Energy-based Confidence Gap Expansion YoLOOD for Multi-label OOD Detection

MENG LI[1], Eric C.C. TSANG [1*], WEI-HUA XU [2], QIANG HE[3]

[1]School of Computer Science and Engineering, Macau University of Science and Technology, Taipa, Macau, China
[2]College of Artificial Intelligence, Southwest University, Chongqing, 400715, China
[3]School of Science, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China
E-MAIL: muzilinxiya@gmail.com[1], cctsang@must.edu.mo [1], chxuwh@gmail.com[2], heqiang@bucea.edu.cn[3]

Abstract:

There are abundant multi-label data in Real-world scenarios, and practical applications frequently face challenges from out-of-distribution (OOD) data that deviate from the training distribution. To address this, we propose a multi-label OOD detection algorithm (Energy-Gap-YoLOOD) based on YOLOOD and energy functions. First, during model training, we propose an energy-based confidence gap expansion function specifically designed to amplify the separation between in-distribution (ID) and OOD characteristics in the feature space. Then for inference phase, we introduce a confidence enhancement mechanism that: selectively amplifies high-confidence predictions and adaptively suppresses low-confidence outputs. Finally, we experimentally validate the effectiveness and robustness of our method.

Keywords:

Multi-label OOD Detection; Energy-based Confidence Gap Expansion Function; Confidence Enhancement Mechanism

1. Introduction

There are many images or videos taken in real scenes in the form of multiple labels[1]. And it is not possible for the datasets to contain all kinds of targets. Thus classification of multi-labeled images faces the challenge of OOD samples outside the distribution of the training datasets. This challenge is often faced, for example, in the field of autonomous driving[2] or in medical diagnosis[3]. Therefore, it is crucial to address the problem of OOD detection in the field of multi-label classification.

However, most previous research has focused on OOD detection in multi-class classification[4][5], where samples and labels follow a one-to-one correspondence. These methods cannot be directly applied to OOD detection in multi-label classification. As the OOD detection problem in multi-label classification has garnered increasing attention, more and more studies have emerged in this area[6].

For instance, the YOLOOD algorithm[7] leverages the inherent capability of the YOLO object detection model[8]—which naturally learns multi-label classification and focuses on target objects while ignoring background—for multi-label OOD detection. The YOLOOD algorithm treats foreground objects as ID information and background objects as OOD information, achieving promising results in multi-label OOD detection. However, YOLOOD employs a binary 0-1 distinction to separate background and foreground information. In reality, OOD-class images also contain both foreground and background elements. Relying solely on ”0” to learn background information from ID data for OOD detection is insufficient. During training, since no OOD data is available, we propose leveraging background information as supplementary learning to establish a clear separation between ID and OOD targets, thereby enhancing multi-label OOD detection. Building upon the YOLOOD algorithm, this paper further introduces the Energy-Gap YOLOOD algorithm.

The main contributions and innovations of this paper are summarised as follows:

- This study proposes an enhanced algorithm for multi-label OOD detection, with its effectiveness being rigorously verified through comprehensive experiments.
- A novel energy-based margin expansion loss function is developed to better separate ID and OOD samples.
- A novel confidence augmentation mechanism is developed to asymmetrically enhance larger confidence val-

ues without substantially altering smaller ones, serving to better reveal the model's intrinsic classification ability.

The rest of this article is organised as follows. Section 2 presents the details of the Related Work and Preliminasies. Section 3 describes the Proposed Method. Section 4 compares the results of our method with other models. Section 5 concludes the paper.

2. Related Work and Preliminasies

2.1 Multi-label OOD Detection

In recent years, numerous multi-label OOD detection algorithms have emerged. Wang et al.[9] proposed the JointEnergy function, which estimates OOD uncertainty by aggregating label energy scores from multiple labels. As a post-processing approach for classification results, this method does not affect the network's learning process. But the JointEnergy is a representative multi-label OOD inference criterion. Sun et al.[10] proposed a multi-label learning model that reshapes the uncertain energy space by incorporating auxiliary outlier exposure based on the JointEnergy function. Wang et al.[11] developed a Sparse Label Co-occurrence Scoring (SLCS) method that leverages label sparsity and co-occurrence information to compute OOD detection scores. Aguilar et al.[12] designed an evidential neural network based on evidential learning principles to quantify uncertainty for OOD detection. The YOLOOD algorithm utilizes YOLO's inherent multi-label classification capability and object-focused attention mechanism for multi-label OOD detection. Furthermore, specialized multi-label graph neural networks[13] have been developed for processing multi-label graph-structured data.

