

SAFE: Sensitivity-Aware Features for Out-of-Distribution Object Detection

Samuel Wilson¹, Tobias Fischer¹, Feras Dayoub², Dimity Miller¹, Niko Sünderhauf¹

¹QUT Centre for Robotics, Queensland University of Technology

²Australian Institute for Machine Learning, University of Adelaide

s84.wilson@hdr.qut.edu.au

Abstract

We address the problem of out-of-distribution (OOD) detection for the task of object detection. We show that residual convolutional layers with batch normalisation produce Sensitivity-Aware FEatures (SAFE) that are consistently powerful for distinguishing in-distribution from out-of-distribution detections. We extract SAFE vectors for every detected object, and train a multilayer perceptron on the surrogate task of distinguishing adversarially perturbed from clean in-distribution examples. This circumvents the need for realistic OOD training data, computationally expensive generative models, or retraining of the base object detector. SAFE outperforms the state-of-the-art OOD object detectors on multiple benchmarks by large margins, e.g. reducing the FPR95 by an absolute 30.6% from 48.3% to 17.7% on the OpenImages dataset.

1. Introduction

Across a variety of tasks, deep neural networks (DNNs) produce state-of-the-art performance when tested on data that closely matches the training data distribution [19, 54]. However, when deployed into the real world, out-of-distribution (OOD) samples that do not belong to the training distribution are likely encountered. Upon encountering OOD samples, DNNs tend to fail silently and produce overconfident erroneous predictions [51, 40, 21, 3, 5, 16]. Especially in safety-critical applications, such as self-driving vehicles or medical robotics, such silent failures present a severe safety risk that must be addressed before the widespread adoption of these systems [59, 2].

OOD detection, where OOD samples are distinguished from in-distribution (ID) samples, is thus an important task. OOD detection has been addressed widely in the image classification setting [31, 35, 22, 53, 66, 6, 57, 68]. In this paper, we expand upon the limited body of work in OOD object detection [9, 8, 7], leveraging recent theoret-

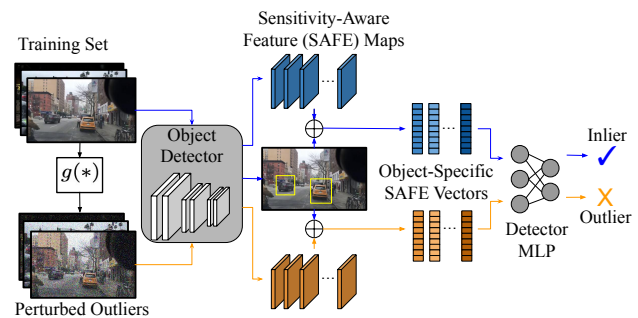


Figure 1. Overview of our proposed SAFE OOD object detector. Feature maps are extracted from sensitivity-aware layers in the backbone of a pretrained object detector. Object-specific SAFE vectors are extracted for the predicted bounding boxes. Pre-deployment, an auxiliary MLP is trained to distinguish the feature vectors of normal ID detections (blue) from adversarially-perturbed ID samples (orange). At test time, the pipeline for the training samples is repeated (blue) for all test samples, with the auxiliary MLP producing OOD detection scores for each object in a test image. Illustrative input images are drawn from BDD100K [70].

ical insights on the behaviour of the feature space of DNNs. Specifically, recent theory [34, 43, 61, 55] has highlighted the importance of ensuring that the feature space of a DNN is *distance-preserving* through the constraints of *sensitivity* and *smoothness*. In particular, *sensitivity*, *i.e.* the preservation of input distances in the output, has been shown to play a crucial role in learning a robust feature set that avoids mapping ID and OOD data to similar feature representations [61].

Furthermore, prior work established the role of adversarial attacks in OOD generalisation [69], increasing the separability of ID samples from OOD [32, 23] and perturbing feature representations [15, 49, 36]. We thus leverage the most *sensitive* layers in a pretrained object detector backbone through targeted input-level adversarial perturbations.

This paper introduces **SAFE**, a new approach to visual OOD object detection that utilises object-specific

Sensitivity-Aware Features (Figure 1). Our approach has three core components, each offering advantages over existing works in this area:

1. We identify a critical subset of layers that are *sensitive* to OOD input variations, *i.e.* these layers preserve differences from the input in their feature space, and trigger abnormally high activations. SAFE layers are residual convolutional layers with batch normalisation within the backbone of an object detector. We empirically validate their superiority for OOD detection in our results. In contrast, previous work only uses features from the classification head of the object detector and do not consider the characteristics of different layers for OOD detection [9, 53, 31].
2. We extract object-specific SAFE vectors and use a multi-layer perceptron (MLP) to classify every detected object as ID or OOD. This allows OOD samples to be detected in a *posthoc* manner, *i.e.* it does not require retraining of the base network [68] and can be applied to any pre-trained object detector with a backbone containing SAFE layers (*e.g.* ResNet [19] and RegNetX [44]).
3. We train this MLP on the surrogate task of distinguishing the SAFE vectors of adversarially-perturbed samples and clean ID training samples. This avoids the necessity of access to real outlier training data [31, 25, 71, 35, 22, 47, 4, 1] or a complex generative process to synthesise such data [30, 64, 56, 52].

SAFE achieves new state-of-the-art results across multiple established benchmarks. We release a publicly available code repository to replicate our experiments at: https://github.com/SamWilso/SAFE_Official

2. Related Work

In this section, we identify the core contributions in the related areas to OOD object detection: i) Many methods attempt to calibrate the confidence of the network utilising available or self-generated outlier data. ii) When access to outlier data is unavailable, deep features of the network are monitored for deviations from known values. iii) Under some regularisation constraints, the feature space of a deep network can be tuned to be *distance-aware*, improving OOD detection performance. iv) Whilst work on OOD object detection is scarce, recent works have been proposed for adjacent tasks (*e.g.* open-set, performance monitoring) in object detection.

Outlier-based OOD Detection A common approach to OOD detection is to calibrate the model confidence by tuning the weights or hyperparameter on an auxiliary validation dataset [31, 25, 71, 35, 22, 47, 4, 1]. These OOD-specific characteristics can be extrapolated from an available set of real outlier data constructed from either the testing OOD set [25, 31, 35] or an entirely separate dataset [47, 4, 22, 1].

While these methods often present impressive performance, the use of these outlier sets is inherently problematic: if the real outlier set does not accurately represent the OOD samples encountered at test time, substantial drops in performance are observed [57].

