This ICCV 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.

Hierarchical Visual Categories Modeling: A Joint Representation Learning and Density Estimation Framework for Out-of-Distribution Detection

Jinglun Li[1] Xinyu Zhou[2] Pinxue Guo[1] Yixuan Sun[1] Yiwen Huang[2]
Weifeng Ge[2†] Wenqiang Zhang[1,2†]
[1]Academy for Engineering and Technology, Fudan University
[2]School of Computer Science, Fudan University
jingli960423@gmail.com, zhouxinyu20@fudan.edu.cn,
{pxguo21, 21210860014, 21210240056}@m.fudan.edu.cn,
weifeng.ge.ic@gmail.com, wqzhang@fudan.edu.cn

Abstract

Detecting out-of-distribution inputs for visual recognition models has become critical in safe deep learning. This paper proposes a novel hierarchical visual category modeling scheme to separate out-of-distribution data from in-distribution data through joint representation learning and statistical modeling. We learn a mixture of Gaussian models for each in-distribution category. There are many Gaussian mixture models to model different visual categories. With these Gaussian models, we design an in-distribution score function by aggregating multiple Mahalanobis-based metrics. We don't use any auxiliary outlier data as training samples, which may hurt the generalization ability of out-of-distribution detection algorithms. We split the ImageNet-1k dataset into ten folds randomly. We use one fold as the in-distribution dataset and the others as out-of-distribution datasets to evaluate the proposed method. We also conduct experiments on seven popular benchmarks, including CIFAR, iNaturalist, SUN, Places, Textures, ImageNet-O, and OpenImage-O. Extensive experiments indicate that the proposed method outperforms state-of-the-art algorithms clearly. Meanwhile, we find that our visual representation has a competitive performance when compared with features learned by classical methods. These results demonstrate that the proposed method hasn't weakened the discriminative ability of visual recognition models and keeps high efficiency in detecting out-of-distribution samples.

1. Introduction

Modern deep neural networks have shown strong generalization ability when training and test data are from the same distribution [44, 18, 51, 10, 33]. However, encountering unexpected scenarios is inevitable in real-world applications. Thus assuring that training and test data share the same distribution becomes problematic. In applications like autonomous driving [3, 4] and medical image analysis [55, 14, 43], it is critical for models to identify inputs beyond their recognition capacity – known as out-of-distribution (OOD) detection. OOD detection algorithms can enable the system to warn humans promptly in many safety-related scenarios. Moreover, it has become an important research topic in the research community of safe artificial intelligence [21, 29, 23, 32, 56].

Many popular OOD detection methods aim to build probability models to describe training distributions [29, 56, 46, 24, 23]. With these probability models, they built a score function that can calculate in-distribution scores for test samples. These in-distribution scores reflect whether these samples fall into a given distribution. Then the test sample can be evaluated by the score function to decide whether it is an OOD sample or not. Thus modeling features of in-distribution data become extremely important. Previous works [29, 16, 5, 12, 38] build a distribution over the whole training data. Since training images may come from various visual categories, the decision boundary between In-Distribution (InD) and OOD data becomes extremely complex. To solve this problem, subsequent studies [24, 7, 15, 56] decomposed the whole dataset into several subgroups to simplify the decision boundary. Although representative algorithms like MOS [24], have gotten impressive performance in identifying OOD samples, they failed to detect near OOD samples. Because when different visual categories are grouped together, the OOD decision boundary will become even more uncertain.

A typical framework for out-of-distribution detection involves two key steps: 1) learning a compact feature rep-

[†]indicates corresponding authors.

resentation that can fit probability models easily; 2) modeling features of in-distribution data in complex distributions accurately. The above two problems are mutually connected because more compact features will make modeling the data distribution easier, and stronger probability modeling techniques will exploit fewer restrictions on representation learning. However, achieving the above goals is difficult, even when there are a lot of breakthroughs in deep learning. Because if training samples from the same category are too close in the feature space, it will usually lead to overfitting. Meanwhile, in-distribution samples may come from different visual categories that have large variations in appearance and semantic information, which makes modeling complex training distributions become challenging even for excellent statisticians.

In this paper, we propose a new out-of-distribution detection framework, called *hierarchical visual category modeling*, to solve the above two issues simultaneously. We hold an assumption that given a training set that contains multiple visual categories, we can learn a probability model for every category independently. The out-of-distribution detection problem can be solved easily by aggregating probability models of known categories. Our motivation is that decomposing the whole dataset into subsets and modeling each category independently can avoid finding common characteristics shared by different categories. However, modeling an individual visual category is still challenging since classical supervised learning won't lead to compact feature representation. Thus, for each input sample, we need to force its feature representation to match the corresponding statistical model. That means we need to jointly conduct density estimation and representation learning. If we can jointly learn visual representations and optimize statistical models end-to-end, we can get good feature representations falling into distributions of the corresponding visual categories. Besides, we exploit knowledge distillation as done in [6] to learn robust feature representation. In this way, we can describe the complex training distribution with multiple Gaussian mixture models while not impairing the generalization ability of visual features.

In practice, to learn visual concepts that are in complex distributions, we build a Gaussian mixture model (GMM [39]) for each visual category. Given input samples, we extract their deep features and project these features into a high-dimensional attribute space. Different from classical Gaussian mixture models that send the same input into K different Gaussian models, we divide the attribute space into multiple groups and build a Gaussian model in each group independently. This strategy can give every attribute group a clear learning target and lead to better convergence. Experimental results indicate that this strategy works quite well. After the visual representation learning and statistical model parameters optimization, we can directly aggregate these statistical models to judge whether a test sample comes from the training distribution or not. To evaluate our OOD detector, we split ImageNet into ten folds randomly and select one of these splits as the training set and all other splits as the OOD dataset to conduct extensive tests. Experiments indicate that the proposed method has a strong ability to identify OOD samples. We also evaluate our method on seven popular OOD benchmarks. Experimental results demonstrate that the proposed method not only can identify OOD samples efficiently but also improves the discriminative ability of learned visual representations.

