Density-driven Regularization for Out-of-distribution Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Detecting out-of-distribution (OOD) samples is essential for reliably deploying deep learning classifiers in open-world applications. However, existing detectors relying on discriminative probability suffer from the overconfident posterior estimate for OOD data. Other reported approaches either impose strong unproven parametric assumptions to estimate OOD sample density or develop empirical detectors lacking clear theoretical motivations. To address these issues, we propose a theoretical probabilistic framework for OOD detection in deep classification networks, in which two regularization constraints are constructed to reliably estimate sample density to identify OOD. Specifically, the density consistency regularization enforces the agreement between analytical and empirical densities of observable low-dimensional categorical labels. The contrastive distribution regularization separates the densities between in distribution (ID) and distribution-deviated samples. A simple and robust implementation algorithm is also provided, which can be used for any pre-trained neural network classifiers. To the best of our knowledge, we have conducted the most extensive evaluations and comparisons on computer vision benchmarks. The results show that our method significantly outperforms state-of-the-art detectors, and even achieves comparable or better performance than methods utilizing additional large-scale outlier exposure datasets. Our code will be open-sourced upon acceptance.

1 Introduction

Deep neural networks have achieved impressive accuracy on many classification tasks when evaluated in a static and closed environment, in which the test samples are sampled from the same distribution as the training data. However, when deployed in real-world applications, deep learning models often encounter anomalous samples different from the training data. Detecting such unexplored samples is referred to as out-of-distribution (OOD) detection [1–13], which is crucial for reliable decision making and essential in safety-critical applications such as autonomous driving, medical diagnosis, secure authentication, and financial risk management [1, 3, 8, 9]. This topic is particularly important for deep learning classifiers because it has been well documented that it can be easily fooled by OOD samples to produce high confidence predictions [4, 7–9, 11, 14–16].

In recent years, most reported and favored strategies for deep-learning based OOD detection are to design the detectors by inspecting the distribution of high-level features extracted by the relevant target task network. The studies are briefly reviewed below.

As a pilot study, Hendrycks et al. proposed to detect OOD samples based on maximal softmax probability (Msp) [16], i.e., $\max_y p(y|x)$, which has become a common baseline for OOD detection [4–7, 9, 10, 12, 13, 17, 18]. Subsequently, Liang et al. proposed the Odin method, which provided temperature scaling and input perturbation strategies, to push the Msp score further apart for ID

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

and OOD data [17]. Hsu et al. generalized the Odin method to avoid access to OOD samples for parameter tuning [5]. Another similar technique, which uses softmax probability for OOD detection, is pNML [3]. In pNML, Bibas et al. detected OOD samples using their proposed regret score $\log \sum_y \overline{p}(y|x)$, where $\overline{p}(y|x)$ denotes online fine-tuned softmax probability by adding the test sample x to the training set with an arbitrary label y. However, these methods relying on softmax probability score $p(y|x)$ can suffer from overconfidence problem since many studies have shown that neural networks can produce arbitrarily high softmax probability for OOD inputs [1, 4, 8, 10, 12, 19, 20].

Another line of research utilizes the distribution of extracted high-level features $f(x)$ to estimate ID data density $p(x)$ for OOD detection as modeling $p(x)$ directly with high-dimensional input x is prohibitive in most cases. For example, Lee et al. [9] proposed using the Mahalanobis distance between the test sample's feature representations and the class-conditional gaussian distributions in the intermediate layers to train a logistic regression OOD detector. Essentially, this Mah method constructed the OOD score using $\max_y p(f(x)|y)$, in which $p(f(x)|y)$ was assumed to obey Gaussian distribution for the ID training data. Afterward, Sehwag et al. [21] and Zhou et al. [19] further extended the Mah model for unsupervised learning (no training labels available) and semi-supervised learning settings (small set of training labels available), respectively. However, the strong parametric assumption for these studies that class-conditional feature representation follows Gaussian distribution often fails in practical applications. Liu et al. [4] interpreted the softmax function from an energy perspective which assumed that $\log(p(x)) \propto \log(\sum_y \exp(f^y(x)/T))$ where T was temperature coefficient and $f^y(x)$ denoted the classification logit for input x and class y. The energy method proposed to use Helmholtz free energy $-T\log(\sum_y \exp(f^y(x)/T))$ as the OOD score. Nevertheless, we will explain in the method section that the underlying assumption of the energy model does not necessarily hold, which can lead to poor estimation of $p(x)$ for OOD detection.

