Classifier-head Informed Feature Masking and Prototype-based Logit Smoothing for Out-of-Distribution Detection

Zhuohao Sun, Yiqiao Qiu, Zhijun Tan, Weishi Zheng, Ruixuan Wang

***Abstract*—Out-of-distribution (OOD) detection is essential when deploying neural networks in the real world. One main challenge is that neural networks often make overconfident predictions on OOD data. In this study, we propose an effective post-hoc OOD detection method, named HIMPLoS, based on a new feature masking strategy and a novel logit smoothing strategy. Feature masking determines the important features at the penultimate layer for each in-distribution (ID) class based on the weights of the ID class in the classifier head and masks the rest features. Logit smoothing computes the cosine similarity between the feature vector of the test sample and the prototype of the predicted ID class at the penultimate layer and uses the similarity as an adaptive temperature factor on the logit to alleviate the network's overconfidence prediction for OOD data. With these strategies, we can reduce feature activation of OOD data and enlarge the gap in OOD score between ID and OOD data. Extensive experiments on multiple standard OOD detection benchmarks demonstrate the effectiveness of our method and its compatibility with existing methods, with new state-of-the-art performance achieved from our method. The source code will be released publicly.**

***Index Terms*—out-of-distribution detection, deep learning, feature masking, logit smoothing.**

I. Introduction

DEEP learning has made extraordinary achievements in various fields in recent years. However, when deployed to the real world, deep learning models often encounter samples of unknown classes that were not seen during training [1]–[3]. These out-of-distribution (OOD) data may compromise the stability of the model, with potentially severe consequences such as in autonomous driving [4]–[6], videos [7], [8], and medical diagnosis [9]. Therefore, deep learning models are expected to have the ability to detect whether any new data is OOD or not.

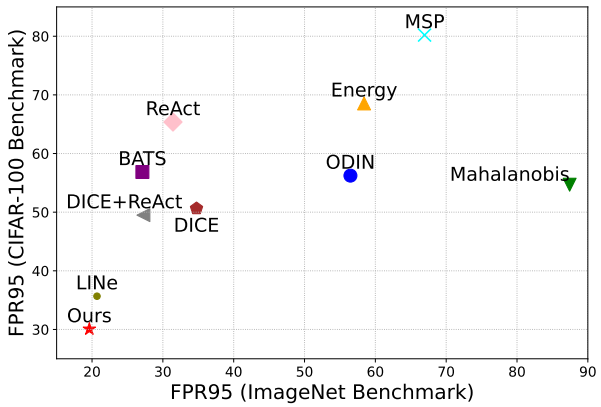

Fig. 1. OOD detection performance from different methods on CIFAR-100 and ImageNet benchmarks. FPR95: the false positive rate when the true positive rate is 95%. Smaller FPR95 values mean better performance.

There have been many explorations of OOD detection [10]–[15]. One line of work is post-hoc, i.e., the model is pre-trained and fixed and the focus is on how to design an effective scoring function [2], [16], [17] to measure the degree of any new input belonging to one of the learned classes. Any data from any learned class is called in-distribution (ID). The aim is to assign higher scores to ID data and lower scores to OOD data. Such post-hoc methods have practical significance when deploying models to the real world as it does not require any additional design of new training modules for OOD detection. However, recent studies reveal that neural networks have overconfident predictions for OOD inputs [18], [19], resulting in small difference in score between ID and OOD data. Therefore, how to solve the overconfidence problem of neural networks and make the activation of networks as small as possible for OOD data is the key to improving OOD detection performance.

In this study, we propose a novel method called classifier-**H**ead **I**nformed feature **M**asking and **P**rototype-based **Lo**git **S**moothing (**HIMPLoS**) for OOD detection. HIMPLoS provides two new strategies to reduce the feature activation of OOD data, while preserving the activation of ID data largely unchanged, thus improving the separability in detection score between ID and OOD data. The first strategy, called

This work is supported in part by the Major Key Project of PCL (No. PCL2023AS7-1), the NSFC (No. 62071502, 12371418), the Guangdong Excellent Youth Team Program (No. 2023B1515040025), the National Key R&D Program of China (No. 2022ZD0117805), the Guangdong NSF (No. 2022A1515010426), and the Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (No. 2020B1212060032). Corresponding author: Ruixuan Wang (wangruix5@mail.sysu.edu.cn)

Zhuohao Sun, Weishi Zheng, and Ruixuan Wang are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China, and the Peng Cheng Laboratory, Shenzhen 518066, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou 510275, China (e-mail: sunzhh8@mail2.sysu.edu.cn; wszheng@ieee.org; wangruix5@mail.sysu.edu.cn).

Yiqiao Qiu is with the University of California, San Diego, La Jolla 92037, United States of America (e-mail: y7qiu@ucsd.edu).

Zhijun Tan is with the School of Computer Science and Engineering, Sun Yat-sen Univerisity, Guangzhou 510275, China (e-mail: tzhij@mail.sysu.edu.cn).

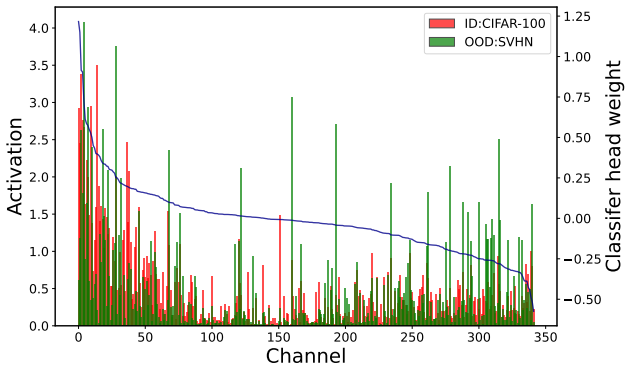

Fig. 2. Distribution of weights (blue curve) in the classifier head associated with a specific ID class, and distribution of activation over feature channels at the penultimate layer from ID (red histogram) and OOD data (green histogram). Activation of each feature is the average over the data of the same ID class in CIFAR-100 or over the OOD data which are predicted as the ID class from the OOD dataset SVHN. The feature channels are sorted in descending order by the weights of the ID class in the classifier head. Decreasing weights from the left to the right is largely correlated with the overall decreasing trend of feature activation from ID data, but not from OOD data.

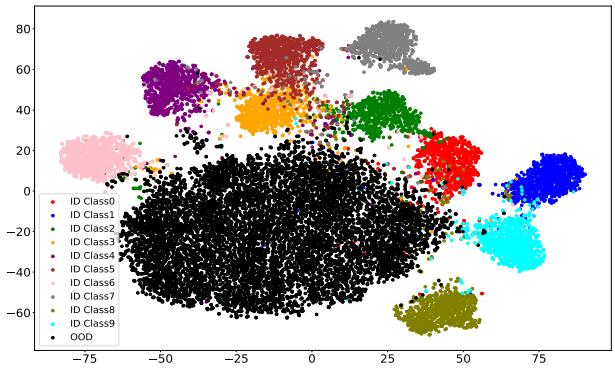

Fig. 3. Visualization of the distribution of ID and OOD feature vectors at the penultimate layer by t-SNE [21]. The model is DenseNet [22] trained on the CIFAR-10 dataset. The ID dataset is CIFAR-10 and OOD dataset is LSUN. Overall OOD data are far away from any ID class.

feature masking, is inspired by the observation that the feature activation at the penultimate layer is positively correlated with the associated weights in the classifier head for each class of ID data, while not for OOD data (Figure 2). This observation is consistent with the efficacy of widely used model interpretation methods CAM [20] which use the weights in the classifier head to represent the importance of feature elements when interpreting the predicted specific class. Based on this observation, we propose to determine the important features at the penultimate layer for each ID class according to the weights in the classifier head associated with the ID class, and then mask the other features which are less important to the ID class. In doing so, most of the high activation units that play an important role in classification can be preserved for ID data, while those high feature activation appearing in the removed units from OOD data are removed, thus largely reducing the feature activation of OOD data. The second strategy, called logit smoothing, is motivated by the observed difference in feature vector distribution between ID and OOD data at the penultimate layer as shown in Figure 3. In particular, ID data is often close to its class center ('prototype') while OOD data is relatively not close to any ID class center in the feature space. With such observed difference, the cosine similarity between new input data and the prototype of the predicted ID class at the penultimate layer is used to tune the logit vector in the output layer. Such a combination further enlarges the OOD score gap between ID and OOD data. In summary, the main contributions are as follows:

- We propose a new post-hoc OOD detection method that reduces feature activation of OOD data to enlarge the gap in OOD score between ID and OOD data.
- We introduce a new class-specific feature masking strategy simply based on weights in the classifier head.
- We propose a novel logit smoothing strategy combining information at the penultimate layer with logits for better OOD detection.
- We extensively evaluate our method on multiple standard benchmarks and show the compatibility with existing methods, with new state-of-the-art performance achieved.

