

# OUT-OF-DISTRIBUTION DETECTION & APPLICATIONS WITH ABLATED LEARNED TEMPERATURE ENERGY

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

As deep neural networks become adopted in high-stakes domains, it is crucial to be able to identify when inference inputs are Out-of-Distribution (OOD) so that users can be alerted of likely drops in performance and calibration (Ovadia et al. 2019) despite high confidence (Nguyen et al. 2015). Among many others, existing methods use the following two scores to do so without training on any apriori OOD examples: a learned temperature (Hsu et al. 2020) and an energy score (Liu et al. 2020). In this paper we introduce Ablated Learned Temperature Energy (or "AbET" for short), a method which combines these prior methods in novel ways with an effective ablation. Due to these contributions, AbET lowers the False Positive Rate at 95% True Positive Rate (FPR@95) by 47.32% in classification (averaged across all ID and OOD datasets measured) compared to state of the art without training networks in multiple stages or requiring hyperparameters or test-time backward passes. We additionally provide empirical insights as to why our model learns to distinguish between In-Distribution (ID) and OOD samples while only being explicitly trained on ID samples via exposure to misclassified ID examples at training time. Lastly, we show the efficacy of our method in identifying predicted bounding boxes and pixels corresponding to OOD objects in object detection and semantic segmentation, respectively - with an AUROC increase of 5.15% in object detection and both a decrease in FPR@95 of 41.48% and an increase in AUPRC of 34.20% on average in semantic segmentation compared to previous state of the art.<sup>2</sup>

## 1 INTRODUCTION

In recent years, machine learning models have shown impressive performance on fixed distributions (Lin et al. 2017; Ren et al. 2015; Girshick 2015; Liu et al. 2016; Redmon et al. 2016; Dai et al. 2016; Radford et al. 2021; Feichtenhofer 2020; Devlin et al. 2018; Dosovitskiy et al. 2020). However, distribution from which examples are drawn at inference time are not always stationary or overlapping with training distributions. In these cases where inference examples are far from the training set, not only does model performance drop, all known uncertainty estimates also become miscalibrated (Ovadia et al. 2019) - i.e. the output probabilities of the model become greatly misaligned with performance. Without OOD detection, users can be completely unaware of these drops in performance and calibration, and often can be fooled into false trust in model predictions due to high confidence on OOD inputs (Nguyen et al. 2015). Thus, identifying the presence of examples which are far from the training set is of utmost importance to AI safety and reliability.

Aimed at OOD detection, existing methods have explored (among many other methods) modifying models via a learned temperature which is dynamic depending on input (Hsu et al. 2020) and an inference-time post-processing energy score which measures the log of the

---

<sup>2</sup>We make our code publicly available at <https://github.com/anonymousoodauthor/abet> with our method requiring only a single line change to the architectures of classifiers, object detectors, and segmentation models prior to training.

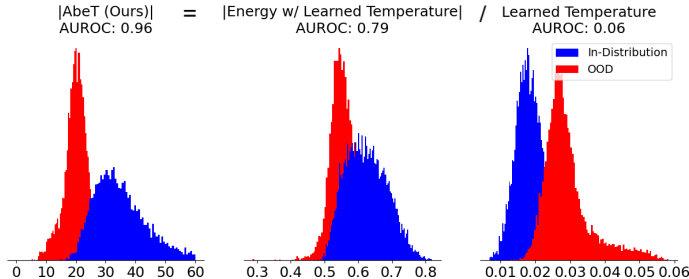


Figure 1: Histograms showing the separability between OOD scores on OOD inputs (red) and ID inputs (blue) for different methods. The goal is to make these red and blue distributions as separable as possible, with scores on OOD inputs (red) close to 0 and scores on ID inputs (blue) of high-magnitude (away from 0). **(Center)** Our first contribution is replacing the Scalar Temperature in the Energy Score (Liu et al., 2020) with a Learned Temperature (Hsu et al., 2020). This infusion leads to Equation 2 with the Learned Temperature showing up in the Exponential Divisor Temperature (overlined in Equation 2) and Forefront Temperature Constant (underlined in Equation 2) forms. **(Right)** The Forefront Temperature Constant contradicts the desired property of scores being close to 0 for OOD points (red) and of high magnitude for ID points (blue). **(Left)** Therefore, our second contribution is to ablate this Forefront Temperature Constant, leading to our final Ablated Learned Temperature Energy (AbeT) score. This ablation increases the separability of the OOD scores vs. ID scores, as can be seen visually and numerically (in terms of AUROC) comparing the center and left plots - where the only difference is this ablation of the Forefront Temperature Constant. Higher AUROC means more separability.