Some other studies reported different empirical methods for OOD detection. For example, Sastry et al. [8] detected OOD samples by defining the deviation of test samples from training data based on a layer-wise Gram matrix that encoded pairwise feature correlations between channels. Inspired by the contrastive learning strategy, Tack et al. [22] contrasted different types of augmented training samples and proposed an empirical OOD score by measuring the distance between the test sample and nearest training samples in contrastive representation. Additionally, Sun et al. [1] observed that OOD data tend to trigger unit activation with larger variations and positive skewness. In light of this, the proposed ReAct model [1] directly truncated high activations in the penultimate layer, which was experimentally confirmed to be helpful in improving the performance of common OOD scores.

In summary, the existing methods have one or more of the following limitations: (1) Due to lack of clear probabilistic definition of OOD detection, some studies identify OOD samples using estimated discriminative probability $p(y|x)$ instead of data distribution $p(x)$, which is susceptible to the influence of overconfident posterior; **(2)** Some researchers estimate $p(x)$ by imposing strong explicit parametric or implicit assumptions of high-level extracted features $f(x)$, which however are not necessarily justified and can impair the OOD detection performance; **(3)** Several empirical methods lack a clear theoretical motivation.

To overcome these limitations, we will explore how to reliably estimate ID data distribution $p(x)$ for OOD detection using the high-level representation features. First, we consider it is inappropriate to estimate the sample distribution $p(x)$ directly from the distribution of latent feature representation learned by the discriminative network since it only concerns the conditional probability $p(y|x)$ without any constraints to guarantee the reliability of the estimated $p(x)$. To address this issue, we propose a novel density-consistency regularization strategy: given that the distribution of low dimensional variable Y can be estimated with high accuracy, this study enforces the consistency between empirical density $\widehat{p}(y)$ and the estimated density $p(y)$=$\int p(x,y)dx$=$\int p(x)p(y|x)dx$ derived from latent feature representation in deep layers, e.g., the last logit layer, on a batch-by-batch basis. Theoretically, with sufficient training samples (batch numbers), combining this consistency regularization with common supervised loss can reliably estimate both $p(y|x)$ and $p(x)$, thus enabling effective OOD detection. Secondly, to ensure OOD samples highly deviated from the ID distribution can also be identified, we further propose contrastive distribution regularization to produce low-density estimates for the distribution deviated samples. These two regularizations are properly combined to further improve the detection accuracy.

The following methods section will investigate how to achieve the above density regularizations. It includes building the consistency criterion with the law of large numbers, analyzing asymptotic prop-

erties for the deviation from consistency with the central limit theorem, developing asymptotic-test based regularization for the consistency hypothesis, and finally incorporating density regularization with contrastive distributions. Our contribution can be summarized as follows:

- We propose a unified probabilistic framework for OOD detection in discriminative networks, which reveals that direct estimation of ID data distribution $p(x)$ via latent feature representations can be biased and introduces density-consistent and contrastive criteria to address this issue. Unlike previous studies, our method does not rely on any unproven assumptions.
- We derive analytical regularization terms by analyzing the asymptotic properties of the density-consistent and contrastive criteria, and provide a simple yet surprisingly effective computational algorithm for OOD detection applicable to any deep networks with a softmax classifier.
- Extensive experiments on various standard OOD detection benchmarks demonstrate that the proposed density-driven OOD detector achieves state-of-the-art performance compared with many other existing methods.

2 Related work

In general, OOD detectors can be broadly categorized into generative and discriminative approaches.

OOD detection with generative models Deep generative models have been explored for OOD detection [2, 6, 7, 11, 23–28] as an intuitive strategy since they enable modeling likelihoods $p_\theta(x)$ for high dimensional input. One contributing aspect of the generative approach is that it directly models the input distribution without using labeled data[11]. However, it requires additional generative networks for classification tasks, which can be prohibitively challenging to optimize [1, 4, 29]. Moreover, many studies have highlighted significant issues that deep generative models can frequently assign spuriously higher likelihoods to OOD than ID samples [1, 2, 4–8, 10, 11, 23–25, 27, 30, 31], and the performance can often lag behind the discriminative approaches [1, 7, 29]. Recent studies revealed that it might be because deep generative models can be confounded by background statistics [11], or because of the discordance between the model's typical set and high-density region [23].

