

Situation Monitor: Diversity-Driven Zero-Shot Out-of-Distribution Detection using Budding Ensemble Architecture for Object Detection

Syed Sha Qutub^{1,2}, Michael Paulitsch¹, Kay-Ulrich Scholl¹, Neslihan Kose Cihangir¹, Korbinian Hagn¹, Fabian Oboril¹, Gereon Hinz², and Alois Knoll²

¹Intel Labs, Munich, Germany

²Technical University of Munich, Munich, Germany

Abstract

We introduce Situation Monitor, a novel zero-shot Out-of-Distribution (OOD) detection approach for transformer-based object detection models to enhance reliability in safety-critical machine learning applications such as autonomous driving. The Situation Monitor utilizes the Diversity-based Budding Ensemble Architecture (DBEA) and increases the OOD performance by integrating a diversity loss into the training process on top of the budding ensemble architecture, detecting Far-OOD samples and minimizing false positives on Near-OOD samples. Moreover, utilizing the resulting DBEA increases the model's OOD performance and improves the calibration of confidence scores, particularly concerning the intersection over union of the detected objects. The DBEA model achieves these advancements with a 14% reduction in trainable parameters compared to the vanilla model. This signifies a substantial improvement in efficiency without compromising the model's ability to detect OOD instances and calibrate the confidence scores accurately.

1. Introduction

In machine learning, models must exhibit effective generalization capabilities that require adaptability beyond their training data. This adaptability is crucial for ensuring the effective performance of models in real-life situations populated with diverse objects. In real-world applications, particularly in safety-critical scenarios such as autonomous driving or medical diagnosis, the capability to detect OOD instances is decisive for ensuring the robust performance of machine learning models. OOD instances refer to situations where the model encounters data patterns or objects that differ significantly from what it was exposed to during training. Detecting OOD instances becomes particularly challenging when models are expected to generalize effectively across

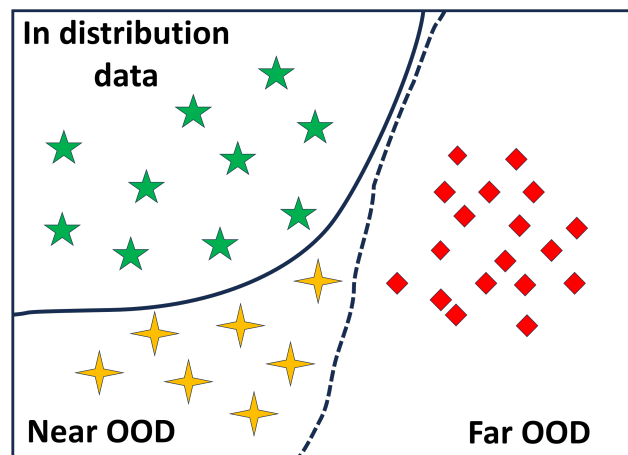


Figure 1. Out-of-Distribution definition, Dotted line represents the decision boundary of an OOD detection model that generalizes effectively.

diverse and unpredictable situations [22]. This adaptability is essential for ensuring the reliability and safety of machine learning models in dynamic and complex environments.

To address this classification challenge, this study explores the two types of OOD conditions, as depicted in Fig. 1: Near-OOD and Far-OOD [9]. In the context of a Near-OOD dataset, there is a notable resemblance in features and characteristics to the training dataset, also referred to as in-distribution (IN) data. The Near-OOD dataset may resemble datasets obtained from diverse acquisition sensors, e.g., when evaluating a model trained with the autonomous driving dataset KITTI [10], other autonomous driving datasets like BDD100K [23], Cityscapes [6] (2D bounding boxes from CityPersons [25]) or Lyft [11] (2D bounding boxes computed from 3D bounding boxes) can be categorized as Near-OOD. The Far-OOD dataset introduces a different paradigm, where the dataset is entirely dissimilar to the In-Distribution (IN) dataset, surpassing the characteristics

of Near-OOD conditions. Given the previous example, the CoCo [15] dataset would suffice for this Far-OOD condition. The OOD module integrated within the Deep Neural Networks (DNNs) for a specific application should avoid misclassifying Near-OOD instances as OOD while correctly flagging instances significantly different from the training dataset, i.e., Far-OOD, as the OOD cases.

With the transition from traditional Convolution Neural Networks (CNNs) to transformer-based models due to their remarkable ability to capture long-range dependencies and contextual information [19], a shift and redesign of OOD methods designed initially for CNNs is in progress.

Therefore, we showcase the OOD detection performance of our proposed Situation Monitor derived by leveraging the deviations of the ensemble predictions of the DBEA model integrated into the DINO-DETR [24] transformer-based vision model. DBEA is derived from the budding ensemble architecture proposed by [18]. The tandem loss function is hereby extended to incorporate a diversity loss function. Given that vision-based transformers commonly feature encoder-decoder structures. As a final stage of fully connected layers, it is generally applicable to integrate the Situation Monitor into various other vision transformer models without loss of generality.

In this work, we propose to:

- Create the DBEA model by incorporating diversity-based loss into the training process of BEA.
- Define the Situation Monitor to detect Far-OOD samples and suppress false positives on Near-OOD samples.
- Minimize the overhead of the transformer model with the Situation Monitor compared to baseline models.

Through comprehensive ablation study experiments, we evaluate the performance of DBEA in comparison to multiple sample-free OOD detection baselines. These baselines were specifically trained on well-established datasets such as KITTI and BDD100K, serving as benchmarks for our analysis.

