SOUND: Unknown Object Detection for Autonomous Driving

Donger Chen[1], Xu Ma[2], Ying He[1], Wang Feng[1], Qing Yang[1], Song Fu[1], Ted Tsao[3]

[1] Department of Computer Science and Engineering, University of North Texas, Denton, TX
[2] Department of Electrical and Computer Engineering, Northeastern University, Boston, MA
[3] STT WebOS, Inc., Fremont, CA
DongerChen@my.unt.edu, ma.xu1@northeastern.edu, {YingHe, WangFeng}@my.unt.edu, {Qing.Yang, Song.Fu}@unt.edu, ted.tsao@sttwebos.com

***Abstract*—Existing perception algorithms for autonomous vehicles are currently limited to recognizing only the objects encountered during their training. As a result, these systems often fail to identify untrained objects encountered in real-world driving scenarios, increasing the risk of accidents. In this paper, we introduce a simple yet efficient method for detecting unknown objects in autonomous driving applications. Our approach, named SOUND, effectively recognizes samples from known classes while rejecting those from previously unseen classes. This capability is critical for autonomous vehicles to navigate complex and unpredictable environments. At the core of SOUND is a novel metric learning framework, projection softmax (p-softmax), which offers a clear geometric interpretation of the evaluation metric and simplifies training compared to existing approaches. To connect the recognition of unknown samples and the classification of known samples, we further extend the feature norm to a smooth maximum projection, bridging the metric learning and energy-based model gap. In addition, we propose a fine-tuning method that enhances feature discrimination between known and unknown classes, further improving the performance. Our experimental evaluation shows that SOUND outperforms existing models and methods by a clear margin. Experimental results on the MNIST, CIFAR, and ImageNet datasets demonstrate its effectiveness and robustness. SOUND offers substantial promise for enhancing the safety and reliability of autonomous driving systems by enabling them to generalize beyond their training data and adapt to new objects and situations encountered on the road.**

***Index Terms*—unknown object detection, metric learning, energy-based model**

I. Introduction

Deep neural networks have shown significant advances in various visual tasks [1]–[3]. However, applying these deep learning algorithms and models to real-world applications remains a challenge. A primary issue is the closed-set assumption, which posits that all training and testing classes originate from the same labeled space. This assumption does not hold in the real world, leading to potentially unexpected results when applying closed-set models. For example, recognition methods may misclassify unknown samples as known classes, while detection methods might ignore them altogether [4]. This discrepancy can be hazardous, particularly in autonomous driving.

Autonomous vehicles, despite being trained on extensive datasets such as KITTI [5], V2V4Real [6], and RCooper [7], are vulnerable to this limitation. Traditional closed-set algorithms in these vehicles may fail to recognize uncommon but potentially dangerous obstacles, such as irregularly shaped tree branches or debris on the road. This failure can result in unknown objects being overlooked by perception models, significantly increasing the risk of accidents.

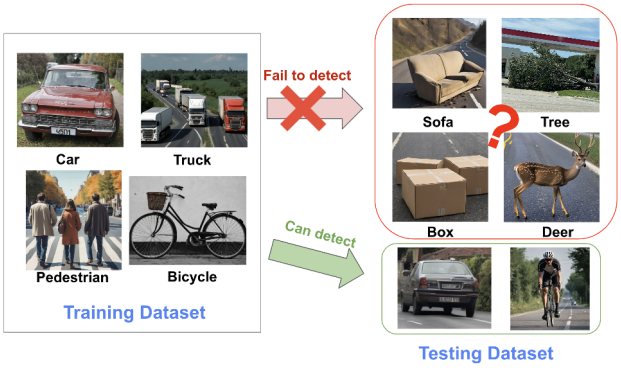


Fig. 1. The distinction between known set and unknown set recognition. Given the same training set, known set recognition is required to recognize the samples from classes in the testing set. In contrast, the unknown set recognition model needs to simultaneously identify samples from known classes while rejecting those belonging to unknown classes.

The real-world driving environment is inherently complex and unpredictable, often presenting objects not covered by training datasets. These datasets typically focus on frequently encountered classes, such as vehicles, pedestrians, and bicycles. When an autonomous vehicle encounters an untrained object or situation, its ability to accurately detect and respond is severely compromised. This recognition gap can result in misinterpreting surroundings, potentially causing incorrect decisions or actions. In the worst-case scenario, this could lead to collisions or other dangerous incidents.

Training models on a predefined set of classes, without accounting for the real world's diversity, exposes a critical vulnerability in the current development of autonomous driving systems. This underscores the urgent need for innovative solutions that allow these systems to generalize beyond their training and adapt to new objects or situations that they will inevitably encounter.

Unknown set recognition aims to relax the closed-set assumption and allow the existence of testing inputs not belonging to any of the training classes. Fig. 1 illustrates this concept.

Whereas classical object detection models handle localization of known objects, here we focus on classifying an object as known vs. unknown. In this paper, we define an "unknown object" as any category that was not included in the training set. The primary challenge centers on identifying the unknown samples under the condition of incomplete knowledge about the open world, while maintaining the classification performance for known classes.

Recent studies on unknown recognition mainly focus on **1)** determining a threshold based on the softmax probability score, *e.g.*, OpenMax [8] and OpenHybird [9], **2)** generating synthetic input images using Generative Adversarial Networks (GANs) to compress the classification boundaries, *e.g.*, G-OpenMax [10] and OpenGAN [11], and **3)** minimizing the reconstruction error between inputs and reconstructed outputs, *e.g.*, C2AE [12] and CROSR [13]. In addition, recent works have provided customized frameworks. For example, OpenSlot [14] leverages object-centric learning and an anti-noise-slot (ANS) technique to tackle mixed scenarios with co-occurring in known and unknown classes. Cas-DC [15] separates unknown detection and known classification to cascading functions. An alternative approach involves employing a contrastive learning-based method that incorporates an unknown score derived from multi-layer features [16].

While the preceding methods addressed certain issues, they have notable limitations. Specifically, in these threshold-based methods, the softmax layer brings challenges to unknown set recognition due to the bounded and closed nature [17], which is not suitable for unknown world scenarios. Similarly, generation model and reconstruction error-based models face two main constraints. First, the extra generation model or added flow-based model doubles the computational overhead during training and testing, which significantly increases the inference time and requires additional computational resources; Second, recent study such as [18] has indicated that both flow-based models and generation models cannot consistently distinguish unseen samples, which compromises the stability of those methods for unknown set recognition.