Discriminative model-based OOD detection The discriminative model-based approach, which derives OOD score by using representations from a predefined classification task, has advantages over generative-based approaches by circumventing the difficult optimization process and obtaining better detection performance [1, 4, 6, 9], especially for complex datasets [22]. This can be because supervised learning effectively learns discriminative representations useful in identifying in-distribution samples by exploiting label information [6, 8, 19]. A brief overview of the baseline Msp model [16] and state-of-the-art discriminative OOD detection methods, including Odin [17], Mah [9], Gram [8], Energy [4], pNML [3] and ReAct [1], is presented in the introduction section.

Discriminative model-based OOD detection with outlier exposure Some researchers have explored using auxiliary datasets of outliers, which allow the model to be explicitly regularized through fine-tuning, to increase the margin between the OOD scores of the ID and OOD data. Such an outlier exposure strategy has been reported by Oe [10], Oecc [32] and Energy [4] models to be helpful in improving OOD detection performance. However, OOD samples are different and infinite. To avoid specific exposure OODs inducing a bias to the detector, researchers typically used a large-scale diverse dataset as the outlier exposure dataset [4, 10, 32], e.g., 80 Million Tiny Images. Nevertheless, it is often not possible to have such a large amount of external data in many real-world tasks, especially those with large ID datasets.

3 Methods

3.1 Probabilistic framework for OOD detection

The goal of a machine learning classifier is to learn the mapping from X to Y, where input X is usually a high-dimensional random vector and Y is a discrete random variable. Let $g(x, y)$ be the probability density function (p.d.f.) of (X, Y) with respect to measure $\nu_X \times \pi_Y$, where ν_X is Lebesgue measure in Euclidean space R^m and π_Y is discrete measure with $\pi_Y(Y = y) = 1, 1 \leq y \leq L$. L is the

number of classification categories. The classification probabilities for discriminant model $P(Y|X)$ can be represented

$$P\left(Y=y \mid X\right)=\frac{g\left(X,y\right)}{\int g\left(X,Y\right)d\pi_{Y}}. \tag{1}$$

For deep learning models, researchers generally considered the following softmax values as $P(Y|X)$.

$$P\left(Y=y \mid X\right)=\frac{\exp\left(h^{y}\left(X\right)\right)}{\sum_{1\leq y\leq L}\exp\left(h^{y}\left(X\right)\right)} \tag{2}$$

where $h^y(X)$ is the classification logits produced by the deep neural networks. By connecting Eq.(1) and Eq.(2), $\exp(h^y(x))$ can be considered as the Boltzmann-Gibbs form for $g(x,y)$ from an energy-based perspective [33, 34], i.e.,

$$g\left(x,y\right)\propto\exp\left(h^{y}\left(x\right)\right). \tag{3}$$

Alternatively, $g\left(x,y\right)=\frac{1}{Z}\exp\left(h^{y}\left(x\right)\right)$ and denominator $Z\triangleq\sum_{1\leq y\leq L}\int\exp\left(h^{y}\left(X\right)\right)d\nu_{X}$.

For OOD detection, we first argue that OOD score should be constructed from $P(X)$ rather than $P(Y|X)$, which is commonly used in previous studies [3, 5, 16, 17], since the estimated $P\left(Y=y|X=x\right)$ can create over-confidence probability [4, 9] even when $P(X=x)$ is relatively small. In terms of $P(X)$, we can obtain the following estimated p.d.f. $k(x)$ based on Eq.(3):

$$k\left(x\right)=\int g\left(x,Y\right)d\pi_{Y}=\frac{1}{Z}\sum_{1\leq y\leq L}\exp\left(h^{y}\left(x\right)\right). \tag{4}$$

$k(x)$ appears to be a reasonable OOD detection score. Similarly, Liu et al. proposed to use Helmholtz free energy $-T\log(\sum_{1\leq y\leq L}\exp(h^y(x)/T))$ as OOD score with temperature T [4], which is similar to $k\left(x\right)$ from a computational perspective, and demonstrated its advantage over OOD scores based on discriminative probability $P(Y|X)$ [16, 17].