To overcome this, many prior works synthesise outliers as a proxy for OOD samples, training a network to distinguish between ID samples and the synthetic outliers [14, 30, 63, 64, 56, 52, 9]. Early works on outlier synthesis focused on Generative Adversarial Networks (GANs) [14], training a model that generated low ID density samples in the image space for calibrating confidence measures [30], training a reject class [63, 64] or encouraging uniform predictions on OOD samples [56]. Scaling input-based generative models becomes complex as the fidelity of images increases, and thus feature-based generative methods have been proposed for OOD object detection [9]. Synthetic outliers have also been created by adding input-level perturbations to the known ID dataset via adversarial generation [32, 23, 69], pixel-level mutation [48] or permutation [6] and additive noise [37, 60]. We adopt a similar approach and generate synthetic outliers via adversarial perturbations with the goal of training an auxiliary MLP to distinguish between ID samples and adversarial samples.

Feature-based OOD Detection Many methods avoid generating sufficiently realistic outliers by directly monitoring the outputs [21, 20, 16, 12, 72, 65, 26] or features [53, 66, 6, 61, 1, 58] of the DNN. These methods are generally more computationally efficient in contrast to alternative OOD detectors, but often rely on fundamental assumptions about the characteristics of the feature space *e.g.* separability of classes in feature space [11, 66, 72, 61, 58]. Following these findings, recent works have proposed methods that enable the assessment of layer-wise performance, subsequently demonstrating that not all layers are equally effective at detecting OOD data [66, 6, 53, 37] and that some layers exhibit abnormal behaviour when presented with OOD data [57]. All these works are applied exclusively to image classification. In this work, we extend the usage of feature-based OOD detectors to object detection by only leveraging the backbone features that are the most *sensitive* to OOD data.

Feature Space Regularisation The beneficial properties of *sensitivity* and *smoothness* have recently been highlighted in the context of OOD detection for classification [34, 43, 61, 55]. *Sensitivity* ensures that differences in the input space (*i.e.* pixels) result in sufficiently different representations in the feature space, preventing the feature collapse problem [61]. *Smoothness* prevents the feature mapping from being *too* sensitive, thus avoiding reduced generalisation and robustness [61]. Both properties constitute the lower and upper bounds of a bi-Lipschitz constraint (Eq. (1)) and can be enforced during training, *e.g.* by train-

ing a network with residual connections [19] with spectral normalisation [42], as applied by [43, 34]. However, recent work [43, Sec. C.4] has shown that residual connections constitute an inductive bias towards sensitivity, even *without* spectral normalisation. We build upon this insight and expand it to the task of OOD for object detection.

Reliable Object Detection Applications of OOD detection methods to object detection is a new field; however, there are existing works in related domains. Akin to OOD detection, open-set error detection [5] commonly relies on the outputs of the final layer of the object detector network [41, 28, 40]. In a similar vein, recent works have sought to explain failures in deep object detectors by analysing the relationship between individual architecture components and unique errors [39] and the influence of the datasets on these errors [38]. Related works in failure monitoring use auxiliary networks trained on backbone object detector features [46, 45]. Few works [9, 8, 7] have explicitly addressed the problem of OOD object detection. These methods require explicit retraining of the base network or fine-tuning of the hyperparameters with an auxiliary outlier dataset [9, 8, 7].

3. SAFE: Sensitivity-Aware Features

We propose SAFE, a post-hoc OOD detector for object detection that leverages the *sensitivity* of residual convolutional layers and abnormal batch normalisation activations to identify OOD object detections. In Section 3.1, we provide the motivation that informs the critical aspects of SAFE, and then in Section 3.2, we introduce the SAFE method in detail.

3.1. Motivation

Fundamental to SAFE, we identify residual convolution layers followed immediately by batch normalisation are consistently *sensitive* and thus powerful layers for OOD object detection. Whilst it has been shown that not all layers perform equally for OOD detection in image classification [66, 6, 53, 1] or LIDAR-based OOD detection [24], we are the first to empirically validate this for OOD object detection and the first to investigate the layer characteristics that induce stronger performance. We select our subset of critical layers based on prior work in the image classification setting demonstrating *sensitivity*-preserving properties of residual connections [43] and abnormal activations of batch normalisation layers [57]. In the following, we detail the theoretical groundwork underpinning our findings.

Sensitivity of Residual Connections We consider a frozen pre-trained base network f that functions as a feature extractor, mapping samples from the input space \mathbb{X} to the hidden feature space $f : \mathbb{X} \rightarrow \mathbb{H}$. To detect OOD samples, the DNN’s feature space needs to be well-regularised [34],

according to the *bi-Lipschitz constraint*:

$$L_1 \cdot \|x - x^*\|_I \leq \|f(x) - f(x^*)\|_F \leq L_2 \cdot \|x - x^*\|_I \quad (1)$$

where x and x^* are two input samples, $\|\cdot\|_I$ and $\|\cdot\|_F$ denote distance metrics in the input and feature space respectively, and L_1 and L_2 are the lower and upper Lipschitz constants [34]. The lower bound, *sensitivity*, ensures that distances in the input space are sufficiently preserved in the hidden space, and the upper bound, *smoothness*, limits the sensitivity of the hidden space to input variations, ensuring that distances in the hidden space have a meaningful correspondence to distances in the input space. Encouraging sensitivity and smoothness is commonly accomplished by applying spectral normalisation [42] to the weight matrices of a DNN with residual connections [34, 43, 55]. In posthoc OOD detection, where f is pretrained, we cannot guarantee that a network is pretrained with spectral normalisation, hence not fulfilling the *smoothness* constraint. However, [43, Sec C.4] has shown that a network solely trained *with residual connections* and no *smoothness* constraint is still sufficiently *sensitive* to changes in the input.

Mismatched BatchNorm Statistics Batch Normalisation [27] (BatchNorm) is a commonly used normalisation technique to help training deep networks. BatchNorm assists the network in learning the designated task on ID data by normalising a given input z with respect to the expected value $\mathbb{E}_{in}[\cdot]$ and variance $\mathbb{V}_{in}[\cdot]$ calibrated over the ID data:

$$\text{BatchNorm}(z; \gamma, \beta, \epsilon) = \frac{z - \mathbb{E}_{in}[z]}{\sqrt{\mathbb{V}_{in}[z] + \epsilon}} \cdot \gamma + \beta. \quad (2)$$

Recent work [57] empirically observed that BatchNorm statistics calibrated on the ID set and directly applied to the OOD set trigger abnormally high activations due to a mismatch of the true parameters between datasets $\mathbb{E}_{in}, \mathbb{V}_{in} \neq \mathbb{E}_{out}, \mathbb{V}_{out}$. Propagation of these abnormal activations throughout the network results in abnormally high logits for an erroneous prediction, resulting in overconfidence of the network on OOD samples. We propose a deep feature-based approach to leverage this characteristic; training an auxiliary network to monitor feature activations from these layers and flagging a sample as OOD when an abnormal activation is detected.