exponential sum of logits on a given input, scaled by a scalar temperature (Liu et al., 2020). In this paper, we combine these methods and introduce an ablation, leading to our method deemed "AbeT." Due to these contributions, we demonstrate the efficacy of AbeT over existing OOD methods. We establish state of the art performance in classification, object detection, and semantic segmentation on a suite of common OOD benchmarks spanning a variety of scales and resolutions. We also perform extensive visual and empirical investigations to understand our algorithm. Our **key results and contributions** are as follows:

- We combine the learned temperature (Hsu et al., 2020) and post-processing energy score (Liu et al., 2020) by simply using the learned temperature in the calculation of the energy score.
- We resolve a contradiction in the energy score (Liu et al., 2020) by ablating one of the learned temperature terms. We deem this "Ablated Learned Temperature Energy" (or "AbeT" for short) and it serves as our ultimate OOD score.
- We provide visual and empirical evidence to provide intuition as to why our method is able to achieve superior OOD detection performance despite not being exposed to explicit OOD inputs at training time via exposure to misclassified ID samples.
- We show the efficacy of AbeT in identifying predicted bounding boxes and pixels corresponding to OOD objects in object detection and semantic segmentation, respectively.

## 2 PRELIMINARIES

Let  $X$  and  $Y$  be the input and response random variables with realizations  $x \in \mathbb{R}^D$  and  $y \in \{1, 2, \dots, C - 1, C\}$  (where  $C$  is the number of output classes), respectively. Typically  $X$  has information about  $Y$  and we'd like to make inferences about  $Y$  given  $X$  using a learned model  $\hat{f} : \mathbb{R}^D \rightarrow \mathbb{R}^C$ . In practice, a learner only has access to a limited amount of training examples in this data-set  $D_{in}^{train} = \{(x_i, y_i)\}_{i=1}^N$  which are realizations of  $(X, Y)$  (or a subset thereof) on which to train  $\hat{f}$ .

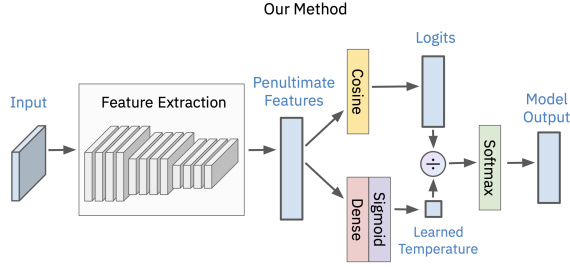


Figure 2: The network architecture of a classification network with a learned temperature.

## 2.1 PROBLEM STATEMENT

We define  $D_{in}^{test}$  identical to  $D_{in}^{train}$  but unseen at training time. And we define  $D_{out}^{test}$  as any dataset that has non-overlapping output classes with those of  $D_{in}^{train}$ , as is standard in OOD detection evaluations (Huang et al. 2021; Hsu et al. 2020; Liu et al. 2020; Djuricic et al. 2022; Sun et al. 2021; Hendrycks & Gimpel, 2016; Liang et al. 2017; Sun et al. 2022; Katz-Samuels et al. 2022). The goal of Out-of-Distribution Detection is to define a score  $S$  such that  $S(x_{out})$  and  $S(x_{in})$  are far from each other  $\forall x_{out} \in D_{out}^{test}, x_{in} \in D_{in}^{test}$ .

## 2.2 STANDARD CLASSIFICATION MODEL OPTIMIZATION

In OOD detection,  $\hat{f}$  serves a dual purpose: its outputs are directly optimized to classify among outputs  $\{1, 2, \dots, C - 1, C\}$  and functions of the network are used as inputs to  $S$ . Though we use  $\hat{f}$  for both purposes, our aim is for AbeT to neither have OOD data at training time nor significantly modify training to account for the ability to detect OOD data. Therefore, we train our classification models in a standard way: to minimize a loss function  $\mathcal{L} = \sum_{i=1}^N \mathcal{L}(\hat{f}(x_i; \theta), y_i)$ , where  $(x_i, y_i) \in D_{in}^{train}$ . In deep learning classification settings, the cross-entropy loss function is normally used for training:  $\mathcal{L}_{CE}(\hat{f}(x_i; \theta), y_i) = -\log \hat{f}_{y_i}(x_i; \theta)$ . Networks can be optimized towards other loss functions, but we did not test our method in conjunction with any other loss functions.

## 2.3 MODEL OUTPUT

To estimate  $\hat{f}_{y_i}(x_i; \theta)$ , models typically use logit functions per class which calculate the activation of class  $c$  on input  $x_i$  as  $L_c(x_i; \theta) = \hat{g}_c(x_i; \theta)$ .