The contributions of this paper are summarized as follows:

- We introduce a new out-of-distribution detection scheme, called *hierarchical visual category modeling* to conduct joint representation learning and density estimation. It provides a new perspective for out-of-distribution detection to learn visual representation and probability models end-to-end.
- We exploit multiple Gaussian mixture models to model visual concepts in complex distributions. Visual attributes are divided into subgroups and modeled by different Gaussian components, which makes parameter learning much more efficient.
- We conducted comprehensive experiments and ablation studies on popular benchmarks to investigate the effectiveness of the proposed method. Experiments demonstrated that our out-of-distribution detection models achieve better performance clearly when compared with previous methods.

2. Related Work

Out-Of-Distribution Detection. Out-of-distribution detection aims to distinguish out-of-distribution samples from in-distribution data. Numerous methods have been proposed for OOD detection. Maximum softmax probability (MSP) [21] has been recognized as a strong baseline by using the maximum score across all classes as an OOD score. ODIN [30] improves MSP by perturbing the inputs and adjusting the logits via rescaling. Gaussian discriminant analysis has been employed in [29, 56] to detect OOD samples. ReAct [45] uses rectified activation to reduce model overconfidence in OOD data. Shama *et al.* in [42] utilized Gram matrices to measure feature correlations for OOD detection. Bibas *et al.* in [2] proposed pNML regret to detect OOD samples with a single-layer neural network. The Generalized-ODIN approach [23] decomposes the confidence of class probability using a dividend/divisor structure to incorporate prior knowledge. Another promising line of work focuses on designing new learning objectives to train

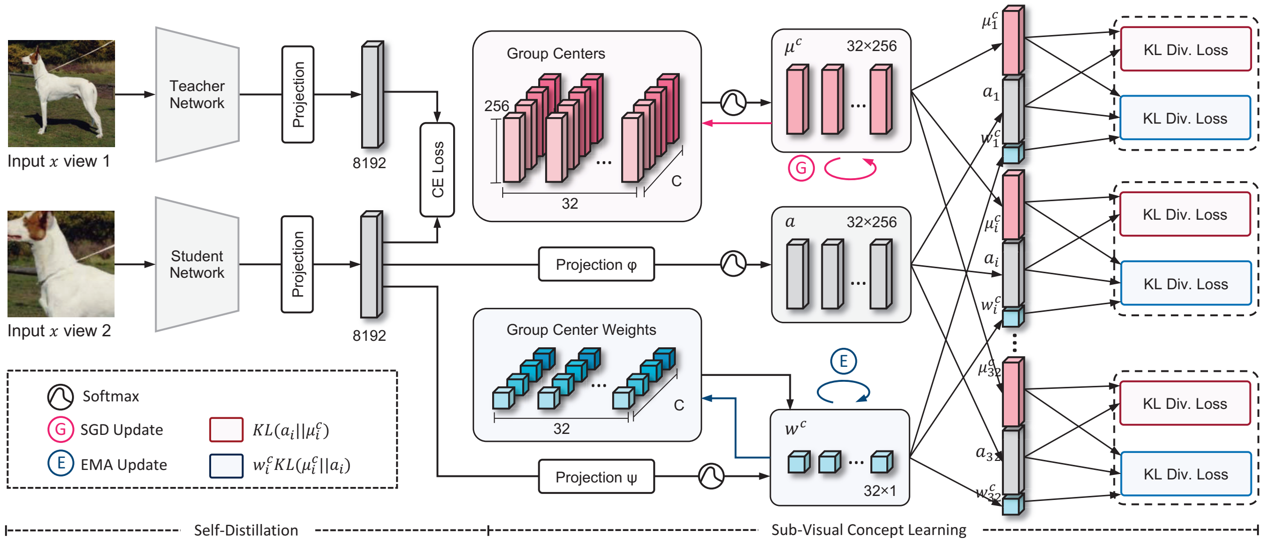

Figure 1. Illustration of the training pipeline of hierarchical visual category modeling, written as HVCM. In HVCM, we jointly learn the visual representation and parameters of probabilistic models. We get two different views of an input image and send the outputs into a knowledge distillation framework (DINO [6]). Image representations are projected into a high-dimensional attribute space. Then these attributes are divided into different groups and pass SoftMax functions to get attribute distributions. We match attributes in each group with stored attribute centers of the target visual category. The whole model is trained in an end-to-end manner.

deep models from scratch. Leave-out Classifiers [52] introduces a margin loss to ensure that InD and OOD samples are separated in the feature space. Lee *et al.* in [28] proposed a novel confidence loss to give lower confidence for OOD samples. There are also some methods [58, 41, 35] designed to conduct OOD detection based on generative models. Unlike these methods, our proposed method jointly conducts representation learning and density estimation to improve the OOD detection performance and keep the learned feature with strong discriminative ability.

Density Estimation in Deep Learning. Density estimation tries to describe the probabilistic density distribution of observed data accurately and has been investigated in deep learning for a long time [47, 11, 31]. In [1], Chong *et al.* employed variational autoencoders for anomaly detection, where the reconstruction probability was used to compute the anomaly score of each sample. Papamakarios *et al.* [36] proposed a novel method for density estimation based on masked autoregressive flow. Zhou *et al.* [62] extended variational autoencoders by incorporating a mixture of Gaussians to model the latent space distribution, allowing for more flexible and expressive representations. Yang *et al.* [13] combined Flow-based generative models with generative adversarial networks for density estimation and sample generation. Zhao *et al.* [11] used discrete latent variables to conduct density estimation, which has been applied to many research topics in natural language processing and image processing. In this paper, we build probabilistic models that can be jointly learned with deep networks. We wish latent representations of the visual categories could easily be modeled by Gaussian mixture models even when complex distributions exist. Besides, we do not use post-hoc methods to conduct feature learning and density estimation separately. We aim to use the probabilistic model to guide the feature-learning process of deep networks.