Considering these limitations, we ask the following question: *Can we develop a simple and stable unknown set model with a minimal amount of manual setting?*

In this paper, we propose SOUND, an effective unknown recognition framework designed for real-world safety-critical tasks like autonomous driving. Starting with softmax, we delve deep into metric learning. Previous metric learning methods [19]–[21] focus on discrimination ability in different tasks. We experimentally show that this ability may not be the desired trait in unknown object detection (see Fig. 3). Our key observation is that the feature norms of known samples typically exceed those of unknown samples. This observation is verified in recent advances [22], [23] and motivates our further study of the unknown world scenarios. We extend the feature norm to the projection along the direction of each class and propose a novel metric learning method, named *projection softmax* (p-softmax in short). P-softmax projects the learned features along the direction of each class, and these projections can be used for both known class recognition and unknown class rejection.

We then reformulate p-softmax in an energy-based model form and leverage the logarithm to disentangle the criteria for both known recognition and unknown rejection. Notably, unknown rejection is achieved through a $\mathrm{LogSumExp}$ function, which aggregates information from all projections and performs a smooth maximum conversion. Mathematically, the smooth maximum conversion of projections is both deterministic and differentiable, enabling fine-tuning to enhance discrimination between known and unknown classes. We demonstrate this intuition through a multi-task loss function that considers both classification accuracy and sensitivity. To train this fine-tuning loss effectively, we introduce a simple yet effective solution for handling negative unknown data. Unlike methods that employ GANs [10], [11] or real auxiliary data [23], [24], we generate negative samples from the training data using the mixup method [25]. A more detailed analysis of negative data generation using data augmentation methods can be found in Section II-C1 and [26]. With synthetic negative data, we explore more discriminative boundaries to push the known classes apart from the potential unknown space.

To demonstrate the effectiveness of SOUND, we have conducted experiments using multiple datasets including MNIST, CIFAR, and ImageNet. SOUND has consistently achieved promising performance and outperformed other approaches by a clear margin. We summarize the main contributions as follows.

- We propose a novel metric learning method, p-softmax, that provides a clear geometric interpretation and simplifies the training process. P-softmax serves as a prerequisite for our SOUND framework for unknown object detection.
- We connect metric learning and energy-based model and develop a new criterion for unknown rejection. Furthermore, it makes fine-tuning efficient.
- We present a fine-tuning loss that learns to push known samples apart by negative data generation and maintains the prominent known set recognition ability.
- Extensive experiments verify the effectiveness and robustness of our proposed SOUND.

II. SOUND Framework for Unknown Object Detection in Autonomous Driving

In this section, we describe the design of our SOUND framework which is shown in Fig. 2. First, we introduce our novel projection softmax (p-softmax) and how it relates to energy-based models. To gather information from all class projections, we employ a smooth maximum conversion, which integrates with the energy-based model and improves performance and robustness. Compared to traditional closed-set recognition models, our approach requires only minimal modifications: replacing the loss function with p-softmax during training and applying the projection-based smooth maximum metric to identify unknown samples. Additionally, we develop

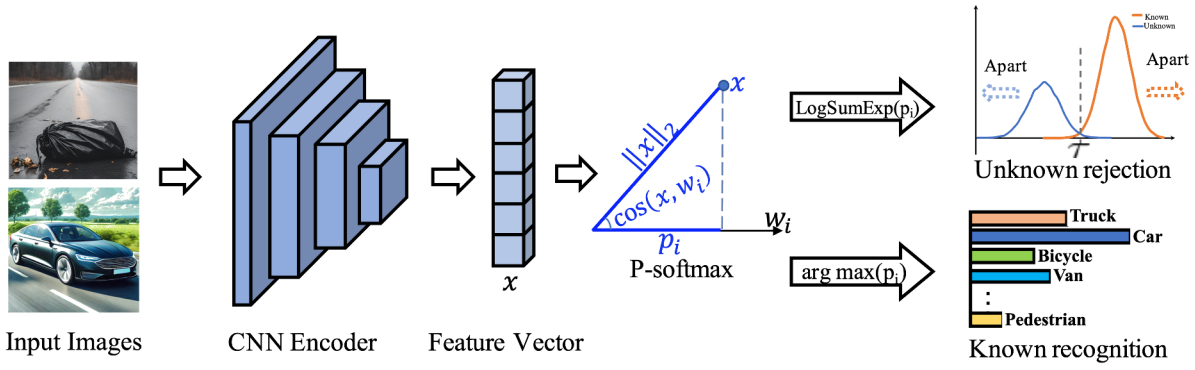


Fig. 2. The structure of **SOUND**. We employ a novel **p-softmax** loss to optimize the network and produce projections p on all known classes. Using these projections, we compute the negative energy of input samples through a smooth maximum function to facilitate unknown rejection. If the value is smaller than a threshold τ, the input is rejected as unknown. Otherwise, it is classified as belonging to one of the known classes based on the maximum projection value. See Section II for the details.

a fine-tuning method that further improves the discrimination between known and unknown samples.

A. Revisiting Metric Learning for Projection-Softmax

We begin by revisiting and analyzing several softmax-based metric learning methods, followed by introducing the proposed projection-softmax. This method offers a clear geometric interpretation and serves as the foundation for our SOUND framework.

In a conventional recognition model, the softmax posterior probability of a learned feature vector x, considering the ith class y_i, can be presented as

$$P\left(y_i|x\right) = \frac{e^{w_i^\intercal x + b_i}}{\sum_{j=1}^{c} w_j^\intercal x + b_j}, \tag{1}$$

where w_i can be considered as a proxy of the ith class. In general, $\{w_1, w_2, \cdots, w_c\}$ indicates the weights in the fully-connected (FC) layer and $\{b_1, b_2, \cdots, b_c\}$ are the biases. While softmax serves as the foundation of recognition tasks, it lacks the ability to provide strong discriminative power. To address this limitation, researchers have developed deep metric learning techniques. One notable approach is center loss [27], which simultaneously learns a center for each class and minimizes the distance between deep features and their respective class centers. The center loss (combined with softmax loss) is expressed as

$$\mathcal{L} = -\log \frac{e^{w_i^\intercal x}}{\sum_{j=1}^{c} e^{w_j^\intercal x}} + \lambda \left\| x - w_i \right\|_2^2, \tag{2}$$

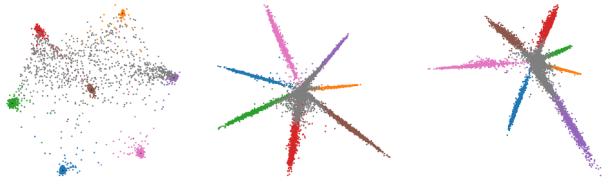
where λ balances the two terms, and the bias is omitted for clarity. Under joint supervision, center loss offers promising discriminative capability while maintaining comparable classification performance. However, it focuses on improving intra-class compactness while neglecting inter-class dispersion.