However, directly calculating $k\left(x\right)$ as the OOD score can lead to poor density estimation since the prerequisite Eq.(3) is not clearly satisfied, which has been overlooked in previous studies. Particularly, classification models generally focus on improving the accuracy of $P(Y|X)$ by classification loss L_{cls} without explicitly enforcing the validity of Eq.(3). Therefore, this study aims to develop an explicit regularization to satisfy the constraint in Eq.(3) based on the empirical density estimation of the ID training data. Specifically, since the empirical density function (EDF) of a low-dimensional random variable (r.v.) can be estimated with high accuracy, we start with estimating the EDF of discrete r.v. Y and enforce the consistency between its EDF and its analytical parametric p.d.f. of the following form based on Eq.(3).

$$v\left(y\right)=\int g\left(X,y\right)d\nu_{X}=\frac{1}{Z}\int\exp\left(h^{y}\left(X\right)\right)d\nu_{X} \tag{5}$$

In the following sections, we will construct the analytical consistency regularization and implementation algorithm for ensuring the constraints of Eq.(3), thereby improving OOD detection performance, using the law of large numbers (LLN), central limit theorem (CLT) and asymptotic test statistics.

3.2 Consistency criterion for low-dimensional density estimation

As noted above, density function $k(x)$, estimated by high layer features, can be a useful OOD score if the underlying assumption Eq.(3) holds. To realize this assumption, we now enforce the consistency between analytical p.d.f $v\left(y\right)$ derived by Eq.(3) and the empirical density function (EDF) for low-dimensional variable Y, which can be reliably estimated by $\widehat{v}\left(y\right)=n_y/N$. $n_y=\sum_{1\leq i\leq N}1_{\{Y_i=y\}}$ represents the number of samples in class y and $N=\sum_{y=1}^{L}n_y$ is the sample size. Namely, with Eq.(3), we derive the following property:

$$1\approx\frac{v\left(y\right)}{\widehat{v}\left(y\right)}=\frac{1}{Z}\frac{\int\exp\left(h^{y}\left(X\right)\right)d\nu_{X}}{\widehat{v}\left(y\right)}=\frac{1}{Z}\int\frac{\exp\left(h^{y}\left(X\right)\right))}{\widehat{v}\left(y\right)}d\nu_{X};\forall 1\leq y\leq L \tag{6}$$

By the law of large numbers, $\widehat{v}(y)$ converges to $v(y)$ almost surely (a.s.) and in probability as $N \to \infty$. Thus, the pivotal variable $\int \frac{\exp(h^y(X)))}{\widehat{v}(y)} d\nu_X$ converges to Z a.s. and in probability. By the Monte Carlo method, Z can be approximated by:

$$Z \approx \int \frac{\exp\left(h^y\left(X\right)\right))}{\widehat{v}\left(y\right)} d\nu_X \approx \frac{\sum_{1 \leq i \leq N} \exp\left(h^y\left(X_i\right)\right)/N}{n_y/N}; \forall 1 \leq y \leq L \tag{7}$$

Obviously, the right-hand side (RHS) of Eq.(7) also converges to Z a.s. and in probability as $N \to \infty$. Thus, we arrive at the proposition that for each class $1 \leq y \leq L$, the exponential of classification logit has a mean value proportional to the class ratio.

Since deep learning models are trained in a batch mode, in which random batch sampling with replacement allows bootstrap estimates, we now proceed to construct batch-wise statistics and analytical regularization for Eq.(7). We first give the following lemma. The proof is in Appendix A.1 of the supplementary material.

Lemma 1. *If Eq.*(3) *holds, then*

$$\frac{\sum_{1 \leq i \leq N} \exp\left(h^y\left(X_i\right)\right)}{\sum_{1 \leq i \leq N} \sum_{1 \leq l \leq L} \exp\left(h^l\left(X_i\right)\right)} - \frac{n_y}{N} \to_{a.s.} 0; \forall 1 \leq y \leq L \tag{8}$$

Lemma 1 gives a theoretical batch-wise constraint for the classification logit. However, the convergence rate is not specified here, making it challenging to design an adaptive regularization subjected to varying class sample n_y and batch size N. To address this issue, in the next section, we will specify the convergence rate of $O(1/\sqrt{N})$ for Eq.(8) using the central limit theorem and accordingly propose the adaptive consistency regularization by our asymptotic test statistics.