II. Related Work

The key to OOD detection is to find potentially different output patterns from the model for ID and OOD data. Reconstruction-based methods show that an encoder-decoder framework trained on ID data usually produces different quantities of reconstruction errors for ID and OOD samples [10], [23], [24]. In particular, a reconstruction model trained only on ID data cannot recover OOD data well. Distance-based methods measure the distance between the input sample and the centroid or prototype of ID class in the feature space to detect OOD samples [25], [26]. For example, Mahalanobis score [27] uses the Mahalanobis distance between the feature vector of input sample and the prototype feature vector of training data for OOD detection. KNN [28] computes the k-th nearest neighbor distance between the feature vector of input sample and the feature vector of the training set. Besides, confidence enhancement methods attempt to enhance the confidence of the network via data augmentation [29] or designing a confidence estimation branch etc. [19], [30], [31]. For example, OpenMix+ [29] mixes ID samples to generate fake OOD samples and combines regularization methods for enhancing the confidence of ID samples. LogitNorm [32] proposes to enforce a constant norm on the logit during training to alleviate the overconfidence of the neural network.

For better distinguishing between ID and OOD data, training-based methods aim to perform OOD detection by training-time regularization [11], [15], [33]–[35]. C-LMCL [33] introduces a centralized large margin cosine loss which constrains intra-class feature representations to be more compact. MoEP-AE [36] proposes to use multiple mixtures of exponential power distributions to encode the features of ID classes and learn discriminative representations of ID classes. Besides, D-pedcc [34] uses PEDCC-Loss [37] to train a network and design a scoring function based on the PEDCC-

Loss classifier and another work [35] trains a second classifier based on the PEDCC-Loss classifier for OOD detection.

Differently, our method belongs to another line of work, post-hoc methods, attempt to perform OOD detection by designing a scoring function for a pre-trained and fixed model, assigning an OOD score to each new input [38]–[40]. For example, MSP [2] provides a simple baseline for OOD detection by using the maximum probability output of the model. ODIN [41] introduces two operations based on MSP called temperature scaling and input perturbation to separate OOD from ID samples. Energy score [16] uses the energy function of the logits (i.e., input to the softmax at the output layer) for OOD detection. To alleviate the overconfidence problem of the model on OOD data, based on the observation that OOD data often cause abnormally high activation at the penultimate layer of the network, ReAct [42] rectifies feature activation at an upper limit and reduces most of the activation values caused by OOD data. DICE [43] reduces the variance of the output distribution by clipping some noise units irrelevant to ID classes, resulting in improved separability in the OOD score distribution between ID and OOD data. LINe [44] employs the Shapley value [45] to measure each neuron's contribution and reduces the effect of less important neurons at the last network layer.

Our method is similar to LINe and DICE, as ours and the two methods all propose a sparsification strategy for the model to reduce feature activation of OOD data. While both LINe and DICE need a complicated process to select important elements using a training dataset, our method simply utilizes the classifier head's weight to select the important feature elements for each ID class. In addition, our method uses a novel logit smoothing operator to combine the feature information at the penultimate layer and the logit information to further improve OOD detection.

III. Preliminary

Consider a neural network classifier trained with a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is the i-th training image and $y_i \in \{1, 2, \ldots, C\}$ is the associated class label. When deploying the neural network classifier in the real world, new data may be from a certain unknown distribution which is different from the distribution of the training data. Such data are out-of-distribution (OOD) and should not be predicted as any of the in-distribution (ID) classes learned during classifier training. The task of OOD detection is to identify whether any new data is ID or OOD.

OOD detection can be considered as a binary classification problem. In particular, a scoring function $S(\mathbf{x}; f)$ can be designed to estimate the degree of any new data $\mathbf{x}$ belonging to any of the ID classes, where the function f denotes the overall feature extraction process whose output is used as the input to the scoring function. With the scoring function, OOD detection can be simply formulated as a binary classifier $g(\mathbf{x}; f)$ as below,

$$g(\mathbf{x}; f) = \begin{cases} 1 & \text{if } S(\mathbf{x}; f) \geq \gamma \\ 0 & \text{if } S(\mathbf{x}; f) < \gamma \end{cases}, \tag{1}$$

where data with higher scores $S(\mathrm{x}; f)$ are classified as ID (with label 1) and lower scores are classified as OOD (with label 0), and γ is the threshold hyper-parameter.

IV. Method

An overview of the proposed post-hoc OOD detection framework is illustrated in Figure 4. Given a pre-trained neural network (Figure 4, components in gray) in this framework, the feature elements at the penultimate layer which are more relevant to each ID class can be identified based on the weight parameters of the classifier head, and then feature masking is performed by masking the less important feature elements. Meanwhile, motivated by the observed differences between OOD and ID samples in the feature space, logit smoothing is applied by combining information from the penultimate layer's features and the output layer's logit before estimating the OOD score. In addition, as in the state-of-the-art method LINe [44], the clipping of feature activation at the penultimate layer called ReAct [42] is also applied here.

A. Feature Masking

For a well-trained network, given any test data $\mathbf{x}$, denote by $h(\mathbf{x}) \in \mathbb{R}^L$ the feature vector from the penultimate layer of the network. The classifier head's weight matrix $\mathbf{W} \in \mathbb{R}^{L \times C}$ together with the bias vector $\mathbf{b}$ transforms the feature vector $h(\mathbf{x})$ to the output logit vector $f(\mathbf{x})$ as follows,

$$f(\mathbf{x}) = \mathbf{W}^\intercal h(\mathbf{x}) + \mathbf{b}\,. \tag{2}$$

Inspired by the observation in Figure 2 and the model output interpretation methods CAM [20], where the weights in the classifier head are correlated with the importance of feature channels from the penultimate layer for each class, we propose selecting the top-k weights for each class based on the k-largest elements from each column in $\mathbf{W}$. Specifically, denote by $\mathbf{M} \in \mathbb{R}^{L \times C}$ the binary feature mask matrix, where 1 is set for the k-largest elements from each column in $\mathbf{W}$ and 0 for the rest of the column, and suppose the neural network classifier predicts the new data $\mathbf{x}$ as class c, then the modulated feature vector by the feature mask of class c from the penultimate layer becomes

$$h_m(\mathbf{x}) = \mathbf{m}_c \odot h(\mathbf{x})\,, \tag{3}$$

where $\mathbf{m}_c \in \mathbb{R}^L$ is the c-th column of $\mathbf{M}$ representing the feature mask of class c, and $\odot$ represents the element-wise multiplication. In the mask-modulated feature vector $h_m(\mathbf{x})$, those feature elements whose activation is masked (i.e., removed) often have smaller activation values and therefore have little effect on the prediction of class c. If the input $\mathbf{x}$ does belong to ID class c, the modulated feature vector $h_m(\mathbf{x})$ would be still quite similar to the original feature vector $h(\mathbf{x})$, and therefore the modulation operator based on the feature mask $\mathbf{m}_c$ would not likely change the original prediction for the input $\mathbf{x}$. In contrast, if the input $\mathbf{x}$ is an OOD data, and it is originally misclassified as ID class c, such misclassification may be partly from relatively stronger activation of the to-be-masked feature elements. In this case, the modulation based on the feature mask $\mathbf{m}_c$ would result in a modulated feature