We now define  $\hat{g}_c$ : let  $w$  and  $b$  represent the weights and biases of the final layer of a network (mapping penultimate space to output space) respectively and  $f^p(x_i)$  represent the penultimate representation of the network on input  $x_i$ . Normally, deep networks use the inner product of the  $w$  and  $x_i$  (plus the bias term) as the logit function:  $\hat{g}_c(x_i; \theta) = w_c^T f^p(x_i) + b_c$ . However, Hsu et al. (2020) found that a logit function based on the cosine similarity between  $w$  and  $f^p(x_i)$

$$\hat{g}_c(x_i; \theta) = \frac{w_c^T f^p(x_i)}{\|w_c^T\| \|f^p(x_i)\|}$$

was more effective when training logit functions that serve this dual-purpose of classifying among outputs  $\{1, 2, \dots, C - 1, C\}$  and as an input to  $S$ . We therefore use this cosine-similarity score as our logit function  $\hat{g}$ .

We now discuss tempering the logit function  $L$ : often employed to increase calibration, Temperature Scaling (Guo et al. 2017) geometrically decreases the logit function  $L$  by a single scalar  $T_{\text{scalar}}$ . That is,  $\hat{f}$  has a logit function that employs a scalar temperature as  $L_c(x_i; \theta, T_{\text{scalar}}) = \hat{g}_c(x_i; \theta) / T_{\text{scalar}}$ . Introduced in Hsu et al. (2020), a learned temperature  $T_{\text{learned}} : \mathcal{X} \rightarrow (0, 1)$  is a temperature that depends on input  $x_i$ . That is,  $\hat{f}$  has a logit function that employs a learned temperature

$$L_c(x_i; \theta, T_{\text{learned}}) = \hat{g}_c(x_i; \theta) / T_{\text{learned}}(x_i) \quad (1)$$

The softmax of this tempered logit serves as the final model prediction:

$$\hat{f}_{y_i}(x_i; \theta) = \frac{\exp(\hat{g}_{y_i}(x_i; \theta)/T_{\text{learned}}(x_i))}{\sum_{c=1}^C \exp(\hat{g}_c(x_i; \theta)/T_{\text{learned}}(x_i))}$$

This is input to the loss  $\mathcal{L}$  during training. In this formulation,  $\hat{g}$  and  $T_{\text{learned}}$  serve disjoint purposes:  $\hat{g}$  reduces loss by selecting  $\hat{g}_{y_i}(x_i; \theta)$  as highest among  $\{\hat{g}_c(x_i; \theta)\}_{c=1}^C$ ; and  $T_{\text{learned}}(x_i)$  reduces loss by modifying the confidence (but not changing the classification) such that the confidence is high or low when the model is correct or incorrect, respectively.

We provide a visual architecture of a forward pass with a learned temperature in Figure 2.

### 2.3.1 LEARNED TEMPERATURE DETAILS

For all AbET models, the learned temperature function is architecturally identical to that in Hsu et al. (2020): a single fully connected layer which takes inputs from the penultimate layer and outputs a single scalar per input, passes these scalars through a Batch Normalization layer, then activates these scalars via a sigmoid which normalizes them to (0, 1). The learned temperature is automatically trained via the gradients induced by the back-propagation of the network’s optimization since the learned temperature modifies the logits used in the forwards pass of the model. In this way, it is updated like any other layer which affects the forwards pass of the network, and thus requires no tuning, training, or datasets other than the cross-entropy-based optimization using the training dataset. The only requirement to train the learned temperature is this one-line architectural change prior to training or fine-tuning. For analysis on the insignificance of the inference time and memory costs due to the learned temperature, see Appendix Section A.2.

## 3 OUR APPROACH: ABET

The following post-processing energy score was previously used for OOD detection in Liu et al. (2020):  $E(x_i; L, T_{\text{scalar}}, \theta) = -T_{\text{scalar}} \log \sum_{c=1}^C e^{L_c(x_i; \theta, T_{\text{scalar}})}$ . This energy score was intended to be highly negative on ID input and close to 0 on OOD inputs via high logits on ID inputs and low logits on OOD inputs.