3.2. Method

Given the observations that residual connections enable sensitivity to input changes [43] and that BatchNorm layers trigger abnormal activations on OOD data [57], we thus hypothesise that residual connections combined with BatchNorm regularisation provide a clear signal for OOD detection. Connections of this variety are not uncommon, with four of these layers in the standard ResNet-50 [19] and RegNetX4.0 [44] backbone architectures. We now detail the

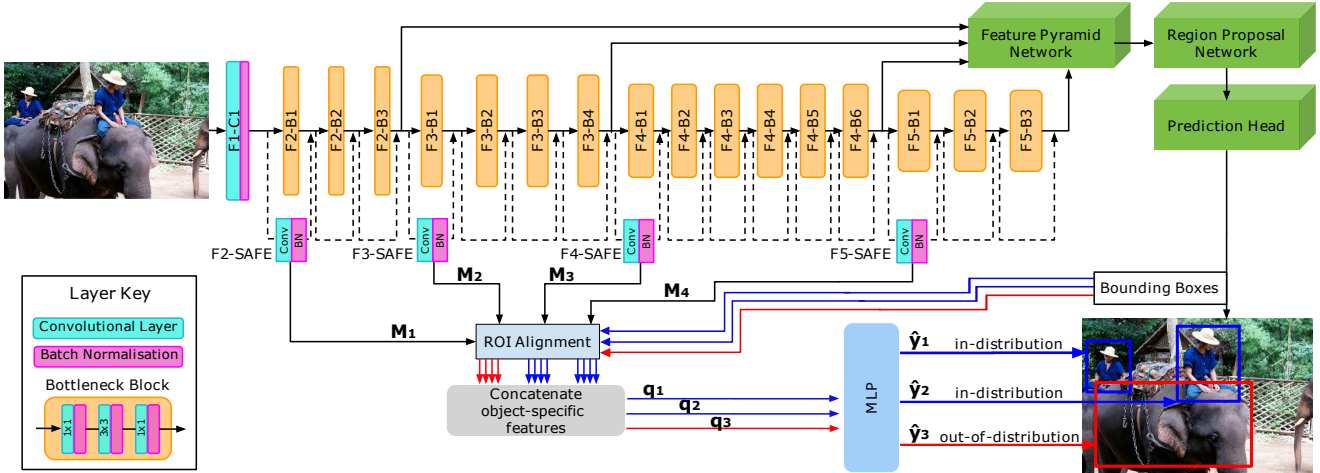


Figure 2. Architecture diagram of our proposed SAFE OOD detector with an example ResNet-50 [19]. We extract object-level feature maps $\{M_1, M_2, M_3, M_4\}$ from the layers that are most sensitive to OOD data. Next, object descriptors $\{q_1, q_2, q_3\}$ are formed by applying region of interest pooling to the feature maps and concatenating the resultant vectors layer-wise. Each object descriptor is passed through the MLP, producing a corresponding OOD score for each object $\{\hat{y}_1, \hat{y}_2, \hat{y}_3\}$ which distinguishes detections with OOD samples (red) from detections with ID samples (blue).

pipeline for SAFE to leverage these critical layers (see Figure 2) and later confirm our hypothesis empirically in Section 4.4.

Preliminaries We consider a pretrained, frozen DNN object detector f , which given an image x will produce a set of D object predictions. Each detection $d \in \{1, \dots, D\}$ has a classification label c_d and bounding box $b_d \in \mathbb{R}^4$. During deployment, we wish to classify each object prediction as being generated from ID or OOD data.

Object-specific SAFE Extraction To distinguish between ID and OOD object predictions, we extract object-specific features from our identified sensitivity-aware layers. In the detector, there are L SAFE layers, *i.e.* residual connection layers with BatchNorm regularisation, that output a set of L feature maps $\{M_1, \dots, M_L\}$. Figure 2 shows an example of an object detector with a ResNet-50 backbone, which contains $L = 4$ SAFE layers. To extract object-specific features, the proposed bounding boxes $\{b_1, \dots, b_D\}$ are used to take cropped regions of each feature map M_l . These object-specific feature maps $O_{l,d}$ are then reduced to a vector representation $\mathbf{p}_{l,d}$ for concatenation via a bilinear interpolation operation along the spatial axis. Finally, the pooled feature vectors $\mathbf{p}_{l,d}$ are concatenated layer-wise to form a single object-specific vector \mathbf{q}_d with a length equal to the sum of the number of channels c for each layer: $|\mathbf{q}_d| = \sum_l c_l$.

Feature Monitoring MLP We instantiate an auxiliary feature monitoring MLP, f_β , to classify detections as ID or OOD. Given an object-specific SAFE vector, the MLP outputs an OOD score $\hat{y}_d = f_\beta(\mathbf{q}_d)$ in the range of $\hat{y} \in [0, 1]$.

These scores can be used in practice to make decisions based on the application and corresponding risk profile of the downstream task, *e.g.* detection scores can be compared to a predefined threshold to classify objects as ID or OOD and achieve a minimum true positive rate.

Surrogate Training with Synthetic Outliers Since OOD samples are inaccessible prior to deployment, the feature monitoring MLP is trained on the *surrogate* task of discriminating between objects within clean ID images and the same objects within adversarially-perturbed ID images. For each image in the training set $x \in \mathbb{X}$, we generate an outlier counterpart of the same image through an adversarial perturbation $x^o = g(x)$. In practice, we utilise the simple Fast Gradient Sign Method [15] (FGSM) adversarial attack that produces a perturbed image x^o by adding noise to the original image x . Given the model parameters θ , the noise for FGSM is computed based on the sign of the gradient $\text{sign}(\nabla_x)$ with respect to the cost function $J(\theta, x, y)$ and then scaled with by magnitude multiplier ϵ :

$$x^o = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)). \quad (3)$$

Next, object-specific feature vectors are extracted from the SAFE layers for both the clean and perturbed images using the bounding boxes b_d predicted on the *clean image*. Finally, the object-specific feature vectors are used to train the auxiliary MLP with clean features corresponding to ID detections and perturbed features corresponding to surrogate OOD detections. We ablate the parameters of our adversarial-perturbation in Section 4.5.

4. Experiments

We conduct a series of experiments to demonstrate the efficacy of our proposed SAFE OOD detector. We first describe our experimental setup in Section 4.1 and detail the implementation of SAFE in Section 4.2. We then compare our method to the state-of-the-art on the challenging task of OOD object detection in Section 4.3. Finally, we demonstrate the unique effectiveness of our identified critical layers in Section 4.4, and we ablate the sensitivity of the auxiliary MLP to the transformation magnitude in Section 4.5. Additional comparisons of SAFE to the state-of-the-art with the transformer-based Deformable DETR object detector [73] are provided in the Supplementary Material.

