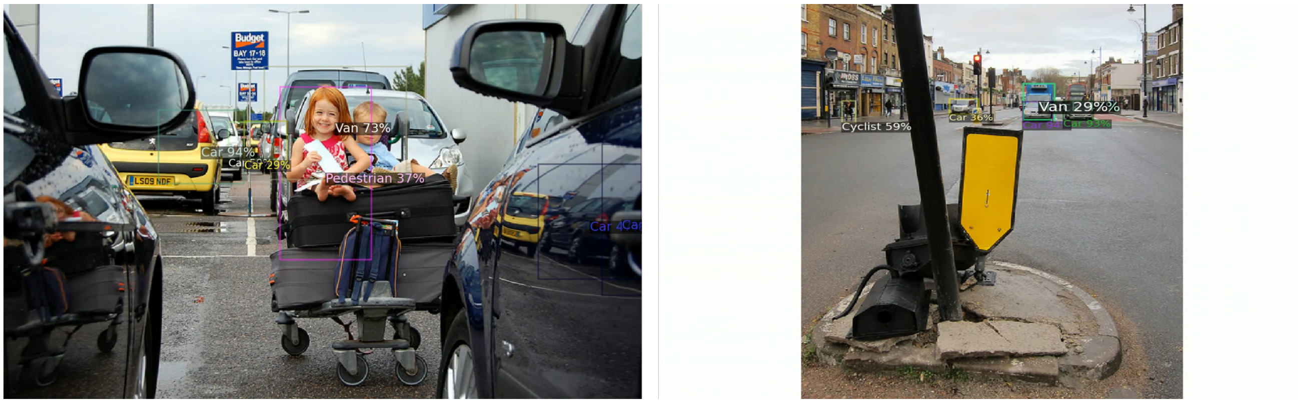
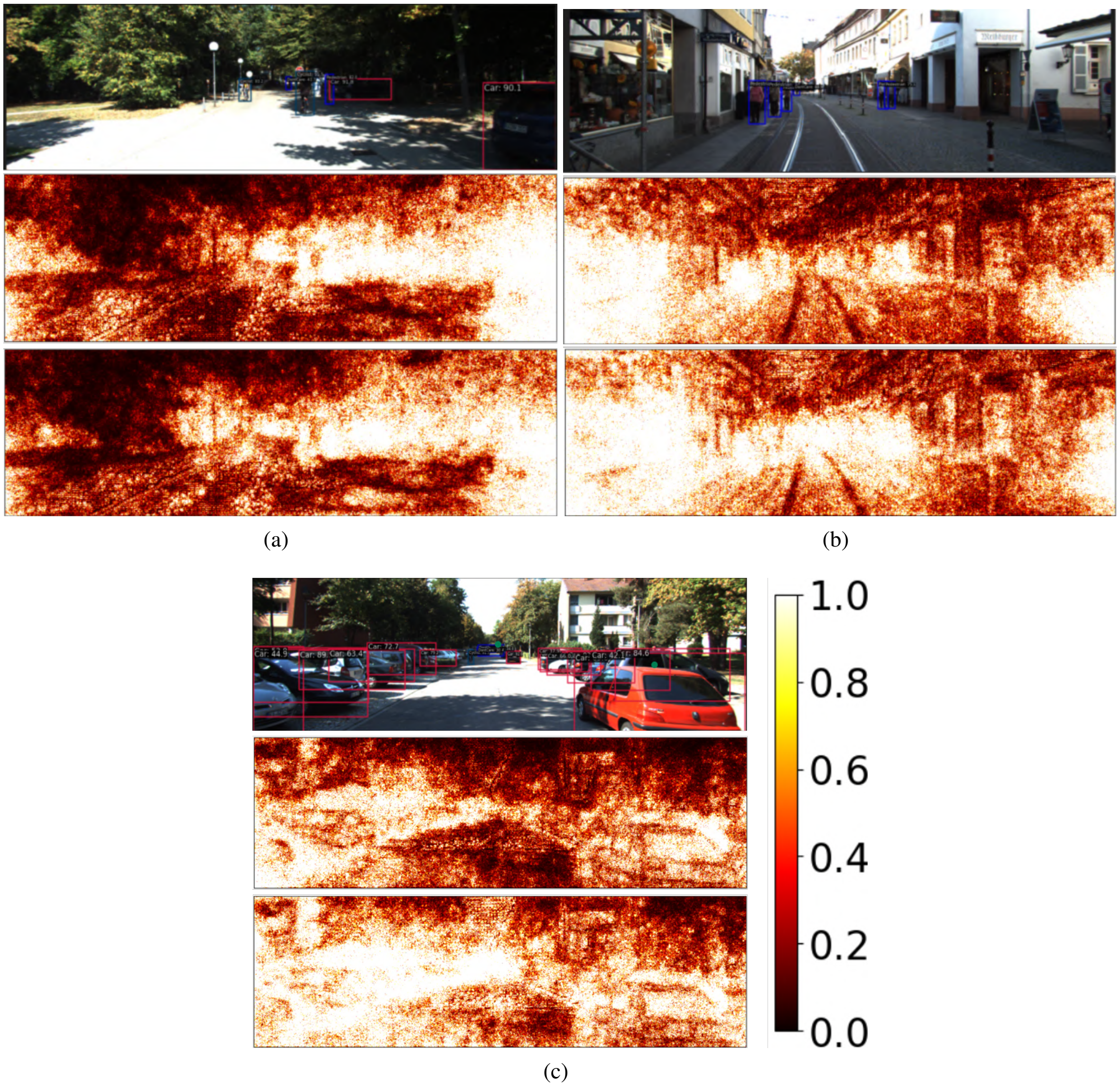