Unlike the Euclidean distance applied in center loss, more recent efforts have been made towards the angular-based softmax that uses a cosine similarity as a metric due to the inherent bounded result. For example, the loss function of NormFace is formulated as

$$\mathcal{L} = -\log \frac{e^{s*cos(w_i, x)}}{\sum_{j=1}^{c} e^{s*cos(w_j, x)}}, \tag{3}$$

where s is a scaling factor that eases the training. In the context of unknown faces or objects, we can easily reject it as unknown if the cosine similarities between x and all the class proxies $\{w_1, w_2, \cdots, w_c\}$ are smaller than a threshold τ. Similarly, SphereFace [28], AM-softmax [19], CosFace [29], and ArcFace [21] also inherit the merits of cosine similarity, and achieve performance improvements.

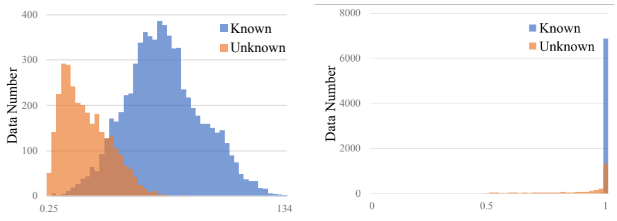
While these metric learning methods have shown promising results in open-world recognition tasks, their effectiveness in unknown set recognition for autonomous driving remains uncertain. To explore this, we designed an illustrative experiment using the MNIST dataset. We train LeNets++ [27] with different loss functions on the first 7 classes (i.e., numbers 0 to 6), and keep the rest as unknown. The results are depicted in Fig. 3.



(a) Center Loss (b) NormFace (c) ArcFace

Fig. 3. Visualization results of learned feature representations in a 2D feature space. Colored points represent the trained classes, and gray points denote unknown classes.

We observe clear discrimination among the known data; however, the unknown data points are unevenly distributed in the feature space. Notably, a significant portion of the unknown data points falls into the known clusters or beams. We refer to this phenomenon as feature collapse, where features in



(a) Feature Norm $\|x\|_2$ (b) Softmax Probability

Fig. 4. Comparison of feature norm and softmax results.

the feature space tend to collapse into conditional regions. This phenomenon is also evident across other datasets and metric learning methods, indicating that feature collapse is not limited to a specific setting. As a result, rejecting the unknown data based on L2 distances or cosine similarities would be defective. Furthermore, we find that overconfident posterior distributions exacerbate feature collapse, as the softmax function creates a label-overfitted output space, leading to recognition failures for unknown inputs [23], [24], [30].

Moreover, two critical issues limit the application of the preceding loss functions. First, the introduced hyper-parameters λ and s require manual tuning, and improper settings can result in significant performance degradation or even training failure. Second, both equations 2 and 3, along with related angular-based loss functions, are challenging to optimize. These limitations highlight the urgent need for a robust metric learning method specifically designed for unknown set recognition tasks.

Inspired by recent advances [22], [31], we leverage the feature norm $\|x\|_2$ for our rejection of unknown samples. Inherently, the feature norms of known samples are typically larger than those of unknown samples (see analysis below), reflecting higher confidence in the recognition. [31]. As shown in Fig. 4, the feature norm $\|x\|_2$ makes it easy to distinguish the distributions of known and unknown data, while softmax always mixes them. Based on the feature norm, we reformulate the softmax-based loss function as follows:

$$\mathcal{L} = -\log \frac{e^{\|x\|_2 * cos(w_i,x)}}{\sum_{j=1}^{c} e^{\|x\|_2 * cos(w_j,x)}} \tag{4}$$

Let $p_i = \|x\|_2 * cos(w_i, x)$, where w_i indicates the proxy of the ith class. Geometrically, p_i denotes the projection of x on the ith class, as shown in Fig. 5. As a result, we name (4) as *projection softmax (p-softmax)*, which can be considered as a modified softmax that is based on the projections p. We emphasize that both terms, feature norm $\|x\|_2$ and cosine similarity $\cos(w_i, x)$, play important roles in the loss function. The feature norm $\|x\|_2$ scales the certainty of x. A larger feature norm indicates more certainty and vice versa. This also explains the phenomenon shown in Fig. 4(a). The cosine similarity $\cos(w_i, x)$ determines the classification results under the closed-set assumption [20]. Compared to NormFace, as shown in (3), p-softmax does not rely on cosine similarity for intra-class compactness. Instead, it incorporates the feature norm to introduce the property of certainty. Moreover, replacing the fixed scalar s with adaptively learned feature norm $\|x\|_2$ significantly facilitates the training process. Unlike the vanilla softmax formulation, p-softmax eliminates the perturbation caused by $\|w_i\|_2$, allowing it to focus more directly on the projection and, consequently, the certainty. Although p-softmax may not significantly enhance traditional classification tasks, it provides a foundational component for our SOUND framework due to its projection properties.

B. From Metric Learning to Energy-Based Model

Using p-softmax, we can reject an input as unknown if the maximum projection $\max(p_i)$ is below a threshold τ. However, this approach has two limitations: **1)** It retains only the maximum value, leaving other valuable information from the feature norm and classification unused; **2)** the maximum function makes further fine-tuning difficult. In order to address the information restriction and enable optimization for fine-tuning, we connect (4) to an energy-based model and design a smooth maximum conversion for unknown rejection.