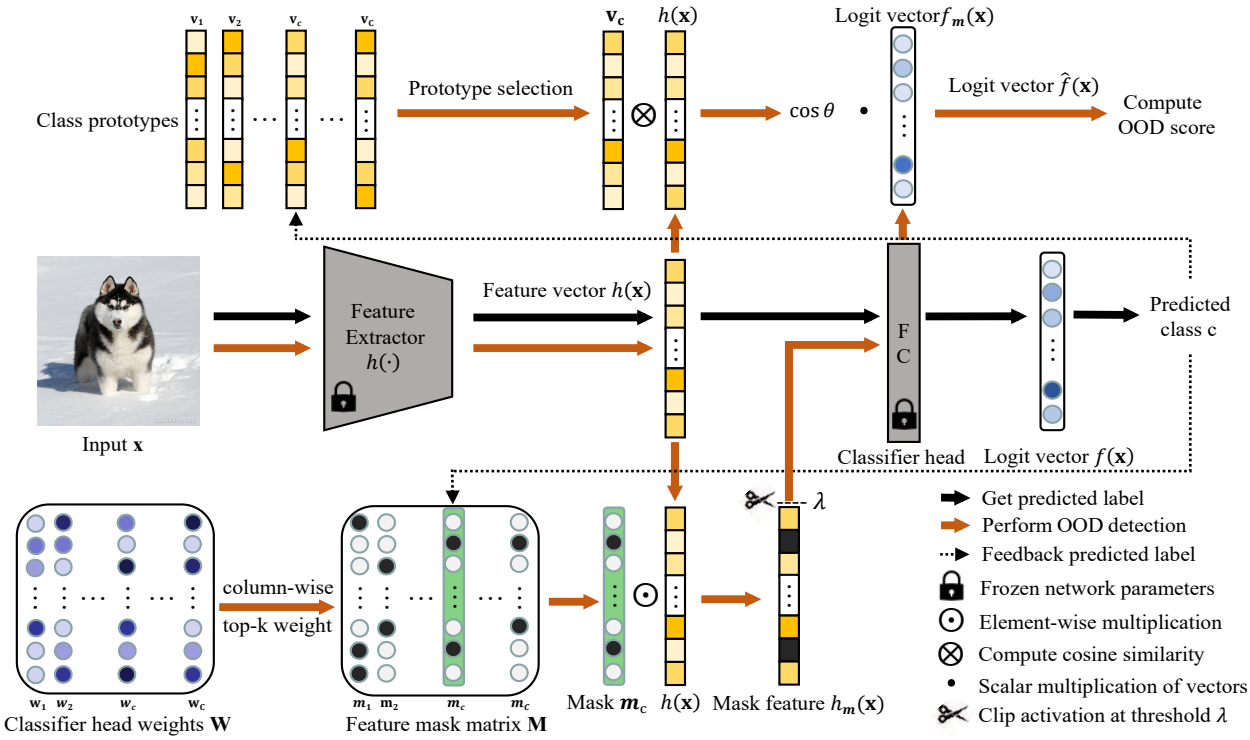

Fig. 4. Overview of the proposed framework. Given a well-trained classifier (middle row) and any test data $\mathbf{x}$, the post-hoc feature masking (bottom row) and logit smoothing (top row) together modulate the original logit vector $f(\mathbf{x})$, and the OOD score based on the modulated logit vector $\hat{f}(\mathbf{x})$ is computed (top right) for OOD detection.

vector $h_m(\mathbf{x})$ whose overall activation (i.e., the norm of feature vector) would be substantially reduced. Considering that feature activation of ID data is often statistically stronger than that of OOD data as observed in previous studies [17], [42], and that scoring functions are often designed by utilizing such difference in feature activation between ID and OOD data [14], [16], the potentially substantial reduction in feature activation for OOD data in the modulated feature vector $h_m(\mathbf{x})$ would further enlarge the gap in feature activation between ID and OOD data, making OOD detection easier.

In addition, on a similar rationale of further reducing feature activation of OOD data, following the previous studies [42], [44], the clipping operator ReAct [42] is applied to the masked feature $h_m(\mathrm{x})$ to reduce substantially higher feature activation which often appears only on OOD data, i.e.,

$$\overline{h}_m(\mathbf{x}) = \texttt{ReAct}(h_m(\mathbf{x}); \lambda), \tag{4}$$

where $\texttt{ReAct}(\mathrm{x}; \lambda) = \min(\mathrm{x}, \lambda)$ is applied element-wise to $h_m(\mathbf{x})$, and λ is a threshold hyper-parameter. Since OOD data often has the abnormal phenomenon of high unit activation in the penultimate layer of the network and ID data often has no such phenomenon [42], by rectifying the activation at an upper limit λ, it can reduce most of the high-activation values for OOD data and largely preserve the activation for ID data. Therefore, by deploying ReAct [42], we can further alleviate the overconfidence of the model on OOD data.

With the mask modulation and the ReAct clipping operators, the output logit vector becomes

$$f_m(\mathbf{x}) = \mathbf{W}^\intercal \overline{h}_m(\mathbf{x}) + \mathbf{b}\,. \tag{5}$$

B. Logit Smoothing

For any test data $\mathbf{x}$ which is predicted as class c, the cosine similarity between the feature vector $h(\mathbf{x})$ of the test data and the prototype of class c in the penultimate layer's feature space can be utilized to help separate OOD data from ID data. Denote by $\mathbf{v}_c \in \mathbb{R}^L$ the prototype of class c which can be estimated in advance by the average of feature vectors over all the training data belonging to class c, i.e.,

$$\mathbf{v}_c = \frac{1}{N_c} \sum_{i:y_i=c} h(\mathrm{x}_i)\,, \tag{6}$$

where N_c is the number of the training data belonging to class c. Simultaneously, denote by $s(h(\mathbf{x}), \mathbf{v}_c)$ the cosine similarity between $h(\mathbf{x})$ and $\mathbf{v}_c$, i.e.,

$$s(h(\mathbf{x}), \mathbf{v}_c) = \frac{h(\mathbf{x}) \cdot \mathbf{v}_c}{\|h(\mathbf{x})\| \|\mathbf{v}_c\|}\,. \tag{7}$$

Then the logit vector $f(\mathbf{x})$ can be modulated by the cosine similarity $s(h(\mathbf{x}), \mathbf{v}_c)$ as follows,

$$f_s(\mathbf{x}) = s(h(\mathbf{x}), \mathbf{v}_c) \cdot f(\mathbf{x})\,. \tag{8}$$

If the input $\mathbf{x}$ is indeed from ID class c, it is expected that the cosine similarity $s(h(\mathbf{x}), \mathbf{v}_c)$ between the feature vector of the input and its class prototype in general is higher. In contrast, if $\mathbf{x}$ is an OOD data, its feature vector $h(\mathbf{x})$ is in general not within the distribution of class c in the feature space, and therefore the cosine similarity would be lower. This has been supported as demonstrated in Figure 5. Such difference in cosine similarity between ID and OOD data can be utilized

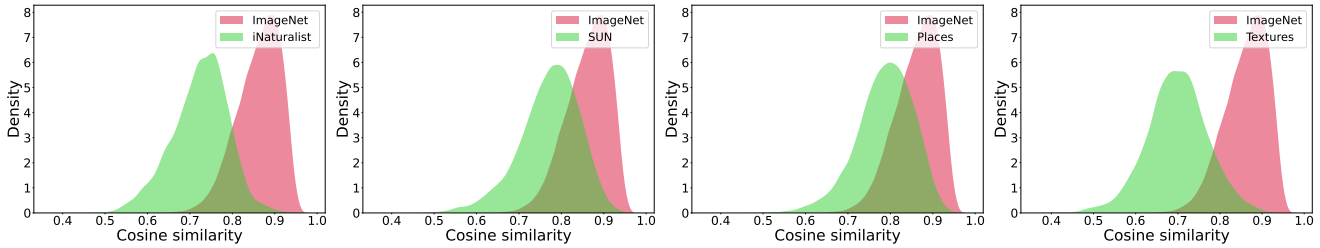

Fig. 5. Distributions of cosine similarities between the feature of ID data (in red) or OOD data (in green) and the the associated class prototypes on ResNet-50. ImageNet-1k is the ID dataset, and iNaturalist, SUN, Places, and Textures are four OOD datasets.

to help design a more effective scoring function for OOD detection by combining with logit vector (see Equation 11 below).

