This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dual Energy-Based Model with Open-World Uncertainty Estimation for Out-of-distribution Detection

Qi Chen Hu Ding*
University of Science and Technology of China
cq991@mail.ustc.edu.cn huding@ustc.edu.cn

Abstract

Out-of-distribution (OOD) detection is crucial for machine learning models deployed in open-world environments. However, existing methods often struggle with model over-confidence or rely heavily on empirical energy value estimation, limiting their scalability and generalizability. This paper introduces DEBO (Dual Energy-Based Model for Out-of-distribution Detection), a novel approach that addresses these limitations through an innovative dual classifier architecture and a unified energy-based objective function. DEBO enhances the standard classification framework by integrating a dual-purpose output space within a single classifier. The primary component classifies in-distribution (ID) data conventionally, while the secondary component captures open-world information and estimates uncertainty. Our method overcomes the dependence of traditional energy model-based OOD detection methods on empirical energy estimation while maintaining theoretical guarantees. Theoretical analysis demonstrates that DEBO promotes low energy and high confidence for ID data, while simultaneously inducing higher energy and decreased confidence for OOD samples. Extensive experiments conducted on benchmark datasets reveal that DEBO achieves state-of-the-art OOD detection performance while maintaining comparable classification accuracy on ID data.

1. Introduction

In open-world environments, machine learning models must demonstrate dual competency: accurately classifying in-distribution (ID) data and reliably identifying out-of-distribution (OOD) samples that fall outside known categories. However, contemporary neural networks often exhibit excessive confidence in their predictions, significantly compromising the effectiveness of OOD detection mechanisms.

To address this over-confidence phenomenon, researchers have extensively explored regularization techniques. Many successful approaches incorporate auxiliary OOD data during training to induce lower confidence predictions for OOD samples [2, 13, 20, 23, 35, 38, 40]. These approaches have shown superior performance when compared to methods that do not leverage additional OOD data. Despite the promising results, existing methods often employ traditional classifiers in OOD detection systems. These classifiers exhibit inherent limitations in mitigating over-confidence for OOD instances, as they must assign inputs to one of K predefined ID classes, even when samples belong to entirely different OOD categories.

The energy model approach [6, 8, 27, 36] presents an alternative paradigm, enhancing OOD detection by focusing on reducing the energy of ID data rather than explicitly optimizing for low OOD confidence. While energy-based models, such as Joint Energy-based Models (JEM) [8], have proven effective for OOD detection, they typically rely on Langevin Monte Carlo during training, resulting in substantial computational overhead and optimization instability. The Energy-Based Out-of-distribution Detection (EBO) method [22], a representative of energy models applicable to OOD detection, addresses these limitations by introducing an innovative energy regularization technique. This approach effectively circumvents the unstable training process inherent in JEM, offering a more efficient solution for OOD detection tasks. Recently, PEBAL [28] has also adopted the energy regularization loss proposed by EBO for anomaly segmentation, further validating the approach's versatility.

Nevertheless, EBO requires empirical estimation of distinct energy values derived from a pre-trained model for both ID and OOD data, attempting to optimize ID data to fall below one threshold while maintaining OOD data above another. **This approach presents two critical challenges**: first, the selection of suboptimal energy values may adversely impact ID classification performance; second, the energy values exhibit inherent dependencies on both the dataset and model architecture, fundamentally limiting the method's scalability and generalizability across different applications.

*Corresponding author.

To address these limitations, we propose the Dual Energy-Based Model for Out-of-distribution Detection (DEBO). DEBO builds upon our proposed Dual Energy-Based Model, which introduces a dual classifier to better represent unseen data and estimate uncertainty. **Our key contributions include:**

- A novel Dual Energy-Based Model featuring a dual classifier architecture that jointly models both label and data distributions. This includes an effective practical solution to address challenges encountered in implementing the Dual Energy-Based Model.
- An objective function based on Dual Energy-Based Model for OOD detection (DEBO) that unifies the learning process by simultaneously optimizing confidence and energy for both ID and OOD data, enhancing the model's discriminative capabilities. We show that DEBO consistently surpasses previous state-of-the-art results on common benchmarks.
- Our proposed DEBO overcomes the limitations of traditional energy-regularized methods for OOD detection by avoiding the dependency on empirical energy estimation and offering a loss function that aligns seamlessly with the cross-entropy loss. This method enables superior OOD detection capabilities while maintaining comparable classification performance on ID data.

The rest of this paper is organized as follows. We begin by providing an overview of the fundamental concepts of energy-based models and OOD detection in Section 2. In Section 3, we introduce our dual energy-based model and elaborate on our methodology for OOD detection in Section 4, which is derived from the proposed dual energy-based model. In Section 5, we present our experimental results, including comprehensive comparisons with several state-of-the-art OOD detection methods. Finally, we discuss related work in Section 6 and offer concluding remarks in Section 7.