2. Related Work

Researchers are continuously exploring innovative methodologies and refining existing techniques to bolster the capabilities of object detectors, particularly in handling out-of-distribution situations. Classifying Deep Neural Networks (DNNs) into deterministic and probabilistic networks provides a foundational understanding.

Deterministic Networks vs Probabilistic Networks: Deterministic networks in DNNs generate consistent outputs for a given input in a deterministic manner [8]. However, these networks cannot model prediction uncertainty. As a result, confidence scores associated with their predictions become crucial for measuring uncertainty, serving as valuable indicators for OOD detection [3]. In contrast, probabilistic DNNs explicitly model uncertainty in their predictions

by outputting probability distributions over possible outcomes [8]. This explicit modelling benefits OOD detection across a wide range of applications [1]. While OOD detection using probabilistic networks can be computationally demanding, researchers have devised various techniques for deterministic networks, primarily focusing on post-hoc methods [13, 14, 21].

Unified Frameworks for OOD Detection: In addition to post-hoc methods, a few unified frameworks incorporate OOD detection seamlessly into the primary task of the DNN. These frameworks utilize zero-shot or few-shot learning strategies to improve OOD detection capabilities [4, 7].

The findings from [12], indicating that wider networks with similar architecture learn more similar features, are integrated into the subsequent work of the sample-free uncertainty estimation method BEA [18]. BEA notably exhibited significant improvements in OOD detection, primarily focusing on anchor-based CNN models for object detection. It is worth noting that most of the related research is connected to CNN-based DNNs, and there is limited exploration into adapting and extending these techniques for transformer-based object detection models.

Sample-Free Uncertainty Estimation Method: Notably, BEA [18], a sample-free uncertainty estimation method, demonstrated considerable performance enhancements in OOD detection. This method primarily focused on anchor-based CNN models for object detection. Similarly, Gaussian-Yolov3 [5] introduced Gaussian parameters to exploit variances and reduce false positives by calibrating them to the IoU.

In summary, researchers are continually exploring new methodologies and refining existing techniques to enhance the capability of object detectors in handling out-of-distribution situations. This includes methods focused on both deterministic and probabilistic networks, as well as unified frameworks that seamlessly incorporate OOD detection into the primary task of the DNN. Additionally, there is potential for adapting and extending these techniques for transformer-based object detection models. However, most of this research is associated with CNN-based DNNs, and there is limited exploration into adapting and extending these techniques for transformer-based object detection models.

3. Problem Statement

In the closed-world assumption, the training dataset (\mathcal{D}) and testing dataset (\mathcal{T}) is from in-distribution dataset \mathcal{I} , i.e., $\mathcal{D} \subset \mathcal{I}$. Therefore, the samples from the testing dataset is $\mathcal{S}(\mathcal{T}) = \mathcal{S}(\mathcal{T}^{\mathcal{I}})$. However, in open-world settings and practical, real-world scenarios, samples are also drawn from OOD data. Therefore, the OOD samples \mathcal{O} are composed of both Near-OOD (\mathcal{O}^{near}) and Far-OOD (\mathcal{O}^{far}), i.e. $\mathcal{O} = \mathcal{O}^{near} + \mathcal{O}^{far}$. Similarly, in the open world setting, the testing data \mathcal{T} consists of known situations and classes originating




OOD Monitor			
	Known situation	Known situation	Unknown situation

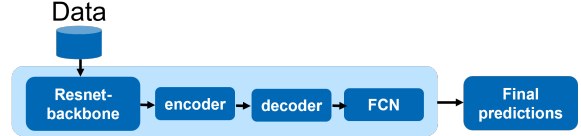
Figure 2. The primary aim of the Situation Monitor is to distinguish between known and unknown situations. For instance, a model trained on datasets such as KITTI for automotive scenarios is categorized as a Near-Out-of-Distribution (Near-OOD) situation. Consequently, encountering an indoor scenario would be labelled as Far-Out-of-Distribution (Far-OOD) situation by the model.

from Near-OOD \mathcal{O}^{near} and unknown situation and unknown classes to the model originating from Far-OOD \mathcal{O}^{far} i.e., $\mathcal{T} = \mathcal{T}^{\mathcal{I}} + \mathcal{T}^{\mathcal{O}^{near}} + \mathcal{T}^{\mathcal{O}^{far}}$ and accordingly the samples from the testing data is $\mathcal{S}(\mathcal{T}) = \mathcal{S}(\mathcal{T}^{\mathcal{I}}) + \mathcal{S}(\mathcal{T}^{\mathcal{O}^{near}}) + \mathcal{S}(\mathcal{T}^{\mathcal{O}^{far}})$.

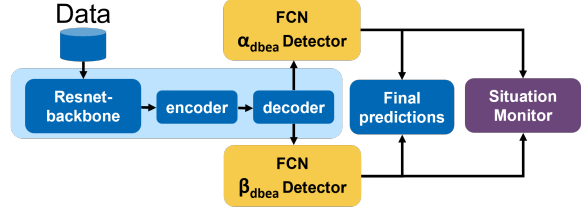
A model trained with \mathcal{D} is not only required to detect objects from $\mathcal{S}(\mathcal{T}^{\mathcal{I}})$ but it is also required to generalize well to $\mathcal{S}(\mathcal{T}^{\mathcal{O}^{near}})$. Therefore, a OOD detection module should not flag $\mathcal{S}(\mathcal{T}^{\mathcal{O}^{near}})$ as an OOD, rather only classify the samples from $\mathcal{S}(\mathcal{T}^{\mathcal{O}^{far}})$ as an OOD sample.

