Operational Readiness for Object Detection

Stefan Becker[1], Jens Bayer[1], Wolfgang Huebner[1], and Michael Arens[1]

Fraunhofer IOSB, Ettlingen, Germany
Fraunhofer Center for Machine Learning
{firstname.lastname}@iosb.fraunhofer.de

Abstract. Reliable object detection in safety-critical applications requires models that can recognize their own failures. This paper introduces a self-assessment approach for monitoring the *operational readiness* (OR) of object detectors under distributional shift. Unlike existing methods that focus on uncertainty estimation or generative models, we propose assessing OR by comparing the *cumulative distribution functions* (CDFs) of feature map activations from multiple layers of a frozen detector. This approach captures both low- and high-level feature information, offering a simple and effective alternative to existing strategies. We formalize operational readiness as estimating the in-distribution probability of a given input image, which serves as a proxy for data uncertainty. Our method is systematically compared to explicit generative modeling with normalizing flows and aggregation of detector uncertainties. Experiments on COCO with synthetic covariate shifts show that our activation-distribution approach outperforms these baselines. The method is detector-agnostic and can be implemented as an add-on module for pre-trained detectors.

Keywords: Operational Readiness · Out-of-Distribution · Object Detection · Image Corruption.

1 Introduction

Object detectors are increasingly deployed in safety-critical applications such as autonomous driving [26,6], where failures can lead to catastrophic outcomes. To mitigate risk, deployed systems need runtime monitors that can flag inputs under which detector outputs are unreliable. We refer to this capability as *operational readiness* (OR). Following the standard setup in recent work [15], we assume no unsafe samples are available during monitor training. For object detection, we write the operational decision using a data-side readiness term:

$$P(\mathbf{y} \mid \mathbf{x}, \mathcal{D})_{\mathrm{OR}} = P(\mathbf{y}|\mathbf{x}, \mathcal{D}) \cdot P_{\mathrm{OR}}(\mathbf{x}) \tag{1}$$

where $P_{\mathrm{OR}}(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to the distribution of safe inputs, $\mathbf{y}$ denotes the detector output, and $\mathcal{D}$ is the training set.

OR is broader than *out-of-distribution* (OOD) detection: an input can be *in-distribution* (ID) yet still cause errors (e.g., extreme occlusion), and some shifted

inputs may remain safe. However, we must decide whether to trust the detector before relying on its outputs, and error-labeled unsafe examples are unavailable. We therefore adopt the ID probability $P_{\text{ID}}(\mathbf{x})$ as a necessary practical, but not sufficient, proxy for the data-side component of OR, i.e., replacing $P_{\text{OR}}(\mathbf{x})$. This focuses the monitor on detecting covariate shift in $p_{\text{ID}}(\mathbf{x})$, complementing model-side uncertainty and task-specific checks.

Most OOD methods were designed for classification and transfer imperfectly to detection due to dense, structured outputs. OOD methods applicable for detection mainly fall into two categories. First, leveraging detector uncertainty (e.g., [17]) and generative approaches that build explicit density models to learn $p_{\text{ID}}(\mathbf{x})$. While detector uncertainty may implicitly capture data uncertainty, it primarily reflects model uncertainty and is therefore not ideal for OOD detection. Generative models, despite empirical improvements for OOD detection [5,23,51], can assign high likelihoods to OOD samples.

Our key idea is a simple, detector-agnostic input gate based on the distribution of neuron activations across multiple layers of a frozen detector. Inspired by [49], we summarize each layer by histograms of channel activations and compare their *cumulative distribution functions* (CDFs) to reference CDFs aggregated from training data. This captures low- and high-level cues and yields an efficient image-level shift score. We systematically compare this activation-distribution monitor to *normalizing flow*-based generative models trained on detector activations and top-k aggregation of detection uncertainties from probabilistic detectors. We evaluate on COCO [38] under synthetic covariate shifts. Despite its simplicity, the activation-distribution approach consistently outperforms flow-based and detection-uncertainty baselines for ID vs. OOD separation. The monitor is lightweight and plug-in, requiring no detector modifications.

Our contributions are: We motivate and formalize using $P_{\text{ID}}(\mathbf{x})$ as a necessary, data-side proxy for OR in detection, under the constraint of no unsafe labels at training time. We propose a detector-agnostic activation-distribution monitor that compares multi-layer activation CDFs to training-set references. We provide a systematic comparison against *normalizing flows* on detector features and top-k detection-uncertainty aggregation, showing consistent gains under covariate shift.

Section 2 situates our work within OOD detection, uncertainty estimation, and OR monitoring. Section 3 details the approaches for $P_{\text{ID}}(x)$ approximations. Section 4 presents experiments and results. Section 5 discusses limitations, and Section 6 concludes.