Figure 15: OOD samples similar to those in the In-Distribution.

The outcomes presented in Figure 14 indicate that the performance of the Situation Monitor is not at a particularly high level in detecting $\mathcal{O}^{far}$ objects, as shown in Table 1. However, despite the observed limitations, these results are encouraging, suggesting that there is potential for enhancement, and one avenue for refinement lies in the application of the Diversity-based Budding Ensemble architecture. Implementing this architecture could further elevate the SM's ability to discern and effectively handle instances of FAR-OOD ($\mathcal{O}^{far}$) objects.

Anomalies are apparent in the evaluated CoCo ($\mathcal{O}^{far}$) datasets, displaying characteristics closely resembling either in-distribution datasets or near-OOD $\mathcal{O}^{near}$ datasets. Notably, images related to cars exhibit a closer resemblance to datasets like KITTI, BDD100K, or Cityscapes. This leads to the overlap of $\mathcal{U}_{SM}$ values in Fig. 10(a). This observation is depicted in Fig. 15.



(a)

(b)

(c)

Figure 16: Visualization of GradCAM on sample images: The top image shows the original detections of DBEA-DINO-DETR. The middle image is the GradCAM image of Vanilla-DINO-DETR. while the bottom image corresponds to DBEA-DINO-DETR. The color-map on the GradCAM image indicates positive relevance as white and negative relevance as black

8.2. More GradCAM examples of DBEA-DINO-DETR

GradCAM visualizations (Selvaraju et al., 2017), depicted in Fig. 16, offer valuable insights into the attention and focus areas of the DBEA-based DINO-DETR model. These visualizations illustrate that the model emphasizes specific regions within the image, providing a detailed view of the areas considered crucial for its decision-making process. This heightened focus contributes to the model's improved detection performance and enhances the correlation of its confidence scores. Anomalies in the assessed CoCo ($\mathcal{O}^{far}$) datasets show features resembling either in-distribution datasets or nearby out-of-distribution ($\mathcal{O}^{near}$) datasets, particularly in images related to cars and roads. Thus resulting in overlapping $\mathcal{U}_{SM}$ values in Fig. 10(a).