As shown in Fig. 2, the primary purpose of the Situation Monitor is to distinguish between known and unknown situations. Leveraging the remarkable generalization capability of deep learning models [17], the monitor adeptly classifies Near-OOD situations as known situations. This classification is grounded in the understanding that Near-OOD instances share significant similarities with the In-Distribution (IN) dataset, aligning with the inherent generalization prowess of deep learning models. Consequently, the Situation Monitor is pivotal in identifying situations based on their familiarity with the model’s learned context. By accurately discerning known and unknown situations, the monitor empowers the model to make informed decisions in real-world applications, even when encountering instances beyond its training data. This enhances the model’s reliability and performance in diverse and dynamic environments. Overall, the Situation Monitor plays a crucial role in ensuring the effectiveness of deep learning models across various situations.

In summary, our goal is to train a transformer-based object detection model on a training set \mathcal{D} and without introducing the samples from \mathcal{O} , the Situation Monitor (OOD detection module) of the model should have ability to classify the samples from $\mathcal{S}(\mathcal{T}^{\mathcal{O}^{far}})$ as an OOD situation and on contrary should not flag the $\mathcal{S}(\mathcal{T}^{\mathcal{O}^{near}})$ samples as an OOD situation.



(a) Abstract DINO-DETR architecture



(b) Abstract DBEA-DINO-DETR architecture

Figure 3. Adaptation of BEA [18] to DINO-DETR [24].

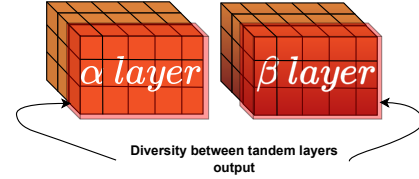


Figure 4. Within the BEA model, tandem detectors undergo further refinement with diversity loss, aiming to enhance the distinction between α and β detectors. This leads to the development of Diversity-based BEA.

4. Our Approach

4.1. DBEA: Diversity based Budding Ensemble Architecture

The recently introduced Budding Ensemble Architecture (BEA) [18] represents a sample-free methodology for detecting OOD instances, showcasing notable performance advancements. Consequently, we employ the Budding Ensemble Architecture approach in the design of our Situation Monitor.

Within the Budding Ensemble Architecture (BEA) framework, a unified backbone and duplicated detectors replace the conventional ensemble setup. This alteration enhances confidence score calibration, diminishes uncertainty errors, and introduces an overlooked advantage: superior OOD detection compared to other state-of-the-art sample-free methods. The **Tandem loss** function ($\mathcal{L}_{\text{tandem}}$) introduced in BEA, devised initially for YOLOV3 [20] and SSD [16] was primarily tailored for anchor-based object detection models. Therefore, the $\mathcal{L}_{\text{tandem}}$ function to be applied to the transformer model requires a series of modifications and adaptations to integrate with this novel ensemble approach seamlessly.

We base our Situation Monitor design on the ResNet-based DINO-DETR transformer model. It typically comprises a CNN-based backbone (ResNet), self-attention layer-based encoders, and decoders, followed by fully connected

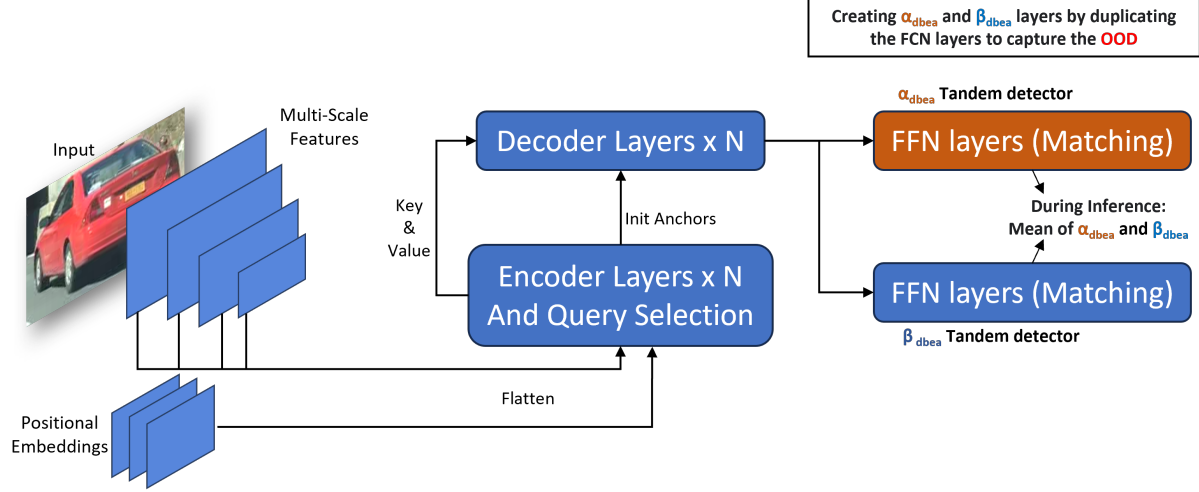


Figure 5. Adapted architecture diagram from DINO-DETR [24] for DBEA-DINO-DETR, illustrating the replication of final layers. For a comprehensive understanding of DINO-DETR, please refer to Figure 2 in [24]. In BEA, it is proposed to duplicate the final layers as α and β to create tandem detectors, which are subsequently utilized in the situation monitoring model.

regression layers for classification and bounding box detection. We leverage the insights from BEA to tailor the DINO-DETR object detection model for our Situation Monitor. Specifically, we introduce **diversity-based tandem detection layers**, as depicted in Fig. 3(b). In the architectural integration of BEA into DINO-DETR, the 3 layer feed-forward neural network layers (FFN) forming the final regression layers immediately following the decoder are duplicated, resulting in two detectors, α and β collectively called as tandem detectors. During inference, confidence scores and bounding boxes are computed from the mean of the tandem detectors.