2 Related Work

Methodologically, OOD approaches can be applied to OR, but most were designed for classification, where each input yields a single output label. Classification-focused OOD methods typically operate on penultimate-layer features or logits and are not transferable to detection, which produces dense, structured outputs and a more complex failure landscape. Notable examples include *maximum*

softmax probability MSP [22], ODIN [36], *GradNorm* [25], *ReAct* [52] or *energy-based* approaches [39]. In object detection, however, classification logits live at the detection level (per anchor/proposal/query). As a result, these OOD methods primarily diagnose missing or novel classes (open-set concerns) (e.g., [8]) for specific detections and do not provide a reliable image-level signal of data-side readiness. We also leave out rule based or outlier monitors which address shape assertions (e.g., [28,4]).

We distinguish OOD from several related areas, such as *anomaly detection* (AD), *novelty detection* (ND), *open set recognition* (OSR), and *outlier detection* (OD), which differ in objectives and evaluation. Our focus is covariate shift in $p_{\text{ID}}(\mathbf{x})$ (e.g., corruptions) that degrades detector reliability [48,47]. For completeness, covariate shift may also arise from adversarial examples [14,42] or style transfer [11]. For a unified framework of *generalized* OOD detection with a taxonomy along classification tasks, see Yang et al. [57].

Among OOD methods applicable to detection, generative models estimate likelihood under a learned data distribution. *Variational auto encoders* (VAEs) [30] use the *evidence lower bound* (ELBO). *Normalizing flows* (NFs) [32] provide exact, tractable likelihoods, which makes them an attractive alternative for the OOD task. While image-level likelihoods can be unreliable for OOD (e.g., NF pitfalls [31]), recent work improves performance by modeling feature activations from multiple layers rather than raw pixels [53,40]. We follow this feature-space direction and adopt NFs as a reference baseline, training them on detector activations to capture multi-scale cues and complement detection-level uncertainty aggregation.

Detection-level uncertainty relies on probabilistic detectors to estimate predictive distributions. Exact Bayesian inference is intractable [41,43]. Practical approximations include MC dropout [10] and variance networks [29]. Although popular, MC dropout is computationally expensive and has been criticized for producing sub-optimal uncertainty estimates in certain settings [46,33]. Another strategy is to extend standard object detectors with variance networks to directly estimate uncertainty [29] by augmenting detectors to predict parameters of a multivariate Gaussian predictive distribution. These networks are typically trained using scoring rules such as *negative log-likelihood* (NLL), *energy score* (ES) [13], and *direct moment matching* DMM [9], and have been successfully employed in works like [20,19,17]. For a broader review of probabilistic methods in deep learning, we refer to these works [12,1], and for a comprehensive overview with a focus on object detection, see [16]. However, detection-level uncertainty alone is often insufficient to capture global anomalies or broader reliability signals. Building on [45], we aggregate uncertainty scores from top-k detections to obtain a coarse image-level uncertainty estimate to approximate $P_{ID}(\mathbf{x})$.

3 Operational Readiness

In this section, we present approaches for estimating $P_{\text{ID}}(\mathbf{x})$ as a proxy for OR in object detection, as formalized in(Eq. (1)). Each method quantifies data

uncertainty from a distinct perspective: modeling global activation statistics, learning a feature-space density, or aggregating detection-level uncertainties. The acceptance decision uses an indicator function to assign a binary score $\hat{or} \in \{0, 1\} : \hat{or} = \mathbb{I}[P_{\text{ID}}(\mathbf{x}) < \tau]$ where τ is a threshold. For distance-based monitors, we use the same convention by interpreting P_{ID} as a shift score.

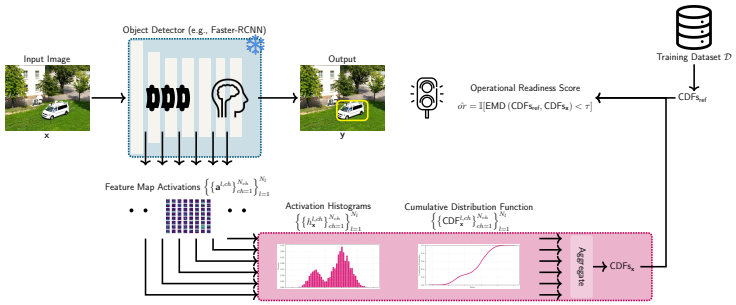


Fig. 1: Architecture of the proposed approach for *operational readiness* (OR) estimation. Feature map activations extracted at different layers l (N_l: number of layers) of the (frozen) object detector are used to capture activation statistics in the form of histograms $h_{\mathbf{x}}^{l,ch}$ per channel ch (N_{ch}: number of channels). Then the *cumulative distribution functions* (CDFs) are calculated and compared to the reference CDFs_{ref} from the training dataset with *Earth Mover's Distance* (EMD) to estimate the *operational readiness* score $\hat{or}$.

Activation-Distribution: The core idea is to represent the ID data by *cumulative distribution functions* (CDFs) of feature map activations across different layers of an object detector. This concept is inspired by Ramtoula et al.[49], who proposed to use histograms of neuron activations as a memory-efficient way to represent datasets.