To illustrate the effect of modulating logit by the cosine similarity, we consider the softmax function of the modulated logit. The cosine similarity $s(h(\mathbf{x}), \mathbf{v}_c)$ in the modulated logit vector $f_s(\mathbf{x})$ can be viewed as an input-adaptive temperature τ on the logit vector $f(\mathbf{x})$, specifically by setting $\tau = 1/s(h(\mathbf{x}), \mathbf{v}_c)$. With the temperature-tuned logit, the output of class c after the softmax operator becomes

$$\text{softmax}_c(f(\mathbf{x}); \tau) = \frac{\exp(f_c(\mathbf{x})/\tau)}{\sum_{k=1}^{C} \exp(f_k(\mathbf{x})/\tau)}, \tag{9}$$

where $f_c(\mathbf{x})$ is the logit element of the predicted class c in the logit vector$f(\mathbf{x})$. Since the cosine similarity $s(h(\mathbf{x}), \mathbf{v}_c)$ is in general smaller for OOD data than for ID data, τ would be larger for OOD data. As a result, the softmax outputs become smoother (i.e., closer to a discrete uniform distribution) for OOD data, and correspondingly the confidence of predicting the OOD data as ID class c becomes lower. Since the modulated logit vector $f_s(\mathbf{x})$ can alleviate overconfidence of predicting OOD data as ID classes, it can be also applied to softmax-based OOD detection methods such as MSP [2] and ODIN [41]. Considering the smoothing effect on the softmax outputs, the modulation of the logit vector by the cosine similarity is named logit smoothing.

C. Scoring Function

Combining feature masking and logit smoothing, the modulated final logit vector is

$$\begin{aligned} \hat{f}(\mathbf{x}) &= s(h(\mathbf{x}), \mathbf{v}_c) \cdot f_m(\mathbf{x}) \\ &= s(h(\mathbf{x}), \mathbf{v}_c) \cdot \left\{ \mathbf{W}^\intercal \overline{h}_m(\mathbf{x}) + \mathbf{b} \right\}. \end{aligned} \tag{10}$$

While various scoring functions can be applied based on $\hat{f}(\mathbf{x})$, the energy scoring function [16] is used by default, i.e.,

$$S(\mathbf{x}; \hat{f}) = \log \sum_{k=1}^{C} \exp(\hat{f}_k(\mathbf{x})), \tag{11}$$

where $\hat{f}_k(\mathbf{x})$ is the k-th element in the modulated final logit vector $\hat{f}(\mathbf{x})$. Since the cosine similarity $s(h(\mathbf{x}), \mathbf{v}_c)$ is positively correlated with the energy score $S(\mathbf{x}; \hat{f})$, lower cosine similarity from OOD data will lead to smaller energy score, while higher cosine similarity from ID data will lead to larger energy score. Similarly for the modulated logit vector $f_m(\mathbf{x})$ by the mask modulation and the ReAct clipping operators (Equations 3 and 4). Overall, feature masking and logit smoothing help enlarge the gap in energy score between ID and OOD data.

D. Differences from existing methods

Our method is partly inspired by the state-of-the-art method LINe [44] in masking features based on estimated important feature elements for the predicted class. However, our method is significantly different from LINe in multiple aspects. First, the masking strategy is different and ours is much simpler and more efficient. LINe employs the Shapley value [45] to measure each feature's contribution for each class, and generates masks based on these values. Notably, LINe needs training data to calculate the Shapley value. In contrast, our method simply uses the weight parameters of the classifier head to select the important features for each class. Second, our method includes a novel logit smoothing operator which can further enlarge the difference in the modulated logit vector between ID and OOD data and also alleviate the overconfidence of OOD predictions. Third, our method outperforms LINe on standard benchmarks with different model backbones. Furthermore, our method is compatible with other OOD detection methods, and even the performance of LINe can be further improved when combined with the proposed logit smoothing.

In addition, compared to other existing methods [32], [42], [43] for alleviating the overconfidence issue of the model prediction on OOD data, our method offers several compelling advantages. First, our method can remove at least some highly activated feature elements for OOD data, without compromising the prediction accuracy for ID data, by simply utilizing the classifier head's weights. Second, our method utilizes the feature information at the penultimate layer as an input-adaptive temperature on the logit vector during inference time, which can further mitigate overconfidence when predicting OOD data as ID classes. Last but not least, our method can be efficiently applied to various model architectures and is robust to the choice of hyper-parameters.

TABLE I
COMPARISON BETWEEN DIFFERENT METHODS IN OOD DETECTION ON CIFAR-10 AND CIFAR-100 BENCHMARKS WITH DIFFERENT MODEL BACKBONES. ↓ INDICATES SMALLER VALUES MEAN BETTER PERFORMANCE AND ↑ INDICATES LARGER VALUES MEAN BETTER PERFORMANCE. BOLD NUMBERS ARE SUPERIOR RESULTS AND UNDERLINED NUMBERS ARE THE 2ND BEST RESULTS. ALL VALUES ARE PERCENTAGES.