Our first contribution is replacing the scalar temperature with a learned one:

$$E(x_i; L, T_{\text{learned}}, \theta) = - \underbrace{T_{\text{learned}}(x_i)}_{\text{Forefront Temperature Constant}} \log \sum_{c=1}^C e^{L_c(x_i; \theta, \overbrace{T_{\text{learned}}}^{\text{Exponential Divisor Temperature}})} \quad (2)$$

By introducing this learned temperature, there become two ways to control the OOD score: by modifying the logits and by modifying the learned temperature. We note that there are two different operations that the learned temperature performs in terms of modifying the energy score. We deem these two operations the "Forefront Temperature Constant" and the "Exponential Divisor Temperature", which are underlined and overlined, respectively, in Equation 2. Our second contribution is noting that only the Exponential Divisor Temperature contributes to the OOD score being in adherence with this property of highly negative on ID inputs and close to 0 on OOD inputs, while Forefront Temperature Constant counteracts that property - we therefore ablate this Forefront Temperature Constant. To see this, we note that the learned temperature is trained to be higher on inputs on which the classifier is uncertain - such as OOD and misclassified ID inputs - in order to deflate the softmax confidence on those inputs (i.e. increase softmax uncertainty); and the learned temperature is trained to be lower on inputs on which the classifier is certain - such as correctly classified ID inputs - in order to inflate the softmax confidence on those inputs (i.e. increase softmax certainty). Thus, the temperature term being high on OOD inputs and low on ID inputs leads to the energy score being closer to 0 on OOD inputs and more highly negative on ID inputs when used in the Exponential Divisor Term, as desired. However, the temperature term being high on OOD inputs and low on ID inputs leads to the energy score being more highly negative on OOD inputs and closer to 0 on ID inputs when being used in the

Forefront Temperature Constant, which is the opposite of what we would like. We therefore (as previously mentioned) simply ablate the Forefront Temperature Constant, leading to the following Ablated Temperature Energy score:

$$\text{AbeT}(x_i; L, T_{\text{learned}}, \theta) = -\log \sum_{c=1}^C e^{L_c(x_i; \theta, T_{\text{learned}})}$$

We visualize the effects of this ablation in Figure 1 using Places365 (Zhou et al. 2018) as the OOD dataset, CIFAR-100 (Krizhevsky, 2009) as the ID dataset, and a ResNet-20 (He et al. 2016a) trained with learned temperature and a cosine logit head as the model.

## 4 CLASSIFICATION EXPERIMENTS

### 4.1 CLASSIFICATION EXPERIMENTAL SETUP

#### 4.1.1 CLASSIFICATION DATASETS

We follow standard practices in OOD evaluations (Huang et al. 2021; Hsu et al. 2020; Liu et al. 2020; Djuricic et al. 2022; Sun et al. 2021; Hendrycks & Gimpel 2016; Liang et al. 2017; Sun et al. 2022; Katz-Samuels et al. 2022) in terms of metrics, ID datasets, and OOD datasets. For evaluation metrics, we use AUROC and FPR@95. We measure performance at varying number of ID classes via CIFAR-10 (Krizhevsky, 2009), CIFAR-100 (Krizhevsky, 2009), and ImageNet-1k (Huang & Li 2021). For our CIFAR experiments, we use 4 OOD datasets standard in OOD detection: Textures (Cimpoi et al. 2014), SVHN (Netzer et al. 2011), LSUN (Crop) (Yu et al. 2015), and Places365 (Zhou et al. 2018). For our ImageNet-1k experiments we also evaluate on four standard OOD test datasets, but subset these datasets to classes that are non-overlapping with respect to ImageNet-1k as is common practice (Huang et al. 2021; Sun et al. 2022, 2021; Sun & Li 2021): iNaturalist (Van Horn et al. 2018), SUN (Xiao et al. 2010), Places365 (Zhou et al. 2018), and Textures (Cimpoi et al. 2014). See Appendix D.1.1 for details about these datasets.

#### 4.1.2 CLASSIFICATION MODELS AND HYPERPARAMETERS

For all experiments with CIFAR-10 and CIFAR-100 as ID data, we use a ResNet-20 (He et al. 2016a). For all experiments with ImageNet-1k as ID data, we use a ResNetv2-101 (He et al. 2016b). We present additional experiments where we retain top OOD performance with ImageNet-1k as the ID data using an alternative architecture, DenseNet-121 (Huang et al. 2017), in Appendix Section B.3. All models are trained from scratch. For more experimental details, see Appendix Section D.2.1.

#### 4.1.3 PREVIOUS CLASSIFICATION APPROACHES

We compare against previous methods with similar constraints, training settings, and testing settings. We note that we do not compare against other methods that are trained on OOD data (Ming et al. 2022; Katz-Samuels et al. 2022; Hendrycks et al. 2018) or methods that require multiple stages of training (Khalid et al. 2022). Principally, we compare against Maximum Softmax Probability (Hendrycks & Gimpel 2016), ODIN (Liang et al. 2017), GODIN (Hsu et al. 2020), Mahalanobis (Lee et al. 2018), Energy Score (Liu et al. 2020), Gradient Norm (Huang et al. 2021), ReAct (Sun et al. 2021), DICE (Sun & Li 2021), Deep Nearest Neighbors (DNN) (Sun et al. 2022), and ASH (Djuricic et al. 2022). For more details about these competitive methods, see Appendix Section E.