To collect the reference CDFs, we use the training data $\mathcal{D}$ from a pre-trained object detector. Let $\mathcal{L}$ be a set of selected layers, each with N_{ch}^l channels. After a forward pass of an image $\mathbf{x}$ through the detector, each layer $l \in \mathcal{L}$ produces an activation feature map $\mathbf{a}^l$ with spatial dimensions $S_h^l \times S_w^l$. For each channel ch, we summarize the distribution of activations by constructing a histogram $h_{\mathbf{x}}^{ch} \in \mathbb{N}_{ch}^{N_b}$ with N_b bins, where bin edges are $\{b_0, b_1, \ldots, b_{N_b}\}$.

The count for bin k is:

$$h_{\mathbf{x}}^{ch}[k] = \sum_{s_h=1}^{S_h^l} \sum_{s_w=1}^{S_w^l} \begin{cases} 1, & \text{if } \mathbf{a}^{l,ch}[ch, s_h, s_w] \in [b_k, b_{k+1}] \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The reference histogram for channel ch is $\mathcal{H}_{\text{ref}}^{ch}[k] = \sum_{\mathbf{x} \in \mathcal{D}} h_{\mathbf{x}}^{ch}[k]$. To make the comparison less sensitive to binning noise, we calculate the normalized empirical *cumulative distribution functions* (CDFs). For channel ch the training-set

aggregated reference $\mathrm{CDF}^{ch}_{\mathrm{ref}}[k]$ is given by:

$$\mathrm{CDF}^{ch}_{\mathrm{ref}}[k] = \frac{1}{\sum_{\mathbf{x}\in\mathcal{D}} S^l_h S^l_w} \sum_{j=1}^{k} \mathcal{H}^{ch}_{\mathrm{ref}}[j] \tag{3}$$

For a test image $\mathbf{x}$, we compute CDF_x analogously using the same fixed per-channel bin edges. We do not select a reference image, instead each test image is compared to the training-set aggregated reference $\mathrm{CDFs}_{\mathrm{ref}}$. The distance between single image CDFs_x and the training-set aggregated reference $\mathrm{CDFs}_{\mathrm{ref}}$ (e.g., using *Earth Mover's Distance* or *Hellinger* distance) is aggregated over all channels and layers, yielding a measure of distributional shift. Here, we use *Earth Mover's Distance* (EMD) as described in [35]. An image is accepted as *in-distribution* with $\mathbb{I}[\mathrm{EMD}\,(\mathrm{CDFs}_{\mathrm{ref}}, \mathrm{CDFs}_{\mathbf{x}}) < \tau]$. A schematic illustration of the proposed approach is depicted in Figure 1.

Generative Models: We train *normalizing flows* (NFs) on detector activations to model $p_{\mathrm{ID}}(\mathbf{x})$ in feature space, following Lofti et al. [40]. Let f_t be an invertible transform from base $p_Z(\mathbf{z})$ (standard Gaussian) to $p_X(\mathbf{x})$:

$$p_X(\mathbf{x}) = p_Z(f_t^{-1}(\mathbf{x})) \left| \det \frac{\partial f_t^{-1}}{\partial \mathbf{x}} \right|. \tag{4}$$

We use RealNVP with four masked affine coupling blocks [7], in two variants: a single NF trained on concatenated multi-layer features and a multi-scale NF (M-NF) with one NF per layer and aggregated scores. Features are globally average pooled. The OR score is $-\log p_{\mathrm{ID}}(\mathbf{x})$, with acceptance $\mathbb{I}[-\log p_{\mathrm{ID}}(\mathbf{x}) < \tau]$.

Top-k Detection Uncertainties: To estimate detection-level uncertainty, we employ variance networks which extend standard object detectors to output the parameters of a multivariate Gaussian predictive distribution. The predicted covariance matrix is represented via its lower-triangular Cholesky factor to ensure positive semi-definiteness.

These networks are trained using *scoring rules* such as the *negative log-likelihood* (NLL), *energy score* (ES) [13], and *direct moment matching* DMM [9]. These rules incentivize truthful uncertainty estimation and can be *local*, *non-local*, *proper*, or *strictly proper* [13]. We adopt the probabilistic detection framework of Harakeh et al. [17], which extends widely used object detectors such as FasterRCNN [50], DETR [2], and RetinaNet [37].

Each detection $\mathbf{y}_i = (s_i, \mathbf{b}_i, \mathbf{c}_i)$ includes the detection confidence score s, the predicted bounding box $\mathbf{b}$, and the predicted class label $\mathbf{c}$. For variance networks, the bounding box is represented as a predictive distribution over the top-left and bottom-right corners given by $\mathcal{N}(\mu_{tl}, \Sigma_{tl}), \mathcal{N}(\mu_{br}, \Sigma_{br})$. To compute $G(\mathbf{x})$ all detector outputs can be used. Related to classification, these include the confidence score $\hat{s}_i$ and the entropy of the predictive classification distribution. For regression output, we consider the determinant, trace, and entropy of the multivariate Gaussian bounding box distributions, as described in [43]. Since the approximation of image-level uncertainty relies on available detections, no confidence threshold is applied to filter detections.