3. Method

3.1. Framework

State-of-the-art methods [21, 30, 29, 24, 46, 32] typically assume all InD data follow the same distribution. However, such an assumption will lead to the decision boundary of an OOD detector being doped with some uncertain space. In this paper, we avoid modeling the whole InD dataset and only focus on independently modeling each object category's distribution. We propose hierarchical visual categories modeling to achieve this goal while maintaining high classification accuracy. In hierarchical visual categories modeling, we first project image features into a high-dimensional attribute space (usually 8192 dimensions). These attributes can be grouped into multiple sub-visual concepts as components of an image category which can be easily modeled by Gaussian distributions. Then combinations of multiple sub-visual concepts (abbreviated as sub-concepts) can be grouped to describe a more complex visual concept of an in-distribution category. Since we define visual categories based on sub-visual concepts, simply modeling distributions of these sub-concepts and describing visual categories with these sub-concepts hierarchically can model complicated training distributions.

Formally, given a visual recognition model f, it maps an input image $\boldsymbol{x}$ with label y to a high-dimensional feature vector $\boldsymbol{z} \in \mathbb{R}^q$. As described above, we project $\boldsymbol{z}$ into an attributes space $\mathcal{S} \in \mathbb{R}^d$ with a higher dimension and get an attribute description $\boldsymbol{a}$. Attributes in $\boldsymbol{a}$ can be grouped into multiple attribute subgroups $\{\boldsymbol{a}_i\}_{i=1}^G$ where $\boldsymbol{a}_i \in \mathcal{S}_i \subset \mathbb{R}^{d/G}$ and $i \in \{1, 2, \dots, G\}$. For images from a visual category c, we assume that attributes in their i-th attribute group follow a simple Gaussian distribution $\mathcal{N}(\mu_i^c, \Sigma_i^c)$, where μ_i^c and Σ_i^c are the mean and variance respectively. Since attributes are divided into G groups, we will have G different Gaussian distributions for each visual category. The benefits of dividing the whole attribute space into several subspaces come from two folds: (1) modeling attribute distributions in these groups becomes easier; (2) assemble of distributions in these attribute groups can describe complex distributions, which will lead to more accurate decision boundaries. To combine all attributes to describe every visual category, our OOD detector learns weights $\{w_i^c\}_{i=1}^G$ of all attribute groups through exponential moving averages. Therefore, for each in-distribution class c, we can model the corresponding visual concept in a probability perspective with a mixture of Gaussian models:

$$p\left(\boldsymbol{x};c\right) = \sum_{i=1}^{G} w_i^c \mathcal{N}(\boldsymbol{a}_i; \boldsymbol{\mu}_i^c, \boldsymbol{\Sigma}_i^c), \tag{1}$$

where $w_i^c \in \mathbb{R}, \boldsymbol{\mu}_i^c \in \mathbb{R}^{d/G}, \boldsymbol{\Sigma}_i^a \in \mathbb{R}^{d/G \times d/G}$ and $\mathcal{N}(\cdot)$ means a normal distribution. We build a Gaussian mixture model for each class and get C different Gaussian mixture models. Then, our proposed HVCM focuses on training deep neural networks to learn image features that follow the above distributions and parameters of these probability models jointly.

With the above Gaussian probability models, given a test sample $\boldsymbol{x}'$, we define the score function $g(\boldsymbol{x}'; \boldsymbol{w}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c)$ to measure whether it belongs to the c-th visual category using the learned probability density function. Here, $\boldsymbol{w}^c = \{w_i^c\}_{i=1}^G$, $\boldsymbol{\mu}^c = \{\boldsymbol{\mu}_i^c\}_{i=1}^G$ and $\boldsymbol{\Sigma}^c = \{\boldsymbol{\Sigma}_i^c\}_{i=1}^G$. This score function can be used as a reliable metric to detect OOD samples:

$$h(\boldsymbol{x}') = \begin{cases} \text{InD}, & \text{if } \min_c g(\boldsymbol{x}'; \boldsymbol{w}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c) \geq \gamma, \\ \text{OOD}, & \text{if } \min_c g(\boldsymbol{x}'; \boldsymbol{w}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c) < \gamma, \end{cases} \tag{2}$$

where γ are thresholds to be determined in subsequent sections. Eq. (2) indicates that we use the minimal InD score among all C categories to make the final decision.

Our framework is illustrated in Figure 1, where there are two steps: (1) jointly learn deep features that fit our probability models and parameters of these models; (2) calculate the InD score based on a set of Gaussian mixture models as the metric for out-of-distribution detection. In the following subsections, we give more details.

3.2. Joint Visual Representation Learning and Parameter Optimization of Probability Models

To learn visual representations that follow Gaussian mixture models and keep their discriminative ability at the same time, we exploit the knowledge distillation framework DINO [6] to perform the joint learning. As shown in Figure 1, for an image $\boldsymbol{x}$, we sample ten different views of $\boldsymbol{x}$ and send them into teacher and student branches simultaneously to perform self-distillation. During knowledge distillation, we project ResNet50 [19] features in 2048 dimensions into an attribute space with dimension $d(= 8192)$. Apart from the learning objective produced by self-distillation, we force the attributes of each class to follow a class-specific Gaussian mixture model. We divide the attributes $\boldsymbol{a} \in \mathbb{R}^d$ of $\boldsymbol{x}$ into G groups to learn the Gaussian mixture model parameters. However, in practice, it's too hard to directly learn the mean and variance of Gaussian models. We follow He *et. al* [20] to learn attribute centers $\{\boldsymbol{\mu}_i^c\}_{i=1}^G$ of the c-th category (the image label y is c). A linear projection layer is exploited to predict the weights $\{w_i^c(\boldsymbol{x})\}_{i=1}^G$ of $\boldsymbol{x}$ on all G attribute groups. Our learning objective can be written as follows:

$$\mathcal{L} = \mathcal{L}_{KD} + \alpha \sum_{i=1}^{G} \mathrm{KL}(\boldsymbol{a}_i || \boldsymbol{\mu}_i^c) + \beta \sum_{i=1}^{T} w_i^c\left(\boldsymbol{x}\right) \mathrm{KL}(\boldsymbol{\mu}_i^c || \boldsymbol{a}_i) \tag{3}$$

where $\mathcal{L}_{KD}$ stands for the cross entropy loss in self-distillation, KL stands for the Kullback-Leibler(KL) divergence, and α and β are hyperparameters. The reason why we exploit two KL divergences is that experiments indicate that the term $\mathrm{KL}(\boldsymbol{a}_i || \boldsymbol{\mu}_i^c)$ will favor learning attribute centers and the term $\mathrm{KL}(\boldsymbol{\mu}_i^c || \boldsymbol{a}_i)$ is more suitable to learn image attribute descriptions and group weights. Note that softmax operations normalize attributes and their learnable centers in each group before calculating learning objectives.

With the learning objective in Eq. (3), student network parameters θ_s, and weights $\{w_i^c\}_{i=1}^G$ and attribute centers $\{\boldsymbol{\mu}_i^c\}_{i=1}^G$ of all groups are learnt at the same time. They are updated using the following equations:

$$\theta_s^{t+1} = \theta_s^t - \gamma_1 \frac{\partial \mathcal{L}}{\partial \theta_s^t}, \tag{4}$$

$$\boldsymbol{\mu}_i^{c,t+1} = \boldsymbol{\mu}_i^{c,t} - \gamma_2 \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_i^{c,t}}, \tag{5}$$

$$\boldsymbol{w}_i^{c,t+1} = (1 - \gamma_3) \boldsymbol{w}_i^{c,t} + \gamma_3 w_i^c(\boldsymbol{x}), \tag{6}$$

where γ_1, γ_2 and γ_3 are the learning rates. γ_1, γ_2 control the gradient update speed, while γ_3 controls the updating speed of the group weights. Following He *et. al* [20], θ_s, $\{\boldsymbol{w}^c\}_{c=1}^C$ and $\{\boldsymbol{\mu}^c\}_{c=1}^C$ are initialized with Gaussian noises. We use

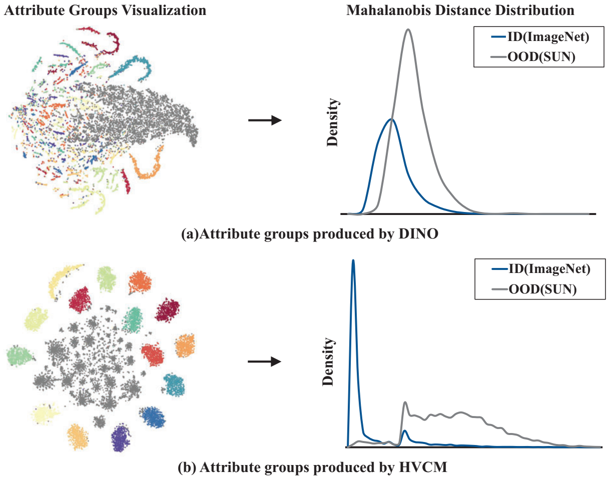

Figure 2. Illustration of attribute group visualization and Mahalanobis distance distribution. Both attribute groups are visualized by t-SNE [49]. The colors encode different in-distribution data(ImageNet), and out-of-distribution features(SUN) marked as gray points. Models are trained on ResNet-50 [18] using DINO(a) and HVCM(b).

Adam optimizer [25] with momentum to update both θ_s and $\{\boldsymbol{\mu}^c\}_{c=1}^{C}$. While the attribute weights are learned through exponential moving averages [17].

3.3. Out-of-Distribution Detection based on Probability Models in Hierarchy

When the training of our hierarchical probability model converges, we will get visual attributes that follow learned Gaussian distributions hierarchically for in-distribution samples. Meanwhile, we also get weights and centers of all attribute groups for each class. However, the mean attributes are updated frequently during the training and are thus not suitable to be used as group centers. So we need to recalculate the attribute centers for each visual category. Given an in-distribution visual category c, we estimate the mean vector and covariation matrix of the i-th attribute group:

$$\boldsymbol{\mu}_i^c = \frac{1}{N_c}\sum_{m=1}^{N_c} \boldsymbol{a}_i^m \tag{7}$$

$$\boldsymbol{\Sigma}_i^c = \frac{1}{N_c - 1}\sum_{m=1}^{N_c} (\boldsymbol{a}_i^m - \boldsymbol{\mu}_i^c)(\boldsymbol{a}_i^m - \boldsymbol{\mu}_i^c)^\top, \tag{8}$$

where N_c notes the number of samples in the c-th class, and $\boldsymbol{a}_i^m$ is the sub-attribute vector of the m-th sample. With these weights, means, and covariances, we can describe each visual category in the hierarchy accurately. We try to use the probability density function in Eq. (1) as the in-distribution function. However, we will encounter the problem of numerical overflow when calculating the determinants of covariance matrices as [48]. Instead, we compute the Mahalanobis distance between the i-th sub-visual attributes $\boldsymbol{a}_i'$ of a test sample $\boldsymbol{x}'$ and the corresponding attribute center $\boldsymbol{\mu}_i^c$ to measure the likelihood of these attributes belongs to the target category:

$$M_i^c(\boldsymbol{x}) = -(\boldsymbol{a}_i' - \boldsymbol{\mu}_i^c)^\top \left(\boldsymbol{\Sigma}_i^c\right)^{-1} (\boldsymbol{a}_i' - \boldsymbol{\mu}_i^c). \tag{9}$$