2. Preliminaries

Energy-based Models. Energy-based models (EBMs) [19] defines a probability distribution on $\mathbb{R}^d$ as follows: for $x \in \mathbb{R}^d$,

$$p_\theta(x) = \frac{\exp\left(-E_\theta(x)\right)}{Z_\theta} \tag{1}$$

where $E_\theta(x)$ is the energy function parameterized by θ, and Z_θ denotes the normalizing constant, known as the partition function: $Z_\theta = \int_{\mathbb{R}^d} \exp\left(-E_\theta(x)\right) dx$. Given the training data distribution $p_{\text{data}}\left(x\right)$ and the model distribution $p_\theta(x)$, the goal of EBMs is to optimize the neural network parameters θ such that the model distribution $p_\theta(x)$ closely approximates the training data distribution $p_{\text{data}}\left(x\right)$. An approach to achieve this is by minimizing the Kullback-Leibler (KL) divergence between $p_{\text{data}}\left(x\right)$ and $p_\theta(x)$:

$$D_{\text{KL}}\left(p_{\text{data}}\left(x\right) \| p_\theta(x)\right) = -\mathbb{E}_{x \sim p_{\text{data}}}\Big[\log p_\theta(x)\Big] + \delta \tag{2}$$

where δ denotes a constant. This objective is known to have the following gradient [16]:

$$\begin{aligned} &\partial_\theta D_{\text{KL}}\left(p_{\text{data}}\left(x\right) \| p_\theta(x)\right) \\ &= \mathbb{E}_{x \sim p_{\text{data}}\left(x\right)}\Big[\partial_\theta E_\theta(x)\Big] - \mathbb{E}_{\tilde{x} \sim p_\theta(x)}\Big[\partial_\theta E_\theta(\tilde{x})\Big] \end{aligned} \tag{3}$$

Equation (3) plays a crucial role in our subsequent theoretical analysis.

Out-of-distribution Detection. Out-of-distribution detection can be conceptualized as a binary classification challenge. The core objective during evaluation is to ascertain whether a given test sample is drawn from the ID data distribution or if it resides outside of it. A widely adopted strategy for this classification task involves computing a score $S(x)$ using a specific OOD detection methodology. Following this computation, a threshold value α is determined, which acts as the decision boundary:

$$D(x) = \begin{cases} \text{ID}, & S(x) \geq \alpha, \\ \text{OOD}, & S(x) < \alpha, \end{cases} \tag{4}$$

Samples with scores above the threshold are classified as ID, while those below are flagged as OOD.

3. Dual Energy-Based Model

In this section, we propose our proposed energy-based model called the Dual Energy-Based Model.

3.1. The Overall Framework

Our dual classifier architecture consists of two complementary parts: the primary part serves the conventional purpose of classifying ID data, while the secondary part is specifically designed to capture and represent open-world data. Building upon this dual classifier framework, we introduce our Dual Energy-Based Model. The Dual Energy-Based Model treats data generated by the Energy model via Stochastic Gradient Langevin Dynamics (SGLD) [33] as OOD, effectively partitioning the generated data into dual negative classes.

In particular, for a training dataset comprising K classes, we define a dual classifier l that maps each input $x \in X$ to a vector in $\mathbb{R}^{2K}$. This mapping is structured such that the first K dimensions correspond to the positive labels associated with the training data (denoted as y^+), while the subsequent K dimensions represent the negative labels attributed to the generated data (denoted as y^-). For a detailed visualization of this architectural design, please refer to Figure 1. More precisely, $l_\theta(x)[y^+]$ denotes the y^+-th element of $l_\theta(x)$, which corresponds to the true ground-truth label

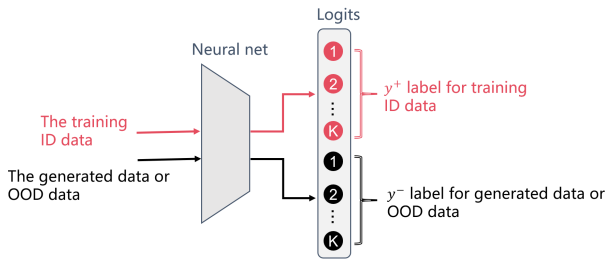


Figure 1. The dual classifier architecture for our proposed Dual Energy-Based Model and the Dual Energy-Based Model for OOD detection (DEBO). The generated data is sampled from the distribution modeled by Dual Energy-Based Model via SGLD. For OOD detection, we classify the OOD samples that are nearest to each class of ID data into a unified negative class. (For further details, please refer to Section 4.)

y^+. Conversely, $l_\theta(x)[y^-]$ signifies the y^--th element of $l_\theta(x)$, associated with the negative ground-truth label y^-. For a given data point $x \in X$, we define $\mathtt{P}_{\mathtt{conf}}(y^+|x)$ as the confidence that x belongs to class y^+. Analogously, $\mathtt{P}_{\mathtt{conf}}(y^-|x)$ represents the confidence of x belonging to the negative class y^-.

Utilizing a softmax assignment based on the dual classifier, we can express the confidence of sample x belonging to either the positive label y^+ or negative label y^-. This probabilistic assignment is formulated as:

$$\mathtt{P}_{\mathtt{conf}}(y^+|x) = \frac{\exp(l_\theta(x)[y^+])}{\sum_{k=1}^{2K}\exp(l_\theta(x)[k])}; \tag{5}$$

$$\mathtt{P}_{\mathtt{conf}}(y^-|x) = \frac{\exp(l_\theta(x)[y^-])}{\sum_{k=1}^{2K}\exp(l_\theta(x)[k])}. \tag{6}$$

We introduce the following loss function for the Dual Energy-Based Model, which is designed to effectively minimize the Kullback-Leibler (KL) divergence between two sets of distributions: the marginal and joint data-label distributions derived from the training data, and their corresponding distributions as represented by the Dual Energy-Based Model. The loss function we propose is defined as:

$$C_{\mathtt{deb}} = C_{\mathtt{x}} + C_{\mathtt{xy}} \tag{7}$$

where

$$\begin{aligned} C_{\mathtt{x}} = {} & \mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[(-\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x))\Big] \\ & + \mathbb{E}_{\tilde{x}\sim p_\theta(x)}\Big[(\log(1-\sum_{y^-}\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})))\Big] \end{aligned} \tag{8}$$

$$\begin{aligned} C_{\mathtt{xy}} = {} & \mathbb{E}_{x,y^+\sim p_{\text{data}}(x,y^+)}\Big[(-\log\mathtt{P}_{\mathtt{conf}}(y^+|x))\Big] \\ & + \mathbb{E}_{\tilde{x},y^+\sim p_\theta(x,y^+)}\Big[(\log(1-\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})))\Big] \end{aligned} \tag{9}$$