Building on the loss function in (4), we reformulate the probability density function $P(y_i|x)$ of p-softmax in an energy-based model formulation as follows

$$P(y_i|x) = \frac{e^{(-E(w_i,x))}}{Z}, \tag{5}$$

where the energy function is $E(w_i, x) = -\|x\|_2 * cos(w_i, x)$, and Z is the normalizing constant with respect to x. Specifically, $Z = \int_x e^{-E(x)} = \sum_{j=1}^{c} e^{\|x\|_2 \cos(x,w_j)}$. The logarithm for a single input can be expressed as:

$$\begin{aligned} \log P(x) &= -E(x) - \log Z \\ &= \|x\|_2 * cos(w_i, x) - \log \sum_{j=1}^{c} e^{\|x\|_2 \cos(x,w_j)} \end{aligned} \tag{6}$$

Equation (6) intuitively separates the criteria for known recognition and unknown rejection. For unknown rejection, we replace the criterion from $\max(p_i)$ with a smooth maximum conversion, $\log \sum_{j=1}^{c} e^{\|x\|_2 \cos(x,w_j)}$, implemented using the built-in function LogSumExp in PyTorch. By doing so, we leverage all the information from the class space. The distinction between known and unknown samples is defined as follows

$$\begin{cases} known & if \ \log \sum_{j=1}^{c} e^{p_i} \leq \tau, \\ unknown & if \ \log \sum_{j=1}^{c} e^{p_i} \geq \tau, \end{cases} \tag{7}$$

where τ is a threshold. For the known object recognition, we still take the operation of $\arg\max(p_i)$.

C. Fine-Tuning to Pull Unknown Objects Apart

SOUND is compatible with any CNN backbone by substituting the traditional softmax with p-softmax. Since p-softmax is initially trained under a closed-set assumption, the network does not directly see unknown examples. To improve separation, we introduce a fine-tuning stage that pushes negative data (mock unknowns) away from known classes.

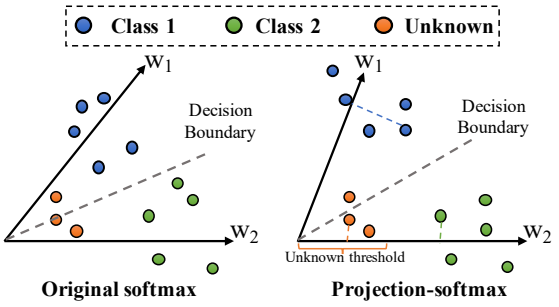


Fig. 5. Unlike the original softmax or angular-based counterparts, our proposed p-softmax is specifically designed for unknown object detection tasks. We demonstrate that $\|x\|_2 * cos(w_i, x)$ represents the projection of the feature norm onto the direction of each class.

1) Negative Data Generation: Some open set approaches generate negative data using generative adversarial networks (GANs) [10], [11] to achieve compact decision boundaries. However, these methods rely on an independent generative model for creating synthetic negative samples, which significantly increases computational cost during training. Additionally, generating effective negative samples remains a challenging task. To address this, we propose a simple and model-free method for negative data generation, specifically designed to fine-tune SOUND.

Unlike GAN-based methods or approaches relying on large-scale auxiliary data, our method generates negative samples using mixup techniques [25], which do not require additional generative models and rely solely on the input training data. While mixup was originally designed for data augmentation and model generalization, we repurpose it for negative data generation, inspired by the concepts proposed in Negative Data Augmentation (NDA) [26]. Effectively, we construct new samples that lie outside known classes but between the class boundaries.

Given a batch of input data $\mathbf{X}$, we first shuffle the samples to create a new ordered batch of samples $\mathbf{X}'$. Next, we construct negative data batch $\tilde{\mathbf{X}}$ by

$$\tilde{\mathbf{X}} = (1-\lambda)\,\mathbf{X} + \lambda \mathbf{X}', \tag{8}$$

where $\lambda \in [0,1] \sim \mathrm{Beta}\,(\alpha, \alpha)$ and $\alpha \in (0, \infty)$. By default, α is set to 1. Importantly, we remove samples from both $\mathbf{X}$ and $\mathbf{X}'$ that have the same labels at the same position to ensure the generated data remain distributed between known classes.

2) Fine-Tuning Loss Function: With the generated negative data $\tilde{X}$, we fine-tune SOUND by pulling $\tilde{X}$ from the original input data X through a loss function, i.e.,

$$\begin{aligned} \mathcal{L}_{open} =& \frac{1}{n}\sum_{i=1}^{n}\left\|\max\left(1-r_i^k, 0\right)\right\|^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left\|\max\left(r_i^u - 1, 0\right)\right\|^2, \\ \text{s.t. } r_i^k =& \frac{\mathrm{LogSumExp}\,(\mathrm{p_i})}{\tau_k} \sim X, \\ r_i^u =& \frac{\mathrm{LogSumExp}\,(\mathrm{p_i})}{\tau_u} \sim \tilde{X}, \end{aligned} \tag{9}$$

Here n represents the number of samples in a batch, and τ_k and τ_u are pre-defined thresholds for known and unknown samples. These thresholds are determined as the middle values of $\mathrm{LogSumExp}\,(p_i)$ over the training set and the generated negative data set, respectively. Two important design considerations are as follows. Firstly, since the $\mathrm{LogSumExp}$ function and the projection length p_i are not bounded, we define r_i^k and r_i^u as ratios to eliminate the impact of amplitude. Secondly, not all samples are involved in the training. Only samples within the range $[r_i^u, r_i^k]$ are included in the training process.

To maintain the closed-set classification performance, we fine-tune SOUND by combining the p-softmax loss with the loss function defined in (9) as follows:

$$\mathcal{L} = \underbrace{\mathcal{L}_{p-softmax}}_{\text{over}X} + \gamma \underbrace{\mathcal{L}_{open}}_{\text{over}X\,\&\,\tilde{X}} \tag{10}$$

Here, γ is a balancing parameter, which is set to 1 by default.

D. Unknown Object Detection

1) Openness: The concept of openness [32] quantifies how open a problem setting is, and is defined as

$$\mathbb{O} = 1 - \sqrt{\frac{2 \times C_{re}}{C_{tr} + C_{te}}}, \tag{11}$$

where C_{re}, C_{tr} and C_{te} denote the classes of recognized, trained, and tested sets, respectively. However, this formulation does not guarantee a positive value for openness. To address this, Geng et al. [33] revised the formulation as

$$\mathbb{O} = 1 - \sqrt{\frac{2 \times C_{tr}}{C_{tr} + C_{te}}}. \tag{12}$$

While this adjustment resolves the positivity issue, neither equation captures the distribution of classes during testing. To address this problem, we introduce a new formulation to quantify openness as

$$\mathbb{O} = \frac{T_{un}}{T_{kn} + T_{un}}, \tag{13}$$

where T_{kn} denotes the number of known classes observed during training, and T_{un} represents the number of unknown classes that are not seen in training. This revised definition has several desirable properties: **1)** it explicitly describes the distributions of known and unknown classes in testing; **2)** it removes dependency on the training set, relaxing the assumption that all training classes must appear during testing; **3)** the value of openness increases smoothly as the number of unknown classes increases.