3.3 Adaptive density-consistency regularization by asymptotic test

To construct the density-consistency regularization adaptive to different class size n_y and batch size N, we now consider the convergence rate with respect to sample size for Eq.(8) by the following lemma 2. The proof is provided in Appendix A.2.

Lemma 2. *If Eq.*(3) *holds, then*

$$\Upsilon_N^y = \sqrt{N}\left(\frac{\sum_{1 \leq i \leq N} \exp\left(h^y\left(X_i\right)\right)}{\sum_{1 \leq i \leq N} \sum_{1 \leq l \leq L} \exp\left(h^l\left(X_i\right)\right)} - \frac{n_y}{N}\right) \to_d \mathcal{N}\left(0, \Delta^y\right) \tag{9}$$

where $\Delta^y = \left[\nabla g\left(\mu^y\right)\right]^\top \Sigma^y \left[\nabla g\left(\mu^y\right)\right]$ *and* $g([x, y, z]^\top) = y/x - z$. μ^y *and* Σ^y *represent the mean and covariance matrix for random vector* $u^y = \left[\sum_{1 \leq l \leq L} \exp\left(h^l\left(X\right)\right), \exp\left(h^y\left(X\right)\right), 1_{\{Y=y\}}\right]^\top$.

Lemma 1 and lemma 2, in fact, show that if implicit assumption Eq.(3) holds, i.e., $\frac{1}{Z}\exp(h^y(x))$ can accurately approximate density $g(x, y)$, we can then construct batch-wise statistics Υ_N^y , in which $\frac{1}{\sqrt{N}}\Upsilon_N^y$ converges to zero at the rate of $O(N^{-1/2})$ by analyzing the distribution of low-dimensional Y. More specifically, Υ_N^y converges to a particular Gaussian distribution whose parameter can be reliably estimated.

Thus to satisfy assumption Eq.(3)**, which is deemed to provide a more accurate estimation of** $P(x)$ **for OOD detection, we propose to construct the asymptotic-test statistics using** Υ_N^y **from the perspective of hypothesis testing, and creatively transform the hypothesis reject criteria into a penalty term to support the assumption Eq.**(3) **can be accepted at any specific level of significance.**

Specifically, we consider the following hypotheses:

$$H_0 : \theta^y = 0 \quad \text{versus} \quad H_1 : \theta^y \neq 0 \tag{10}$$

where

$$\theta^y = \frac{E\left[\exp\left(h^y\left(X\right)\right)\right]}{E\left[\sum_{1 \leq l \leq L} \exp\left(h^l\left(X\right)\right)\right]} - E1_{\{Y=y\}} \tag{11}$$

First, assumption Eq.(3) implies null hypothesis H_0, and accordingly leads to the following convergence in distribution by lemma 2:

$$[\Delta_N^y]^{-1/2}\Upsilon_N^y \to_d \mathcal{N}(0,1) \quad \text{and} \quad [\Upsilon_N^y]^\top [\Delta_N^y]^{-1} [\Upsilon_N^y] \to_d \chi^2(1) \tag{12}$$

where $\Delta_N^y \triangleq [\nabla g(\mu_N^y)]^\top \Sigma_N^y [\nabla g(\mu_N^y)]$ approximates and converges to Δ^y in probability. μ_N^y and Σ_N^y denote sample mean and sample covariance for μ^y and Σ^y, respectively. From the perspective of asymptotic test theory [35], we can construct a χ^2-test with the following reject region which has asymptotic significance level α.

$$[\Upsilon_N^y]^\top [\Delta_N^y]^{-1} [\Upsilon_N^y] > \chi^2_{1,\alpha} \tag{13}$$

where $\chi^2_{1,\alpha}$ is the $(1-\alpha)$th quantile of chi-squared distribution $\chi^2(1)$. When Eq.(13) is observed, the asymptotic test rejects H_0, as well as assumption Eq.(3). **Hence, to ensure the assumption Eq.**(3) **is rightly accepted, we propose to transform the hypothesis test rejection criteria into the following regularization term:**

$$\lambda \triangleq \gamma \frac{1}{L} \sum_{1 \le y \le L} \max\left(\log\left([\Upsilon_N^y]^\top [\Delta_N^y]^{-1} [\Upsilon_N^y] / \chi^2_{1,\alpha}\right), 0\right). \tag{14}$$

The regularization term λ will be added to the classification loss L_{cls}, where hyperparameter $\alpha \in [0,1]$ controls the significance level and γ is the loss weight balancing the regularization term λ and classification loss L_{cls}.