In the following section, we will conduct an in-depth analysis of the loss functions presented in Equation (8) and Equation (9). Our examination will reveal that both loss functions operate by minimizing the first term, thereby reducing the energy associated with it. Conversely, the minimization of the second term results in an increase in energy. We will also discuss the challenges associated with implementing these loss functions in practical scenarios and propose an efficient solution.

3.2. Theoretical Insight on Dual Energy-Based Model

To begin with, we define a new energy function associated with a data point x and a positive label y^+, which we denote as $E_\theta(x, y^+)$. Subsequently, we establish an energy-based model for the joint distribution as follows:

$$p_\theta(x, y^+) = \frac{\exp\left(-E_\theta(x, y^+)\right)}{Z(\theta)} \tag{10}$$

where $Z(\theta) = \int_x \exp(-E_\theta(x, y^+))$ is the unknown normalizing constant. In the context of Energy-Based Models (EBMs), the complex task of density estimation is elegantly simplified to a nonlinear regression problem [16]. This framework allows for remarkable flexibility in model design, as the energy function can be parameterized using any scalar-valued function that accepts x as its input [8]. Consequently, we can formulate the joint energy $E_\theta(x, y^+)$ as follows:

$$E_\theta(x, y^+) = -\log\mathtt{P}_{\mathtt{conf}}(y^+|x) \tag{11}$$

By marginalizing out the variable y^+, an unnormalized density model for x can be derived as well:

$$\begin{aligned} p_\theta(x) &= \sum_{y^+} p_\theta(x, y^+) = \frac{\sum_{y^+}\exp\left(-E_\theta(x, y^+)\right)}{Z(\theta)} \\ &= \exp(\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x))\cdot\frac{1}{Z(\theta)} \end{aligned} \tag{12}$$

Analogous to the formulation presented in Equation (1), we can define an energy function associated with the data point x, which we refer to as the marginal energy. This function can be expressed as:

$$E_\theta(x) = -\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x) \tag{13}$$

Finally, when calculating $p_\theta(y^+ \mid x)$ by dividing Equation (10) by Equation (12), the normalization constant cancels out, resulting in the standard K-ways Softmax parameterization:

$$\begin{aligned} p_\theta(y^+ \mid x) = \frac{p_\theta(x, y^+)}{p_\theta(x)} &= \frac{\mathtt{P}_{\mathtt{conf}}(y^+|x)}{\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x)} \\ &= \frac{\exp(l_\theta(x)[y^+])}{\sum_{k=1}^{K}\exp(l_\theta(x)[k])} \end{aligned} \tag{14}$$

We propose the following theorem to demonstrate that our loss functions of Dual Energy-Based Model defined in Equation (7) can minimize the Kullback-Leibler (KL) divergence between the true data distribution and the distribution modeled by the Dual Energy-Based Model.

Theorem 3.1 *Given the training data distribution $p_{data}(x)$ and the energy model distribution $p_\theta(x)$ defined by the marginal energy function $E_\theta(x)$ as detailed in Equation (13), along with the training data joint distribution $p_{data}(x, y^+)$ and the energy model joint distribution $p_\theta(x, y^+)$ defined by the joint energy function $E_\theta(x, y^+)$ as specified in Equation (11), optimizing the loss of the Dual Energy-Based Model defined in Equation (7) can minimize the following upper bound of the learning objective:*

$$D_{\mathrm{KL}}(p_{data}(x)\|p_\theta(x))+D_{\mathrm{KL}}\left(p_{data}(x, y^+)\|p_\theta(x, y^+)\right) \tag{15}$$

The sketch of proof. To prove Theorem 3.1, we proceed in two steps, given that Equation (7) comprises two distinct loss functions: Equation (8) and Equation (9). First, we demonstrate that optimizing the loss function presented in Equation (8) is equivalent to minimizing the KL divergence between the data distribution and the model distribution for the marginal case. Subsequently, we establish that optimizing the loss function delineated in Equation (9) effectively minimizes the KL divergence for the joint distribution of inputs and positive labels. First, based on Equation (3), we have

$$\begin{aligned}&\partial_\theta D_{\mathrm{KL}}\left(p_{\mathrm{data}(x)}\|p_\theta(x)\right)\\&=\mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[\partial_\theta E_\theta(x)\right]-\mathbb{E}_{\tilde{x}\sim p_\theta(x)}\left[\partial_\theta E_\theta(\tilde{x})\right]\end{aligned} \tag{16}$$

This formulation provides a tractable method to optimize the KL divergence through gradient descent. Specifically, we can leverage this gradient to guide the model towards minimizing the discrepancy between the data distribution $p_{\mathrm{data}}(x)$ and the model distribution $p_\theta(x)$. To facilitate our analysis, we can focus on the expectation terms in Equation (17) without the gradient symbols:

$$\mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[E_\theta(x)\right]-\mathbb{E}_{\tilde{x}\sim p_\theta(x)}\left[E_\theta(\tilde{x})\right] \tag{17}$$

$$\begin{aligned}&=\mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[(-\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x))\right]\\&\quad-\mathbb{E}_{\tilde{x}\sim p_\theta(x)}\left[(-\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|\tilde{x}))\right]\\&=\mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[(-\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x))\right]\\&\quad+\mathbb{E}_{\tilde{x}\sim p_\theta(x)}\left[(\log(1-\sum_{y^-}\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})))\right]\end{aligned} \tag{18}$$