4 Evaluation

We evaluate the proposed and reference OR monitors on COCO using image-level OOD detection under synthetic covariate shift. Because dense error labeling under shift is impractical, we use ID vs. OOD separability as a proxy. First, we verify that corruptions significantly degrade detector performance on increasing corrupted data, then measure OOD performance on the same inputs.

4.1 Dataset

For evaluation, we build on COCO [38] as an established dataset for object detection. The training set $\mathcal{D}$ is used to train detectors and NF-based models, and to collect the reference CDFs, while the unaltered validation set serves as ID set. To simulate covariate distribution shift in $p(\mathbf{x})$, we apply various image corruptions (19 corruption types defined by [21]), thus synthetically altering external conditions and simulating sensor failures [48]. Each corruption is applied across ten severity levels (extending the standard five by intensifying parameters), yielding progressively degraded OOD splits with unchanged labels.

4.2 Models

To collect activation distributions and extract features for the NF models, we use FasterRCNN [50] with a ResNet-50 *backbone* [18]. Features are tapped at five points in the *backbone*: after the initial convolution and max-pooling layer (C1), and at the outputs of the four residual stages (C2–C5). Activations from C4 and C5 are considered high-level, while those from C1 to C3 are low-level. From the detector *head*, we extract activations before the classification and box-regression *heads* at different pyramid levels (*region of interest* (ROI) alignment layers), referred to as *head* activations. This selection captures hierarchical information from low- to high-level cues.

For variance networks, we use the probabilistic extensions by Harakeh et al. [17] with FasterRCNN as the basis. For OOD score estimation, we use the top-3 predictions. The COCO [38] training set is used with default hyperparameters and data augmentation. For further details, we refer to the original publication.

NF model training and CDFs collection use a frozen, pre-trained detector with the same data and augmentation settings. This process repeats the training to generate detector activations on the ID data. To handle varying activation ranges, each channel's minimum and maximum values are tracked during one epoch, and a margin of 20% is applied for normalization. Following [49], we use 1000 bins for the activation histograms.

4.3 Metrics

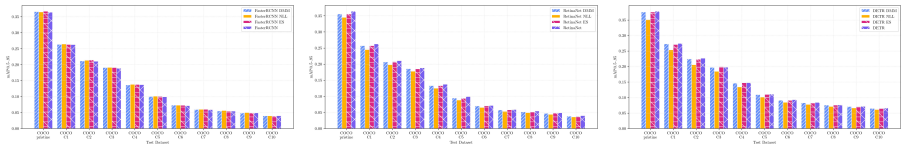
We evaluate detection with standard metrics. *Mean average precision* (mAP) is computed over ten IoU thresholds from 0.5 to 0.95 [38] (mAP@.5:.95). We also

report *mean squared error* (MSE) and the scoring rules from Section 3 (NLL, DMM, ES). These scoring rules assess the predictive distributions of variance networks, require ground truth, and are not used as OR scores.

To dissect errors, we follow [24] and group predictions by IoU with ground truth. the detections groups are: *true positives* (IoU ≥ 0.5; only the highest-confidence match retained), *duplicates* (additional matches to the same instance), *localization errors* ($0.1 \leq \text{IoU} < 0.5$), and *false positives* (IoU ≤ 0.1).

OOD detectability is measured by *area under receiver operating characteristic* (AUROC) between ID and OOD images. Scores are the negative log-likelihood $-\log p_{\text{ID}}(x)$ for NFs, detector-derived signals (e.g., confidence, entropy), and EMD to CDFs_{ref} for activation-distribution approach. Final AUROC scores are computed after z-score normalization [53].

4.4 Results



(a) Faster RCNN (b) RetinaNet (c) DETR

Fig. 2: Detection performance (mAP@.5-.95) of baseline (FasterRCNN, RetinaNet, DETR) and probabilistic object detectors (DMM NLL ES) under increasing corruption severity on the COCO dataset.

Figure 2 illustrates mAP degradation under increasing corruption severity of the provided baseline detectors and their probabilistic variants from [17]. performance drops monotonically but remains nonzero. Since our analysis relies on available detections, no confidence threshold is applied. Thus, even at the highest corruption levels, objects still receive detection assignments. Probabilistic variants have similar mAP, suggesting scoring-rule differences primarily result from covariance estimates, not mean predictions. Figure 3 shows the results for classifying and regressing predictive distributions for *true positives*, *localization errors*, under dataset shift and using various scoring rules (ES, NLL, DMM) for model training. (Results for DETR [2] and RetinaNet [37] are in the supplementary material.)

The results show a consistent degradation in the quality of both classification and box predictive distributions under dataset shift, aligning with findings from [17] and [47]. The degree of degradation, however, varies across models and backbones, suggesting model-specific sensitivities. No scoring rule consistently yields optimal predictive distributions for both classification and regression tasks.