3.4 Density regularization with contrastive distribution

The above density-consistency regularization (DCR) can calibrate classification logit based on observable low-dimensional density $P(Y)$, which offers a more reliable $P(X,Y)$ estimate for ID-like samples. Nevertheless, the real test OOD samples can come from distribution deviating far from the ID distribution. To ensure these samples with large distribution deviations can also be accurately identified, we further propose the density-contrasting scheme, which enforces that random distribution-deviated (DD) samples produce low-density estimates.

We first define the distribution-deviating operator $\mathcal{S}$, which can be applied to ID data X to generate DD samples $\widetilde{X} = \mathcal{S}(X)$. An ideal density estimator should yield the likelihood ratio $\frac{P(\widetilde{X})}{P(X)} \ll 1$ and $[\frac{P(\widetilde{X})}{P(X)}]^r \approx 0$ for large constant r, i.e.,

$$\psi_{\mathcal{S}}(X) \triangleq \log\left[\frac{P(\widetilde{X})}{P(X)}\right]^r = r \cdot \log \frac{\sum_{1 \le l \le L} \exp\left(h^l(\mathcal{S}(X))\right)}{\sum_{1 \le l \le L} \exp\left(h^l(X)\right)} \to -\infty. \tag{15}$$

We incorporate this constraint into the above DCR by defining $\widetilde{\Upsilon}_N^y$:

$$\widetilde{\Upsilon}_N^y \triangleq |\Upsilon_N^y| + \exp\left(\frac{1}{N|\mathbb{S}|} \sum_{\mathcal{S} \in \mathbb{S}, 1 \le i \le N} \psi_{\mathcal{S}}(X_i) + q\right) \tag{16}$$

Here a constant q is added to ensure numerical stability. Since the second term in RHS is considered close to zero, we can obtain that $[\widetilde{\Upsilon}_N^y]^\top [\Delta_N^y]^{-1} [\widetilde{\Upsilon}_N^y]$ also converges to $\chi^2(1)$ approximately. Thus, our final generalized density regularization with contrastive distribution can be defined:

$$\widetilde{\lambda} \triangleq \gamma \frac{1}{L} \sum_{1 \le y \le L} \max\left(\log\left([\widetilde{\Upsilon}_N^y]^\top [\Delta_N^y]^{-1} [\widetilde{\Upsilon}_N^y] / \chi^2_{1,\alpha}\right), 0\right) \tag{17}$$

Algorithm 1 in Appendix A.3 presents the pseudo-code for the training of our density driven regularization method. It is worth noting that to increase numerical stability, we normalize the classification logit by subtracting the maximal logit before calculating $\widetilde{\lambda}$ in Algorithm 1. This normalization procedure is also widely used to avoid overflow when calculating softmax probability [36]. We will show, in Appendix A.4, that the normalized logit essentially yields the same results as the original logit. Finally, after fine-tuning, a monotonic transformation $\log(Z \cdot k(x)) = \log(\sum_{1 \le y \le L} \exp(h^y(x)))$ of the sample density $k(x)$ is used as the OOD score in this study for any testing sample x, which is theoretically equivalent to directly adopting $k(x)$ as the OOD score.

4 Experiments

We demonstrate the effectiveness of the proposed method using diverse deep convolutional neural network architectures on multiple OOD detection benchmark datasets, and compare it against various state-of-the-art methods.

Datasets and experimental setup Following existing literature [1, 3, 4, 8–10, 16, 17, 19, 32], we train OOD models on the training set of CIFAR-10 [37], CIFAR-100 [37] and ImageNet [38], respectively, and evaluate OOD detectors in identifying OOD test samples from other datasets, including iSUN [39], iNaturalist [40], LSUN(R) [17], LSUN(C) [17], Gaussian [17], Places365 [41], Textures [42] and SVHN [43]. ImageNet(R) [3] and Imagenet(C) [3] are also used as OOD sets for the CIFAR experiment. For the ImageNet experiment, CIFAR datasets are considered as OOD sets. These datasets are described and provided in previous studies [1, 4, 17].