### 4.2 PERFORMANCE ON STANDARD OOD IN CLASSIFICATION SUITE

In Table 1 we compare against the aforementioned methods outlined in Section 4.1.3. All results are averaged across the four previously mentioned OOD test datasets per ID dataset outlined in Section 4.1.1, with the standard deviations calculated across these same 4 OOD datasets. All OOD methods keep accuracy within 1% of their respective baseline methods without any modifications to account for OOD. Results from MSP (Hendrycks & Gimpel

$D_{in}^{test}$	CIFAR-10		CIFAR-100		ImageNet-1k	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
Method						
MSP	60.5 ± 15	89.5 ± 3	82.7 ± 11	71.9 ± 7	63.9 ± 8	79.2 ± 5
ODIN	39.1 ± 24	92.4 ± 4	73.3 ± 31	75.3 ± 13	72.9 ± 7	82.5 ± 5
Mahalanobis	37.1 ± 35	91.4 ± 6	63.9 ± 16	85.1 ± 5	81.6 ± 19	62.0 ± 11
Gradient Norm	28.3 ± 24	93.1 ± 6	56.1 ± 38	81.7 ± 13	54.7 ± 7	86.3 ± 4
DNN	49.0 ± 11	83.4 ± 5	66.6 ± 13	78.6 ± 5	61.9 ± 6	82.9 ± 3
GODIN	26.8 ± 10	94.2 ± 2	47.0 ± 7	90.7 ± 2	52.7 ± 5	83.9 ± 4
Energy	39.7 ± 24	92.5 ± 4	70.5 ± 32	78.0 ± 12	71.0 ± 7	82.7 ± 5
Energy + ReAct	39.6 ± 15	93.0 ± 2	62.8 ± 17	86.3 ± 6	31.4*	92.9*
Energy + DICE	20.8 ± 1	95.2 ± 1	49.7 ± 1	87.2 ± 1	34.7*	90.7*
Energy + ASH	20.0 ± 21	95.4 ± 5	37.6 ± 34	89.6 ± 12	16.7 ± 13	96.5 ± 2
AbeT	<b>12.5 ± 2</b>	<b>97.8 ± 1</b>	<b>31.1 ± 12</b>	<b>94.0 ± 1</b>	40.0 ± 11	91.8 ± 3
AbeT + ReAct	<b>12.2 ± 1</b>	<b>97.8 ± 1</b>	<b>26.2 ± 7</b>	<b>94.1 ± 2</b>	38.1 ± 11	92.2 ± 3
AbeT + DICE	<b>11.6 ± 2</b>	<b>97.9 ± 1</b>	<b>31.3 ± 13</b>	<b>94.3 ± 2</b>	30.7 ± 15	93.2 ± 3
AbeT + ASH	<b>10.9 ± 5</b>	<b>97.9 ± 1</b>	<b>30.6 ± 12</b>	<b>94.4 ± 2</b>	<b>3.7 ± 3</b>	<b>99.0 ± 1</b>

\* Results where a ResNet-50 is used as in their corresponding papers instead of ResNet-101 as in our experiments. This is due to our inability to reproduce their results with ResNet-101

Table 1: **Comparison with other competitive OOD detection methods in classification.** OOD detection results on a suite of standard datasets compared against competitive methods which are trained with ID data only and require only one stage of training. All results are averaged across 4 OOD datasets, with the standard deviations calculated across these same 4 OOD datasets. ↑ means higher is better and ↓ means lower is better.

[2016], ODIN (Liang et al., 2017), Energy (Lee et al., 2018), and Gradient Norm (Huang et al., 2021) are taken from Huang et al. (2021) and results of Mahalanobis using ImageNet-1k as the ID dataset are taken from Huang & Li (2021), as their models, datasets, and hyperparameters are identical to ours. We provide detailed results for each OOD test dataset in Appendix Section B.

We note that our method achieves an average reduction in FPR@95 of 45.50% on CIFAR-10, 18.61% on CIFAR-100, and 77.84% on ImageNet. We additionally note that not only is the mean performance of our method superior in all settings, but the standard deviation of our performance across OOD datasets is relatively low in nearly all cases, meaning our method is **consistently** performant.

We additionally present a study of our Forefront Temperature Constant ablation in Appendix Section B.1 and show that this singular ablation contribution leads to a reduction in FPR@95 of 28.76%, 59.00%, and 24.81% with CIFAR-10, CIFAR-100, and ImageNet as the ID datasets respectively (averaged across their 4 respective OOD datasets) compared to AbeT without the Forefront Temperature Constant ablation.

In Appendix Section B.4 we present Gradient Input Perturbation (Liang et al., 2017) in conjunction with our method and show that it harmed our method.