Despite the duplication of layers, as indicated in the BEA paper [18], no advantages were observed for the Situation Monitor during the training of the DINO-DETR model. This lack of advantage stems from the tandem layers learning similar representations. Therefore, the incorporation of the tandem loss function, $\mathcal{L}_{\text{tandem}}$, alongside the original base loss function, $\mathcal{L}_{\text{base}}$, becomes crucial. The adaptation of BEA with the diversity-based $\mathcal{L}_{\text{tandem}}$ function is called Diversity-based Budding Ensemble Architecture (DBEA). This serves a dual purpose: creating an effective Situation Monitor and enhancing the calibration of confidence scores. The adapted tandem loss is now defined as follows:

$$\mathcal{L}_{\text{ta}}(\phi) = \mathbb{1}_i^{\text{obj}} \sqrt{(\phi_i^\alpha - \phi_i^\beta)^2}, \quad (1)$$

$$\mathcal{L}_{\text{ta}} = \mathcal{L}_{\text{ta}}(\mathbf{x}) + \mathcal{L}_{\text{ta}}(\mathbf{y}) + \mathcal{L}_{\text{ta}}(\mathbf{w}) + \mathcal{L}_{\text{ta}}(\mathbf{h})$$

$$\mathcal{L}_{\text{tq}}(\phi) = \mathbb{1}_i^{\text{noobj}} \frac{1}{\sqrt{(\phi_i^\alpha - \phi_i^\beta)^2}}, \quad (2)$$

$$\mathcal{L}_{\text{tq}} = \mathcal{L}_{\text{tq}}(\mathbf{x}) + \mathcal{L}_{\text{tq}}(\mathbf{y}) + \mathcal{L}_{\text{tq}}(\mathbf{w}) + \mathcal{L}_{\text{tq}}(\mathbf{h})$$

$$\mathcal{L}_{\text{tandem}} = \lambda_{\text{ta}} \mathcal{L}_{\text{ta}} + \lambda_{\text{tq}} \mathcal{L}_{\text{tq}} \quad (3)$$

where $\mathbb{1}_i$ denotes whether the object prediction overlaps with the ground-truth. The variables \mathbf{x} and \mathbf{y} signify the predicted centre points of the bounding box, while w and h represent the predicted height and width of the object, respectively. The $\mathcal{L}_{\text{tandem}}$ operates on positive predictions and negative predictions, which is possible due to access to ground truth during training. The **Tandem-Aiding loss** function \mathcal{L}_{ta} diminishes the errors associated with the positive predictions between α and β detector. Similarly, the **Tandem-Quelling loss** function \mathcal{L}_{tq} amplifies the errors related to negative predictions between α and β detectors. Therefore promoting agreement and disagreement concerning positive and negative predictions.

Within the BEA paper, the \mathcal{L}_{ta} and \mathcal{L}_{tq} loss functions are applied independently to specific components of both classification and bounding box regression outputs. In the context of the transformer model, the $\mathcal{L}_{\text{tandem}}$ is exclusively applied to the bounding box regression layers as shown in Eq. (1) and (2). This selective application is adopted because introducing this loss to the classification layer decreases the performance of the Situation Monitor. To support $\mathcal{L}_{\text{tandem}}$ loss function for better calibration of the confidence scores, we introduce diversity at classification regression layer output between the tandem layers as shown in Fig. 4. To this end, we use similarity score as a diversity measure to ensure that the tandem detectors capture distinct representations from the decoders. Specifically, we use the cosine similarity score to introduce the diversity as depicted in Eq. (4).

$$\begin{aligned} \text{Cosine_Similarity}(\alpha, \beta) &= \frac{\alpha \cdot \beta}{\|\alpha\| \cdot \|\beta\|}; \alpha, \beta \in Z \\ \mathcal{L}_{\text{diversity}} &= \frac{1}{n} \sum_{i=1}^n \text{Cosine_Similarity}(\alpha_i, \beta_i) \end{aligned} \quad (4)$$

where n represents total number of predictions and Z represents the classification logits denoted as $Z = (z_1, z_2, \dots, z_k)$, z_i is the unnormalized score for class i .

Incorporating diversity introduces a dynamic range of classification outputs at the tandem layers. This diversity is instrumental in fostering a broader spectrum of values, thereby providing the tandem loss function with a larger space to operate upon. Specifically, the tandem loss function leverages this diversity to mitigate errors associated with positive predictions effectively. By encouraging divergence in classification values, the model can better discern and refine its understanding of positive instances. Simultaneously, this diversity amplifies errors in the context of negative predictions. To this end, the diversity and tandem-based loss function is constructed by incorporating both $\mathcal{L}_{\text{tandem}}$ and $\mathcal{L}_{\text{diversity}}$ into the base loss function of the DINO-DETR model, denoted as $\mathcal{L}_{\text{base}}$. This integration is illustrated in Equation (5).

$$\mathcal{L}_{\text{dbea}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{tandem}} + \lambda_{\text{div}} \mathcal{L}_{\text{diversity}} \quad (5)$$

The training hyper-parameters of the DBEA-DINO-DETR remain unchanged except the λ parameters introduced in Eq. (3) and (5).