4.1. Experimental Setup

We follow the evaluation protocol defined by [9] with the accompanying benchmark repository¹.

Datasets We use the predefined ID/OOD splits for the object detection task defined in [9]. The two ID datasets are constructed from the popular PASCAL-VOC [10] and Berkley DeepDrive-100K [70] (BDD100K) datasets. For the OOD datasets, subset versions of the MS-COCO [33] and OpenImages [29] datasets are provided where classes that appear in the custom ID datasets are removed.

Evaluation Metrics We consider the standard AUROC and FPR95 metrics defined in [9] extensively used across the image classification literature [72, 32, 66, 53]. **AUROC:** The Area Under The Receiver Operating Characteristic curve (AUROC) is defined by the area under the ROC curve with true positive rate (TPR) on the y-axis and false positive rate (FPR) on the x-axis; higher is better. An AUROC score of 50% indicates a method that is as effective as random guessing. **FPR95** reports the false positive rate when the true positive rate is at 95%; lower is better. For real-world deployment, a binary classifier based on a threshold of the confidence scores determines if a detection is ID or OOD; under these conditions, FPR95 provides better insight into how an OOD detector will perform. **AP:** Our SAFE OOD detector is a *posthoc* addition to a pre-trained network and does not affect the on-task performance of the base model under the average precision (AP) metric, as such, we do not report AP as in [9].

Importantly, under the benchmark setting by [9], the AUROC and FPR95 metrics are computed *after* low-confidence objects are suppressed according to a confidence threshold determined by [17]. Our comparisons in Table 1 implement this suppression for fair comparisons while our *qualitative* visuals in Figure 3 visualise some low-confidence detections.

Random Seeds As there is inherent randomness in the

¹<https://github.com/deeplearning-wisc/vos>

initialisation of the auxiliary MLP, we report the mean μ and standard deviation σ of each metric over five seeds; the default benchmark seed [9] (0) and four randomly generated seeds in the range of $[1, 10^5]$ for replicability, in the format of $\mu \pm \sigma$.

Baselines We compare against the following state-of-the-art methods: MSP [21], ODIN [32], Mahalanobis Distance [31], Energy Score [35], Gram Matrices [53], ViM [65], KNN [58], Generalized ODIN [23], CSI [60], GAN-Synthesis [30] and Virtual Outlier Synthesis (VOS) [9]. Performance metrics for ViM and KNN are reported from implementations based on public code. Performance metrics for all other methods are reported from [9].

4.2. Implementation

Base Network Architecture Following the evaluation protocol defined in [9], we implement the Faster-RCNN [50] detector with either a ResNet-50 [19] or RegNetX4.0 [44] backbone using the Detectron2 library [67]. All compared methods, excluding VOS [9], are evaluated exclusively using the ResNet-50 backbone consistent with [9]; VOS and SAFE are compared on both the ResNet-50 and RegNetX4.0 backbones. Of the compared methods, Generalized ODIN [23], CSI [60], GAN-Synthesis [30] and VOS [9] all require the base object detector to be retrained with a custom loss objective, we identify these methods with a checkmark \checkmark in Table 1. For a fair comparison, we report the results of SAFE and VOS [9] using both the ResNet-50 and RegnetX4.0 backbones to ensure that the differing on-task performance, which has been shown to affect open-set recognition performance [62], does not bias the results.

Feature Extraction During feature extraction, hooks are applied to the output of the critical residual + BatchNorm layer combinations within the ResNet-50 and RegNetX4.0 backbones of the Faster-RCNN model. Object-specific features $\mathbf{p}_{l,d}$ are retrieved using the ROIAlign [18] operation with the predicted bounding boxes b . Appropriate spatial scaling factors in ROIAlign are set so that features are pooled to a channels length c_l vector per layer l .

MLP Architecture Following previous works on auxiliary network feature monitoring [6], the auxiliary MLP is constructed as a 3-layer fully connected MLP with a single output neuron fed into a Sigmoid activation with a dropout connection before the final layer. The size for each fully connected layer is progressively halved with each consecutive layer. The MLP, initialised with Xavier initialisation [13], is trained for 5 epochs using binary cross entropy loss optimised by SGD with a learning rate of 10^{-3} , momentum of 0.9, dropout rate of 50% and batch size of 32 images².

²The size of each individual batch for the MLP is determined by the number of predicted boxes within the 32 images.

Method	Retrain?	ID: PASCAL-VOC				ID: Berkley DeepDrive-100K			
		OpenImages		MS-COCO		OpenImages		MS-COCO	
		AUROC \uparrow	FPR95 \downarrow	AUROC \uparrow	FPR95 \downarrow	AUROC \uparrow	FPR95 \downarrow	AUROC \uparrow	FPR95 \downarrow
MSP [21]		81.91	73.13	83.45	70.99	77.38	79.04	75.87	80.94
ODIN [32]		82.59	63.14	82.20	59.82	76.61	58.92	74.44	62.85
Mahalanobis [31]		57.42	96.27	59.25	96.46	86.88	60.16	84.92	57.66
Energy Score [35]		82.98	58.69	83.69	56.89	79.60	54.97	77.48	60.06
Gram Matrices [53]		77.62	67.42	79.88	62.75	59.38	77.55	74.93	60.93
ViM [65]		68.73	88.40	71.94	83.47	86.49	53.80	87.17	54.58
KNN [58]		85.08	55.73	86.07	54.50	88.37	44.50	87.45	47.28
Generalized ODIN [23]	✓	79.23	70.28	83.12	59.57	87.18	50.17	85.22	57.27
CSI [60]	✓	82.95	57.41	81.83	59.91	87.99	37.06	84.09	47.10
GAN-Synthesis [30]	✓	82.67	59.97	83.67	60.93	81.25	50.61	78.82	57.03
VOS-ResNet50 [9]	✓	85.23 \pm 0.6	51.33 \pm 1.6	88.70 \pm 1.2	47.53 \pm 2.9	88.52 \pm 1.3	35.54 \pm 1.7	86.87 \pm 2.1	44.27 \pm 2.0
VOS-RegNetX4.0 [9]	✓	87.59 \pm 0.2	48.33 \pm 1.6	89.00 \pm 0.4	47.77 \pm 1.1	92.13 \pm 0.5	27.24 \pm 1.3	89.08 \pm 0.6	36.61 \pm 0.9
SAFE-ResNet50 (ours)		92.28 \pm 1.0	20.06 \pm 2.3	80.30 \pm 2.4	47.40 \pm 3.8	94.64 \pm 0.3	16.04 \pm 0.5	88.96 \pm 0.6	32.56 \pm 0.8
SAFE-RegNetX4.0 (ours)		94.38 \pm 0.2	17.69 \pm 1.0	87.03 \pm 0.5	36.32 \pm 1.1	95.97 \pm 0.1	13.98 \pm 0.3	93.91 \pm 0.1	21.69 \pm 0.5