2.2 Preliminasies Definition

OOD Detection. We regard OOD detection as a binomial classification problem, where $\mathcal{D}_{\text{in}}$ indicates that samples are the ID data, and $\mathcal{D}_{\text{out}}$ indicates the OOD data. The OOD detection of multi label classification is to judge whether a multi-label sample x belongs to a known distribution domain through detector $G(x)$. The detector $G(x)$ is determined by the following formula:

$$G(x) = \begin{cases} 1 & \text{if } x \sim \mathcal{D}_{\text{in}} \\ 0 & \text{if } x \sim \mathcal{D}_{\text{out}} \end{cases} \tag{1}$$

Energy Function. Our energy function adopts the JointEnergy function proposed in[9]. This function is formally expressed as:

$$E(f) = \sum_{i=1}^{C} \log(1 + e^{f_i}), \tag{2}$$

where f_i represents the network's logits output for the i-th class. This function enlarges the confidence gap between ID and OOD data by summing the confidence of all labels.

3. Proposed Method

The key to multi-label classification lies in capturing the similarities between objects. To facilitate OOD detection, we also need to learn the differences between ID objects and other objects. Thus we extend the YolOOD framework and propose an Energy-Gap-YoLOOD (EG-YolOOD) algorithm.

In this section, we will introduce the EG-YolOOD algorithm for multi-label OOD detection.The EG-YolOOD algorithm inherits YolOOD's modified YOLOv5 architecture for multi-label OOD detection. The training procedure (top section of Figure 1) processes ID images through the network to extract multi-scale features, generating three levels of logits outputs. During testing, the network processes both ID and OOD data samples. The three-tiered logits from the output head are then transformed into final OOD scores through the computational framework depicted in the bottom section of Figure 1.

3.1 EG-YolOOD Classifier

The EG-YolOOD algorithm mainly consists of the following three components.

3.1.1 Multi-label Confidence Scores Learning

The network outputs three parallel logits layers with identical channel dimensions. Each layer contains $n + 1$ channels, where n corresponds to the number of object categories and the additional channel represents the confidence score - indicating the probability of a spatial location being an object center point. To optimize this confidence prediction, we employ the Binary Cross-Entropy (BCE) loss function, formally defined as:

$$\mathcal{L}_{\text{conf}} = \sum_{c \in C} \mathcal{L}_{\text{BCE}} \left(c_{\text{conf}}, \hat{c}_{\text{conf}} \right). \tag{3}$$