ID Dataset Model	Method	OOD Datasets												Average	
		SVHN		LSUN-C		LSUN-R		iSUN		Textures		Places365			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10 ResNet-18	MSP [2]	57.52	92.05	43.53	94.05	55.65	91.88	56.10	91.74	61.58	89.60	61.28	88.96	55.94	91.38
	ODIN [41]	35.06	94.00	11.37	97.89	25.82	**95.58**	27.53	**95.26**	47.57	89.86	42.28	90.66	31.60	93.87
	Mahalanobis [27]	26.10	93.74	47.22	77.33	**10.98**	90.73	**11.80**	91.26	**37.50**	90.38	86.54	64.60	36.69	84.67
	Energy [16]	36.14	93.95	11.97	97.80	29.40	94.98	31.26	94.66	48.62	89.72	42.33	90.75	33.29	93.64
	BATS [46]	35.81	93.99	11.87	97.82	28.37	95.19	30.13	94.88	47.96	90.24	42.23	90.76	32.73	93.81
	DICE [43]	34.04	94.04	8.89	98.43	33.15	94.55	35.57	94.03	52.22	88.56	44.71	90.30	34.76	93.32
	ReAct [42]	40.06	93.65	15.02	97.48	28.41	95.22	30.25	94.98	45.25	91.36	**40.98**	**91.60**	33.33	94.05
	DICE+ReAct [43]	36.39	93.74	9.86	98.27	31.58	94.75	34.01	94.31	48.37	90.43	43.04	91.03	33.87	93.75
	LINe [44]	**18.63**	**96.94**	8.41	98.46	42.01	93.31	45.14	92.61	42.77	92.10	62.44	84.67	36.57	93.01
	HIMPLoS (Ours)	22.30	96.26	**7.17**	**98.70**	31.65	94.79	35.15	94.19	39.04	**92.79**	50.26	89.17	**30.92**	**94.32**
CIFAR-10 DenseNet	MSP [2]	60.03	90.84	32.25	95.83	39.52	94.69	42.58	94.24	63.48	87.94	62.41	88.74	50.04	92.05
	ODIN [41]	14.52	97.14	2.98	99.25	**3.39**	**99.02**	**4.64**	**98.83**	57.85	81.18	49.44	89.60	22.14	94.17
	Mahalanobis [27]	35.28	88.44	18.24	93.18	7.34	97.88	8.72	97.76	30.41	89.53	87.52	64.83	31.25	88.60
	Energy [16]	42.73	93.27	3.46	99.18	9.44	98.21	12.49	97.89	59.20	85.75	42.35	91.55	28.28	94.31
	BATS [46]	15.42	97.35	3.16	99.25	8.17	98.38	10.22	98.05	45.80	92.03	**41.04**	91.57	20.63	96.10
	DICE [43]	34.72	93.61	0.39	99.89	5.48	98.85	6.33	98.75	46.35	87.33	45.65	90.25	23.15	94.78
	ReAct [42]	18.94	96.84	4.33	99.05	10.78	97.99	13.70	97.57	51.52	90.19	41.53	**91.60**	23.47	95.54
	DICE+ReAct [43]	4.98	98.89	**0.27**	**99.90**	8.43	98.33	9.19	98.22	28.30	94.30	51.50	88.70	17.11	96.39
	LINe [44]	6.18	98.78	0.48	99.84	8.87	98.34	10.17	98.20	28.56	94.53	47.44	89.86	16.95	96.59
	HIMPLoS (Ours)	**4.19**	**99.12**	0.56	99.83	6.71	98.79	7.63	98.62	**25.18**	**95.53**	43.42	91.19	**14.62**	**97.18**
CIFAR-100 ResNet-18	MSP [2]	83.67	75.41	72.36	83.31	81.84	75.21	83.39	73.90	85.62	73.58	82.78	74.97	81.61	76.06
	ODIN [41]	85.97	80.86	37.97	93.72	65.83	85.35	69.39	83.52	83.85	75.53	80.10	77.53	70.52	82.75
	Mahalanobis [27]	58.59	75.92	84.92	49.51	59.04	61.12	52.25	67.38	76.31	47.72	93.68	58.13	70.80	59.96
	Energy [16]	84.07	81.97	35.48	94.01	68.88	84.19	72.29	82.24	84.66	75.14	80.28	77.39	70.95	82.49
	BATS [46]	68.28	88.67	32.67	94.00	56.96	88.10	59.45	87.28	55.12	89.16	77.91	78.33	58.40	87.59
	DICE [43]	64.56	87.52	**7.14**	**98.63**	65.24	85.47	66.48	84.41	74.59	76.54	78.22	77.93	59.37	85.08
	ReAct [42]	72.27	87.04	41.37	92.15	**49.46**	**89.98**	**52.04**	**89.22**	51.21	89.18	**75.53**	**79.41**	56.98	87.83
	DICE+ReAct [43]	39.63	93.04	7.99	98.32	74.78	85.52	71.79	86.64	45.64	88.81	82.29	75.44	53.69	87.96
	LINe [44]	43.26	92.98	10.25	98.06	63.80	87.10	63.76	87.21	44.10	90.71	78.48	78.00	50.61	89.01
	HIMPLoS (Ours)	**28.21**	**95.32**	14.41	97.29	64.16	85.74	63.54	85.89	**31.91**	**93.27**	77.81	77.59	**46.67**	**89.18**
CIFAR-100 DenseNet	MSP [2]	82.03	75.19	60.54	85.60	85.30	69.17	86.05	70.17	85.14	71.21	82.15	74.78	80.20	74.35
	ODIN [41]	45.14	91.75	9.54	98.20	59.14	86.34	61.46	85.87	81.72	72.21	80.38	77.90	56.23	85.38
	Mahalanobis [27]	52.30	89.16	86.21	68.78	30.47	94.19	30.22	93.81	34.10	89.52	94.79	57.42	54.68	82.15
	Energy [16]	88.04	81.30	14.78	97.43	70.72	80.14	74.60	78.95	85.16	70.98	**77.93**	**78.25**	68.54	81.18
	BATS [46]	91.80	76.53	21.19	95.55	41.56	92.87	44.10	92.27	63.17	82.02	79.29	75.65	56.85	85.82
	DICE [43]	60.07	88.20	**0.93**	**99.74**	51.76	89.32	49.59	89.51	61.24	77.22	80.20	77.65	50.63	86.94
	ReAct [42]	82.70	82.22	19.49	96.33	63.85	85.98	68.11	84.96	78.33	79.08	79.76	76.21	65.37	84.13
	DICE+ReAct [43]	55.54	88.03	7.57	98.61	54.44	89.84	44.32	91.45	41.29	86.11	93.75	56.48	49.49	85.09
	LINe [44]	31.13	91.89	5.76	98.85	**25.36**	94.54	**24.15**	94.76	39.18	87.85	88.44	64.17	35.67	88.68
	HIMPLoS (Ours)	**14.04**	**96.90**	3.27	99.28	29.97	**94.89**	25.25	**95.51**	**25.05**	**93.90**	82.78	72.87	**30.06**	**92.23**

V. EXPERIMENTS

A. Experimental Setup

1) Datasets: Our method is extensively evaluated on the widely used CIFAR benchmarks [47] and the large-scale OOD detection benchmark based on ImageNet [1]. For CIFAR benchmarks, CIFAR-10 and CIFAR-100 [47] are respectively used as in-distribution datasets, with 50,000 training images and 10,000 test images per dataset. Six OOD datasets are used during testing, including SVHN [48], LSUN-Crop [49], LSUN-Resize [49], iSUN [50], Textures [51], and Places365 [52]. For the ImageNet benchmark, ImageNet-1k [53] is used as the in-distribution dataset and four OOD datasets are used during testing, including Places365 [52], Textures [51], iNaturalist [54], and SUN [55].

2) Implementation details: Following the common experimental setting [1], [42], [43], ResNet-18 [56] and DenseNet [22] are used as the backbones on CIFAR benchmarks. During model training, each CIFAR image is randomly cropped to 32×32 pixels and then randomly flipped horizontally. The models are trained with batch size 128 for 100 epochs, weight decay 0.0001, and momentum 0.9. The initial learning rate is 0.1 and decays by a factor of 10 at epochs 50, 75, and 90. On the ImageNet benchmark, the pre-trained ResNet-50 [56] and MobileNetV2 [57] provided by Pytorch are adopted. During testing, all images are resized to 32×32 pixels on CIFAR benchmarks, and resized to 256×256 and then center-cropped to size of 224×224 pixels on the ImageNet benchmark. All training images are used to compute class prototypes for each model on both CIFAR and ImageNet benchmarks.

3) Evaluation metrics: Based on the scoring function (Equation 11 by default), we measure the performance of OOD detection using the two most widely used evaluation metrics FPR95 and AUROC. FPR95 is the false positive rate when the true positive rate is 95%, with lower FPR95 indicating better OOD detection performance and vice versa. AUROC is the area under the receiver operating characteristic curve, with larger AUROC indicating better performance.

4) Competitive methods for comparison: Our method is compared with multiple competitive post-hoc OOD detection methods, including MSP [2], ODIN [41], Mahalanobis [27], Energy [16], BATS [46], DICE [43], ReAct [42], DICE+ReAct [43], and LINe [44].

B. Evaluation on CIFAR Benchmarks

Table I summarizes the comparisons between our method and competitive post-hoc OOD detection methods on CIFAR-10 and CIFAR-100 benchmarks. Our method achieves state-of-the-art performance with both ResNet-18 and DenseNet backbones. For example, on the CIFAR-10 benchmark, our

TABLE II
PERFORMANCE COMPARISON ON THE IMAGENET BENCHMARK WITH RESNET-50 AND MOBILENET BACKBONES.

Model	Method	OOD Datasets								Average	
		iNaturalist		SUN		Places		Textures			
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ResNet-50	MSP [2]	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
	ODIN [41]	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
	Mahalanobis [27]	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47
	Energy [16]	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
	BATS [46]	12.57	97.67	22.62	**95.33**	34.34	91.83	38.90	92.27	27.11	94.28
	DICE [43]	25.63	94.49	35.15	90.83	46.49	87.48	31.72	90.30	34.75	90.77
	ReAct [42]	20.38	96.22	24.20	94.20	33.85	91.58	47.30	89.80	31.43	92.95
	DICE + ReAct [43]	18.64	96.24	25.45	93.94	36.86	90.67	28.07	92.74	27.25	93.40
	LINe [44]	12.26	97.56	**19.48**	95.26	**28.52**	**92.85**	22.54	94.44	20.70	95.03
	HIMPLoS (Ours)	**8.91**	**98.18**	23.08	94.78	32.50	92.17	**13.99**	**96.97**	**19.62**	**95.53**
MobileNet	MSP [2]	64.29	85.32	77.02	77.10	79.23	76.27	73.51	77.30	73.51	79.00
	ODIN [41]	55.39	87.62	54.07	85.88	57.36	84.71	49.96	85.03	54.20	85.81
	Mahalanobis [27]	62.11	81.00	47.82	86.33	52.09	83.63	92.38	33.06	63.60	71.01
	Energy [16]	59.50	88.91	62.65	84.50	69.37	81.19	58.05	85.03	62.39	84.91
	BATS [46]	31.56	94.33	41.68	90.21	52.43	86.26	38.69	90.76	41.09	90.39
	DICE [43]	43.09	90.83	38.69	90.46	53.11	85.81	32.80	91.30	41.92	89.60
	ReAct [42]	42.40	91.53	47.69	88.16	51.56	86.64	38.42	91.53	45.02	89.47
	DICE + ReAct [43]	32.30	93.57	**31.22**	92.86	46.78	88.02	16.28	96.25	31.64	92.68
	LINe [44]	24.95	95.53	33.19	92.94	47.95	88.98	**12.30**	**97.05**	29.60	93.62
	HIMPLoS (Ours)	**20.43**	**96.41**	33.51	**93.32**	**45.64**	**89.97**	15.28	96.94	**28.72**	**94.16**