The above equation gives the in-distribution measure for one attribute group. While, for every category, we have multiple attribute groups and need to assemble related in-distribution measures to get the class-level in-distribution score. Since we have gotten the weights of attribute groups for each category, we can easily assemble them and get the class-level score function:

$$g(\boldsymbol{x}'; \boldsymbol{w}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c) = \sum_{i=1}^{G} w_i^c M_i^c(\boldsymbol{x}'). \tag{10}$$

With this score function, we can easily get the in-distribution score of a test sample on each visual category. Since there are C categories in the whole in-distribution dataset, we get the maximal in-distribution score across different visual categories as the in-distribution score on the whole dataset:

$$g(\boldsymbol{x}') = \max_c g(\boldsymbol{x}'; \boldsymbol{w}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c). \tag{11}$$

A high in-distribution score $g(\boldsymbol{x}')$ indicates that the semantic attributes of a test sample lie very close to one or multiple in-distribution visual categories(as shown in Figure 2). On the contrary, if a sample does not belong to the previously modeled categories, it will get a low in-distribution score. We follow Eq. (2) to set thresholds to judge whether a sample is an out-of-distribution sample. In the experiment section, we discuss how to set these thresholds.

4. Experiments

4.1. Experimental Setup

In-distribution Datasets. We use ImageNet-1K [40] and CIFAR10 [27] as our in-distribution datasets. ImageNet-1K is a large-scale visual recognition dataset containing 1000 object categories and 1281167 images. We split it into 10 folds randomly and ensured each fold contain 100 object categories. Since our computation resources are limited, we randomly select one fold as the in-distribution dataset. There other nine folds are used as OOD datasets as other popular benchmarks to test the performance of the proposed method in detecting near OOD samples. For CIFAR10 [27], there are 60000 color images in 10 classes. We conduct OOD algorithm evaluation as previous methods [30, 9, 46, 32, 45, 54].

Table 1. OOD detection performance comparison of HVCM and existing methods. All comparison methods rely on ResNet-50 trained with cross-entropy loss. * indicates that the method is fine-tuned on InD data. ↑ indicates larger values are better, and ↓ is the opposite. **Bold** numbers are superior results. All values are percentages.

Method	OOD Datasets								Average		InD Acc
	iNaturalist		SUN		Places		Textures				
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
MSP [21]	68.12	87.48	58.59	89.76	59.53	89.18	72.66	82.71	64.73	87.28	85.74
ODIN [30]	55.78	85.92	60.47	83.83	60.94	88.06	64.06	81.28	60.31	84.77	
Maha [29]	97.00	55.12	98.80	51.93	97.20	50.52	**20.00**	**94.99**	78.25	63.14	
Energy [32]	58.91	86.45	27.03	93.44	38.75	91.51	56.25	87.26	45.24	89.67	
GODIN* [23]	72.00	79.86	60.09	84.58	64.94	82.30	39.50	89.28	59.13	84.01	
MOS* [24]	52.94	91.41	67.78	86.82	71.31	84.38	73.65	80.56	66.42	85.79	
ReAct [45]	58.48	82.60	78.18	69.11	86.33	59.84	50.53	87.03	68.38	74.65	
HVCM(Ours)	**21.56**	**92.19**	**17.20**	**94.44**	**19.98**	**93.62**	29.22	90.68	**21.99**	**92.73**	**88.28**

Table 2. Evaluation on more challenging detection tasks. * indicates that the method is fine-tuned on InD data. ↑ indicates larger values are better, and ↓ is the opposite. **Bold** numbers are superior results. All values are percentages.

Method	ImageNet-O		OpenImage-O	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP [21]	72.50	83.33	89.11	57.99
ODIN [30]	73.44	71.03	64.06	79.88
Maha [29]	54.40	79.30	59.20	77.62
Energy* [32]	66.56	81.06	62.34	84.61
GODIN* [23]	71.55	79.89	73.57	77.27
MOS* [24]	87.40	64.87	66.78	81.42
ReAct [45]	87.05	64.15	84.17	64.30
HVCM(Ours)	**42.86**	**86.72**	**28.58**	**90.26**

Out-of-distribution Dataset. On ImageNet, we follow Huang *et al.* [24] to test our methods and use Texture [8], iNaturalist [50], Places365 [61], and SUN [57]) as OOD test sets. To further explore the limitation of our approach, we evaluate our method on another two OOD datasets, OpenImage-O [26] and ImageNet-O [22]. For CIFAR10, as in [46, 32, 45, 54], we selected eight widely-used datasets, including Texture [8], SVHN [34], Places365 [61], iSUN [59], LSUN-Crop [60], LSUN-Resize [60],ImageNet-Resize [40], and ImageNet-Fix [40] as our test sets. Besides, to test the ability of HVCM to identify near OOD datasets, we use the remaining nine ImageNet subsets as the OOD test sets. *Note that since our evaluation on ImageNet differs from previous methods [24, 45], we implement these algorithms with open source provided by authors and follow standard experimental settings.*

Evaluation Metrics. We employ the commonly used metrics in OOD detection [24, 45] to evaluate our approach, including AUROC, FPR95 and InD Acc. AUROC stands for the area under the receiver operating characteristic curve, FPR95 is short for TPR@FPR95 and represents the false positive rate when the true positive rate is 95%, and InD Acc is the classification accuracy of in-distribution data.

Training Details. We utilize ResNet-50 [19] as the feature backbone for ImageNet and the dimension of the attribute space is set to 8192. The training is finished in 300 epochs. On CIFAR10, we use ResNet-18 [19] as our feature backbone, and the dimension of the attribute space is set to 1024. The training on CIFAR10 is finished in 200 epochs. The number of attribute groups is set to 32, and α, β, γ_1, γ_2, and γ_3 are set to 1, 0.1, 1, 1 and 1×10^{-4}, respectively. Updating μ too fast can lead to the oscillation of group centers, negatively influencing the precision of probabilistic modeling. So we utilize a smaller β than α to update μ. All the hyperparameters are tuned according to the experimental results. We employ SGD with a momentum of 0.9, an initial learning rate of 0.1, and a batch size of 128. The learning rate is reduced by a factor of 10 at 50% and 75% of the total training epochs. We train all backbones from scratch using random initialization. All experiments are performed using PyTorch [37] with default parameters on four NVIDIA GeForce RTX 3090.