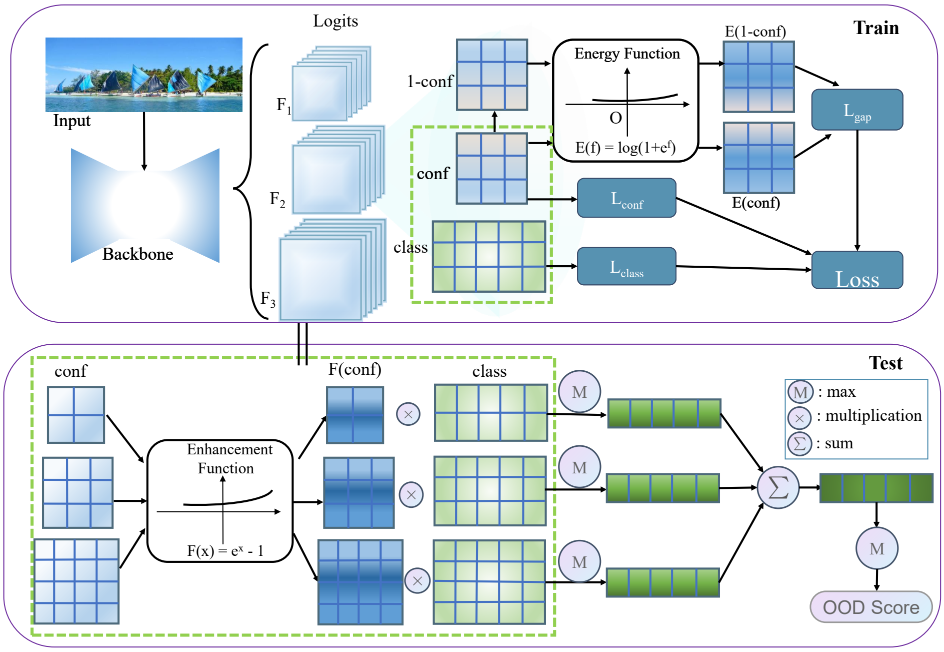


FIGURE 1. Diagram of the EG-YolOOD classifier structure.The upper part is the training process. The lower portion is the testing phase.

Here, L_{BCE} represents the binary cross-entropy loss function, c_{conf} is the network's predicted confidence score, and $\hat{c}_{\text{conf}}$ is the ground-truth label composed of binary values (0 or 1). The value is set to 1 within the minimal enclosing square (infimum bounding box) centered on the target's center point, and 0 at all other positions.As shown in the following equation:

$$\begin{aligned}\varphi_u &= i \geq x_{k,\text{center}} - p_k \cdot \frac{W_r}{2},\ \varphi_l = j \geq y_{k,\text{center}} - p_k \cdot \frac{H_r}{2}.\\ \varphi_d &= i \leq x_{k,\text{center}} + p_k \cdot \frac{W_r}{2},\ \varphi_r = j \leq y_{k,\text{center}} + p_k \cdot \frac{H_r}{2}.\end{aligned} \tag{4}$$

$$\hat{c}_{\text{conf}}(i,j) = \begin{cases} 1, & \varphi_u \wedge \varphi_d \wedge \varphi_l \wedge \varphi_r, \\ 0, & \text{else.} \end{cases} \tag{5}$$

where p_k represents the ratio between the size of the lower bounding box and the size of the original label box. The values are set to 0.0, 0.1, and 0.5 on the three scale logits, respectively.

3.1.2 Multi-label Classification Scores Learning

During the training process, the network's output logits are activated using the sigmoid function. For multi-label classification score learning, we employ the conventional BCE loss function, which is formulated as follows:

$$\mathcal{L}_{\text{cls}} = \sum_{c \in \mathcal{C}} \sum_{n \in \{1,\ldots,N_c\}} \hat{c}_{\text{obj}} \cdot \mathcal{L}_{\text{BCE}}(c_{\text{cls}\,n}, \hat{c}_{\text{cls}\,n}) \tag{6}$$

Among them, $c_{\text{cls}\,n}$ represents the predicted class score for each category. $\hat{c}_{\text{cls}\,n}$ is the conceptual value indicating whether each position in the lower bounding box belongs to each category. If a position is responsible for predicting class i, the value at that position is 1; otherwise, it is 0. Additionally, each position can predict multiple categories. The specific values are as shown in the equation:

$$\hat{c}_{\text{cls}\,n}(i,j) = \begin{cases} 1, & \text{class } n \text{ is in cell } (i,j) \\ 0, & \text{else} \end{cases} \tag{7}$$

3.1.3 Energy-based Confidence Extend Gap Learning

To widen the gap between the foreground values (ID) and background values (OOD) in the predicted confidence scores, a function is designed using the energy function to amplify the confidence disparity. This function first computes two differences. The first is the gap between the energy value of the predicted confidence scores and the energy value of 1 minus the predicted confidence scores. The second is the difference between the corresponding labels. Then, the function also calculates the euclidean distance between these two differences. As shown in the equation below, the learning of this function aims to separate ID and OOD data, thereby enhancing OOD detection.

$$\mathcal{L}_{gap} = \sum_{c \in C} \frac{1}{w * h} \sum_{i \in w} \sum_{j \in h} |E(c_{conf(ij)}) - E(1 - c_{conf(ij)})| - |E(\hat{c}_{conf(ij)}) - E(1 - \hat{c}_{conf(ij)})| \tag{8}$$

Where w and h are the width and height of the output feature map, respectively.