4.2. Situation Monitor

The Situation Monitor is a component of the BEA-DINO-DETR that serves as an Out-of-Distribution (OOD) detection module. It identifies Far-OOD situations by analyzing disparities in the predictions of tandem layer bounding boxes. The transformer model undergoes end-to-end training through a zero-shot approach, which, in this case, means an explicit OOD dataset is not shown during the training process. This methodology distinguishes explicit situations by highlighting the errors in variance between the tandem layer predictions, specifically in Far-OOD situations. The $\mathcal{L}_{\text{dbea}}$ loss function, incorporating cosine diversity during training, compels tandem detection layers to consider objects from diverse feature maps and perspectives.

The introduction of diversity loss ($\mathcal{L}_{\text{diversity}}$) during training prompts a change in perspective, compelling tandem detection layers to examine objects from diverse feature maps and viewpoints. Driven by the $\mathcal{L}_{\text{tandem}}$ loss function, tandem layers generate distinct predictions for bounding box centre points (Fig. 6). In instances where the set \mathcal{T} is a subset of \mathcal{I} within the specific category $\mathcal{O}^{\text{near}}$, it is observed that the predicted widths and heights tend to be closely aligned

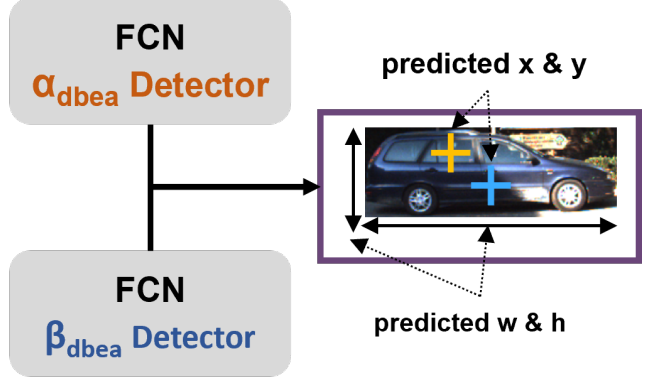


Figure 6. Situation Monitor: The variance between α_{DBEA} and β_{DBEA} detector predictions is interpreted as the prediction uncertainty \mathcal{U}_{SM} . A high uncertainty means Far-OOD, whereas low uncertainty means Near-OOD or in-distribution.

with no large variances. Similarly, we should expect large variances when \mathcal{O}^{far} is encountered. Therefore, to capture Far-OOD situations effectively at the image level, a methodology is applied to heighten the variances of predicted centre points (x and y) and heights and widths (w and h). These bounding box-related attributes are referred to as b in the following equation.

$$\begin{aligned} xy_{\text{var}} &= \sqrt{\sum_{i \in (x,y)} (b_i^\alpha - b_i^\beta)^2}, \\ xy_{\text{centered_var}} &= xy_{\text{var}} * \mu(xy_{\text{var}}), \\ wh_{\text{var}} &= \sum_{i \in (w,h)} (b_i^\alpha - b_i^\beta)^2, \\ wh_{\text{centered_var}} &= wh_{\text{var}} * \mu(wh_{\text{var}}), \\ \mathcal{U}_{SM} &= \mu(\sqrt{xy_{\text{centered_var}} * wh_{\text{centered_var}}}) \end{aligned} \quad (6)$$

To heighten sensitivity to deviations and pinpoint errors at an image-specific level, these variances undergo additional processing as shown in Eq. (6). This involves centring variances by multiplication with their respective means. The final OOD value, denoted as \mathcal{U}_{SM} , is computed as the mean of these variances at the image level, as depicted in Eq. (6). This approach effectively addresses global trends across the entire dataset while localizing errors to individual images. By doing so, anomalies unique to each image are identified, enhancing performance in capturing Far-OOD situations.

5. Evaluation Metrics for Situation Monitor

The impact of adding $\mathcal{L}_{\text{tandem}}$ and $\mathcal{L}_{\text{diversity}}$ to the $\mathcal{L}_{\text{base}}$ loss function is evaluated using **mean average precision (mAP)** metrics. The calibration is evaluated using **Pearson correlation (PCorr)** between confidence scores and the intersection over union of objects and the ground truths. Pearson correlation is evaluated on two sets **PCorr-all** and

PCorr-tp which is correlation on all the predictions and separately on true positive samples. The Situation Monitor in a DBEA model uses \mathcal{U}_{SM} values and it is assessed using four standard metrics to evaluate its performance in detecting OOD situations. Whereas, the vanilla baseline model’s OOD detection is evaluated using their confidence scores. In this study, we do not assess the object detection performance (mAP/AP) on Near-OOD datasets. Instead, we simply demonstrate that the Situation Monitor does not identify Near-OOD samples as OOD.

- **AUROC** calculates the area under the receiver operating characteristic curve, a metric utilized to assess the performance of OOD detection. The OOD ($\mathcal{T}(\mathcal{T}^{O^{far}})$) samples are considered as positive samples. A higher AUROC value indicates superior performance.
- **AUPR(In/Out)** is the area under the receiver operating characteristic curve and is a key metric for evaluating OOD detection performance. It assesses how well a model distinguishes positive instances, with $\mathcal{T}(\mathcal{T}^{O^{far}})$ samples considered as positives. AUPR comprises of AUPR-In and AUPR-Out. It considers the in-distribution ($\mathcal{T}(\mathcal{T}^{\mathcal{I}})$) samples as either positive or negative respectively. A higher AUPR value indicates superior model performance.
- **FPR@95** expressed as FPR (false positive rate) at a fixed TPR (true positive rate) point represents the rate of falsely identified positive instances among all negative instances when the true positive rate is held at a specific percentage which in this case at 95%. The lower value indicates superior performance.
- **DE@95** is the detection error at 95% TPR quantifies the detection error (miss-classification probability) when TPR is set at 95%. A lower value indicates superior performance.