We also present experiments in Appendix Section B.2 where we replace the Cosine Logit Head with the standard Inner Product Head which reaffirms the finding of Hsu et al. (2020) that the Cosine Logit Head is preferable over the Inner Product Head in OOD Detection.

## 5 UNDERSTANDING ABET

In Section 5 we provide intuition-building evidence to suggest that the superior OOD performance of AbeT despite not being exposed to explicit OOD samples at training time is due to exposure to misclassified ID examples during training. Towards building intuition we provide visual evidence in Figure 3 based on TSNE-reduced (Van der Maaten & Hinton, 2008) embeddings to support the following two hypotheses:



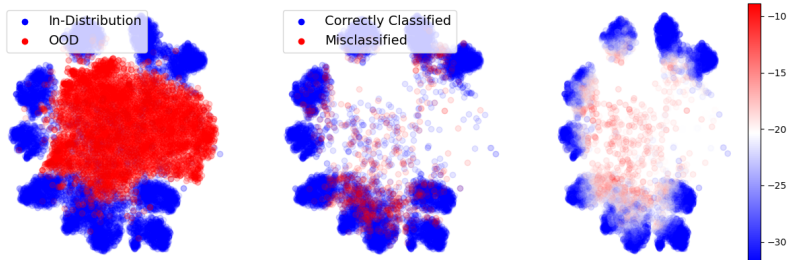


Figure 3: **(Left)** Scatter plot of OOD LSUN examples (red) and ID test CIFAR-10 examples (blue). **(Center)** ID test CIFAR-10 examples correctly classified (blue) and incorrectly classified (red). **(Right)** ID test CIFAR-10 examples colored by their  $\text{AbE}$  score. Red is estimated to be more OOD. The learned temperature increasing on misclassified points (in order to deflate softmax confidence when incorrect) leads our score to inflate towards 0 on misclassified points, as can be seen in the center plot. The presence of a comparatively higher proportion of points in the center of penultimate representation space which are misclassified therefore leads to the relationship that our score inflates towards 0 as distance to the center decreases (as can be seen on ID points in the right plot). In combination with OOD points lying in the center of penultimate representation space (as can be seen on the left plot), this means that our scores are close to 0 on OOD points - thus providing intuition (but not proof) as to why our method is able to achieve superior OOD detection performance.

1. Our method learns a representation function such that OOD points are closer to misclassified ID points than correctly classified ID points, in general.
2. Our OOD scores are comparatively higher (closer to 0) on misclassified ID examples.

These combined hypotheses suggest that our score responds to OOD points similarly to the way it responds on misclassified ID points. Our scores are therefore close to 0 on OOD points due to the OOD points learning this inflation of our score (towards 0) from the (sparse) misclassified points near them, while this inflation of our score doesn't apply as much to ID points overall - as desired. This is learned without ever being exposed to OOD points at training time.

In Appendix Section [C.1](#), we present empirical evidence to support these two hypothesis which does not utilize dimensionality reduction.

## 6 APPLICATIONS IN SEMANTIC SEGMENTATION & OBJECT DETECTION

### 6.1 SEMANTIC SEGMENTATION

In Table [2](#), we compare against competitive OOD Detection methods in semantic segmentation that predict which pixels correspond to object classes not found in the training set (i.e. which pixels correspond to OOD objects). For evaluations, we report ID metric mIOU and OOD metrics FPR@95, AUPRC, and AUROC. For our experiments with  $\text{AbE}$ , we replace the Inner Product per-pixel in the final convolutional layer with a Cosine Logit head per-pixel and a learned temperature layer per-pixel. Further details about model training can be found in Appendix Section [D.2.2](#). We compare against methodologies which follow similar training, inference, and dataset paradigms. Notably, similar to classification, we do not compare with methods which fine-tune or train on OOD (or OOD proxy data like COCO [\(Lin et al., 2014b\)](#)) [\(Chan et al., 2021\)](#) [\(Tian et al., 2022\)](#) or with methods which significantly change training [\(Mukhoti & Gal, 2018\)](#). We additionally do not compare with methods which involve multiple stages of OOD score refinement by leveraging the geometric location of the scores [\(Jung et al., 2021\)](#) [\(Chan et al., 2021\)](#), as these refinement processes could be performed on top of any given OOD score in semantic segmentation. We compare with Standardized Max