2) Evaluation Metrics: To evaluate the performance of unknown object detection, we use standard metrics such as the F1-measure and AU-ROC (area under the ROC curve). In addition, we compute a weighted macro-F1 score in our experiments, as suggested in [33]. For all these metrics, higher values indicate better performance.

III. Performance Evaluation

We evaluate our proposed SOUND for unknown object detection by conducting experiments on multiple datasets. For comparison, we select five existing methods including OpenMax [8], OLTR [22], Center Loss [27], Normface [20] , and RPL [34]. All implementations are based on the PyTorch framework [35].

A. Datasets for Unknown Object Detection

Benchmark datasets for evaluating unknown object detection methods are limited. A common approach is to combine datasets from different domains to simulate the open space. However, this is not ideal for unknown set recognition [18]. Following the approach of OLTR [22], which splits a benchmark into separate training and testing sets, we construct unknown object detection datasets by randomly sampling classes from the original datasets. This setup simulates various open scenarios and enables the exploration of relationships between known and unknown classes during testing. We generate our unknown object datasets based on MNIST, CIFAR, and ImageNet. To differentiate these datasets from the originals, we denote them as MNIST-O, CIFAR-O, and ImageNet-O, where the suffix "-O" indicates they are open scenarios for unknown object detection.

B. Experimental results on MNIST and CIFAR

1) MNIST Experiments: To set up an open scenario, we select the first 7 classes (i.e., numbers 0-6) as the known classes and the remaining three classes (i.e., numbers 7-9) as the unknown classes. According to our defined openness equation, the openness is 0.3. We employ LeNet++ [27] as the backbone for the tested methods. Each experiment is repeated five times using different random seeds. For all threshold-based methods, we use the default thresholds provided by each method, determined through a validation set (i.e., unselected data in the original training set). Table I presents the results.

TABLE I
Performance on MNIST-O (openness = 0.3).

Methods	F1	macro-F1	AU-ROC
SoftMax-close	0.694± 0.00	0.583±0.00	0.998±0.00
SoftMax	0.818±0.01	0.800±0.01	0.998±0.00
OpenMax	0.838±0.01	0.826±0.01	0.984±0.00
CenterLoss	0.860±0.02	0.854±0.02	**0.999±0.00**
OLTR	0.784±0.01	0.752±0.02	**0.999±0.00**
RPL	0.860±0.00	0.854±0.00	0.998±0.00
SOUND	**0.896±0.01**	**0.885±0.01**	0.997±0.00

Note: The mean and standard deviation of F1-measure, weighted macro-F1, and AU-ROC achieved by the evaluated methods are compared.

We use "Softmax-close" to represent the results obtained by directly employing closed-set classification output, where no unknown samples are predicted. This serves as a lower boundary for unknown object detection performance, as any method designed for unknown set recognition should outperform this baseline. "SoftMax" denotes the results obtained using softmax with the threshold-based approach. The results in Table I show that our SOUND achieves state-of-the-art performance and outperforms the other methods by a clear margin. Specifically, compared to OpenMax, we achieve an absolute F-1 improvement of 5.8% and a 3.6% improvement over CenterLoss.

2) CIFAR Experiments: We have also conducted experiments on the CIFAR10 dataset. To create CIFAR10-O, we randomly selecte five classes as the known classes for the training set, while all ten classes were included in the testing set, resulting in five unknown classes during testing. The openness of CIFAR10-O is 0.5. For all of the evaluated methods, we employ ResNet18 as the backbone. Each network was trained for 100 epochs. The results from five runs for each method are presented in Table II.

The results show that SOUND consistently outperforms the other methods. SOUND achieves an F1 score of 66% and a macro-F1 score of 65.3%, outperforming the baseline and other methods by margins of 0.011-0.291 (F1) and 0.004-0.399 (macro-F1). We observed that the AU-ROC values for all methods and datasets have reached saturation. Therefore, we focus on the F1 and macro-F1 metrics in the subsequent experiments.

TABLE II
Performance on CIFAR10-O (openness = 0.5).

Methods	F1	macro-F1	AU-ROC
SoftMax-close	0.369±0.01	0.254±0.00	0.976±0.00
SoftMax	0.636±0.00	0.636±0.01	0.976±0.00
OpenMax	0.622±0.00	0.621±0.00	0.917±0.00
CenterLoss	0.649±0.01	0.649±0.01	0.975±0.00
OLTR	0.638±0.00	0.637±0.00	**0.977±0.00**
SOUND	**0.660±0.00**	**0.653±0.00**	**0.977±0.00**

C. Experimental Results on ImageNet

We further test on the large-scale ImageNet [36] dataset. We randomly select 500 classes as known classes in the training set and randomly select the rest 100, 200, ..., and 500 classes as unknown classes in the testing sets. Note that all known classes are also included in the testing. Consequently, the openness of ImageNet-O varies from 0.167 (100 unknown classes) to 0.5 (500 unknown classes). For all methods, we use ResNet18 as the backbone. We employ probabilistic methods for comparison.

As shown in Table III, our SOUND framework achieves substantial improvements on the ImageNet dataset. Specifically, it achieves a 2.02% gain in F1-measure for 100 unknown classes, and a 6.69% gain on 500 unknown classes. Probabilistic methods perform comparably when openness is low (e.g., 0.167 for 100 unknown classes). However, as openness increases, the performance of these methods drops significantly. In contrast, SOUND exhibits relatively stable performance as openness increases. This trend is consistent across other datasets, as shown in Fig. 6(a).

TABLE III
UNKNOWN OBJECT DETECTION PERFORMANCE ON THE IMAGENET-O DATASET (THE NUMBER OF KNOWN CLASSES IS 500).