Upon examination of Equation (17) and Equation (18), we know that the term $\log(1-\sum_{y^-}\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x}))$ in Equation (18) exhibits a negative correlation with marginal energy $E_\theta(\tilde{x})$. This relationship is essential as it allows us to optimize the model to consistently assign low energy values to ID data while assigning high energy values to the out-of-distribution data. Next, we only need to prove that optimizing the loss function of Equation (9) can effectively minimizes $D_{\mathrm{KL}}\left(p_{\mathrm{data}}(x, y^+)\|p_\theta(x, y^+)\right)$. Similar to the derivation of Equation (18), we have

$$\mathbb{E}_{x,y^+\sim p_{\mathrm{data}}(x,y^+)}\left[E_\theta(x, y^+)\right]-\mathbb{E}_{\tilde{x},y^+\sim p_\theta(x,y^+)}\left[E_\theta(\tilde{x}, y^+)\right] \tag{19}$$

$$\begin{aligned}&\leq\mathbb{E}_{x,y^+\sim p_{\mathrm{data}}(x,y^+)}\left[(-\log\mathtt{P}_{\mathtt{conf}}(y^+|x))\right]\\&+\mathbb{E}_{\tilde{x},y^+\sim p_\theta(x,y^+)}\left[(\log(1-\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})))\right]\end{aligned} \tag{20}$$

The derivation of Equation (20) is provided in the supplementary materials due to space constraints in the main text. This derivation provides an upper bound for the difference between the expectations under the data joint distribution and the model joint distribution. Overall, we can obtain the proof for Theorem 3.1 based on the above analysis.

Therefore, we can assert that the loss function presented in Equation (7) serves as an effective mechanism for minimizing the Kullback-Leibler divergence between two sets of distributions: the marginal and joint data-label distributions of the training data, and their corresponding distributions as modeled by the Dual Energy-Based Model.

Upon examination of Equation (19) and Equation (20), we know that the term $\log(1-\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x}))$ exhibits a negative correlation with joint energy $E_\theta(\tilde{x}, y^+)$. This relationship is a key aspect of how we subsequently apply the Dual Energy-Based Model for OOD detection.

3.3. Pratical Modification for Dual Energy-Based Model

To implement our proposed Dual Energy-Based Model as delineated in Section 3.2, it is imperative to address the challenges that emerge when translating the theoretical objective loss functions Equation (8) and Equation (9) into practical applications. A critical consideration is the term $\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})$ in Equation (9). In the early stages of optimization, this term typically assumes small values due to the classifier's initial suboptimal performance. Consequently, $-\log(1-\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x}))$ approaches zero, potentially impeding the model's learning process by hindering effective parameter updates.

To mitigate this issue, we propose a strategic modification inspired by the approach employed in Generative Adversarial Networks (GANs) [9, 10]. In GANs, a practical non-saturating loss $\mathbb{E}_{\tilde{x}\sim p_\theta(x)}[-\log D(\tilde{x})]$ is utilized instead of the theoretically derived loss $\mathbb{E}_{\tilde{x}\sim p_\theta(x)}[\log(1-D(\tilde{x}))]$,

where $D(\tilde{x})$ denotes the discriminator's probability output. Analogously, we introduce modifications to Equation (8) and Equation (9), resulting in the following practical formulations:

$$C'_{\mathtt{x}} = \mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[-\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x)\Big] + \mathbb{E}_{\tilde{x}\sim p_\theta(x)}\Big[-\log\sum_{y^-}\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})\Big] \quad (21)$$

$$C'_{\mathtt{xy}} = \mathbb{E}_{x,y^+\sim p_{\text{data}}(x,y^+)}\Big[-\log\mathtt{P}_{\mathtt{conf}}(y^+|x)\Big] + \mathbb{E}_{\tilde{x},y^+\sim p_\theta(x,y^+)}\Big[-\log\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})\Big] \quad (22)$$

These modifications offer several key advantages: **(i)** Optimization Consistency: The terms $\sum_{y^-}\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})$ and $\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})$, which incorporate the neural network parameters, maintain consistent optimization directions in both the original and modified formulations. This ensures that the essential learning objectives are preserved. **(ii)** Convergence Properties: Both Equation (21) and Equation (22) converge to zero if and only if the classifier achieves optimal performance in classifying both x and $\tilde{x}$. This property ensures that the optimization process drives the model towards the desired performance. **(iii)** Practical Efficacy: The proposed formulations effectively address the challenges raised in Equation (8) and Equation (9), offering a more robust and trainable objective for real-world applications.