Table 1. OOD detection results comparing SAFE to state-of-the-art OOD detectors. Comparison metrics are FPR95 and AUROC, directional arrows indicate if higher (\uparrow) or lower (\downarrow) values indicate better performance. **Best** results are shown in **red and bold**, **second** best results are shown in **orange**. Methods that require retraining are indicated with a checkmark \checkmark . Mean and standard deviation over 5 seeds is shown for SAFE. We observe that SAFE provides strong performance across almost all benchmarks and metrics, achieving the highest performance across 7 out of 8 of the benchmark permutations. Notably, we observe substantial reductions in FPR95, particularly when OpenImages is the OOD set, with a greater than 30% reduction for both backbones under the PASCAL-VOC setting.

Transform Implementation We implement FGSM [15], parameterised by a scalar magnitude multiplier ϵ , as our adversarial-perturbation for the surrogate MLP training task. During comparisons in Section 4.3, we set $\epsilon = 8$ when ResNet-50 is the backbone and $\epsilon = 1$ for RegNetX4.0. We ablate the sensitivity to ϵ on ResNet-50 in Section 4.5.

4.3. Results and Discussion

Table 1 compares the performance of our SAFE detector to the current state-of-the-art in OOD object detection. SAFE sets a new state-of-the-art across 7 out of the 8 benchmark permutations. We observe substantial reductions to the FPR95 metric, with the OpenImages as OOD setting improving by more than 30% when PASCAL-VOC is ID and 20% when BDD100K is the ID set, with the most significant differences when comparing directly between ResNet-50 models. These observations are further substantiated when considering SAFE in contrast to other posthoc OOD detectors with substantial performance improvements across the majority of metrics, exemplified by improvements of \sim 35% in FPR95 under the OpenImages setting for both datasets. In summary, SAFE, which does not require retraining, outperforms OOD detectors *that do require retraining*, and significantly outperforms other posthoc OOD detectors.

Robustness We further note that the results from Table 1 demonstrate the robustness of SAFE to varying model architectures (*i.e.* ResNet-50 and RegNetX4.0), given that the target models contain the specified critical layers as discussed in Section 3.1. We reiterate that SAFE does not

require a specified training regime and thus both networks are trained without a specialised loss. Directly comparing between SAFE and VOS [9] on the same ResNet-50 backbone, we observe that SAFE outperforms VOS across all metrics under the BDD100K setting and the majority of metrics when PASCAL-VOC is ID. Under the architectural shift towards the RegNetX4.0 backbone, we observe that SAFE still retains high performance, outperforming VOS under the majority of metrics under the PASCAL-VOC setting and providing higher AUROC and FPR95 results for both OOD sets under the BDD100K setting.

Qualitative Results Figure 3 visualises the object predictions of the base network (Top) and subsequent OOD detections from SAFE (Middle) or VOS [9] (Bottom) on a set of MS-COCO test images when the ID dataset is BDD100K. We observe that SAFE successfully identifies many of the OOD objects within the scenes, reducing the impact of these erroneous predictions during deployment.

Consistent with the quantitative results from Table 1, we observe that SAFE is more reliable at detecting OOD samples than VOS [9]. In particular, we observe that in some instances VOS generates additional erroneous predictions (Figure 3, Columns 2 & 4), flagging only a subset of these instances as OOD. In contrast, SAFE correctly detects all of the object instances predicted by the vanilla network as OOD in these images.

We note that SAFE is susceptible to some failures where an object may have similar features to an ID class. The right-most column of Figure 3 provides two examples of this where the base network predicts vehicle labels