4.2. Comparison with State-of-the-Art Algorithms

Standard evaluation on ImageNet. We compare our HVCM with seven popular OOD detection methods, including MSP [21], ODIN [30], GODIN [23], Maha [29], Energy [32], MOS [24], and ReAct [45]. For datasets that describe objects or scenes, such as SUN, Places, and iNaturalist, HVCM achieves better AUROC and FPR95 metrics. When we summarize the results of all four datasets, HVCM achieves 21.99% on FPR95 and 92.73% on AUROC, which outperforms the previous best method Energy [32] by 23.25% and 3.06%. This is a significant improvement, which demonstrates that end-to-end training is very important to get good results. When compared to Maha [29], our proposed method exhibits inferior performance in terms of both FPR95 and AUROC. This observation suggests that our method is less effective in describing textures. We attribute this limitation to the fact that textures often encompass numerous repeated patterns, which differ from the characteristics required for general object

Table 3. Comparison of OOD detection performance of HVCM and existing methods on CIFAR10 dataset. All compared methods use ResNet-18 trained with cross-entropy loss except our proposed method, which uses HVCMLoss. The performance is evaluated based on AUROC (A) and FPR95 (F). ↑ indicates larger values are better and ↓ indicates the opposite. **Bold** numbers indicate superior results. All values are expressed in percentages.

Method		**OOD Dataset**																**Average**	
		Texture		**SVHN**		**Places365**		**iSUN**		**LSUN(C)**		**LSUN(R)**		**ImageNet(R)**		**ImageNet(F)**			
		F↓	**A↑**	**F↓**	**A↑**	**F↓**	**A↑**	**F↓**	**A↑**	**F↓**	**A↑**	**F↓**	**A↑**	**F↓**	**A↑**	**F↓**	**A↑**	**F↓**	**A↑**
CELoss	MSP	56.47	90.20	58.40	90.56	51.86	91.98	53.84	91.64	46.22	92.74	49.10	92.48	59.65	89.35	59.90	89.40	54.43	91.04
	ODIN	40.37	91.98	27.82	93.28	31.80	94.23	17.59	96.60	25.23	95.31	14.74	97.21	27.18	94.65	45.26	90.62	28.75	94.24
	Gram	10.81	97.73	2.58	99.39	30.15	93.37	1.11	99.76	13.91	97.03	0.52	99.86	1.61	98.38	66.94	82.02	15.95	95.94
	Gram+pNML	**7.18**	98.50	1.63	**99.60**	23.21	95.13	0.83	**99.80**	9.42	98.00	**0.42**	99.88	**1.24**	98.76	57.90	85.53	12.73	96.90
HVCMLoss	**HVCM(ours)**	**1.88**	**99.31**	**1.32**	99.47	**0.95**	**99.54**	**0.65**	99.71	**0.77**	**99.66**	0.51	**99.96**	1.80	**99.20**	**4.77**	**98.19**	**1.58**	**99.38**

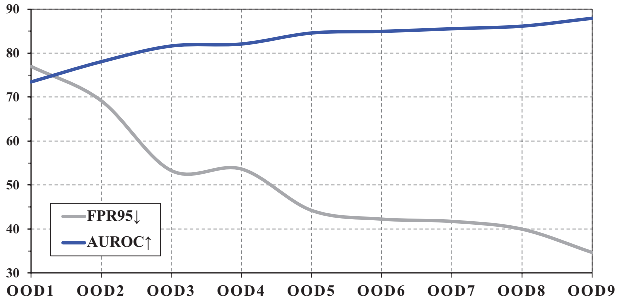

Figure 3. HVCM performance comparison as the increasing distances between InD and OOD data.

recognition. It is important to note that our method only utilizes features from the last layer of the network, while Maha [29] leverages features from both intermediate and deep network layers. However, although our results are worse than Maha [29], we are better than all the other methods. This still indicates that the proposed method is very robust in identifying different types of outliers. Furthermore, we build a cosine classifier with the learned attribute centers for image classification. We got 88.28% accuracy which is 2.57% higher than our supervised learning baseline. This is solid evidence to show that our proposed method can model the InD data accurately and ensure the learned features keep high discriminative ability simultaneously.

Standard evaluation on CIFAR10. We performed a more conventional OOD detection task on CIFAR10. This experiment aims twofold: firstly, to demonstrate that our method is not dependent on self-distillation (DINO [6]), and secondly, to validate the effectiveness and robustness of our HVCM. Table 3 compares our method with several classic and top-performing algorithms. All comparison methods use ResNet-18 as the main backbone network and are trained with cross-entropy loss, while our method uses only the loss function in Eq. (3). As shown in Table 3, the proposed method outperforms the previous best methods Gram [42] and pNML [2] on both average FPR95 and AUROC obviously. These results demonstrate the proposed method can perform well even on small datasets. Meanwhile, there is no self-distillation exploited, which indicates our joint representation learning and statistical modeling is independent of self-supervised learning algorithms [6].

Table 4. The performance of HVCM with varying numbers of group centers G. Results are averaged across four standard OOD datasets, consistent with the main results.