	Unkown:100		Unkown:200		Unkown:300		Unkown:400		Unkown:500	
	F1	macro-F1	F1	macro-F1	F1	macro-F1	F1	macro-F1	F1	macro-F1
SoftMax-close	0.6243	0.6687	0.5354	0.5723	0.4683	0.4990	0.4163	0.4405	0.4163	0.4403
SoftMax	0.6261	**0.6731**	0.5389	0.5800	0.4730	0.5089	0.4219	0.4523	0.4217	0.4515
OpenMax	0.5846	0.6014	0.5302	0.5877	0.4972	0.5327	0.4512	0.5214	0.4497	**0.5179**
SOUND	**0.6463**	0.6485	**0.6229**	**0.6268**	**0.5407**	**0.5751**	**0.5208**	**0.5480**	**0.5149**	0.5157

D. Ablation Studies

1) Comparison with Metric Learning Methods: Our SOUND framework is designed to address the limitations of existing metric learning methods. To evaluate its effectiveness, we compare SOUND against several metric learning approaches across multiple datasets. Unlike previous settings, in the ablation studies, we determine the optimal thresholds for each method using grid search. This ensures that the presented performance reflects the best possible results for each method. We also explore various types of thresholds for each method, including probability-based and cosine similarity-based thresholds. For our proposed p-softmax, we evaluated additional threshold types: **1)** Feature norm-based threshold: Defined using the norm $\|x\|_2$. **2)** Energy-based threshold: Defined using $\mathrm{LogSumExp}\,(p_i)$. Additionally, we investigate the impact of fine-tuning, as described in (10), on improving performance.

Table IV presents the results. We measure the mean and standard deviation of the last five epochs in each training script. The baseline, i.e., softmax with the threshold method, starts from 86.41% F1 for MNIST-O and 74.31% for CIFAR10-O. While all methods demonstrate promising results, SOUND consistently outperforms the others by a significant margin, specifically 95.82% (ours) vs. 93.36% (ArcFace) and 79.68% (ours) vs. 75.32% (CenterLoss). With p-softmax, we observe that the norm-based threshold consistently outperforms the probability-based threshold, and is slightly inferior to the energy-based threshold. Fine-tuning with (10) further enhances p-softmax performance, leading to substantial gains. With our fine-tuning method, even the best-performing energy-based p-softmax can be further improved by 1.35% on MNIST-O and 1.08% on CIFAR10-O.

2) Openness vs. Performance: We further investigate the impact of openness on performance by conducting experiments on the CIFAR10-O dataset. The number of known classes is fixed at 5, while the number of unknown classes is varied from 0 to 5. We compare the performance of softmax and our p-softmax with different metrics, including classification probability, feature norm, and the $\mathrm{LogSumExp}$ energy metric.

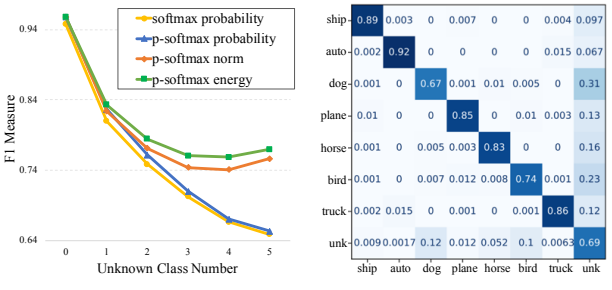
Fig. 6(a) illustrates the results. By modifying softmax to p-softmax, we achieve a slight but consistent improvement in the F1 measure across varying numbers of unknown classes. Specifically, p-softmax outperforms softmax by a margin ranging from 0.5% (unknown classes = 4) to 1.6% (unknown classes = 1). Additionally, we observed that norm-based and energy-based metrics are more robust in high-openness scenarios. For example, with 2 unknown classes, the p-softmax energy-based metric achieves an F1 measure that is 2.4% higher than the p-softmax probability-based metric. This gap widens to 11.6% when the number of unknown classes increases to 5.

3) Confusion matrix analysis.: To better understand the performance of SOUND, we present the confusion matrix for the CIFAR10-O dataset, where 7 classes are randomly

TABLE IV
COMPARE P-SOFTMAX WITH OTHER LOSSES ON THE MNIST-O AND CIFAR10-O DATASETS.

		MNIST-O	CIFAR10-O
Metric	threshold	F1(%)	F1(%)
softmax	probability	86.41±0.01	74.31±0.03
CenterLoss	probability	89.15±0.02	75.32±0.04
NormFace	probability	69.76±0.01	67.08±0.02
	cosine	92.86±0.03	71.17±0.04
ArcFace	probability	70.44±0.02	67.15±0.02
	cosine	93.36±0.08	68.59±1.12
p-softmax	probability	85.09±0.01	75.93±0.04
	feature norm	93.65±0.00	76.26±0.02
	energy	94.47±0.00	78.60±0.02
	fine-tune	**95.82±0.08**	**79.68±0.03**

Note: 7 classes are included in training, and the rest is set as unknown in testing.



(a) Openness vs. F1 (b) Confusion Matrix

Fig. 6. **Left**: Test results of F1 under different numbers of unknown classes. **Right**: Confusion matrix of SOUND. The x-axis indicates the prediction and the y-axis indicates the ground truth.

selected as known and the remaining 3 classes are treated as unknown. The pre-trained model is taken from Table IV. From Fig. 6(b), we observe that SOUND effectively recognizes unknown samples and achieves an appealing result on the unknown samples which close to the known set, despite the accuracy of 69% on the unknown set is somewhat lower than the mean accuracy of 82% on the known set. An important observation is that misclassifications among known classes are rare. Most errors stem from misclassifying unknown samples as belonging to one of the known classes. This finding suggests that SOUND can be further enhanced by addressing subtle challenges specific to unknown set recognition tasks.

4) Hyper-parameter γ: We have also investigated the influence of the hyper-parameter γ in (10) on fine-tuning. We compare different values of γ and present the mean values of five runs in Table V. The results indicate that SOUND performes better when γ is within a small range (e.g. 0.01 to 5), with the highest performance observed at $\gamma = 1$ (79.68% F1). However, increasing the value of γ further (e.g. 10 to 100) leads to a decline in performance, with results falling below the baseline.

TABLE V
FINE-TUNING USING DIFFERENT VALUES OF γ IN (10).

γ	0.01	0.1	1.0	3.0	5.0	10.0	50	100
F1	78.6	79.2	**79.7**	79.6	78.6	77.9	77.1	71.5

Note: The pre-tained model is from Table 3 (i.e., p-softmax, energy, CIFAR10-O), baseline is 78.60% without fine-tuning.

IV. RELATED WORKS

A. Object Detection

We have witnessed great advancements in the field of known-set object detection, with models like AlexNet [37], ResNet [1], DenseNet [38], and the YOLO series, spanning from YOLOv1 [39] to the latest YOLOv8 [40]. In the field of autonomous vehicles, perception is the foundation for other downstream tasks such as tracking, prediction, planning, and control. Moreover, object detection has a wide range of applications, including pedestrian detection [41]–[43] congestion detection [44], detection under complex road or weather conditions [45]. A common assumption in this field is that data from the testing and training phases belong to the same closed set.