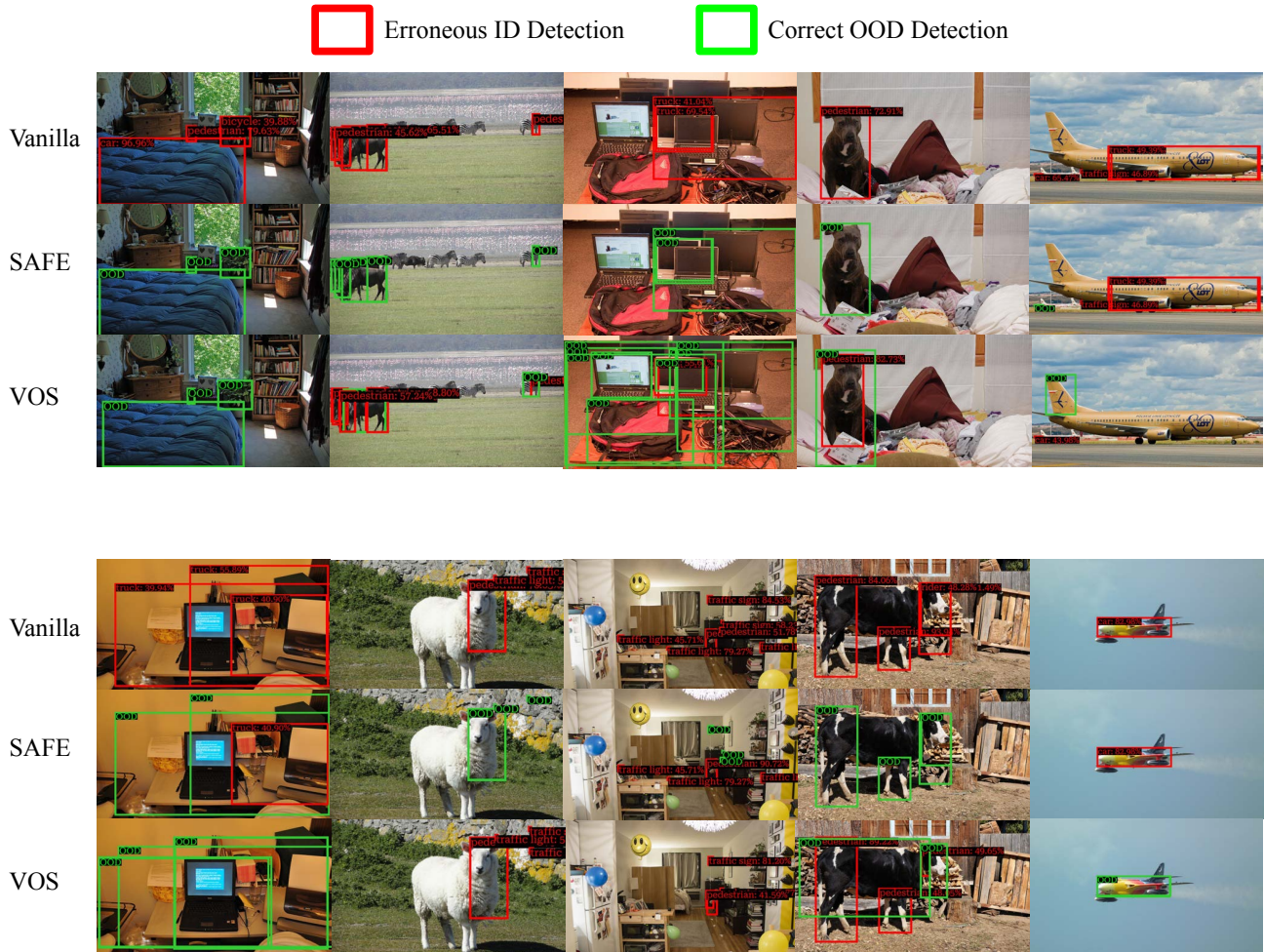


Figure 3. Qualitative visualisation of object detections from the ResNet-50 Faster-RCNN on samples from the MS-COCO OOD dataset. BDD100K is the ID dataset. Binarisation of OOD detection scores are achieved by thresholding the OOD scores using the same threshold used to compute the FPR95 metrics. **Green** bounding boxes signify detections that were *correctly* flagged as OOD and **red** detections are *incorrectly* considered ID. **Top**: Vanilla predictions from the object detector. **Middle**: OOD detections by SAFE. **Bottom**: OOD detections by VOS [9].

(truck/car) onto an airplane which SAFE does not detect as erroneous.

4.4. Layer Importance

Fundamental to the theory of our proposed SAFE detector (Section 3.1) is the importance of residual and BatchNorm layers. Critically, we leverage theoretical and empirical foundations for residual connections enabling *sensitivity* of the network [43] and BatchNorm layers triggering abnormal activations on OOD data [57] to address OOD object detection.

We expand upon these foundations by considering residual convolution + BatchNorm combinations which we expect to leverage the characteristics of both; triggering abnormal activations on OOD inputs which the auxiliary MLP consequently detects. Therefore, we expect that layers that

do not satisfy *both* the residual convolution and BatchNorm combinations will not perform as effectively as those layers that do. We empirically verify this hypothesis by ablating the performance of individual layers (Figure 4) and sampling random layer subsets with increasing size (Table 2). We provide expanded versions of these ablations with an additive noise input perturbation in the Supplementary Material.

Individual Layer Performance Figure 4 ablates the performance of individual Conv2d layers of the ResNet-50 backbone as the average over both OOD datasets under the AUROC (Figure 4, top) and FPR95 (Figure 4, bottom) metrics when PASCAL-VOC is the ID set. Our identified residual convolution + BatchNorm (SAFE) layers are among the highest performing layers in the network. The majority of the other Conv2d layers report low performance,

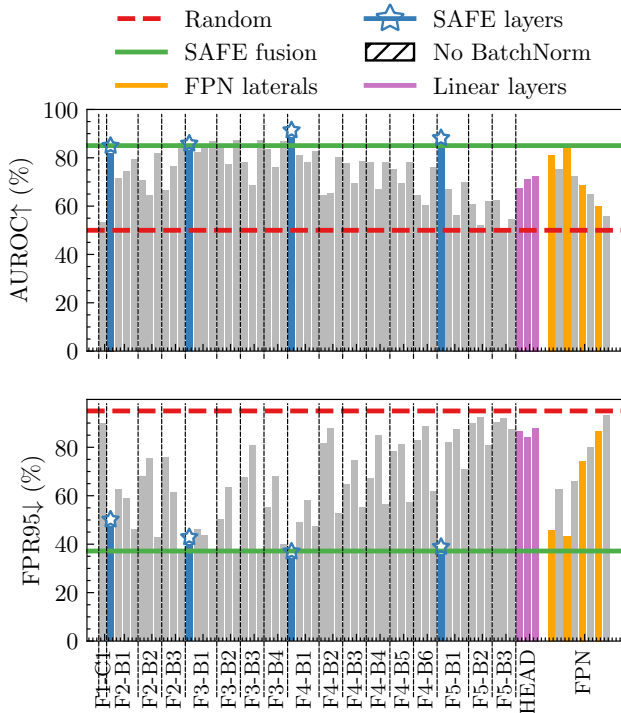


Figure 4. OOD detection performance of individual Conv2d layers in the standard ResNet-50 backbone (see Figure 2) when PASCAL-VOC is the ID set. **Top:** Comparison metric is AUROC, higher is better. **Bottom:** Comparison metric is FPR95, lower is better. Results are reported as averages over both OOD datasets. Layers in blue with a star are the identified critical layers for SAFE. Striped layers belong to the Feature Pyramid Network (FPN) and are the only Conv2d layers that *do not have BatchNorm applied immediately after*. Purple layers are the fully-connected layers of the Faster-RCNN object detector head. The SAFE critical layers consistently provide among the highest performance across all layers within the ResNet-50 backbone.

with very few layers having comparable performance to the SAFE critical layers. Residual connections alone are insufficient as there is no consistently high performance at the beginning of ConvBlocks (separated by vertical dashed lines) which take the added residual from the previous ConvBlock as input or the Feature Pyramid Network lateral connections. Similarly, all of the layers outside of the Feature Pyramid Network are followed immediately by a Batch-Norm layer, but this alone is insufficient since many of these layers produce poor performance. Figure 4 thus provides further empirical evidence, compounding the foundational works in image classification supporting our hypothesis [57, 55, 61, 43, 34], that residual convolution + Batch-Norm layer combinations provide powerful OOD detection performance.