Among the scoring rules, ES produces lower-entropy predictive distributions for bounding boxes on *true positive* detections, outperforming both NLL and

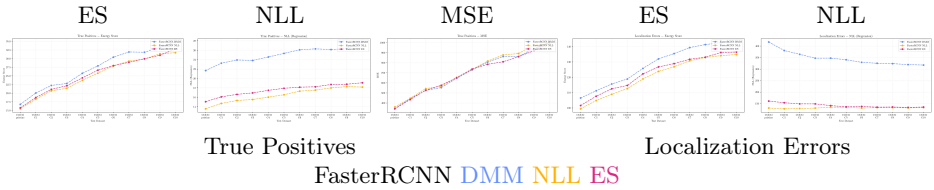


Fig. 3: NLL, ES, and MSE of the bounding box predictive distributions across object categories. These metrics are computed for probabilistic detectors based on FasterRCNN and evaluated on the synthetically shifted variants of the COCO dataset.

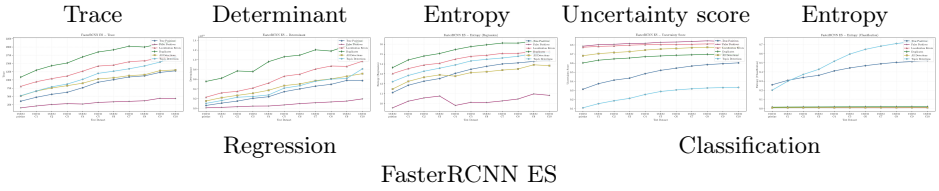


Fig. 4: Mean detection scores of FasterRCNN ES evaluated for the synthetically shifted COCO datasets. The regression-related scores are the determinant, trace, and the differential entropy. Scores related to classification are the uncertainty score (1- confidence score) and the Shannon entropy. (Color coding: True Positives, Localization Errors, Duplicates, All Detections, Top-k Detections False Positives)

DMM across evaluated detection models. In contrast, models trained with NLL tend to produce overly diffuse (high-entropy) distributions even for correct detections. This behavior can result in misleading evaluations when relying solely on NLL, as it may underestimate confidence even in accurate predictions. For predictions with moderate error (i.e., *localization errors*), models trained with NLL outperform those trained with the ES. Higher predictive entropy is appropriate here, as the predicted bounding boxes diverge substantially from ground truth. However, this trade-off yields better calibration for true positives with ES.

Overall, ES-trained models offer a more balanced uncertainty profile. Confident when predictions are accurate and uncertain otherwise. In contrast, NLL-trained models often assign high uncertainty across the board, which can distort evaluation outcomes if judged solely by NLL. DMM-trained models frequently exhibit overconfidence, leading to poor NLL but acceptable ES values.

Importantly for OR, variance networks are sensitive to dataset shift. Across all detections and loss functions, predictive entropy consistently increases as test data moves from ID to synthetically shifted OOD. The separation between *true positives* and *localization errors* indicates the model's ability to reflect prediction quality, crucial for detecting both input uncertainty and shift.

Since models trained with ES demonstrate to be effective in reflecting distributional shift, showing both high entropy on unfamiliar inputs and calibrated low

Model	Severity Levels AUROC ↑ 3	5	8	10
Normalizing Flows (FasterRCNN)				
NF *head*	53.55 ± 1.12	63.12 ± 1.17	71.34 ± 1.26	74.63 ± 1.31
NF *backbone* (C4, C5)	51.12 ± 1.15	61.53 ± 1.21	71.21 ± 1.21	76.53 ± 1.24
NF *backbone* (C1, C2, C3)	55.12 ± 1.21	73.56 ± 1.02	75.51 ± 1.16	76.61 ± 1.21
NF *backbone*	56.12 ± 1.16	73.45 ± 0.99	76.25 ± 1.14	77.72 ± 1.13
NF *backbone* & *head*	56.43 ± 1.24	73.32 ± 1.15	76.87 ± 1.22	77.89 ± 1.21
M-NF *head*	54.76 ± 1.31	64.41 ± 1.05	72.11 ± 1.20	74.93 ± 1.13
M-NF *backbone* C4, C5)	55.76 ± 1.22	67.54 ± 1.18	73.71 ± 1.11	76.73 ± 1.12
M-NF *backbone* (C1, C2, C3)	58.61 ± 1.30	68.94 ± 1.11	75.81 ± 1.12	77.73 ± 1.11
M-NF *backbone*	58.76 ± 1.24	69.12 ± 1.01	73.91 ± 1.09	78.73 ± 1.07
M-NF *backbone* & *head*	61.75 ± 1.26	76.54 ± 1.34	77.10 ± 1.15	79.28 ± 1.27
Top-3 Detection Uncertainties (FasterRCNN ES)				
Trace	52.32 ± 1.27	60.63 ± 1.09	64.84 ± 1.11	68.79 ± 1.00
Determinant	46.47 ± 1.31	47.94 ± 1.11	53.23 ± 1.17	57.16 ± 0.94
Entropy (Regression)	56.70 ± 1.15	60.72 ± 1.12	63.01 ± 1.21	68.72 ± 0.97
Confidence Score	58.83 ± 1.28	72.61 ± 1.09	75.56 ± 1.20	78.30 ± 1.01
Entropy (Classification)	62.84 ± 1.11	77.60 ± 1.10	81.21 ± 1.07	81.70 ± 1.03
Activation-Distribution (Ours) (FasterRCNN)				
CDFs *head*	49.02 ± 1.35	48.26 ± 1.31	56.22 ± 0.87	69.19 ± 1.04
CDFs *backbone* (C4, C5)	48.59 ± 1.29	62.17 ± 1.28	71.61 ± 1.01	78.03 ± 0.96
CDFs *backbone* (C1, C2, C3)	71.90 ± 1.33	79.46 ± 0.96	84.62 ± 0.85	86.73 ± 0.79
CDFs *backbone*	68.02 ± 1.31	80.26 ± 1.29	86.82 ± 1.16	87.39 ± 0.82
CDFs *backbone* & *head*	66.22 ± 1.35	77.66 ± 1.27	83.54 ± 1.11	86.52 ± 0.95