Concepts Number	**FPR95↓**	**AUROC↑**
G=8	24.37	91.73
G=16	23.07	92.24
G=**32**	**21.99**	**92.73**

Results on more challenging OOD datasets. To overcome the limitations of current OOD benchmarks [53] and evaluate the robustness of our approach against adversarial attacks, we conducted experiments on two challenging datasets, namely OpenImage-O [26] and ImageNet-O [22]. As shown in Table 2, HVCM achieves the highest AUROC and lowest FPR95 among all methods on the OpenImage-O dataset. Although ImageNet-O contains adversarial examples and is more challenging, HVCM still outperforms other methods on this dataset.

Results on near-to-far OOD datasets. To investigate the ability of the proposed method to detect near OOD samples, we construct 9 OOD test sets with the remaining ImageNet images. We rank the semantic distance between the remaining 900 visual categories with the 100 categories in the InD dataset. We use the average cosine distance as the measure and construct 9 different OOD test sets. Details are introduced in the supplementary material. For convenience, we denote these datasets from OOD 1 to OOD 9. The experimental results are depicted in Figure 3, and several samples are displayed in Figure 4. We can find that the proposed method achieves good AUROC even when the test set is very close to the InD dataset. When the test sets become farther, FPR95 decreases quickly, which indicates the proposed method is very sensitive to the semantic distances of OOD datasets.

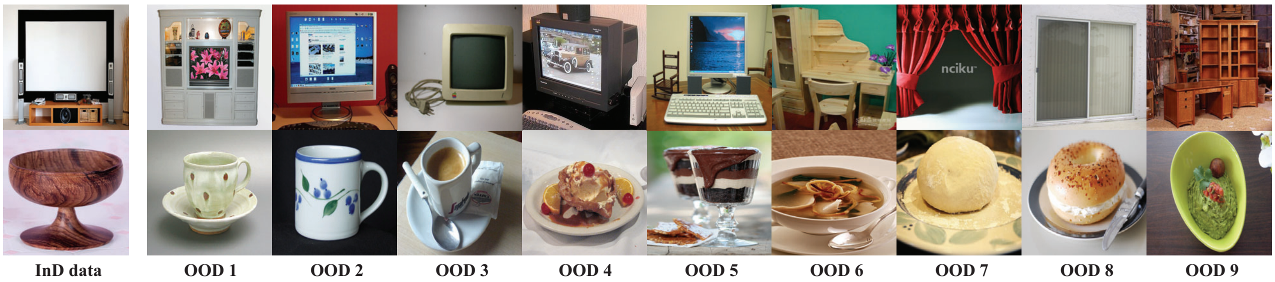

Figure 4. Each dataset exhibits displayed two categories of images. The leftmost samples belong to the InD dataset, while the categories on the right correspond to nine OOD datasets arranged in ascending order of distance. It can be observed that the gap between the OOD samples and InD samples gradually widens as the distance increases.

Table 5. A set of ablation results about HVCM. The top row investigates the effect of using MSE, KL divergence, or JS divergence on the model performance; the bottle row compares the performance of different OOD detection methods. Results are averaged across four standard OOD datasets following the main results.

Strategy Ablation		FPR95↓	AUROC↑
Learning Objective	L2	23.59	92.26
	KL	22.99	92.58
	JS	**21.99**	**92.73**
InD Distance	Cosine	66.37	84.06
	Linear	36.77	86.59
	Maha	**21.99**	**92.73**

4.3. Ablation Study

Number of attribute groups. We varied the number of attribute groups from 8 to 32 to analyze the components in Gaussian mixture models. In Table 4, we find a positive correlation between the number of attribute groups and model performance, with the best performance achieved when G is set to 32. We also try to set G to bigger numbers and list the results in the supplementary material. However, more attribute groups will lead to bigger correlation matrices when computing the InD score. Thus, we finally set G to 32 to balance performance and inference speed.

Choice of learning objectives. The first row of Table 5 investigates how the choice of the learning objective influence the performance of HVCM. We test three learning objectives, including the L2 loss, JS divergence loss, and KL divergence loss. The results show that compared with the L2 loss and KL divergence loss, the JS divergence loss achieves the lowest FPR95, demonstrating its superiority for statistical modeling. We attribute this to the fact that the group centers and features need to learn from each other, and the JS divergence loss is symmetrical in enclosing them.

Different InD distance metrics. In Table 5, we also compared our Maha metric with two different InD distance metrics. The cosine distance metric directly measures the distance by calculating the cosine similarity between the input feature and the mean of Gaussian distribution models. The linear distance metric is used to calculate the distance with trainable linear layers. The results show that our Maha metric is an effective metric compared with its counterparts. We attribute this to the Maha distance space can better fit the training distribution in realistic scenes.

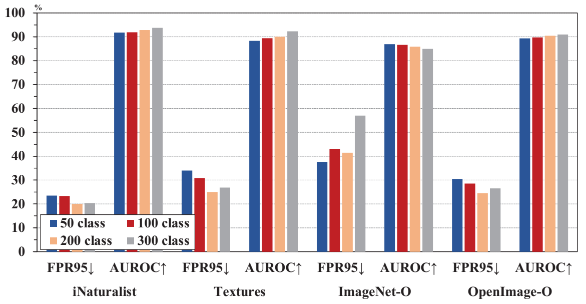

Figure 5. The performance of HVCM is evaluated as the number of InD classes increases across four OOD datasets.

Increasing numbers of InD Categories in ImageNet. We test how the OOD detection performance varies with the increasing of object categories in the in-distribution dataset. Following Wang et al. [53], we test on four popular benchmarks and set C to $\{50, 100, 200, 300\}$ respectively. As depicted in Figure 5, the performance of HVCM fluctuates on different datasets with the increasing of the InD object categories, which suggests that the number of categories has little impact on our approach. These results validate our assumption that we only need to model in-distribution image categories, and the out-of-distribution samples can be detected easily then.

Different thresholds for OOD detection. Figure 6 illustrates the accuracy of OOD detection for our method across various datasets. On most datasets, our method exhibits the same trend for accuracy variation with thresholds. This indicates that our approach has strong generalization and ideal performance to domains with significant differences. We explain the performance of Imagenet-O as its task difficulty with adversarial samples.