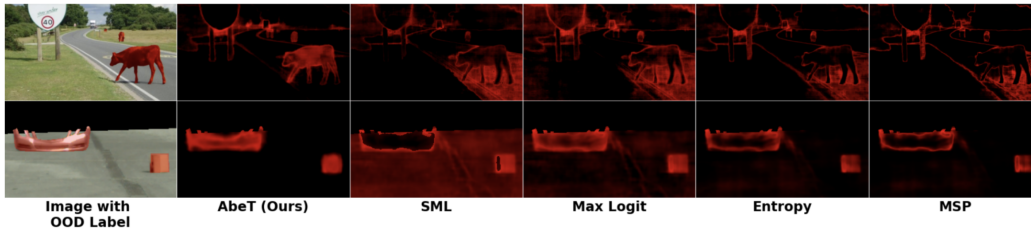


Figure 4: Qualitative comparison of OOD scores for semantic segmentation. The top row and bottom row contain examples from the datasets RoadAnomaly (Lis et al. 2019) and LostAndFound (Pinggera et al. 2016), respectively. Pixels corresponding to OOD objects are highlighted in red in each image in the leftmost column, which are cropped to regions where we have ID/OOD labels. Scores for each example (row) and technique (column) are thresholded at their respective 95% True Positive Rate and then normalized  $[0, 1]$  in the red channel, with void pixels (which have no ID/OOD label) set to 0. Bright red pixels represent high OOD scores, which should cover the same region as the pixels which correspond to OOD objects in the leftmost column. We invert the scores of Standardized Max Logit, Max Logit, and MSP to allow these methods to highlight OOD pixels in red.

Method	$D_{in}^{test}$		LostAndFound			RoadAnomaly		
	ID	mIOU	FPR@95 ↓	AUPRC ↑	AUROC ↑	FPR@95 ↓	AUPRC ↑	AUROC ↑
Entropy	81.39		35.47	46.01	93.42	65.26	16.89	68.24
MSP	81.39		31.80	27.49	91.98	56.28	15.24	65.96
SML	81.39		44.48	25.89	88.05	70.70	17.52	75.16
ML	81.39		15.56	65.45	94.52	70.48	18.98	72.78
MHLBS	81.39		27	48	-	81.09	14.37	62.85
AbeT	80.56		<b>3.42</b>	<b>68.35</b>	<b>99.09</b>	<b>53.5</b>	<b>31.12</b>	<b>81.55</b>
PEBAL	80		0.81	78.29	99.76	44.58	45.10	87.63

Table 2: **Comparison with other competitive OOD detection methods in semantic segmentation.** OOD detection results on a suite of standard datasets compared against competitive methods which are trained with similar constraints. We present PEBAL (Tian et al. 2022) which trains on OOD data to provide context as to current performance of methods which **do** have access to OOD samples at training time, rather than presenting it as a competitive method against which we directly compare. Cityscapes is used as the ID dataset. ↑ means higher is better and ↓ means lower is better.

Logit (SML) (Jung et al. 2021), Max Logit (ML) (Hendrycks et al. 2022), and PEBAL (Tian et al. 2022), with results taken from Tian et al. (2022). We also compare with Mahalanobis (MHLBS) (Lee et al. 2018), with results for LostAndFound and RoadAnomaly taken from (Chan et al. 2021) and (Tian et al. 2022), respectively. Additionally, we compare to entropy (Chan et al. 2021) and Max Softmax Probability (MSP) (Hendrycks & Gimpel 2016). We use Mapillary (Neuhold et al. 2017) and Cityscapes (Cordts et al. 2016) as the ID datasets and LostAndFound (Pinggera et al. 2016) and RoadAnomaly (Lis et al. 2019) as the OOD datasets. Details about these OOD Datasets can be found in Appendix Section D.1.2

Notably, our method reduces FPR@95 by 78.02% on LostAndFound and increases AUPRC by 63.96% on RoadAnomaly compared to competitive methods.

We also present visualizations of pixel-wise OOD predictions for our method and methods against which we compare on a selection of images from OOD datasets in Figure 4

## 6.2 OBJECT DETECTION

In Table 3 we compare against competitive OOD Detection methods in object detection. For evaluations we use ID metric AP and OOD metrics FPR@95, AUROC, and AUPRC. These evaluation metrics are calculated without thresholding detections based on any of the ID or OOD scores. For our experiments with AbeT, the learned temperature and Cosine



Method	ID AP	↑ FPR@95 ↓	AUROC	↑ AUPRC
Baseline	40.2	91.47	60.65	88.69
VOS (Du et al., 2022)	40.5	<b>88.67</b>	60.46	88.49
AbeT (Ours)	<b>41.2</b>	88.81	<b>65.34</b>	<b>91.76</b>

Table 3: **Comparison with other competitive OOD detection methods in object detection.** ID model performance and OOD performance of baseline model, state of the art OOD object detection method Virtual Outlier Synthesis (Du et al., 2022), and our method, all of which do not have access to OOD at training time. ↑ means higher is better and ↓ means lower is better.