4. Dual Energy-Based Model for Out-of-distribution Detection

Outliers positioned at the margins of ID data have been demonstrated to be particularly instrumental in regularizing the ID classifier and refining its decision boundary [24]. To fully capitalize on the role of boundary samples in enhancing the decision boundary for OOD detection, we classify the OOD samples that are nearest to each class of ID data into a unified negative class. Building upon this insight, we designate these OOD classes as K negative classes, each corresponding to a class of ID data. By doing so, we leverage the informative nature of boundary cases, thereby enhancing the model's ability to discriminate between ID and OOD samples more effectively. Specifically, as illustrated in Figure 1, the K positive classes and K negative classes can be seamlessly incorporated into the Dual Energy-Based Model framework, with additional K dimensions allocated to represent the K negative classes, thereby bolstering its OOD detection capabilities. To accomplish this, we simply substitute $p_\theta(x)$ with $p_{\text{ood}}(x)$ in Equation (21) and Equation (22). In this context, $p_{\text{ood}}(x)$ signifies the distribution of OOD data. This adjustment is in harmony with our dual objectives: **(i)** For ID data, our aim is to assign low energy values associated with the marginal distribution, while simultaneously attributing high energy values to OOD data. This objective is achievable through the optimization of Equation (21). **(ii)** Concurrently, we strive to enhance confidence for ID data by assigning low energy values associated with the joint distribution. Conversely, for OOD data, we seek to diminish confidence by assigning higher energy values. This complementary goal is attainable through the optimization of Equation (22).

Considering that auxiliary OOD data typically lacks labels in real-world scenarios, or that acquiring labels can be prohibitively expensive, we modify the second term of Equation (22) into an unsupervised form. Specifically, given the prediction distribution $P(x) = q_1(x), q_2(x), \cdots, q_K(x), p_1(x), p_2(x), \cdots, p_K(x)$ generated by the dual classifier, we define $\mathtt{Pr}(y^- = k^-|x) = p_k(x)$ to represent the probability of example x belonging to one of the K negative classes, where k^- denotes the negative label corresponding to the k-th negative class. While the classification loss in Equation (22) inherently promotes low entropy for OOD samples with respect to the probability distribution $\mathtt{Pr}(y^- = k^-|x)$, we propose to enhance this approach by explicitly minimizing the entropy of OOD samples in relation to the same probability distribution:

$$C''_{\mathtt{xy}} = \mathbb{E}_{x,y^+\sim p_{\text{data}}(x,y^+)}\Big[-\log\mathtt{P}_{\mathtt{conf}}(y^+|x)\Big] + C_{\mathtt{en}} \quad (23)$$

Here, $C_{\mathtt{en}}$ is defined by

$$C_{\mathtt{en}} = \mathbb{E}_{\tilde{x}\sim p_{\text{ood}}(x)}\Big[\sum_{j=1}^{K}-p_j(\tilde{x})\log p_j(\tilde{x})\Big] \quad (24)$$

where $p_{\text{ood}}(x)$ denotes the unlabeled auxiliary OOD training data. Note that Equation (23) is modified on the basis of Equation (22) to adapt to OOD detection scenarios. Let $C_{\mathtt{cse}} = \mathbb{E}_{x,y^+\sim p_{\text{data}}(x,y^+)}\Big[-\log\mathtt{P}_{\mathtt{conf}}(y^+|x)\Big]$ (representing the first term of Equation (23)). The final objective functions for "Dual Energy-Based Model for OOD detection (DEBO)" can be defined as follows:

$$C_{\mathtt{debo}} = C_{\mathtt{cse}} + \lambda_1\cdot C_{\mathtt{en}} + \lambda_2\cdot C''_{\mathtt{x}} \quad (25)$$

where $\lambda_1, \lambda_2 > 0$ are two given coefficients and $C''_{\mathtt{x}}$ is defined as

$$C''_{\mathtt{x}} = \Big(\mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[-\log\sum_{y^+}\mathtt{P}_{\mathtt{conf}}(y^+|x)\Big] + \mathbb{E}_{\tilde{x}\sim p_{\text{ood}}(x)}\Big[-\log\sum_{y^-}\mathtt{P}_{\mathtt{conf}}(y^-|\tilde{x})\Big]\Big) \quad (26)$$

which is modified on the basis of Equation (21) to adapt to OOD detection scenarios.

Score functions for Out-of-distribution Detection. Given that our loss function design assigns low joint energy

values to ID data and higher joint energy values to OOD data, we can leverage the maximum negative energy value associated with the joint distribution over ID categories as a score for OOD detection. This score is defined as:

$$S_{\text{joint}}(x) = \max_{y^+} -E_\theta(x, y^+) \tag{27}$$

Indeed, the joint energy is positively correlated with the joint distribution density function. We can utilize this property to introduce a new OOD detection score. Specifically, this joint distribution can be articulated as:

$$\begin{aligned} \log p_\theta(x, y^+) &= -E_\theta(x, y^+) - \underbrace{\log Z(\theta)}_{\text{constant}} \\ &= \log p_\theta(y^+ \mid x) + \log p_\theta(x) \\ &= \log p_\theta(y^+ \mid x) + (-E_\theta(x)) - \underbrace{\log Z(\theta)}_{\text{constant}} \end{aligned} \tag{28}$$

where $p_\theta(y^+ \mid x)$ is defined in Equation (14). Based on Equation (28), we observe that the joint distribution can be decomposed into conditional and marginal distributions. To allow for more flexible adjustment and balancing of the influence of these two distributions on OOD detection, we introduce a hyperparameter γ. By incorporating this hyperparameter, we obtain:

$$S_{\text{dec}}(x) = (1-\gamma) \cdot \max_{y^+} \log p_\theta(y^+ \mid x) + \gamma \cdot (-E_\theta(x)) \tag{29}$$

Rather than directly employing $\max_{y^+} \log p_\theta(x, y^+)$ as the OOD detection score, we utilize Equation (29) as the OOD detection score instead. The hyperparameter γ enables us to fine-tune the relative contributions of the conditional probability (represented by $\max_{y^+} \log p(y^+ \mid x)$) and the negative energy function ($-E_\theta(x)$) to the joint score. This approach provides a more adaptable method for OOD detection, allowing for optimal performance across various scenarios. By adjusting γ, we can strike an appropriate balance between the discriminative power of the conditional probability and the density estimation capabilities captured by the marginal energy function, thereby enhancing the overall effectiveness of OOD detection.