method with DenseNet outperforms the strongest baseline by 2.33% in FPR95. On the CIFAR-100 benchmark, our method with DenseNet outperforms the competitive method ReAct [42] by 35.31% in FPR95 and 8.1% in AUROC. Compared to the state-of-the-art method LINe [44], our method reduces FPR95 by 5.61% and improves AUROC by 3.55% on DenseNet while reducing FPR95 by 3.94% on ResNet-18. These results consistently support the effectiveness of our method on different model backbones for OOD detection. However, to further support the superiority of our method compared to the best baseline (LINe [44]) in AUROC, we train 10 models by standard CE-loss under 10 different random seeds and perform paired t-test on both CIFAR-10 and CIFAR-100 benchmarks. The p-value obtained is less than 0.005 for both CIFAR-10 and CIFAR-100 benchmarks with ResNet-18 and DenseNet, indicating that our method demonstrates significantly higher AUROC performance compared to LINe.

C. Evaluation on ImageNet Benchmark

Table II shows the OOD detection performance of our method and competitive baselines with ResNet-50 and MobileNetV2 backbones on the ImageNet benchmark. The detailed performance on four OOD datasets and the average over the four datasets are reported. It shows that our method achieves state-of-the-art performance on average with both backbones. For example, with ResNet-50, our method outperforms Energy [16] by 38.79% in FPR95 and 9.36% in AUROC. Our method reduces FPR95 by 11.81% compared to ReAct [42], which confirms the importance of feature masking and logit smoothing for OOD detection. Also, our method outperforms DICE+ReAct [43] and LINe [44] on ResNet-50 by 7.63% and 1.08% in FPR95, respectively. This further confirms the superiority of our method particularly considering that the two methods and ours all use ReAct to improve performance. Note that our method does not achieve the best performance on some individual OOD datasets like SUN and Places with the ResNet-50 backbone, probably because (OOD) images in SUN and Places are similar to those (ID) images in ImageNet in the feature space (also see Figure 5) such that logit smoothing is hard to further improves OOD detection performance. Particularly, some studies [58], [59] have found that Places have some overlap with ImageNet-1K resulting in the worst performance over all the methods. Besides, we split test data (both test ID and OOD data) into 10 folds without intersection to perform paired t-test and the p-value is $2.17e^{-10}$ for ResNet-50 and $1.4e^{-11}$ for MobileNet which shows the superiority of our method with different models compared to LINe [44] in AUROC.

TABLE III
ABLATION STUDY OF DIFFERENT COMPONENTS IN OUR METHOD. DENSENET IS USED ON THE CIFAR-100 BENCHMARK, AND RESNET-50 ON THE IMAGENET BENCHMARK. 'FM' AND 'LS' DENOTE FEATURE MASKING AND LOGIT SMOOTHING RESPECTIVELY. ALL PERCENTAGE VALUES ARE AVERAGED OVER MULTIPLE OOD DATASETS.

ReAct	FM	LS	CIFAR-100 FPR95↓ / AUROC↑	ImageNet FPR95↓ / AUROC↑
			68.54 / 81.18	58.41 / 86.17
✓			65.37 / 84.13	31.43 / 92.95
✓	✓		36.74 / 90.01	26.33 / 93.38
✓		✓	48.42 / 88.77	23.10 / 95.09
	✓	✓	40.34 / 89.89	32.51 / 92.36
✓	✓	✓	**30.06 / 92.23**	**19.62 / 95.53**

D. Ablation and Sensitivity Studies

1) Ablation of different components in our method: Our method includes feature masking ('FM') and logit smoothing ('LS') as well as the ReAct clipping [42] to improve OOD detection performance. Table III shows an ablation study of each component in our method. Compared to ReAct [42] only (second row), ReAct+FM (third row) reduces FPR95 by

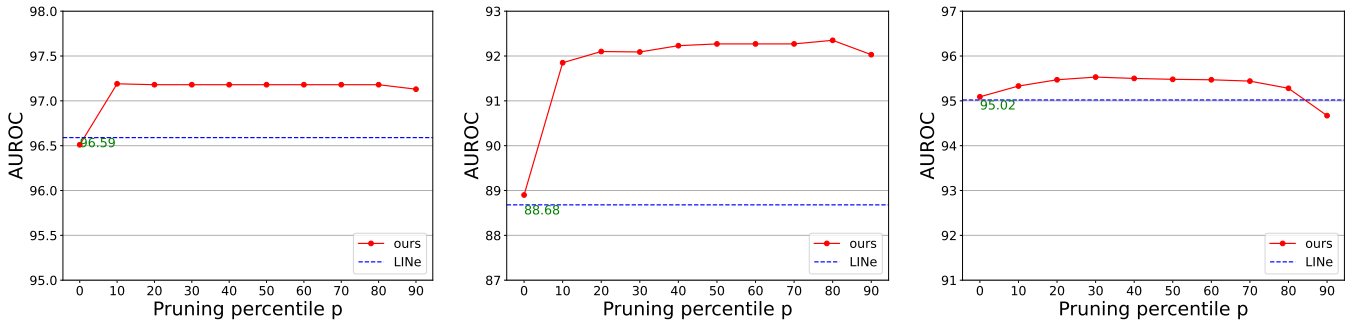

Fig. 6. Sensitivity study of masking percentile p on the CIFAR-10, CIFAR-100, and ImageNet benchmarks. DenseNet is used on CIFAR benchmarks, and ResNet-50 on the ImageNet benchmark. All AUROC values are averaged over multiple OOD datasets.

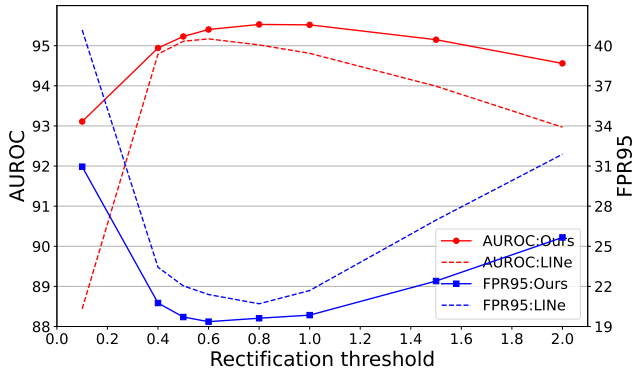

Fig. 7. Sensitivity study of rectification thresholds λ on the ImageNet benchmark with ResNet-50. All values are percentages and averaged over multiple OOD datasets.

TABLE IV
RESULTS OF APPLYING HIMPLOS AND REACT TO DIFFERENT OOD DETECTION METHODS. DENSENET IS USED ON CIFAR BENCHMARKS, AND RESNET-50 ON THE IMAGENET BENCHMARK. ALL VALUES ARE AVERAGED OVER MULTIPLE OOD DATASETS.