Logit Head are directly attached to a FasterRCNN classification head’s penultimate layer as described in the above sections. Further training details can be found in Appendix Section D.2.3. Our method is compared with a baseline FasterRCNN model and a FasterRCNN model using the state of the art VOS method proposed by Du et al. (2022)<sup>3</sup>. For datasets we use PASCAL VOC dataset (Everingham et al., 2010) as the ID dataset and COCO (Lin et al., 2014a) as the OOD dataset.

We note that our method shows improved performance on ID AP (via the learned temperature decreasing confidence on OOD-induced false positives), AUROC, and AUPRC with comparable performance on FPR@95. Our method provides the added benefit of being a single, lightweight modification to detectors’ classification heads as opposed to significant changes to training with additional Virtual Outlier Synthesis, loss functions, and hyperparameters as in Du et al. (2022).

## 7 LIMITATIONS & FAILURE CASES OF OUR METHOD

Because our method uses misclassified ID examples as surrogates for OOD samples (as is shown in Section 5): our method does not perform well in cases where there are few misclassified ID examples during training; and most of our method’s failures are on misclassified ID examples. That being said, in our experiments, *all* tested OOD detection methods’ failures were concentrated on misclassified ID examples. For more information and experimental expositions of these statements, see Appendix Section A.1

## 8 CONCLUSION

Inferences on examples far from a model’s training set tend to be significantly less performant than inferences on examples close to its training set. Moreover, even if a model is calibrated on a holdout ID dataset, the confidence scores of these inferences on OOD examples are typically miscalibrated (Ovadia et al., 2019). In other words, not only does performance drop on OOD examples - users are often completely unaware of these performance drops. Therefore, detecting OOD examples in order to alert users of likely miscalibration and performance drops is one of the biggest hurdles to overcome in AI safety and reliability. Towards detecting OOD examples, we have introduced AbeT which mixes a learned temperature (Hsu et al., 2020) and an energy score (Liu et al., 2020) in a novel way with an effective ablation. We have established the superiority of AbeT in detecting OOD examples in classification, detection, and segmentation. We have additionally provided visual and empirical evidence as to why our method is able to achieve superior performance via exposure to misclassified ID examples during training time. Future work will explore if such exposure drives the performance of other OOD methods which do not train on OOD samples - as is suggested by our finding shown in Appendix Section A.1 that *all* tested OOD detection methods’ failures were concentrated on misclassified ID examples.

<sup>3</sup>When computing OOD metrics on VOS, we use the post-processing Energy Score (Liu et al., 2020) as the OOD score, as in their paper.

## REFERENCES

- Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5128–5137, 2021.
- Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.
- Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pp. 379–387, 2016.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Andrija Djuricic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don’t know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.
- Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *International Journal of Computer Vision*, pp. 303–338, 2010.
- Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213, 2020.
- Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016a. doi: 10.1109/CVPR.2016.90.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.
- Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

- Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, pp. 8759–8773, 2022. URL <https://proceedings.mlrpress/v162/hendrycks22a.html>
- 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*, pp. 10951–10960, 2020.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, 2021.
- Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34: 677–689, 2021.
- Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15425–15434, October 2021.
- Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.
- Umar Khalid, Ashkan Esmaeili, Nazmul Karim, and Nazanin Rahnavard. Rodd: A self-supervised approach for robust out-of-distribution detection. *arXiv preprint arXiv:2204.02553*, 2022.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- 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*, pp. 7167–7177, 2018.
- Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Belongie Perona, Deva Ramanan, Piotr Dollar, , and C Lawrence. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755, 2014a.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014b.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.
- Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 2152–2161, 2019.

- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.
- Yifei Ming, Ying Fan, and Yixuan Li. POEM: Out-of-distribution detection with posterior sampling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15650–15665. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/ming22a.html>.
- Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.
- Samuel G Müller and Frank Hutter. Trivialaugmtent: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 774–782, 2021.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL [http://ufldl.stanford.edu/housenumbers/nips2011\\_housenumbers.pdf](http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf).
- Gerhard Neuhold, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.
- Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1099–1106, 2016.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.
- Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 718–733, 2018.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.
- 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*, pp. 91–99, 2015.
- Yiyou Sun and Yixuan Li. On the effectiveness of sparsification for detecting the deep unknowns. *arXiv preprint arXiv:2111.09805*, 2021.

- Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.
- Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022.
- Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. pp. 246–263, 10 2022.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.
- Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.
- Fitsum A. Reda Kevin J. Shih Shawn Newsam Andrew Tao Bryan Catanzaro Yi Zhu, Karan Sapra. Improving semantic segmentation via video propagation and label relaxation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. URL <https://nv-adlr.github.io/publication/2018-Segmentation>.
- Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.
- Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.