6. Experiments

In this section, we conduct an ablation study to analyze the impact of each component. We then compare our method with the previous state-of-the-art sample-free OOD detection approach.

6.1. Experiment Setup

The Situation Monitor (SM) is integrated into the DBEA-DINO-DETR object detection model, allowing end-to-end training without freezing the backbone or any particular layer. This ensures tandem detectors learn based on the \mathcal{L}_{dbea} loss function. Evaluations are conducted on the KITTI [10] and BDD100K [23] datasets, widely used computer vision datasets for autonomous driving situations. These datasets are divided into training, validation, and testing sets with ratios of 7.5:1:1.5 and 8.5:0.75:0.75, respectively. Eight out of nine usable classes are evaluated for the KITTI dataset, while all ten classes are considered for the BDD100K dataset.

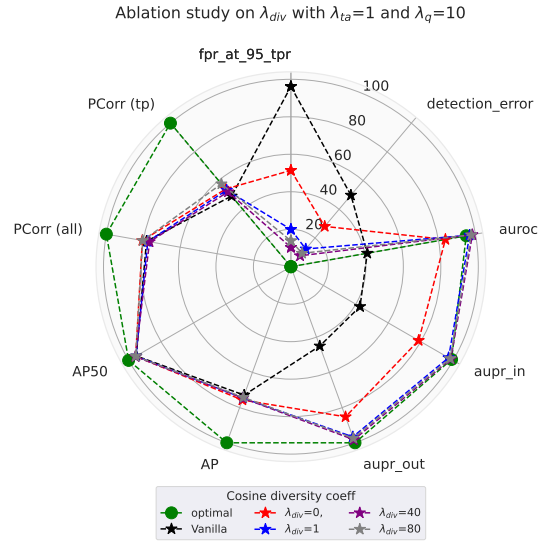


Figure 7. Analysis of the impact of varying parameter \mathcal{L}_{div} in the ablation study.

Additionally, CoCo’s evaluation dataset is utilized. A consistent 416×416 input image size is used for training across all models. For YoloV3 [20] and SSD [16] models, SGD optimizer is employed with a learning rate of 0.001, momentum of 0.95, and weight decay of 0.001, trained for 300 epochs with batch-size of 28. DINO-DETR-based models utilize 900 queries, generating 900 predictions per image, with the top 100 predictions considered. ResNet-50-based transformer models use an AdamW optimizer with a weight decay of 0.0001 and a learning rate 0.0002, trained for 50 epochs with a batch size of 5. Both the baseline and our approach share identical training hyperparameters, ensuring consistency.

6.2. Ablation Study on λ_{div} , λ_{ta} and λ_{tq} on KITTI trained DBEA-DINO-DETR

In this ablation study section, multiple DBEA-DINO-DETR models are trained on the KITTI dataset wherein the parameters outlined in equations 5 and 3 are varied. The primary objective is systematically showcasing and analysing each component’s distinctive impact. In all ablation experiments, the parameter λ_{ta} is consistently maintained at either one when necessary or at zero. This is because \mathcal{L}_{ta} is designed to minimize errors in positive predictions, and an increase in the λ_{ta} factor beyond λ_{tq} adversely affects calibration and the fundamental object detection capability of the model.

Fig. 7 illustrates the impact of introducing \mathcal{L}_{div} with varied λ_{div} parameters on the Situation Monitor performance and its OOD detection. The radar plots presented in this study will include the outcomes of a baseline model (the “vanilla model”) and reference optimal values against which the metrics will be compared. Without enabling diversity, the

Model	mAP (%) ↑	AP50 (%) ↑	PCorr(all) (%) ↑	PCorr(tp) (%) ↑	Out-of-distribution detection on COCO dataset			
					AUROC (%) ↑	AUPR (In/Out) (%) ↑	FPR@95 (%) ↓	DE@95 (%) ↓
KITTI trained								
Vanilla-YOLOv3	45.3	87.4	80.3	45.5	28.5	63.4/17.2	95.4	74.4
BEA-YOLOv3 [18]	54.8	89.3	80.8	45.6	91.7	90.5/92.5	33.6	18.7
Vanilla-SSD	62.6	88.6	76.2	56.2	35.7	44.3/37.6	96.5	55.3
BEA-SSD [18]	63.1	89.6	74.5	54.4	91.6	92.6/90.7	57.6	30.6
Vanilla-DINO-DETR	72.9	95.2	78.2	49.2	41.4	42.6/45.0	96.1	49.8
BEA-DINO-DETR	73.8	95.8	79.1	52.0	92.4	93.1/94.4	15.1	9.6
DBEA-DINO-DETR (ours)	74.6	95.8	79.6	54.2	98.3	98.5/98.3	10.3	7.6
BDD100K trained								
Vanilla-YOLOv3	25.0	53.7	69.3	38.7	22.0	27.4/43.6	98.7	40.9
BEA-YOLOv3 [18]	27.6	58.1	69.0	36.0	96.5	97.5/91.8	12.3	8.5
Vanilla-DINO-DETR	47.3	84.0	71.1	37.4	25.3	36.5/35.9	99.8	49.6
DBEA-DINO-DETR (ours)	47.9	84.6	78.9	48.3	99.6	99.7/99.6	1.2	2.2

Table 1. Comparison of the performance of our sample-free method with that of previous state-of-the-art sample-free method. Our method is trained on the KITTI and BDD100K datasets, and the out-of-distribution detection evaluation is conducted on the COCO dataset instances.