We observe two further characteristics when inspecting Figure 4: (1) A cluster of relatively high-performing

Layers	OpenImages		MS-COCO	
	AUROC \uparrow	FPR95 \downarrow	AUROC \uparrow	FPR95 \downarrow
1	75.31	69.21	68.09	82.63
4	68.82	77.85	65.19	83.91
8	71.75	67.21	65.40	78.28
16	73.02	66.10	67.31	75.83
Residuals	91.33	24.82	81.87	48.45
All (60)	89.88	26.73	81.30	48.57
SAFE	92.28	20.06	80.30	47.40

Table 2. Comparison of varied-size layer combinations detecting OOD data when PASCAL-VOC is the ID set. All compared layer subsets *do not contain the identified sensitive layers used in SAFE*. Colour coding and metrics follow those from Table 1. Mean over 5 seeds is shown. We observe that the sensitive layers utilised by SAFE provide disproportionately high performance for OOD detection, outperforming all layer subsets, of which many have access to more than 2x the number of layers as SAFE. Residual layers produce strong performance, comparable to the fusion of all layers, but are still inferior to the fusion of all layers.

layers between block F3-B1 through to F4-B1 and (2) The highest performing layer in most blocks is the last Conv2d layer. It is not unexpected that there are other high-performing layers other than the SAFE critical layers. Prior works [66, 53, 6, 1] in image classification have established that individual layer performance varies dependent on the ID and OOD data distributions. However, with no theoretical foundation for the selection of these layers, *a priori* selection, *i.e.* selection prior to testing, is infeasible. Furthermore, prior works [66, 53, 6, 1] suggest that the performance of the non-SAFE layers will vary as the surrogate outlier data distribution shifts; we discuss this with the additive noise outliers in the Supplementary Material.

Layer Subsets Table 2 compares performance of randomly selected subsets of layers *that do not contain any of the identified critical layers* against our SAFE detector with only the four critical layers. As expected from observations of Figure 4, using the SAFE layers significantly outperforms the randomly sampled subsets, even when the subsets contain more layers than SAFE. Subsets of layers perform worse than randomly sampled individual layers, with the performance gap tapering off as the subsets get larger. We attribute this characteristic to the large prevalence of poorly performing layers in the backbone, where the signal from high-performing layers is lowered due to noise from poor-performing layers in the smaller subsets. Using all 60 layers produces better performance than any of the subsets, but still underperforms when compared to our four SAFE layers.

Consistent with the theory described in Section 3.1, Table 2 demonstrates that the 12 residual connections produce strong performance, performing comparable to the fusion of all Conv2d layers. Whilst the residual connections do provide strong performance, they are outperformed by the

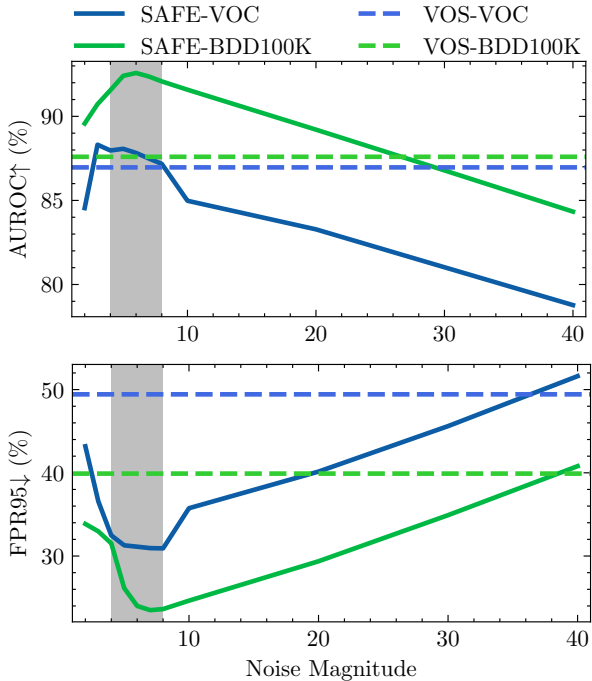


Figure 5. OOD detection performance of SAFE with the ResNet-50 backbone as the gradient sign magnitude ϵ is varied. **Top:** Comparison metric is AUROC, higher is better. **Bottom:** Comparison metric is FPR95, lower is better. Individual lines correspond to the average performance over both OOD sets for the given ID set. Dashed lines correspond to the performance of VOS [9] for the respective datasets. A region of consistent high performance for all ID and OOD permutations exists between $\epsilon \in [4, 8]$ (grey region), suggesting that values in and near this range will generalise well to additional datasets.

SAFE critical layers across both metrics under the OpenImages setting and FPR95 under the MS-COCO setting.

We note that the size of the auxiliary MLP input scales with the number of layers, and hence feature dimensionality, in the subset. This entails $O(n^2)$ scaling in the weight matrices of the auxiliary MLP, making direct inclusion of large subsets (*e.g.* the 12 residual connections) or all layers computationally expensive.

4.5. Gradient Magnitude Sensitivity

Figure 5 ablates the sensitivity of the auxiliary MLP to varying values of the gradient sign magnitude ϵ when PASCAL-VOC is the ID set. In general, the performance curves reported match expectations where we observe initial low relative performance due to the MLP being unable to effectively discriminate between the perturbed ID and clean ID features, which improves up to a peak and is followed by a drop in performance as the weighting parameter ϵ becomes too large, destroying too much of the input content. Critically, we make the observation that a region

of high performance exists across all ID, OOD and metric permutations, residing approximately within $\epsilon \in [4, 8]$. The consistently high performance across both ID and OOD dataset permutations suggests that values in this range generalise well to unseen data.

We further note that Figure 5 shows that SAFE generally performs well under a wide range of perturbation magnitudes. Comparing the performance under the FPR95 metric, we observe that only the edge cases of very large values of ϵ result in worse performance than the previous state-of-the-art. This argument holds particularly true for BDD100K, where a random ϵ value could be selected in the range of $\epsilon \in [1, 20]$ and SAFE would retain better performance under both AUROC (Figure 5, top) and FPR95 (Figure 5, bottom) than the state-of-the-art, VOS [9].

5. Conclusion

In this paper, we propose SAFE, a novel OOD detection framework that leverages the layers in an object detector’s backbone that are most *sensitive* to OOD inputs. Unlike previous feature-based OOD object detectors, SAFE leverages the backbone of an object detector network, identifying that the subset of residual convolutions followed by batch normalisation are consistently among the most powerful layers in the network at detecting out-of-distribution samples.

To take advantage of these powerful layers, SAFE trains an auxiliary MLP on the *surrogate* task of distinguishing minimally perturbed adversarial ID samples to clean ID samples using only the features from this subset of layers. We provide a theoretical grounding for the disproportionate power of these layers from image classification literature, expanding upon it to the challenging task of OOD object detection, where we are the first to demonstrate these characteristics. We provide empirical evidence supporting our theory, demonstrating that our identified SAFE layers are among the most powerful layers individually and outperform the fusion of much larger subsets of layers.

SAFE is the first method that considers the *sensitivity* and the impact of individual layers under the setting of OOD object detection. We are optimistic for future work expanding upon our findings through further leveraging our identified sensitive layers, integration of backbone features into OOD object detection, and further theoretical analysis on *sensitivity* and *smoothness* in object detection. We believe that SAFE represents an important step forward in our understanding of OOD object detection and offers a promising avenue for future research.