Method	CIFAR-10		CIFAR-100		ImageNet	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP	50.04	92.05	80.20	74.35	66.95	81.99
MSP+ReAct	50.07	92.32	80.19	74.71	58.28	87.06
MSP+HIMPLoS	**18.99**	**96.62**	**43.49**	**89.25**	**22.49**	**94.66**
Energy	28.28	94.31	68.54	81.18	58.41	86.17
Energy+ReAct	23.47	95.95	65.37	84.13	31.43	92.95
Energy+HIMPLoS	**14.62**	**97.18**	**30.06**	**92.23**	**19.62**	**95.53**
ODIN	22.14	94.17	56.23	85.38	56.48	85.41
ODIN+ReAct	22.08	95.63	49.18	88.85	44.10	90.70
ODIN+HIMPLoS	**14.60**	**97.10**	**28.42**	**92.15**	**21.16**	**95.05**
LINe	16.95	96.59	35.67	88.68	20.70	95.03
LINe+HIMPLoS	**16.35**	**96.71**	**33.37**	**89.43**	**17.84**	**95.97**

TABLE V
COMPARISON OF LOGIT SMOOTHING (HIMPLOS) WITH LOGIT ENHANCING IN DIFFERENT WEIGHT COEFFICIENT α. DENSENET IS USED ON THE CIFAR-100 BENCHMARK, AND RESNET-50 ON THE IMAGENET BENCHMARK. BOLD NUMBERS ARE SUPERIOR RESULTS. ALL PERCENTAGE VALUES ARE AVERAGED OVER MULTIPLE OOD DATASETS.

Method	Weight coefficient	CIFAR-100		ImageNet	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
logit enhancing	α=1	35.77	90.28	24.63	93.99
	α=5	33.93	90.89	21.52	94.83
	α=10	32.05	91.43	21.25	95.31
	α=20	30.64	92.13	20.72	95.39
	α=50	31.61	92.01	27.44	94.25
	α=100	34.59	91.68	32.36	92.77
HIMPLoS (Ours)	-	**30.06**	**92.23**	**19.62**	**95.53**

28.63% on the CIFAR-100 benchmark and reduces FPR95 by 5.1% on the ImageNet benchmark, which supports the effect of feature masking on performance improvement. ReAct+LS (fourth row) outperforms ReAct by 16.95% and 8.33% in FPR95 on the CIFAR-100 benchmark and the ImageNet benchmark respectively, supporting the effectiveness of logit smoothing for OOD detection. Besides, when we remove ReAct in our method (fifth row), the FPR95 increases by 10.28% and 12.89% compared to our method (last row) on the CIFAR-100 benchmark and the ImageNet benchmark respectively, suggesting that ReAct is also important for performance improvement. The inclusion of all these three components (last row) achieves the best OOD detection performance on both benchmarks, supporting that the three components are complementary to each other and all play important roles in our method for performance improvement.

2) Sensitivity of rectification threshold: As ReAct is one of the important components in our method, we perform a sensitivity study to show the effect of rectification threshold λ and compare the performance between our method and LINe [44] which also uses ReAct. As shown in Figure 7, when the threshold λ is too large (e.g., 2.0), both methods perform relatively worse because ReAct plays little role in clipping large feature activation. As the threshold λ decreases, the performance improves and our method always outperforms the state-of-the-art method LINe for the same rectification threshold λ. Our method performs stably well in the range $[0.5, 1.5]$, suggesting that our method is robust to the choice of the hyper-parameter λ. The performance of both methods drops when λ approaches 0, because most of the feature activation values are rectified to a small threshold which in turn leads to poorer logit separability between test ID and OOD images.

3) Sensitivity of masking percentile: Since feature masking is an important part of our method, we also perform a sensitivity study of masking percentile $p = \frac{L-k}{L} \cdot 100\%$,

TABLE VI
PERFORMANCE COMPARISON ON THE IMAGENET BENCHMARK WITH VIT-B/16 MODEL.

Method	OOD Datasets								Average	
	iNaturalist		SUN		Places		Textures			
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP [2]	19.15	96.13	57.00	86.13	59.97	85.09	51.49	85.10	46.90	88.11
ODIN [41]	6.53	98.57	38.99	91.50	46.66	89.24	33.99	91.29	31.54	92.65
Mahalanobis [27]	**2.13**	**99.54**	51.26	89.14	59.87	86.24	28.32	92.60	35.39	91.88
Energy [16]	6.04	98.66	37.11	91.75	45.30	89.30	31.90	91.70	30.09	92.85
BATS [46]	6.23	98.59	41.51	90.41	49.62	87.85	33.16	91.23	32.63	92.02
DICE [43]	91.86	65.22	88.70	64.47	93.85	61.53	74.27	74.95	87.17	66.54
ReAct [42]	4.19	99.01	39.06	91.54	47.38	89.11	32.27	91.54	30.72	92.80
DICE + ReAct [43]	96.13	55.25	89.42	63.32	95.03	59.73	78.21	71.03	89.70	62.33
LINe [44]	4.36	98.93	**33.79**	92.19	**45.09**	89.59	47.46	90.73	32.68	92.86
HIMPLoS (Ours)	3.00	99.12	37.56	**92.25**	46.70	**90.63**	**28.30**	**93.15**	**28.89**	**93.54**

where L is the total number of elements in the feature vector at the penultimate layer and k is the number of feature elements selected for un-masking. Figure 6 demonstrates the performance of our method (red curves) on the CIFAR-10 and CIFAR-100 benchmarks with DenseNet and on the ImageNet benchmark with ResNet-50 when varying the masking percentile p. The performance from the state-of-the-art method LINe [44] (dashed blue lines) is also included for comparison. Note that $p = 0$ corresponds to the case of using the original feature vector (i..e, no feature masking). Significant performance improvement is observed when varying p from 0% to 10%, clearly supporting the importance of feature masking for OOD detection. The performance of our method remains stably well between the large range $[10\%, 80\%]$ and is consistently better than that of the strong baseline LINe on both benchmarks, confirming that our method is insensitive to the choice of hyper-parameter in feature masking. When p gets extremely large (e.g., 90%), the performance drops rapidly on the ImageNet benchmark as expected, because multiple feature elements which are crucial to ID classes have been masked at such high masking percentile.

E. Compatibility with other methods

We have shown the effectiveness of applying our method to Energy score [16]. Actually, our method is also compatible with other existing methods. In Table IV, we compare the results of applying our method and ReAct [42] to different OOD detection methods, including MSP [2], ODIN [41], LINe [44], and Energy [16]. Note that MSP, ODIN, and Energy are associated with different scoring functions, therefore our method and ReAct can simply be applied to these methods. As for LINe, we only apply logit smoothing to it, since it already masks features and applies ReAct. As we can see, our method can improve the performance of all these methods and outperform all the methods with ReAct, which confirms the flexibility and effectiveness of our method. Notably, when applying logit smoothing to LINe, the best performance is achieved on the ImageNet benchmark with ResNet-50, further confirming the importance of logit smoothing in OOD detection.

F. Comparison with other combination strategies

We have shown that logit smoothing significantly improves the OOD detection performance by combining feature information with logit information. However, considering the alternative strategy (i.e., addition) to combine the feature information with the logit information, we add the cosine similarity to logit as follows and named it logit enhancing,

$$\hat{f}(\mathbf{x}) = f_m(\mathbf{x}) + \alpha \cdot s(h(\mathbf{x}), \mathbf{v}_c), \tag{12}$$

where α is the weight coefficient, and it is applied element-wise to the logit vector $f_m(\mathbf{x})$.

In Table V, we compare the performance of logit smoothing (HIMPLoS) with logit enhancing in different weight coefficient α. When α gets too extreme, we can't balance the effects of logit and feature information for OOD detection. When α is properly chosen (e.g., 20), the performance of logit enhancing gets better. However, our method (logit smoothing) outperforms logit enhancing for all values of α which shows the effectiveness and convenience of logit smoothing.

G. Evaluation on Transformer-based ViT models

Following the general experimental setting, we have shown the effectiveness of our method on various CNN-based models. To further explore the generalizability of our method to more model architectures, we provide a comprehensive evaluation on the ImageNet benchmark with the transformer-based ViT model [60]. ViT [60] is a transformer-based image classification model which treats each image as a sequence of image patches. We use the ViT-B/16 model [60] which is pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K.

Table VI summarizes the performance of OOD detection from various post-hoc detection methods. Note that transformer-based models are quite different from CNN-based models, and the observations on CNN-based models might not always apply to ViT models. For example, the performance of DICE [43] and ReAct [42] drop compared to Energy [16]. However, it shows that our method achieves the best performance on average with ViT backbone, which demonstrates the effectiveness and generalizability of our method with the ViT backbone.