5. Experiments

We compare our proposed DEBO with several OOD evaluation benchmarks, including MSP [12], EBO [22], CIDER [25], OE [13], UDG [35], EBO with auxiliary OOD data (w. D_{aux}) [22], MCD [38], MIXOE [40]. Our code is available at https://github.com/cs911/DEBO.

5.1. Experimental settings

In-distribution Datasets. CIFAR-10 and CIFAR-100 [17]: These widely-adopted datasets serve as our primary in-distribution benchmarks, offering varying levels of complexity with their respective 10 and 100 classes. ImageNet-200 [41]: We further incorporate a more challenging dataset by utilizing a 200-class subset of ImageNet-1K. This subset is specifically curated to align with the categories present in ImageNet-R.

Out-of-distribution Datasets for Evaluation. For the out-of-distribution (OOD) evaluation of models trained on CIFAR-10 and CIFAR-100, we employ a comprehensive suite of widely-recognized benchmarks: MNIST [4], Textures [3], SVHN [26], Places365 [43], LSUN-Crop [37], LSUN-Resize [37], iSUN [34]. To further assess the models' generalization capabilities, we implement a cross-dataset evaluation strategy: CIFAR-100 serves as an OOD test set for CIFAR-10-trained models, and vice versa. For the evaluation of models trained on ImageNet-200, we utilize a set of challenging and diverse OOD benchmarks: SSB-hard [30], NINCO [1], iNaturalist [29], Textures [3], OpenImage-O [31].

Auxiliary Outliers for Training. CIFAR-10 and CIFAR-100: We employ the TIN-597 dataset, introduced by Zhang et al. [41]. This dataset comprises 597 classes from Tiny ImageNet [18], meticulously curated to ensure no overlap with any categories in our OOD test sets. ImageNet-200: For this subset, we leverage the remaining 800 categories from ImageNet-1K that are not included in our 200-class in-distribution set.

Evaluation Metrics. We evaluate the performance of our method using five key metrics: (1) False Positive Rate at 95% True Positive Rate (**FPR95**); (2) Area Under the Receiver Operating Characteristic curve (**AUROC**); (3) Area Under the Precision-Recall curve (**AUPR**): **AUPR-In:** Designates in-distribution examples as the positive class; **AUPR-Out:** Designates out-of-distribution examples as the positive class; (4) In-Distribution Classification Accuracy (**ID-ACC**).

Some experimental details. In our experiments on the CIFAR-10/100 and ImageNet-200 datasets, we employed a ResNet-18 architecture [11], training it for 100 epochs with a consistent batch size of 128 across all datasets. We utilized Stochastic Gradient Descent (SGD) as our optimizer, incorporating a momentum of 0.9 to ensure stable convergence. During the training phase, we conducted a comprehensive hyperparameter search, exploring values for λ_1 and λ_2 in Equation (25) from the set $\{0.1, 0.5, 1.0\}$. After extensive experimentation, we determined the optimal values to be: CIFAR-10: $\lambda_1 = 0.1$, $\lambda_2 = 1.0$; CIFAR-100: $\lambda_1 = \lambda_2 = 1.0$; ImageNet-200: $\lambda_1 = \lambda_2 = 0.1$. In the validation phase, we employ an automatic hyperparameter search strategy for post-hoc methods, including our proposed scoring functions as detailed in Equation (29). This approach leveraged both the validation subset of the in-distribution dataset and the auxiliary outliers dataset mentioned previously. Guided by the automatic hyperparameter search, we selected the hyperparameter configuration that yielded the optimal AUROC score for the final evalu-

ation across various test sets. The hyperparameter search space for the hyperparameter γ in Equation (29) is defined as $\{0.0, 0.3, 0.5, 0.7, 1.0\}$. Due to space limitations, the detailed analysis of the impact of the hyperparameter γ on OOD detection performance is placed to the supplement. The threshold α in Equation (4) is typically chosen to ensure that a high fraction of ID data (e.g., 95%) is correctly classified (We also conducted extensive experiments to evaluate the performance of DEBO under various conditions. These experiments included varying the values of λ_1 and λ_2, the number of negative classes, the model architecture, and the use of different auxiliary outliers for training. Due to space limitations, the results are presented in the full paper).

5.2. Experimental results

Comparison with existing OOD detection methods. Our method demonstrates exceptional performance across various datasets and evaluation metrics. As shown in Table 1, we achieve superior results on all four OOD indicators for CIFAR-10/100 compared to benchmark methods. For ImageNet-200, we outperform other approaches on three out of four OOD indicators. Notably, this enhanced OOD detection capability is achieved while maintaining comparable classification accuracy on ID data. On the CIFAR-100 dataset, our method attains an impressive average FPR95 index of 33.73, substantially surpassing the second-best result (50.37) obtained by the EBO (w. D_{aux}) method.

It is worth noting that we excluded results for EBO (w. D_{aux}) on the ImageNet-200 dataset from our analysis due to the absence of ImageNet-200 experiments in the original EBO paper. A significant advantage of our proposed approach lies in its simplicity and computational efficiency. In contrast to the EBO (with D_{aux}) method, which necessitates the introduction of two additional hyperparameters associated with the mean marginal energy scores of ID and OOD samples (estimated from a pre-trained model), our method entirely circumvents this step. Upon examination of the loss $C_{\mathbf{x}}^{''}$, as defined in Equation (26), which is associated with marginal energy, we observe that it naturally optimizes the first and second terms to converge towards zero. As a result, our approach circumvents the complexity of estimating mean energy scores from a pre-trained model and avoids introducing additional hyperparameters to bound the energy values of ID and OOD samples. This inherent simplicity not only enhances the method's efficiency but also improves its scalability and ease of implementation across various domains.