However, such a strong assumption may not be held in the real world. Open set recognition aims to relax this assumption. Prior to deep learning, a naive approach is based on the support vector machine (SVM) [32], [46]. In [47], two heterogeneous SVMs were employed: one was designed to reject unknown samples using the extreme value theory (EVT), and the other was to classify known classes. Based on EVT, Rudd *et al.* proposed an extreme value machine [48] which provides a well-grounded theoretical interpretation for both online incremental learning and offline open set decisions. Additionally, methods based on sparse representation [49] and nearest neighbors [50] have also been studied.

Recently, deep neural networks have been applied to open set recognition due to their promising feature extraction capability. OpenMax [8] is a pioneering example that replaces the traditional softmax layer with a Weibull distribution-based score derived from mean activation vectors. With a pre-defined threshold, OpenMax rejects a sample as unknown when the score is below the threshold. Otherwise, the sample is classified to a known class. G-OpenMax [10] extends OpenMax by using generative adversarial networks (GANs) for unknown image synthesis, while providing explicit probability estimation for unknown classes. Following this direction, works in [18], [51], [52] also leveraged generative models to synthesize unseen images and further compress decision boundaries. In addition, the concept of variational auto-encoder was investigated extensively for open set recognition, such as C2AE [12], CROSR [13], OpenHybrid [9], OpenGAN [53], counterfactual reasoning-based methods [54], and CA-VAE [55].

Our proposed SOUND framework leverages the inherent properties of feature norms to provide a clear geometric interpretation. Unlike generative models that synthesize new images, SOUND constructs negative samples directly from input images for fine-tuning, which introduces almost no extra computation overhead.

B. Deep Metric Learning for Object Detection

Deep metric learning (DML) focuses on learning representations where semantically similar samples are closer together, and dissimilar samples are farther apart in the feature space. Early DML methods, such as the Siamese Network [56], measured the embedding feature distance between pairs of images. In contrast to the early methods which primarily focused on pairwise [57] or triplet relationships [58], recent DML methods [59]–[62] incorporate more efficient training schemes by learning distances among all samples in a training batch.

In the context of object detection, particularly for autonomous driving, DML enables robust feature representations to distinguish between diverse objects such as vehicles, pedestrians, and road signs under varying conditions. Modern loss functions, such as center loss [27], improve discriminative capabilities by clustering features of the same category around a centroid. Extensions like L-softmax [19] enforce larger inter-class margins, while A-softmax [28] enables networks to learn angularly discriminative features on a hyperspherical manifold. These methods not only enhance class separability but also provide a geometric interpretation of feature embeddings, which is crucial for reliable object detection in safety-critical scenarios like autonomous driving. Other applications include anomaly detection, system management, computer vision, and more [63]–[68].

In this paper, we revisit metric learning loss functions for unknown object detection and propose a novel projection softmax loss. By incorporating feature norms and offering a clear geometric interpretation, this approach significantly

enhances object detection performance, particularly for recognizing unknown objects.

C. Energy-Based Models

Energy-based models explore data dependencies by a scalar metric named energy to each configuration of the variables [69]. In recognition tasks, an input sample x is assigned to a variable y such that the corresponding energy is minimized. Specifically, the Energy-based models strive to learn an energy function $E(x, y)$ that maps an input x into a scalar space $\mathbb{R}$, capturing the level of dependency between the (x, y) pairs. In contrast to the probabilistic approaches, e.g. softmax-based classification models, energy-based models do not require proper normalization. The energy scalar is not constrained to a probabilistic space. Energy-based model eliminates the need to estimate a normalization constant as required in probabilistic models. This property provides energy-based models with more flexibility in model design, enabling them to address a broader range of tasks effectively.

Recently, energy-based models have gained significant attention in the machine learning community for applications including generative modeling [70], [71], recognition [72], and out-of-distribution detection [23], [24], [73], [74]. These advancements motivate further exploration of energy-based models for unknown object detection, leveraging their unique properties and flexibility.

V. Conclusion and Future Direction

Perception forms the foundation of downstream tasks in autonomous vehicles. However, existing closed-set learning algorithms often fail to recognize uncommon yet potentially hazardous obstacles on the road, significantly increasing the risk of accidents.This paper presents a simple yet effective method for unknown object detection, named SOUND. SOUND is powered by a novel metric learning approach based on feature norms and reformulated within the framework of energy-based models. In addition, we introduce a fine-tuning method that further enhances SOUND's performance.

For autonomous driving, even a coarse classification of an object as unknown can be highly valuable. Such a label informs the planning and control modules that the detected object does not belong to any of the predefined categories, prompting the system to exercise caution. While SOUND does not assign novel labels to unknown objects, its ability to detect and flag unfamiliar instances serves as a crucial first step in an open-world perception pipeline. In addition, our model focuses on the recognition (classification) task as a foundational module that can be incorporated into a broader open-world detection framework. In a complete perception pipeline, bounding-box proposal mechanisms can first detect potential objects, while SOUND then classifies them as either known or unknown. This modular approach allows SOUND to complement existing detection systems, contributing to a more robust and reliable perception framework for autonomous driving.

Experimental results demonstrate that SOUND effectively recognizes unknown objects, achieving results on unknown objects that are comparable to those on known objects. These findings underline SOUND's potential to improve the safety and reliability of autonomous systems by addressing the critical challenges posed by unknown object detection.

Acknowledgment

This work has been supported in part by U.S. NSF grants CNS-2231519, CNS-2113805, OAC-2017564, CNS-2037982, and DUE-2225229. The authors thank the reviewers for their constructive comments.

References

[1] 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*, pp. 770–778, 2016.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[3] R. Wang, X.-J. Wu, Z. Chen, C. Hu, and J. Kittler, "Spd manifold deep metric learning for image set classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[4] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, IEEE, 2012.

[6] R. Xu, X. Xia, *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[7] R. Hao, S. Fan, *et al.*, "Rcooper: A real-world large-scale dataset for roadside cooperative perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

[9] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," *European Conference on Computer Vision (ECCV)*, 2020.

[10] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," in *British Machine Vision Conference 2017*, British Machine Vision Association and Society for Pattern Recognition, 2017.

[11] L. Ditria, B. J. Meyer, and T. Drummond, "Opengan: Open set generative adversarial networks," *arXiv preprint arXiv:2003.08074*, 2020.

[12] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2307–2316, 2019.