Therefore, the total loss function is a weighted linear combination of the three aforementioned loss functions, expressed as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{conf}} + \lambda_2 \mathcal{L}_{\text{cls}} + \lambda_3 \mathcal{L}_{\text{gap}} \tag{9}$$

where λ_1,λ_2,λ_3 are the weighting factors.

3.2 Multi-label OOD Scores

In the test phase, we perform two operations for the prediction confidence scores and prediction classification scores of the network output. We first enhance the confidence prediction value. Secondly, we aggregate the enhanced confidence scores and classification scores.

3.2.1 Confidence Enhancement Strategy

For the confidence scores predicted by the network, we first apply the sigmoid function for activation and then combine them with classification scores for OOD detection. Since the sigmoid function constrains the confidence scores to the range [0, 1], it reduces the gap between the distributions of predicted values. To ensure that low values remain almost unchanged while amplifying high values, we design an enhancement function as follows:

$$F(x) = e^x - 1 \tag{10}$$

This function satisfies the following properties: As $x \to 0$, $f(x) \approx x$(small values remain fewer changes). For $x > 0$, $f(x) > x$(larger values are amplified).

3.2.2 Aggregate OOD Scores

Finally, the enhanced confidence scores and classification scores from the three detection heads of different scales are aggregated to obtain the final OOD scores. The specific aggregation process is as follows:

$$\mathrm{S}(x) = \max_{n \in \{1,\ldots,N_c\}} \sum_{\mathcal{C}_k \in f_{\mathrm{S}}(x)} \max_{c \in \mathcal{C}_k} \{\sigma(F(c_{\text{conf})}) \cdot \sigma(c_{\text{cls }n})\} \tag{11}$$

$$G(x,t) = \begin{cases} 1 & \mathrm{S}(x) \geq t \\ 0 & \mathrm{S}(x) < t \end{cases} \tag{12}$$

Here, t is the threshold determined such that the function $G(x,t)$ correctly classifies a specified proportion (e.g., 95%) of ID data. Mathematically, t satisfies:

$$\mathbb{P}_{x \sim \mathcal{D}_{\text{ID}}} (G(x,t) \text{ classifies } x \text{ correctly}) \geq 95\% \tag{13}$$

4. Experiment and Analysis

This section describes the experiment.

4.1 Datasets

ID Datasets. We used three datasets as ID datasets. (1) PASCAL-VOC[14]: containing 20 categories, divided into 5717 training images, 5823 validation images and 10991 test images. (2) MS-COCO[15]: containing 80 categories, divided into 117266 training images, 4952 validation images and 40670 test images. (3) Object365$_{\text{in}}$[16]: This dataset is a subset of the Object365 dataset[16] containing 20 common categories, divided into 68723 training images, 5000 validation images and 10000 test images. Among them, the training and validation sets were used to train the network. The test sets were used for testing as ID datasets.

OOD Datasets. We employed two datasets as OOD datasets: (1) Object365$_{\text{out}}$: A subset of the Object365 dataset comprising 200 categories distinct from those in Object365$_{\text{in}}$, containing 11162 images.(2) NUS-WIDE$_{\text{out}}$[17]: A subset of the NUS-WIDE dataset formed by removing overlapping categories, resulting in 54 remaining classes and 9415 images. These datasets were only used in the testing phase for OOD detection.

TABLE 1. Comparison of the multi-label OOD detection performance of EG-YolOOD vs. state-of-the-art methods on Object365$_{\text{out}}$ OOD datasets. ↓ indicates that smaller values are better, ↑ indicates that larger values are better. Bold indicates optimal results. These indicators have the same meaning as in Table 2.