Ablation study. To rigorously assess the efficacy of each component within our proposed loss function (Equation (25)), we conducted a systematic analysis of its impact on OOD detection performance. Given that the term $C_{\texttt{cse}}$ represents a cross-entropy loss function for in-distribution data classification, our investigation focused specifically on evaluating the necessity of incorporating the terms $C_{\texttt{en}}$ and $C_{\mathbf{x}}^{''}$ into the overall loss function. The results of this comprehensive analysis are presented in Table 2. Our findings provide compelling evidence for both the individual necessity and the synergistic effect of all components incorporated in our carefully designed loss function.

In the second phase of our ablation study, we delved deeper into the effectiveness of our proposed score functions for OOD detection based on our proposed DEBO training method. To rigorously assess the OOD detection performance of our methods, we conducted a comparative analysis of the score functions S_{joint} defined in Equation (27) and S_{dec} defined in Equation (29). To establish a robust benchmark for comparison, we implemented a baseline method analogous to the widely-used Maximum Softmax Probability (MSP) approach, which is based on conditional probability. This baseline, which we denote as Maximum Conditional Probability (MCP), utilizes the maximum conditional probability (as defined in Equation (14)) as the OOD detection score function. By incorporating this baseline, we aim to accentuate the efficacy and superiority of our proposed score functions. The results, as detailed in Table 3, demonstrate that our proposed score function S_{dec} achieved superior performance across all OOD indicators, consistently outperforming both the baseline MCP and the alternative approach S_{joint}.

6. Related Work

Energy-based Model. Energy-based models (EBMs) boast a rich and extensive history in machine learning, with their origins tracing back to seminal works such as Hopfield networks [15] and Boltzmann machines [14]. These versatile models have demonstrated remarkable adaptability, finding widespread application across diverse domains. Their utility extends to generative tasks [7, 42], domain adaptation [39], and model calibration [32], showcasing their broad impact on the field. A particularly noteworthy development is the Joint Energy-based Model [8], which unifies classification and generation tasks within the EBM framework.

Out-of-distribution detection. The field of OOD detection has evolved significantly since Hendrycks et al. [12] introduced their seminal baseline method using maximum softmax probability. Subsequent research has produced several advanced approaches, including ODIN [21], Energy scores [22], and CIDER [25], each offering unique improvements to detection capabilities. However, a notable observation in the field is that methods relying solely on ID data typically underperform compared to those that incorporate auxiliary OOD data during training. This has led to a significant research direction focusing on the integration of OOD data into the training process. These auxiliary OOD samples can be either synthesized data [5, 20] or sourced from naturally occurring unlabeled data [13, 24, 38, 40]. Several notable approaches have emerged in this direction.

Datasets	Methods	FPR95↓	AUROC↑	AUPR-In↑	AUPR-Out↑	ID-ACC↑
CIFAR10	MSP [12]	31.62	90.96	85.50	91.76	94.72
	EBO [22]	39.41	91.61	83.29	93.61	94.72
	CIDER [25]	19.51	95.01	92.74	95.56	N/A
	OE [13]	20.42	93.90	92.29	93.98	**94.93**
	EBO (w. D_{aux}) [22]	16.99	93.74	92.59	92.69	90.40
	UDG [35]	22.44	94.13	90.36	95.00	93.47
	MCD [38]	25.75	92.36	87.11	93.03	94.28
	MIXOE [40]	38.15	91.56	83.53	92.91	94.30
	S_{dec} (ours)	**14.41**	**96.67**	**94.53**	**97.19**	94.86
CIFAR100	MSP [12]	56.23	79.04	74.17	81.21	**76.83**
	EBO [22]	52.49	81.34	76.33	83.28	**76.83**
	CIDER [25]	60.98	79.58	72.82	82.65	N/A
	OE [13]	51.16	83.54	78.48	85.50	76.60
	EBO (w. D_{aux}) [22]	50.37	84.31	77.54	86.04	72.12
	UDG [35]	59.31	79.22	74.14	80.23	71.52
	MCD [38]	56.02	75.50	72.69	74.84	76.69
	MIXOE [40]	63.27	75.83	68.94	79.18	74.76
	S_{dec} (ours)	**33.73**	**89.76**	**86.30**	**89.97**	76.26
ImageNet-200	MSP [12]	43.94	87.15	81.74	87.04	**86.49**
	EBO [22]	45.81	87.14	80.97	**87.70**	**86.49**
	CIDER [25]	43.36	86.26	80.47	87.40	N/A
	OE [13]	41.72	86.96	82.30	85.08	86.10
	UDG [35]	66.67	78.20	71.09	79.89	68.61
	MCD [38]	40.07	86.51	82.32	81.93	86.18
	MIXOE [40]	48.10	85.71	80.14	85.26	85.42
	S_{dec} (ours)	**41.12**	**88.10**	**83.24**	87.03	86.26

Table 1. Main results comparing competitive OOD detection methods, all trained using a ResNet-18 backbone. Values are presented as percentages and represent averages across the OOD test datasets described in Section 5.1. The best results are shown in bold. Arrows "↓" indicates lower is better, and "↑" indicates higher is better.