[13] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4016–4025, 2019.

[14] X. Yin, F. Pan, G. An, Y. Huo, Z. Xie, and S.-E. Yoon, "Openslot: Mixed open-set recognition with object-centric learning," *arXiv preprint arXiv:2407.02386*, 2024.

[15] D. Brignac and A. Mahalanobis, "Cascading unknown detection with known classification for open set recognition," *arXiv preprint arXiv:2406.06351*, 2024.

[16] Y. Zhou, S. Fang, S. Li, B. Wang, and S.-Y. Kung, "Contrastive learning based open-set recognition with unknown score," *Knowledge-Based Systems*, vol. 296, p. 111926, 2024.

[17] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, "Learning open set network with discriminative reciprocal points,"

[18] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?," in *International Conference on Learning Representations*, 2018.

[19] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks.," in *International Conference on Machine Learning (ICML)*, vol. 2, p. 7, 2016.

[20] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049, 2017.

[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

[23] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *NeurIPS*, 2020.

[24] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *International Conference on Learning Representations (ICLR)*, 2018.

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.

[26] A. Sinha, K. Ayush, J. Song, B. Uzkent, H. Jin, and S. Ermon, "Negative data augmentation," in *International Conference on Learning Representations (ICLR)*, 2021.

[27] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*, pp. 499–515, Springer, 2016.

[28] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

[29] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] B. Liu, H. Kang, H. Li, G. Hua, and N. Vasconcelos, "Few-shot open-set recognition using meta-learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *International Conference on Computer Vision (ICCV)*, 2019.

[32] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[33] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[34] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, "Learning open set network with discriminative reciprocal points," in *The European Conference on Computer Vision (ECCV)*, August 2020.

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[38] 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*, pp. 4700–4708, 2017.

[39] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[40] R. Varghese and M. Sambath, "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6, IEEE, 2024.

[41] S. Gilroy, D. Mullins, E. Jones, A. Parsi, and M. Glavin, "The impact of partial occlusion on pedestrian detectability," *arXiv preprint arXiv:2205.04812*, 2022.

[42] M. Liu, J. Jiang, C. Zhu, and X.-C. Yin, "Vlpd: Context-aware pedestrian detection via vision-language semantic self-supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6662–6671, 2023.

[43] I. Logeswaran, H. Eissa, and R. Almadhoun, "Autonomous vehicle pedestrian detection traffic sign recognition using yolov8," in *2024 29th International Conference on Automation and Computing (ICAC)*, pp. 1–6, 2024.

[44] Y. Li, H. Wang, and B. Buckles, "Traffic congestion assessment based on street level data for on-edge deployment," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pp. 289–291, 2019.

[45] A. Wibowo, B. R. Trilaksono, E. M. I. Hidayat, and R. Munir, "Object detection in dense and mixed traffic for autonomous vehicles with modified yolo," *IEEE Access*, vol. 11, pp. 134866–134877, 2023.

[46] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in neural information processing systems*, pp. 582–588, 2000.

[47] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2317–2324, 2014.

[48] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 762–768, 2017.

[49] H. Zhang and V. M. Patel, "Sparse representation-based open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1690–1696, 2016.

[50] P. R. M. Júnior, R. M. De Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. Penatti, R. d. S. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Machine Learning*, vol. 106, no. 3, pp. 359–386, 2017.

[51] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 613–628, 2018.

[52] Y. Yu, W.-Y. Qu, N. Li, and Z. Guo, "Open-category classification by adversarial sample generation," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3357–3363, 2017.

[53] S. Kong and D. Ramanan, "Opengan: Open-set recognition via open data generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 813–822, 2021.

[54] Z. Yue, T. Wang, Q. Sun, X.-S. Hua, and H. Zhang, "Counterfactual zero-shot and open-set visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15404–15414, 2021.

[55] R. Wang, J. Guo, R.-W. Zhao, L. Su, Y. Ye, X. Zhang, Y. Zhang, and R. Feng, "Class-aware variational auto-encoder for open set recognition," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 264–269, IEEE, 2023.

[56] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, pp. 737–744, 1994.

[57] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, IEEE, 2006.

[58] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[59] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2593–2601, 2017.

[60] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.

[61] I. Elezi, J. Seidenschwarz, L. Wagner, S. Vascon, A. Torcinovich, M. Pelillo, and L. Leal-Taixe, "The group loss++: A deeper look into group loss for deep metric learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 2505–2518, 2022.

[62] W. Zheng, J. Lu, and J. Zhou, "Deep metric learning with adaptively composite dynamic constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8265–8283, 2023.

[63] X. Ma, J. Guo, A. Sansom, M. McGuire, A. Kalaani, Q. Chen, S. Tang, Q. Yang, and S. Fu, "Spatial pyramid attention for deep convolutional

neural networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 3048–3058, 2021.

[64] E. Baseman, S. Blanchard, Z. Li, and S. Fu, "Relational synthesis of text and numeric data for anomaly detection on computing system logs," in *IEEE International Conference on Machine Learning and Applications*, 2016.

[65] Q. Guan, Z. Zhang, and S. Fu, "Ensemble of bayesian predictors for autonomic failure management in cloud computing," in *IEEE International Conference on Computer Communications and Networks*, 2011.

[66] H. S. Pannu, J. Liu, and S. Fu, "A self-evolving anomaly detection framework for developing highly dependable utility clouds," in *IEEE Global Communications Conference*, 2012.

[67] S. Fu and C.-Z. Xu, "Service migration in distributed virtual machines for adaptive grid computing," in *IEEE International Conference on Parallel Processing*, 2005.

[68] Z. Zhang and S. Fu, "Characterizing power and energy usage in cloud computing systems," in *IEEE International Conference on Cloud Computing Technology and Science*, 2011.

[69] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting structured data*, 2006.

[70] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[71] Y. Zhu, J. Xie, Y. N. Wu, and R. Gao, "Learning energy-based models by cooperative diffusion recovery likelihood," in *The Twelfth International Conference on Learning Representations*, 2024.

[72] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," in *International Conference on Learning Representations (ICLR)*, 2020.

[73] J. Piland, C. Sweet, P. Saboia, C. Vardeman, and A. Czajka, "Non-generative energy based models," in *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2023.

[74] Z. Cao, Y. Li, and B.-S. Shin, "Attention-guided energy-based model for out-of-distribution data detection," in *International Conference on Pattern Recognition*, pp. 1–15, Springer, 2025.