Table 1: AUROC scores of detecting OOD covariate shift for synthetically corrupted variants of the COCO dataset. Best results are highlighted as first, second, and third.

entropy on *in-distribution*, high-quality detections, we select the FasterRCNN ES variant for comparison. For OOD detection during inference, the above metrics cannot be applied since they require ground truth. To address this, Figure 4 presents the mean score values directly calculated from FasterRCNN ES outputs across the synthetically corrupted COCO dataset.

Across all shown scores, the impact of the distributional shift is clearly reflected. Moreover, the differences between detection types are clearly visible in the results. When approximating image-level uncertainty, these results highlight why using only the top-k detections is preferable to including all detections. The full detection set includes many noisy and poorly calibrated outputs, which can obscure meaningful uncertainty signals.

Due to IoU-based loss functions used in detector training, predictive distributions are better calibrated for *true positives*. This is especially evident for *false positives*, which yield low determinant and trace values. As long as score distributions remain separable between detection types and exhibit a consistent trend across severity levels, these scores can serve to assess OOD behavior.

The results comparing OOD detection performance across Section 3 approaches are summarized in Table 1. The AUROC results are reported as the mean and standard deviation over three independent runs. Despite its simplicity, the proposed approach using CDFs of channel activations achieves the best result, especially when using all backbone layers. Adding *head* layers does not further improve performance, and backbone selections outperform head-only variants. This suggests that *backbone* layers already provide sufficient comple-

mentary information due to their hierarchical structure. Although diversity in activation cues increases sensitivity to shift, the strongest cues come from the lower layers.

Among NF-based approaches, the best AUROC is achieved by the multi-scale variant with both *backbone* and *head* layers, but the performance gap compared to *backbone*-only is not significant. Most distortions are best recognized from lower-layer activations, but cues from mid- to high-levels can help.

Head activations are spatially localized, and as with top-k detections, filtering may be needed to better extract data uncertainty. In general, NF-based models achieved better AUROC scores than detection counterparts. Nevertheless, the entropy of the classification distribution or the confidence score sometimes yielded similar or better results than the best NF-based models. Classification-related uncertainty measures outperformed localization-based ones, likely because regression outputs are optimized for positive detections only, while classification is exposed to both object and background proposals [45].

None of the evaluated methods completely separate ID from OOD samples. This is plausible, since even under high-severity corruptions, the detector can still identify plausible objects. For specific corruptions like *contrast*, mAP drops sharply and AUROC increases across models.

To illustrate the detector-agnostic capabilities of our activation-distribution approach, we evaluated its OOD detection performance across multiple detection architectures. Specifically, we considered recent models from the *You Only Look Once* (YOLO) family, including YOLOv9 [55], YOLOv10 [54], and YOLOv11 [27], as well as the *transformer*-based RT-DETR [58] with an HGNet *backbone* [3].

Table 2 summarizes the AUROC results for each detector evaluated on synthetically corrupted versions of the COCO dataset. For all architectures, feature map activations were systematically extracted from key architectural components—specifically, the initial *convolutional* layers (to capture low-level features) and the *feature pyramid network* (FPN) stages (to capture multi-scale, higher-level representations). For further details regarding the selection of layers and model-specific architectural components, please refer to the original publications and the supplementary material.

Consistent with our previous findings, incorporating activations from the detector *head* did not improve OOD detection performance, and in some cases, even reduced it. *Backbone*-only configurations consistently achieved higher AUROC scores across all detectors, highlighting the importance of hierarchical *backbone* features for robust distribution shift detection. Notably, compared to Faster-RCNN, the activation-distribution method exhibited even greater improvements in OOD detection performance for these alternative detectors, particularly at lower corruption severity levels. This confirms our approach generalizes across detection architectures and is not limited to a specific model family. All results emphasize the importance of going beyond detection-level uncertainty and considering more diverse cues. All analyzed concepts approximating $P_{\text{ID}}(\mathbf{x})$ show promising results, yet offer room for improvement.