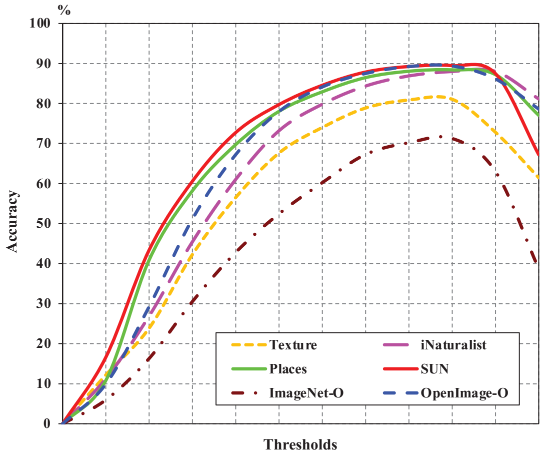

Figure 6. The accuracy of HVCM varies with different thresholds across all OOD datasets.

5. Conclusion

In this paper, we introduce a hierarchical visual category modeling scheme for out-of-distribution detection, which combines visual representation learning and parameter optimization of probability models. It provides a novel perspective for OOD detection by conducting representation learning and density estimation end-to-end. By modeling visual categories with mixtures of Gaussian models, we describe visual categories in very complex distribution and don't rely on outlier training data to perform OOD detection. Experiments demonstrate that the proposed method outperforms state-of-the-art algorithms clearly and does not hinder the discriminative ability of deep features.

Limitations. However, our method needs to map deep features into high-dimensional attribute spaces and build plentiful Gaussian mixture models. These Gaussian mixture models bring a lot of computational costs and make the inference of the OOD detector become inefficient. Thus, simplifying the probability models and accelerating the inference process will be the future direction.

6. Acknowledgment

This work was supported by National Natural Science Foundation of China (No.62072112), Scientific and Technological innovation action plan of Shanghai Science and Technology Committee (No.22511102202), National Key R&D Program of China (2020AAA0108301), National Natural Science Foundation of China under Grant (No. 62106051), Shanghai Pujiang Program (No. 21PJ1400600), and National Key R&D Program of China (2022YFC3601405).

References

[1] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015.

[2] Koby Bibas, Meir Feder, and Tal Hassner. Single layer predictive normalized maximum likelihood for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:1179–1191, 2021.

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[4] Daniel Bogdoll, Jasmin Breitenstein, Florian Heidecker, Maarten Bieshaar, Bernhard Sick, Tim Fingscheidt, and Marius Zöllner. Description of corner cases in automated driving: Goals and challenges. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1023–1028, 2021.

[5] Senqi Cao and Zhongfei Zhang. Deep hybrid models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2022.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[7] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020.

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[9] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[12] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In *International Conference on Machine Learning*, pages 3122–3132. PMLR, 2021.

[13] Aditya Grover, Manik Dhar, and Stefano Ermon. Flow-gan: Combining maximum likelihood and adversarial learning in generative models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[14] Xiaoyuan Guo, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. Cvad: A generic

medical anomaly detector based on cascade vae. *arXiv preprint arXiv:2110.15811*, 2021.

[15] Jiaming Han, Yuqiang Ren, Jian Ding, Xingjia Pan, Ke Yan, and Gui-Song Xia. Expanding low-density latent regions for open-set object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9591–9600, 2022.

[16] Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A statistical framework for efficient out of distribution detection in deep neural networks. In *International Conference on Learning Representations*, 2021.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[18] 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.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[20] Yangji He, Weihan Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, and Wenqiang Zhang. Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9119–9129, 2022.

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

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[23] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017.

[27] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[28] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[29] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[30] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[31] Qiao Liu, Jiaze Xu, Rui Jiang, and Wing Hung Wong. Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, 118(15):e2101344118, 2021.

[32] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[34] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[35] Philipp Oberdiek, Gernot A Fink, and Matthias Rottmann. Uqgan: A unified model for uncertainty quantification of deep classifiers trained via conditional gans. *arXiv preprint arXiv:2201.13279*, 2022.

[36] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[38] Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, and Stephan Mandt. Latent outlier exposure for anomaly detection with contaminated data. *arXiv preprint arXiv:2202.08088*, 2022.

[39] Carl Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12, 1999.

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[41] Murat Sensoy, Lance Kaplan, Federico Cerutti, and Maryam Saleki. Uncertainty-aware deep classifiers using generative

models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5620–5627, 2020.

[42] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *arXiv e-prints*, pages arXiv–1912, 2019.

[43] Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V. Dylov. Anomaly detection in medical imaging with deep perceptual autoencoders. *IEEE Access*, 9:118571–118583, 2021.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[45] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021.

[46] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022.

[47] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[48] Shengjin Tang, Chuanqiang Yu, Xiaoyan Sun, Hongdong Fan, and Xiaosheng Si. A note on parameters estimation for nonlinear wiener processes with measurement errors. *IEEE Access*, 7:176756–176766, 2019.

[49] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[50] 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.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[52] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore 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)*, pages 550–564, 2018.

[53] 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*, 2022.

[54] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022.

[55] Qi Wei, Yinhao Ren, Rui Hou, Bibo Shi, Joseph Y. Lo, and Lawrence Carin. Anomaly detection for medical images based on a one-class classification. In Nicholas Petrick and Kensaku Mori, editors, *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 375 – 380. International Society for Optics and Photonics, SPIE, 2018.

[56] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[57] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[58] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational autoencoder. *Advances in neural information processing systems*, 33:20685–20696, 2020.

[59] 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.

[60] 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.

[61] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

[62] Cangqi Zhou, Hao Ban, Jing Zhang, Qianmu Li, and Yinghua Zhang. Gaussian mixture variational autoencoder for semi-supervised topic modeling. *IEEE Access*, 8:106843–106854, 2020.