D$_{\text{out}}$	Object365$_{\text{out}}$			NUS-WIDE$_{\text{out}}$		
D$_{\text{in}}$	MS-COCO	PASCAL-VOC	Object365$_{\text{in}}$	MS-COCO	PASCAL-VOC	Object365$_{\text{in}}$
Algorithms	FPR95 ↓ /AUROC ↑ /AUPR ↑					
MaxLogit	18.09/96.55/99.06	20.80/96.25/96.54	28.86/94.56/94.68	12.46/97.41/99.39	18.75/96.86/97.60	41.60/91.97/92.98
MSP	47.13/85.00/94.84	40.32/90.66/90.76	66.10/83.57/84.14	39.45/87.55/96.37	38.05/91.47/92.87	81.98/76.36/80.12
JointEnergy	16.35/96.95/99.22	19.99/96.63/97.08	20.07/96.26/96.61	8.59/97.84/99.53	16.06/97.12/97.94	25.29/95.20/96.08
YOLOOD	10.35/97.47/99.28	16.44/96.46/96.33	16.68/96.06/95.49	5.20/98.43/99.66	18.16/96.94/97.69	10.49/97.78/97.99
EG-YolOOD	9.73/97.50/99.30	15.89/97.15/97.38	16.16/96.07/95.51	4.68/98.49/99.67	21.63/96.62/97.52	12.77/97.54/97.85

4.2 Experimental Details

During the training phase, we trained the model for 30 epochs using the training and validation sets from the ID Datasets. We employed the Adam optimizer with an initial learning rate of 10-5. If the validation mAP did not improve for 2 consecutive epochs, the learning rate was reduced by a factor of 0.1. Our experiments were conducted on RTX 4090 GPUs using PyTorch 1.11.0.

4.3 Evaluation Indicators

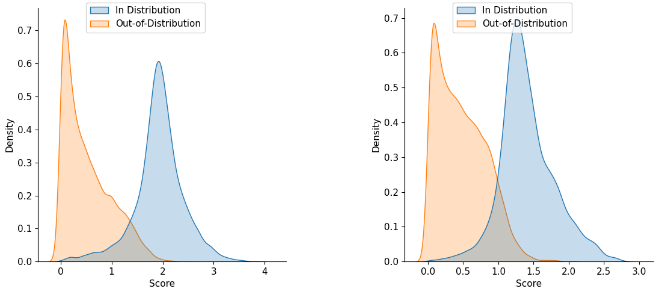
We evaluated OOD detection performance using three metrics: (1)FPR95: the false positive rate at 95% true positive recall; (2) AUROC: the area under the receiver operating characteristic curve; (3) AUPR: the area under the precision-recall curve.

4.4 Analysis Results

The experimental results of our algorithm on the three ID datasets are presented in Table 1. We conducted comprehensive comparisons with existing state-of-the-art OOD detection methods for multi-label scenarios under identical experimental settings and datasets. We implemented the following representative OOD detection algorithms baseline network YOLOOD and using YOLO-cls as the base classifier: MaxLogit[18], Maximum Softmax Probability(MSP)[19], JointEnergy.

From the detection results on the OOD detection dataset Object365$_{\text{out}}$, it can be seen that the EG-YolOOD algorithm performs optimally in the training results on the MS COCO dataset and the PASCAL-VOC dataset compared to other algorithms. It improves by 0.62/0.03/0.02 and 0.55/0.69/1.05 on the three metrics, respectively, compared to the suboptimal algorithm. Secondly, our algorithm achieves the best FPR95 value in the detection results on the Object365out dataset when trained on the Object365$_{\text{in}}$ dataset. Although the AUPR and AUPR values are slightly inferior, they still show improvement over the baseline network.

From the detection results on the OOD detection dataset NUS-WIDE$_{\text{out}}$, it can be seen that our method performs best on the model trained with the MS-COCO dataset, improving by 0.52/0.06/0.01 on the three metrics compared to the suboptimal algorithm. Although the performance is slightly inferior on models trained with the other two datasets, it is still not the worst. Therefore, the above results demonstrate the effectiveness and robustness of our algorithm.



(a) YolOOD (FPR95=15.89%)

(b) YOLO (FPR95=16.44%)

FIGURE 2. Score distribution when using PASCAL-VOC as the ID dataset and Objects365$_{\text{out}}$ as the OOD dataset.