For fair comparison and to aid reproducibility, we use the pretrained model for standard ResNet-34 [44], WideResNet-40 [45] and DenseNet-BC-100 [46] as the backbone network for the CIFAR experiment since they are widely used in previous OOD detection studies [1, 3, 4, 8–10, 17, 32]. Following [1], we adopt the lightweight MobileNet-v2 [47] for the ImageNet experiment. It is noteworthy that since some of the following compared models, e.g., Gram [8] and Mah [9], require prohibitively large memory for large networks with high input dimensions, we only conduct experiments on lightweight MobileNet-v2 for the ImageNet dataset. The pretrained weights for the above ResNet, WideResNet, DenseNet and MobileNet are provided by [1, 9, 17].

When training our OOD detector, we use the same normalization, augmentation and training setting as in [4, 10], where the fine-tuning epoch is 10, and the initial learning rate is 0.001 with cosine decay [48]. Other hyperparameters are set as follows: batch size of 256, regularization weight γ=10^{-3}, constant r=100, and fixed statistical significance level α=0.05. Constant q is a detached tensor dynamically determined to linearly map the exponent from $[-200r, 0]$ to $[-3, 3]$ for ensuring numerical stability. **It is worth noting that unlike many previous studies [4, 8–10, 17, 32], which adjusted hyperparameters for different datasets or networks, our method can achieve state-of-the-art performance across various experiments with fixed hyperparameters.** Moreover, we will also demonstrate the robustness of our approach to hyperparameter changes in our ablation studies. Our distribution-deviating operators $\mathbb{S}$ include random resized crop and random affine transforms realized by TorchVision library [49], and uniform noise generator (see Appendix A.5 for details).

The proposed approach is compared to the state-of-the-art methods, including Msp [16], Odin [17], Mah [9], Energy [4], Gram [8], ReAct [1] and pNML [3] for OOD detection without outlier exposure, as well as Energy [4], Oe [10] and Oecc [32] for OOD detection with outlier exposure. All compared methods are evaluated with their open-source codes and trained weights. Following existing studies [1, 3, 4, 8–10, 16, 17, 32], commonly used metrics including AUROC, AURP and FPR (at 95% TPR) are used to evaluate different OOD detectors. The evaluation code is adopted from [4, 10]. All experiments run on the PyTorch framework [50] with Nvidia A100 and GeForce RTX 3090 GPUs, with a maximum RAM of 512GB.

Experimental results The experimental results on CIFAR-100 in terms of FPR95% are shown in Table 1. Compared to other state-of-the-art detectors without outlier exposure, our method achieves the best detection accuracy for diverse OOD datasets under different network architectures. The averaged FPR95% of the proposed method is significantly lower than that of other detectors without outlier exposure, and even lower than the four state-of-the-art detectors that use an external large-scale outlier exposure dataset with 80 million images. Other evaluation results in terms of AUROC and AUPR also reveal a consistent performance trend, which can be found in Appendix A.6. It is worth noting that the proposed method only utilizes classification logits for OOD detection. In contrast, the best-performed existing method, i.e., Gram, uses features across all layers to detect OODs and requires calculating huge gram matrices dynamically using all latent features of training samples, preventing end-to-end real-time prediction and becoming intractable for deep networks trained on large datasets. Due to the page limit, the reader is referred to Appendix A.7 for detailed results of the CIFAR-10 experiment, in which our method also achieves the best detection accuracy in both FPR95%, AUROC and AUPR compared to other existing methods without outlier exposure. The performance of the proposed method is also comparable to the state-of-the-art OOD detectors that require large-scale additional outlier exposure datasets. Tables in Appendix A.8 also provide

Table 1: Performance comparison (in terms of FPR95%) of various methods in CIFAR-100 (ID dataset) under three commonly-used networks. The best and second-best models without accessing external outlier datasets are coloured red and blue, respectively.