	Detector	Model	Severity Levels AUROC ↑ 3	5	8	10
Activation-Distribution (Ours)	RT-DETR-l [58]	CDFs *backbone*	69.06 ± 1.25	80.09 ± 0.79	87.68 ± 1.15	90.41 ± 1.12
		CDFs *backbone* & *head*	51.02 ± 1.22	68.01 ± 1.21	77.87 ± 1.31	82.11 ± 1.11
	YOLOv9-m [55]	CDFs *backbone*	68.28 ± 1.33	81.02 ± 0.93	89.68 ± 1.15	92.11 ± 1.22
		CDFs *backbone* & *head*	60.51 ± 1.32	76.00 ± 1.07	85.28 ± 1.21	87.15 ± 1.19
	YOLOv10-m [54]	CDFs *backbone*	67.12 ± 1.36	76.23 ± 1.12	86.62 ± 1.11	87.10 ± 1.14
		CDFs *backbone* & *head*	58.76 ± 1.31	74.85 ± 1.08	83.25 ± 1.25	85.74 ± 1.21
	YOLOv11 [27]	CDFs *backbone*	69.02 ± 1.35	78.44 ± 0.94	85.68 ± 1.15	91.20 ± 1.03
		CDFs *backbone* & *head*	61.41 ± 1.29	77.23 ± 1.31	84.40 ± 1.28	88.50 ± 1.26

Table 2: AUROC scores for the activation-distribution approach across several object detectors on synthetically corrupted versions of the COCO dataset. Best results are highlighted as first, second, and third.

5 Discussion and Limitations

Detection-based methods could benefit from improved assignment strategies for regression targets to better calibrate predictive distributions. A central limitation is their behavior on background-only images. When no valid detections exist, aggregated confidence is low (high uncertainty), so the image is rejected. This is benign on COCO (every image contains objects), but problematic for OR in real deployments where inputs may contain no known objects.

NF-based approaches, in turn, are known to sometimes assign higher likelihoods to OOD samples than to the *in-distribution* data they were trained on [31,44]. Recent research has proposed mitigation strategies for this issue [23,51]. However, this fundamental limitation regarding likelihood assignments remains, as discussed in Section 2.

Compared to classification tasks, where generative models are often applied to low-resolution data like MNIST (28×28 pixels) [34]. Our COCO setting (640×640 pixels) required aggressive *global average pooling* to keep NF training tractable, which may limit expressiveness. Future work could explore attention-based or learnable pooling to retain richer statistics.

Alternative generative scoring functions may improve OOD detection. For instance, *likelihood regret* [56] shows promise but requires retraining at inference, making it impractical for OR.

Our activation-distribution monitor, aggregating multi-layer per-channel CDFs, outperformed reference methods in our experiments. Still, combining many channels is likely *polysemantic* (channels respond to multiple, unrelated inputs), which can entangle signals and hinder separability. The design discards inter-channel dependencies and uses heuristic layer selection. More principled weighting and modeling cross-channel structure could strengthen representations.

Despite these limitations, the proposed approach is a simple, detector-agnostic add-on OR monitor. It operates on feature-map activations from any frozen detector without architectural changes or reliance on ad hoc filtering of confident detections.

6 Conclusion

In this paper, we studied *operational readiness* (OR) for object detection and formalized it via estimating the *in-distribution* probability $P_{\mathrm{ID}}(x)$ as a practical proxy. We introduced a detector-agnostic monitor that summarizes multi-layer activation statistics as *cumulative distribution functions* (CDFs) and compares them to training-set references.

For comparison, we evaluated feature-space generative modeling with *normalizing flows* and aggregation of detector uncertainties, thus covering bottom-up (image/feature-level) and top-down (detection-level) strategies. Despite its simplicity, the activation-distribution approach consistently improved detection of covariate shift, is lightweight, requires no architectural changes, and can be attached to arbitrary frozen detectors.

While each approach has limitations, our results indicate that combining cues from different abstraction levels increases sensitivity to diverse corruption types, with backbone features providing the strongest and most generalizable signals.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (2021)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). p. 213–229. Springer-Verlag, Berlin, Heidelberg (2020)
3. Chen, Z., Tang, C., Xiong, L.: Hgnet: A hierarchical feature guided network for occupancy flow field prediction (2024)
4. Cheng, C.H.: Provably-robust runtime monitoring of neuron activation patterns. In: Design, Automation and Test in Europe Conference and Exhibition (DATE). pp. 1310–1313 (2021)
5. Choi, H., Jang, E., Alemi, A.A.: Waic, but why? generative ensembles for robust anomaly detection (2019)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real NVP. In: International Conference on Learning Representations (ICLR). OpenReview.net (2017)
8. Du, X., Wang, Z., Cai, M., Li, S.: Vos: Learning what you don't know by virtual outlier synthesis. In: International Conference on Learning Representations (2022)
9. Feng, D., Rosenbaum, L., Glaeser, C., Timm, F., Dietmayer, K.: Can we trust you? on calibration of a probabilistic object detector for autonomous driving. The IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020)