Acknowledgements: The authors acknowledge continued support from the Queensland University of Technology (QUT) through the Centre for Robotics. TF was partially supported by funding from ARC Laureate Fellowship FL210100156 and Intel Research via grant RV3.290.Fischer.

References

- [1] Vahdat Abdelzad, Krzysztof Czarnecki, Rick Salay, Taylor Denouden, Sachin Vernekar, and Buu Phan. Detecting out-of-distribution inputs in deep neural networks using an early-layer output. *arXiv preprint arXiv:1910.10307*, 2019. 2, 3, 8
- [2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. 1
- [3] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572, 2016. 1
- [4] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *Pattern Recognition*, pages 33–47, 2019. 2
- [5] Akshay Raj Dhamija, Manuel Günther, Jonathan Ventura, and Terrance E. Boult. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1010–1019, 2020. 1, 3
- [6] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and H.T. Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19217–19227, 2022. 1, 2, 3, 5, 8
- [7] Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems*, 2022. 1, 3
- [8] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don’t know from videos in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13678–13688, 2022. 1, 3
- [9] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don’t know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 3, 5, 6, 7, 9
- [10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 5
- [11] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2
- [12] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 2
- [13] Xavier Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9:249–256, 2010. 5
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2
- [15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 1, 4, 6
- [16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017. 1, 2
- [17] Ali Harakeh and Steven L. Waslander. Estimating and evaluating regression predictive uncertainty in deep object detectors. In *International Conference on Learning Representations*, 2021. 5
- [18] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 5
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 2, 3, 4, 5
- [20] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022. 2
- [21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 5, 6
- [22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 1, 2
- [23] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10951–10960, 2020. 1, 2, 5, 6
- [24] Chengjie Huang, Van Duong Nguyen, Vahdat Abdelzad, Christopher Gus Mannes, Luke Rowe, Benjamin Therien, Rick Salay, and Krzysztof Czarnecki. Out-of-distribution detection for lidar-based 3d object detection. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, page 4265–4271, 2022. 3
- [25] Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. In *Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI)*, 2021. 2
- [26] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

- [27] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456, 2015. 3
- [28] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5830–5840, 2021. 3
- [29] Ivan Krasin et al. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://storage.googleapis.com/openimages/web/index.html>*, 2017. 5
- [30] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2, 5, 6
- [31] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 7167–7177, 2018. 1, 2, 5, 6
- [32] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 5, 6
- [33] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 5
- [34] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7498–7512, 2020. 1, 2, 3, 8
- [35] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 5, 6
- [36] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021. 1
- [37] Ahsan Mahmood, Junier Oliva, and Martin Andreas Styner. Multiscale score matching for out-of-distribution detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2
- [38] Dimity Miller, Georgia Goode, Callum Bennie, Peyman Moghadam, and Raja Jurdak. Why object detectors fail: Investigating the influence of the dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4823–4830, 2022. 3
- [39] Dimity Miller, Peyman Moghadam, Mark Cox, Matt Wildie, and Raja Jurdak. What’s in the black box? the false negative mechanisms inside object detectors. *IEEE Robotics and Automation Letters*, 7(3):8510–8517, 2022. 3
- [40] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3243–3249, 2018. 1, 3
- [41] Dimity Miller, Niko Sünderhauf, Michael Milford, and Feras Dayoub. Uncertainty for identifying open-set errors in visual object detection. *IEEE Robotics and Automation Letters*, 7(1):215–222, 2022. 3
- [42] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 3
- [43] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarín Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. In *International Conference on Machine Learning (ICML) Workshops*, 2021. 1, 2, 3, 7, 8
- [44] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollar. Designing network design spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10428–10436, 2020. 2, 3, 5
- [45] Quazi Marufur Rahman, Niko Sünderhauf, and Feras Dayoub. Online monitoring of object detection performance post-deployment. *arXiv preprint arXiv:2011.07750*, 2020. 3
- [46] Quazi Marufur Rahman, Niko Sünderhauf, and Feras Dayoub. Did you miss the sign? A false negative alarm system for traffic sign detectors. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3748–3753, 2019. 3
- [47] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to Mahalanobis distance for improving near-OOD detection. In *International Conference on Machine Learning (ICML) Workshop on Uncertainty and Robustness in Deep Learning*, 2021. 2
- [48] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [49] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020. 1
- [50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 5
- [51] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 1
- [52] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class

- classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3379–3388, 2018. [2](#)
- [53] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8491–8501, 2020. [1](#), [2](#), [3](#), [5](#), [6](#), [8](#)
- [54] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. [1](#)
- [55] Lewis Smith, Joost van Amersfoort, Haiwen Huang, Stephen J. Roberts, and Yarin Gal. Can convolutional resnets approximately preserve input distances? A frequency analysis perspective. *CoRR*, abs/2106.02469, 2021. [1](#), [2](#), [3](#), [8](#)
- [56] Kumar Sricharan and Ashok Srivastava. Building robust classifiers through generation of confident out of distribution examples. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Bayesian Deep Learning*, 2018. [2](#)
- [57] Yiyao Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. [1](#), [2](#), [3](#), [7](#), [8](#)
- [58] Yiyao Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022. [2](#), [5](#), [6](#)
- [59] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, and Peter Corke. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018. [1](#)
- [60] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [2](#), [5](#), [6](#)
- [61] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. [1](#), [2](#), [8](#)
- [62] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. [5](#)
- [63] Sachin Vernekar. Training reject-classifiers for out-of-distribution detection via explicit boundary sample generation. Masters, University of Waterloo, Waterloo, 2020. [2](#)
- [64] Sachin Vernekar, Ashish Gaurav, Vahdat Abdelzad, Taylor Denouden, Rick Salay, and K. Czarnecki. Out-of-distribution detection in classifiers via generation. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Safety and Robustness in Decision Making*, 2019. [2](#)
- [65] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4921–4930, 2022. [2](#), [5](#), [6](#)
- [66] Samuel Wilson, Tobias Fischer, Niko Sünderhauf, and Feras Dayoub. Hyperdimensional feature fusion for out-of-distribution detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. [1](#), [2](#), [3](#), [5](#), [8](#)
- [67] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. [5](#)
- [68] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyao Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [1](#), [2](#)
- [69] Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. Improved OOD generalization via adversarial training and pretraing. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 11987–11997, 2021. [1](#), [2](#)
- [70] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020. [1](#), [5](#)
- [71] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9518–9526, 2019. [2](#)
- [72] Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9452–9461, 2021. [2](#), [5](#)
- [73] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. [5](#)