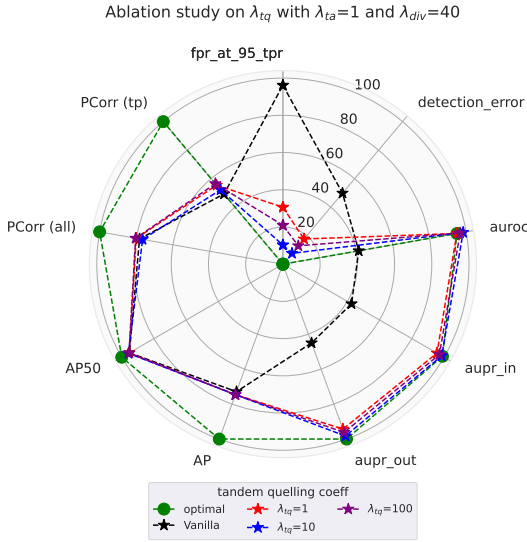


Figure 8. Analysis of the impact of varying parameter \mathcal{L}_{tandem} in the ablation study.

performance on the Far-OOD CoCo dataset experiences a significant decline compared to its enabled counterpart. The optimal configuration is identified when the λ_{div} parameter is set to 40, resulting in elevated Average Precision metrics, as well as improved AUROC and AUPR OOD metrics, while concurrently maintaining lower FPR@95 and detection error (DE@95) values. A subsequent increase of λ_{div} from 40 to 80 brings about enhancements across all metrics, with only a minor impact on detection error and a slight decrease in FPR@95. Hence, choosing 40 as the parameter proves to be the most effective, achieving a balance across all metrics. The selection of λ_{div} , incremented by a factor of 2, is influenced by the consideration that both the classification and GIOU loss [2] components of \mathcal{L}_{base} are scaled by a factor

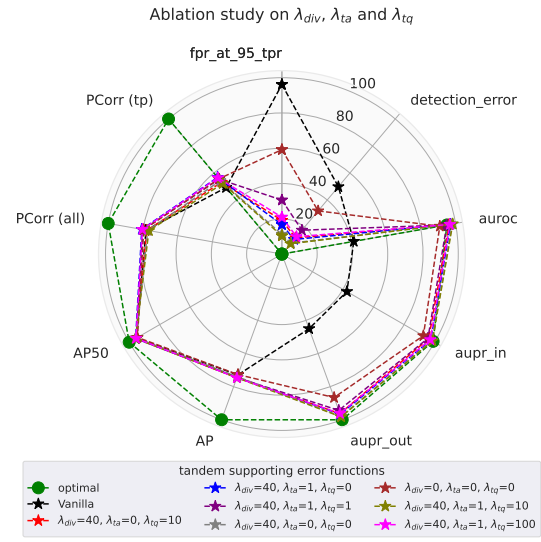


Figure 9. Analysis of the impact of varying parameter \mathcal{L}_{tq} in the ablation study.

of 2, making factors of 2 more effective to the effectiveness of λ_{div} .

Similarly, the influence of λ_{tq} is illustrated in Fig. 8, with λ_{ta} and λ_{div} held at 1 and 40, respectively. This specific radar plot highlights that setting λ_{tq} to 10 produces the most well-balanced results in comparison to other factors. While the transformer model exhibits commendable calibration when trained with a λ_{tq} factor of 100, there is a slight decline in the Situation Monitor OOD performance.

Fig. 9 illustrates the results of comprehensive ablation experiments encompassing all three factors: λ_{div} , λ_{ta} , and λ_{tq} . In the absence of diversity and tandem loss, the model exhibits comparable Average Precision (AP) performance but experiences a notable decline in the FPR@95 and DE@95

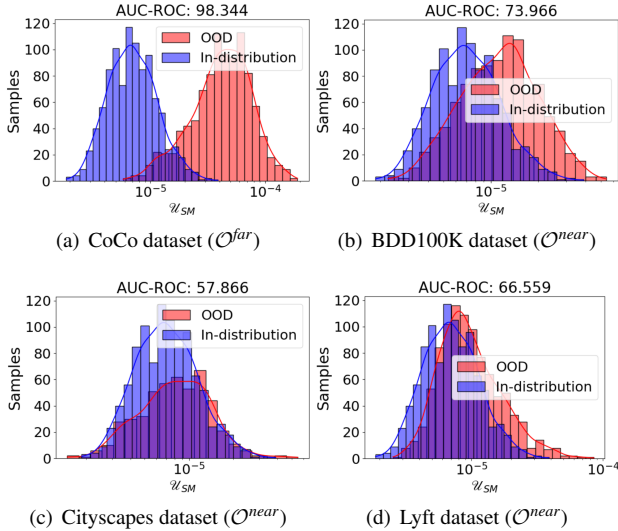


Figure 10. OOD detection performance of KITTI trained DBEA-DINO-DETR model on \mathcal{O}^{far} and \mathcal{O}^{near} datasets. The Situation Monitor of the DBEA-DINO-DETR model can effectively flag Far-OOD situations.

metrics. Notably, when the model is trained with specific parameters, namely $\lambda_{div} = 40$, $\lambda_{ta} = 1$, and $\lambda_{tq} = 10$, it achieves optimal performance, manifesting in heightened accuracy and superior OOD detection capabilities. The configuration resulting in lower FPR@95 and DE@95 metrics, indicates enhanced OOD detection performance. The improved threshold associated with these parameter values contributes to superior classification, thereby leading to better OOD detection.