In addition, we also show the distribution of scores for EG-YolOOD and YOLOOD when using PASCAL-VOC as the ID dataset and Object365$_{\text{out}}$ as the OOD dataset, as shown in Fig. 2. It can be intuitively found that the EG-YolOOD algorithm (a) widens the gap between ID and OOD compared to the YOLOOD algorithm (b) to achieve our original intention, which again proves the robustness of our improved algorithm.

5. Conclusion

To further widen the gap between ID and OOD information in multi-label OOD detection algorithms, we have proposed a confidence gap expansion function based on the energy function. Additionally, during the testing phase, we have introduced an enhancement function to amplify the effect of high-confidence predictions while suppressing the influence of low-confidence values. Finally, we have validated the effectiveness and robustness of our approach.

Moving forward, we will further explore how multi-label OOD detection algorithms can better capture data similarity and uncertainty.

Acknowledgements

This paper is supported by the Macao Science and Technology Development Fund (0036/2023/RIA1, 0037/2024/RIA1) and Jiangsu Kaite Auto Parts Co., Ltd. (MUST Code: 1330).

References

[1] Yang J, Zhou K, Li Y, et al. Generalized out-of-distribution detection: A survey[J]. International Journal of Computer Vision, 2024, 132(12): 5635-5662.

[2] Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D Object Detection for Autonomous Driving: A Comprehensive Survey. International Journal of Computer Vision, 131(8): 1909–1963.

[3] Zhou, H.; Lu, C.; Chen, C.; Yang, S.; and Yu, Y. 2023. A Unified Visual Information Preservation Framework for Self-supervised Pre-Training in Medical Image Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7): 8020–8035.

[4] Amit G, Levy M, Rosenberg I, et al. Food: Fast out-of-distribution detector[C] 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[5] Huang R, Geng A, Li Y. On the importance of gradients for detecting distributional shifts in the wild[J]. Advances in Neural Information Processing Systems, 2021, 34: 677-689.

[6] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings.arXiv preprint arXiv:1911.11132, 2019.

[7] Zolfi A, Amit G, Baras A, et al. YolOOD: Utilizing Object Detection Concepts for Multi-Label Out-of-Distribution Detection[C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5788-5797.

[8] Jocher G, Stoken A, Chaurasia A, et al. ultralytics/yolov5: v6. 0-YOLOv5n'Nano'models[J]. Roboflow integration, TensorFlow export, OpenCV DNN support, 2021, 10.

[9] Wang H, Liu W, Bocchieri A, et al. Can multi-label classification networks know what they don't know?[J]. Advances in Neural Information Processing Systems, 2021, 34: 29074-29087.

[10] Sun Y, Xu Q, Wang Z, et al. EDGE: Unknown-aware Multi-label Learning by Energy Distribution Gap Expansion[C] Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(12): 12613-12621.

[11] Wang L, Huang S, Huangfu L, et al. Multi-label out-of-distribution detection via exploiting sparsity and co-occurrence of labels[J]. Image and Vision Computing, 2022, 126: 104548.

[12] Aguilar E, Raducanu B, Radeva P. Multi-label out-of-distribution detection via evidential learning[C] European Conference on Computer Vision. Springer, Cham, 2025: 202-218.

[13] Cai T, Jiang Y, Li M, et al. ML-GOOD: Towards Multi-Label Graph Out-Of-Distribution Detection[C] Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(15): 15650-15658.

[14] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015, 111: 98-136.

[15] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C] Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. Springer International Publishing, 2014: 740-755.

[16] Shao S, Li Z, Zhang T, et al. Objects365: A large-scale, high-quality dataset for object detection[C] Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8430-8439.

[17] Chua T S, Tang J, Hong R, et al. Nus-wide: a real-world web image database from national university of singapore[C] Proceedings of the ACM international conference on image and video retrieval. 2009: 1-9.

[18] Hendrycks D, Basart S, Mazeika M, et al. Scaling out-of-distribution detection for real-world settings[J]. arXiv preprint arXiv:1911.11132, 2019.

[19] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[J]. arXiv preprint arXiv:1610.02136, 2016.