10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (2016)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265
12. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(1), 1513–1589 (2023)
13. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**(477), 359–378 (2007)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
15. Guerin, J., Delmas, K., Ferreira, R., Guiochet, J.: Out-of-distribution detection is not all you need. Proceedings of the AAAI Conference on Artificial Intelligence **37**(12), 14829–14837 (2023)
16. Hall, D., Dayoub, F., Skinner, J., Zhang, H., Miller, D., Corke, P., Carneiro, G., Angelova, A., Sünderhauf, N.: Probabilistic object detection: Definition and evaluation. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 1031–1040 (2020)
17. Harakeh, A., Waslander, S.L.: Estimating and evaluating regression predictive uncertainty in deep object detectors. In: International Conference on Learning Representations (2021)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. He, Y., Wang, J.: Deep mixture density network for probabilistic object detection. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10550–10555 (2020)
20. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2883–2892 (2019)
21. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
22. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
23. Ho, J., Chen, X., Srinivas, A., Duan, Y., Abbeel, P.: Flow++: Improving flow-based generative models with variational dequantization and architecture design. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2722–2730. PMLR (2019)
24. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) European

Conference on Computer Vision (ECCV). pp. 340–353. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
25. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
26. Janai, J., Güney, F., Behl, A., Geiger, A.: Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art. Foundations and Trends in Computer Graphics and Vision (2020)
27. Jocher, G., Qiu, J.: Ultralytics YOLO11 (2024), `https://github.com/ultralytics/ultralytics`
28. Kang, D., Raghavan, D., Bailis, P., Zaharia, M.: Model assertions for monitoring and improving ml models. In: Dhillon, I., Papailiopoulos, D., Sze, V. (eds.) Proceedings of Machine Learning and Systems. vol. 2, pp. 481–496 (2020)
29. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5574–5584. Curran Associates, Inc. (2017)
30. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations (ICLR)
31. Kirichenko, P., Izmailov, P., Wilson, A.G.: Why normalizing flows fail to detect out-of-distribution data. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 20578–20589. Curran Associates, Inc. (2020)
32. Kobyzev, I., Prince, S.J., Brubaker, M.A.: Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(11), 3964–3979 (2021)
33. Le Folgoc, L., Baltatzis, V., Desai, S., Devaraj, A., Ellis, S., Manzanera, O.E.M., Nair, A., Qiu, H., Schnabel, J., Glocker, B.: Is mc dropout bayesian? (2021)
34. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist **2** (2010)
35. Levina, E., Bickel, P.: The earth mover's distance is the mallows distance: some insights from statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 251–256 vol.2 (2001)
36. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018)
37. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007 (2017)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision (ECCV). pp. 740–755. Springer (2014)
39. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 21464–21475. Curran Associates, Inc. (2020)
40. Lotfi, D., Mahani, M.A.N., Koohi-Moghadam, M., Bae, K.T.: Enhancing out-of-distribution detection in medical imaging with normalizing flows (2025), `https://arxiv.org/abs/2502.11638`

41. MacKay, D.J.C.: A practical bayesian framework for backpropagation networks. Neural Computation **4**(3), 448–472 (1992). https://doi.org/10.1162/neco.1992.4.3.448
42. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
43. Murphy, K.P.: Probabilistic Machine Learning: An introduction. MIT Press (2022)
44. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: International Conference on Learning Representations (2019)
45. Oksuz, K., Joy, T., Dokania, P.K.: Towards building self-aware object detectors via reliable uncertainty quantification and calibration. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
46. Osband, I.: Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. Workshop on Bayesian Deep Learning, NIPS (2016)
47. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Curran Associates Inc., Red Hook, NY, USA (2019)
48. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset Shift in Machine Learning. The MIT Press (2009)
49. Ramtoula, B., Gadd, M., Newman, P., De Martini, D.: Visual dna: Representing and comparing images using distributions of neuron activations. CVPR (2023)
50. Ren, S., He, K., Girshick, R.B., Sun, J.: In: Advances in Neural Information Processing Systems (NIPS)
51. Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J.F., Luque, J.: Input complexity and out-of-distribution detection with likelihood-based generative models. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020)
52. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: Advances in Neural Information Processing Systems (2021)
53. Viviers, C., Valiuddin, A., Caetano, F., Abdi, L., Filatova, L., de With, P., van der Sommen, F.: Can your generative model detect out-of-distribution covariate shift? In: European Conference on Computer Vision (ECCV) Workshops (2024)
54. Wang, A., Chen, H., Liu, L., CHEN, K., Lin, Z., Han, J., Ding, G.: YOLOv10: Real-time end-to-end object detection. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
55. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: Yolov9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision (ECCV). pp. 1–21. Springer Nature Switzerland, Cham (2024)
56. Xiao, Z., Yan, Q., Amit, Y.: Likelihood regret: an out-of-distribution detection score for variational auto-encoder. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
57. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision (2024)
58. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: DETRs beat YOLOs on real-time object detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16965–16974 (2024)