Datasets	C_{cse}	FPR95↓	AUROC↑	AUPR-In↑	AUPR-Out↑
CIFAR10	$+ C_{\text{en}}$	26.56	91.47	88.08	91.50
	$+ C'_{\text{x}}$	14.43	96.46	94.45	96.87
	$+ C_{\text{en}} + C'_{\text{x}}$	**14.41**	**96.67**	**94.53**	**97.19**
CIFAR100	$+ C_{\text{en}}$	58.46	78.30	72.80	80.22
	$+ C'_{\text{x}}$	41.62	86.28	82.89	84.71
	$+ C_{\text{en}} + C'_{\text{x}}$	**33.73**	**89.76**	**86.30**	**89.97**
ImageNet-200	$+ C_{\text{en}}$	42.97	87.21	82.28	86.92
	$+ C'_{\text{x}}$	42.18	87.55	82.74	85.79
	$+ C_{\text{en}} + C'_{\text{x}}$	**41.12**	**88.10**	**83.24**	87.03

Table 2. The ablative results about the components of our proposed DEBO loss (Equation (25)).

Datasets	Scores	FPR95↓	AUROC↑	AUPR-In↑	AUPR-Out↑
CIFAR10	MCP	19.74	94.24	92.16	94.53
	S_{joint}	16.06	96.23	94.11	96.81
	S_{dec}	**14.41**	**96.67**	**94.53**	**97.19**
CIFAR100	MCP	48.38	83.57	79.47	84.48
	S_{joint}	48.13	85.67	79.56	87.60
	S_{dec}	**33.73**	**89.76**	**86.30**	**89.97**
ImageNet-200	MCP	41.56	87.78	82.93	**87.48**
	S_{joint}	41.23	87.73	83.04	85.67
	S_{dec}	**41.12**	**88.10**	**83.24**	87.03

Table 3. The ablative results about different score functions based on our proposed DEBO training method.

Outlier Exposure (OE) [13] maintains a uniform distribution of auxiliary OOD data throughout training, providing a straightforward yet effective approach to OOD detection. Energy-based Regularization [22] employs a carefully designed loss function based on an energy model, simultaneously minimizing the energy of ID data while maximizing that of OOD samples, thereby creating a more distinct separation between the two distributions. The Unsupervised Dual Grouping (UDG) framework [35] jointly models ID and OOD data, leveraging unlabeled external datasets and unsupervised dual grouping to enhance semantically coherent OOD detection. Maximum Classifier Discrepancy (MCD) [38] introduces a two-head deep convolutional neural network architecture designed to enhance OOD detection by explicitly maximizing the predictive discrepancy between its two classifiers. More recently, Zhang et al. introduced MixOE [40], which incorporates mixup operations to enhance OOD detection performance, particularly in fine-grained contexts.

7. Conclusion and Outlook

Energy-based models have the potential to be extended to a multitude of application areas, such as generative tasks and domain adaptation. For future work, we aim to explore the applications of our proposed Dual Energy-Based Model in other domains. Additionally, we plan to investigate the extension of our DEBO method to non-vision tasks.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (No. 2021YFA1000900), the National Natural Science Foundation of China (No. 62272432, No. 62432016), and the Natural Science Foundation of Anhui Province (No. 2208085MF163).

References

[1] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 2471–2506. PMLR, 2023. 6

[2] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. ATOM: robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part III*, pages 430–445. Springer, 2021. 1

[3] Mircea Cimpoi, Subhransu 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*, pages 3606–3613, 2014. 6

[4] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 6

[5] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: learning what you don't know by virtual outlier synthesis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 7

[6] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 1

[7] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[8] David Duvenaud, Jackson Wang, Jorn Jacobsen, Kevin Swersky, Mohammad Norouzi, and Will Grathwohl. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR 2020*, 2020. 1, 3, 7

[9] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016. 4

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 4

[11] 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*, pages 770–778, 2016. 6

[12] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6, 7, 8

[13] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 1, 6, 7, 8

[14] Geoffrey E Hinton and Terrence J Sejnowski. Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 448–453. Citeseer, 1983. 7

[15] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. 7

[16] Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*, 2016. 2, 3

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[18] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[19] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 2

[20] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1, 7

[21] 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. 7

[22] 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. 1, 6, 7, 8

[23] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 1

[24] Yifei Ming, Ying Fan, and Yixuan Li. POEM: out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 15650–15665. PMLR, 2022. 5, 7

[25] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 6, 7, 8

[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS*

workshop on deep learning and unsupervised feature learning, page 4. Granada, 2011. 6

[27] Tobias Schröder, Zijing Ou, Jen Lim, Yingzhen Li, Sebastian Vollmer, and Andrew Duncan. Energy discrepancies: a score-independent loss for energy-based models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[28] 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. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022. 1

[29] 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*, pages 8769–8778, 2018. 6

[30] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? 2021. 6

[31] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022. 6

[32] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9282–9291, 2021. 7

[33] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011. 2

[34] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015. 6

[35] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 1, 6, 8

[36] Xiulong Yang and Shihao Ji. Jem++: Improved techniques for training jem, 2021. 1

[37] 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. 6

[38] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9517–9525. IEEE, 2019. 1, 6, 7, 8

[39] Yige Yuan, Bingbing Xu, Liang Hou, Fei Sun, Huawei Shen, and Xueqi Cheng. Tea: Test-time energy adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23901–23911, 2024. 7

[40] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 5520–5529. IEEE, 2023. 1, 6, 7, 8

[41] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 6

[42] Junbo Jake Zhao, Michaël Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 7

[43] 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, 2017. 6