6.3. Benchmark Results

Tab. 1 comprehensively compares baseline models (CNN and transformer-based) with state-of-the-art sample-free OOD detection methods. Using various metrics discussed in Section 5, we assess the Situation Monitor’s overall object detection and OOD capabilities integrated into the DBEA-DINO-DETR model. Yolov3, SSD and DINO baseline models are trained using original hyperparameters, and the reported results in Tab. 1 reflect their peak performance. Applying our approach to the DINO-DETR model resulted in the DBEA-DINO-DETR model. Training it on datasets KITTI and BDD100K showcases enhanced detection accuracy and improved correlation between predicted confidence scores and intersection over union. On the BDD100K dataset, our DBEA-based DINO-DETR model demonstrates improved correlation in overall and true positive predictions. Additionally, there is a substantial enhancement in the Out-of-Distribution (OOD) performance for detecting CoCo images. Specifically, the OOD detection performance for CoCo images is higher when the model is trained with BDD100K than the KITTI model. This highlights our approach’s effectiveness in advancing object detection and

OOD performance, especially on larger datasets. Illustrated in Fig. 10(a), the Situation Monitor adeptly identifies \mathcal{O}^{far} OOD samples. Furthermore, the Situation Monitor demonstrates strong generalization to \mathcal{O}^{near} datasets, evident in overlapping histograms of U_{SM} values Fig. 10(b), 10(c) and 10(d). In contrast to the Situation Monitor introduced in this study, the OOD values of BEA-based CNN models [18] tended to miss-classify even \mathcal{O}^{near} as \mathcal{O}^{far} in their OOD detection.

6.4. Overhead Analysis

Given the computational intensity of transformers relative to the CNNs, the DBEA adaptation of transformers incurs additional costs due to the duplication of final regression layers. To mitigate this overhead, we limit the feed-forward channels in both the encoder and decoder layers (N) of DBEA-DINO-DETR from 2048 to 1024, offsetting the overhead of duplicating the final regression layers to create the tandem layers. The Vanilla model trained on the KITTI dataset with N=2048 has a size of 48M, while our DBEA-DINO-DETR is only 42M, representing a 14% reduction compared to the Vanilla model. We also observed that there is little but negligible advantage in maintaining the same number of feed-forward channels of 2048 as the vanilla model. The results of the Situation Monitor presented in this section are based on the DBEA model trained with 1024 feed-forward channels.

7. Conclusion

In summary, this paper introduces a Situation Monitor driven by zero-shot learning, which integrates a novel Diversity-based Budding Ensemble Architecture (DBEA) loss function. The incorporation of the DBEA loss function empowers the object detection model to not only identify Far-OOD samples but also to generalize over similar Near-OOD instances effectively. This prevents miss-classification as OOD, enhancing the model’s adaptability and robustness. Through an extensive ablation study with the parameters of DBEA’s loss functions, we show significant improvement in detection accuracy and superior OOD detection outcomes compared to baseline models and other existing methods. Additionally, our research demonstrates the scalability of the DBEA-based model, validated through successful training with KITTI and BDD100K datasets. The Situation Monitor is suitable for safety-critical applications, being 14% less computationally intensive than the baseline DINO-DETR model.

Acknowledgements

This work was partially funded by the Federal Ministry for Economic Affairs and Energy of Germany as part of the research project SafeWahr (Grant Number: 19A21026C).

References

- [1] Nilesh A Ahuja, Ibrahim Ndiour, Trushant Kalyanpur, and Omesh Tickoo. Probabilistic modeling of deep features for out-of-distribution and adversarial detection. *arXiv preprint arXiv:1909.11786*, 2019. [2](#)
- [2] Alexandre B. Araujo, Allan G. Oliveira, and Claudia R. Jung. Giou loss for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4894–4902, 2019. [7](#)
- [3] Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. Confidence-based out-of-distribution detection: a comparative study and analysis. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*, pages 122–132. Springer, 2021. [2](#)
- [4] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European conference on computer vision*, pages 572–588. Springer, 2020. [2](#)
- [5] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 502–511, 2019. [2](#)
- [6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. [1](#)
- [7] Sepideh Esmailpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6568–6576, 2022. [2](#)
- [8] Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, 2021. [2](#)
- [9] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. [1](#)
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. [1](#), [6](#)
- [11] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Woven planet perception dataset 2020. <https://woven.toyota/en/perception-dataset>, 2019. [1](#)
- [12] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. [2](#)
- [13] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018. [2](#)
- [14] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. [2](#)
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. [2](#)
- [16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. [3](#), [6](#)
- [17] Francesco Pinto, Philip HS Torr, and Puneet K. Dokania. An impartial take to the cnn vs transformer robustness contest. In *European Conference on Computer Vision*, pages 466–480. Springer, 2022. [3](#)
- [18] Syed Sha Qutub, Neslihan Kose, Rafael Rosales, Michael Paulitsch, Korbinian Hagn, Florian Geissler, Yang Peng, Gereon Hinz, and Alois Knoll. Bea: Revisiting anchor-based object detection dnn using budding ensemble architecture. 2023. [2](#), [3](#), [4](#), [7](#), [8](#)
- [19] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. [2](#)
- [20] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. [3](#), [6](#)
- [21] Samuel Wilson, Tobias Fischer, Feras Dayoub, Dimity Miller, and Niko Sünderhauf. Safe: Sensitivity-aware features for out-of-distribution object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23565–23576, 2023. [2](#)
- [22] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. [1](#)
- [23] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. [1](#), [6](#)
- [24] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. [2](#), [3](#), [4](#)

- [25] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. [1](#)