VI. Conclusion

In this study, a new post-hoc OOD detection method is proposed based on feature masking and logit smoothing. Feature masking is expected to remove those high activation features caused by OOD samples, while preserving most of the high activation features caused by ID samples. Logit smoothing can further enlarge the difference in OOD score between ID and OOD samples, therefore helping improve OOD detection performance. Extensive experiments confirm that our method establishes new state-of-the-art performance on multiple benchmarks with different model and is robust to the choice of hyper-parameters. Moreover, the flexible combination of our method with existing OOD detection methods suggests the high extensibility of our method. We expect the combination of the information in the feature space with the logit information in our method can help motivate new research on solving the overconfidence issue in OOD detection.

References

[1] R. Huang and Y. 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*, 2021, pp. 8710–8719.

[2] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proceedings of International Conference on Learning Representations*, 2017.

[3] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, 4, pp. 1092–1108, 2020.

[4] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *International Conference on Machine Learning*, 2020, pp. 3145–3153.

[5] J. Fang, J. Qiao, J. Xue, and Z. Li, "Vision-based traffic accident detection and anticipation: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[6] J. Cen, Z. Jiang, L. Xie, D. Jiang, W. Shen, and Q. Tian, "Consensus synergizes with memory: A simple approach for anomaly segmentation in urban scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[7] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, 9, pp. 3694–3706, 2021.

[8] Y. Zhong, X. Chen, Y. Hu, P. Tang, and F. Ren, "Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, 12, pp. 8285–8296, 2022.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[10] Y. Zhou, "Rethinking reconstruction autoencoder-based out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7379–7387.

[11] Y.-C. Hsu, Y. Shen, H. Jin, and Z. 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*, 2020, pp. 10 951–10 960.

[12] X. Du, Z. Wang, M. Cai, and Y. Li, "VOS: learning what you don't know by virtual outlier synthesis," in *Proceedings of International Conference on Learning Representations*, 2022.

[13] Z. Lin, S. D. Roy, and Y. Li, "Mood: Multi-level out-of-distribution detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 313–15 323.

[14] A. Djurisic, N. Bozanic, A. Ashok, and R. Liu, "Extremely simple activation shaping for out-of-distribution detection," in *Proceedings of International Conference on Learning Representations*, 2023.

[15] Y. Ming, Y. Sun, O. Dia, and Y. Li, "How to exploit hyperspherical embeddings for out-of-distribution detection?" in *Proceedings of International Conference on Learning Representations*, 2023.

[16] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21 464–21 475, 2020.

[17] Y. Yu, S. Shin, S. Lee, C. Jun, and K. Lee, "Block selection method for using feature norm in out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 701–15 711.

[18] A. Nguyen, J. Yosinski, and J. 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*, 2015, pp. 427–436.

[19] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 41–50.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, 11, 2008.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[23] Y. Yang, R. Gao, and Q. Xu, "Out-of-distribution detection with semantic mismatch under masking," in *European Conference on Computer Vision*. Springer, 2022, pp. 373–390.

[24] M. Cai and Y. Li, "Out-of-distribution detection via frequency-regularized generative models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5521–5530.

[25] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," *Advances in neural information processing systems*, vol. 33, pp. 11 839–11 852, 2020.

[26] V. Sehwag, M. Chiang, and P. Mittal, "SSD: A unified framework for self-supervised outlier detection," in *Proceedings of International Conference on Learning Representations*, 2021.

[27] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, pp. 7167–7177, 2018.

[28] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *International Conference on Machine Learning*, 2022, pp. 20 827–20 840.

[29] G. Jiang, P. Zhu, Y. Wang, and Q. Hu, "Openmix+: Revisiting data augmentation for open set recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, 11, pp. 6777–6787, 2023.

[30] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.

[31] A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke, "Out-of-distribution detection using an ensemble of self supervised leave-out classifiers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 550–564.

[32] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *International Conference on Machine Learning*, 2022, pp. 23 631–23 644.

[33] D. Zhong and J. Zhu, "Centralized large margin cosine loss for open-set deep palmprint recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, 6, pp. 1559–1568, 2020.

[34] Q. Zhu, G. Zheng, and Y. Yan, "Effective out-of-distribution detection in classifier based on pedcc-loss," *Neural Processing Letters*, vol. 55, 2, pp. 1937–1949, 2023.

[35] Q. Zhu, G. Zheng, J. Shen, and R. Wang, "Out-of-distribution detection based on feature fusion in neural network classifier pre-trained by pedcc-loss," *IEEE Access*, vol. 10, pp. 66 190–66 197, 2022.

[36] J. Sun, H. Wang, and Q. Dong, "Moep-ae: Autoencoding mixtures of exponential power distributions for open-set recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, 1, pp. 312–325, 2023.

[37] Q. Zhu and R. Zhang, "A classification supervised auto-encoder based on predefined evenly-distributed class centroids," *arXiv preprint arXiv:1902.00220*, 2019.

[38] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," *Advances in Neural Information Processing Systems*, vol. 34, pp. 677–689, 2021.

[39] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4921–4930.

[40] J. Zhang, Q. Fu, X. Chen, L. Du, Z. Li, G. Wang, S. Han, D. Zhang *et al.*, "Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy," in *Proceedings of International Conference on Learning Representations*, 2023.
[41] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proceedings of International Conference on Learning Representations*, 2018.
[42] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.
[43] Y. Sun and Y. Li, "Dice: Leveraging sparsification for out-of-distribution detection," in *European Conference on Computer Vision*, 2022, pp. 691–708.
[44] Y. H. Ahn, G.-M. Park, and S. T. Kim, "Line: Out-of-distribution detection by leveraging important neurons," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 852–19 862.
[45] L. S. Shapley *et al.*, *A value for n-person games*. Princeton University Press Princeton, 1953.
[46] Y. Zhu, Y. Chen, C. Xie, X. Li, R. Zhang, H. Xue, X. Tian, Y. Chen *et al.*, "Boosting out-of-distribution detection with typical features," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 758–20 769, 2022.
[47] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," University of Toronto, Technical Report TR-2009, 2009.
[48] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.
[49] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.
[50] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," *arXiv preprint arXiv:1504.06755*, 2015.
[51] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.
[52] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, 6, pp. 1452–1464, 2017.
[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.
[54] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778.
[55] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*, 2010, pp. 3485–3492.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[57] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[58] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, K. Zhou, W. Zhang, Y. Li, Z. Liu, Y. Chen, and H. Li, "Openood v1.5: Enhanced benchmark for out-of-distribution detection," *arXiv preprint arXiv:2306.09301*, 2023.
[59] J. Bitterwolf, M. Müller, and M. 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*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 2471–2506.
[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of International Conference on Learning Representations*, 2021.

Zhuohao Sun received the bachelor's degree in computer science and technology from Sun Yat-sen University in 2022, where he is currently pursuing the master's degree with the School of Computer Science and Engineering. His research interests include computer vision and machine learning.

Yiqiao Qiu received the bachelor's degree in computer science and technology from Sun Yat-sen University in 2022. He is currently pursuing the master's degree at the University of California, San Diego. His research interests include lite models optimization, computer vision, and machine learning.

Zhijun Tan received the Ph.D. degree in computational mathematics from Hong Kong Baptist University in 2005. His current research interests include scientific computing, computational fluid dynamics, interface problems, fractional-order PDE, AI for PDEs, artificial intelligence, and deep learning. In these areas, he has made some outstanding research achievements and has published more than 70 papers in SIAM Journal on Scientific Computing, Journal of Scientific Computing, Computer Methods in Applied Mechanics and Engineering, Journal of Computational Physics, International Journal of Mechanical Sciences, Chemical Engineering Science and other international authoritative journals. He has received many projects funded by the National Natural Science Foundation of China and the Natural Science Foundation of Guangdong Province.

Weishi Zheng is now a full Professor with Sun Yat-sen University. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. He has ever served as area chairs of ICCV, CVPR, ECCV, BMVC, IJCAI and etc. He is associate editors of IEEE-TPAMI, Pattern Recognition. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He is a Cheung Kong Scholar Distinguished Professor, a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of the Royal Society-Newton Advanced Fellowship of the United Kingdom.

Ruixuan Wang received the Ph.D. degree from the National University of Singapore in 2008. He was a postdoctoral researcher with the University of Dundee, UK. He is currently an associate professor with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include computer vision, medical image analysis, and machine learning.