	OOD dataset	w/o outlier exposure (FPR95% ↓)	w/ outlier exposure (FPR95% ↓)
		Msp/Odin/Mah/Energy/Gram/ Energy+ReAct/pNML/Odin+pNML/Ours	Oe/Oecc/Energy/ Oecc+pNML
CIFAR-100 (ResNet34)	iSUN	83.10 / 55.65 / 82.55 / 79.80 / 5.01 / 80.00 / 73.70 / 55.10 / 4.87	69.80 / 2.70 / 54.76 / 2.14
	LSUN(R)	80.75 / 53.15 / 81.05 / 77.05 / 2.92 / 77.75 / 70.75 / 54.90 / 2.99	63.88 / 1.72 / 54.07 / 1.01
	LSUN(C)	81.30 / 55.35 / 67.55 / 81.50 / 33.76 / 80.25 / 70.60 / 47.75 / 8.88	53.24 / 19.46 / 37.20 / 10.36
	ImageNet(R)	79.00 / 52.55 / 78.05 / 75.75 / 5.17 / 75.95 / 67.95 / 53.75 / 3.63	74.19 / 4.60 / 62.30 / 4.12
	ImageNet(C)	77.40 / 57.90 / 71.35 / 73.10 / 11.68 / 72.85 / 67.00 / 53.15 / 3.77	70.38 / 9.32 / 59.27 / 8.53
	Guassian	100.0 / 95.35 / 96.05 / 100.0 / 0.00 / 100.0 / 85.25 / 32.50 / 0.00	81.49 / 0.00 / 0.01 / 0.00
	Places365	84.55 / 84.10 / 72.10 / 85.85 / 46.89 / 85.00 / 71.25 / 61.80 / 20.73	57.33 / 31.76 / 41.05 / 19.74
	Textures	80.80 / 80.15 / 67.25 / 80.35 / 33.43 / 78.50 / 61.35 / 53.80 / 20.43	57.17 / 24.66 / 43.64 / 16.38
	iNaturalist	73.00 / 71.50 / 65.70 / 73.05 / 44.68 / 71.75 / 57.85 / 57.85 / 34.07	75.77 / 46.22 / 68.57 / 47.49
	SVHN	79.70 / 38.60 / 55.70 / 80.45 / 19.99 / 78.75 / 48.40 / 26.30 / 1.85	58.41 / 13.09 / 40.57 / 10.23
	average	81.96 / 64.43 / 73.73 / 80.69 / **20.35** / 80.08 / 67.41 / 49.69 / **10.12**	66.17 / 15.35 / 46.14 / 12.00
CIFAR-100 (WideResNet)	iSUN	86.75 / 63.60 / 36.40 / 78.80 / 5.23 / 82.70 / 73.45 / 73.45 / 1.70	61.29 / 7.93 / 55.97 / 3.01
	LSUN(R)	85.75 / 60.50 / 33.75 / 79.40 / 2.97 / 82.95 / 77.55 / 77.55 / 0.70	57.45 / 5.41 / 53.28 / 1.90
	LSUN(C)	80.85 / 52.10 / 77.95 / 71.85 / 37.29 / 57.85 / 86.95 / 61.85 / 4.92	35.74 / 18.20 / 29.25 / 6.70
	ImageNet(R)	84.90 / 63.95 / 38.25 / 78.95 / 6.01 / 83.60 / 71.55 / 71.55 / 2.24	68.92 / 12.04 / 66.33 / 7.72
	ImageNet(C)	82.55 / 65.30 / 50.95 / 71.05 / 12.98 / 76.80 / 79.40 / 79.90 / 2.44	58.14 / 15.84 / 56.36 / 7.81
	Guassian	99.60 / 0.05 / 0.00 / 67.25 / 0.00 / 47.55 / 0.00 / 0.00 / 0.00	87.00 / 0.00 / 0.00 / 0.00
	Places365	83.80 / 78.80 / 73.00 / 81.05 / 46.27 / 72.50 / 82.45 / 62.60 / 11.03	44.58 / 23.26 / 36.49 / 9.68
	Textures	81.60 / 74.05 / 54.45 / 78.45 / 35.98 / 66.00 / 75.40 / 58.75 / 12.47	46.91 / 19.62 / 38.03 / 8.10
	iNaturalist	76.15 / 70.35 / 68.45 / 71.40 / 47.86 / 70.75 / 85.40 / 73.40 / 16.93	62.02 / 38.38 / 62.70 / 24.92
	SVHN	77.30 / 42.45 / 42.85 / 85.40 / 19.26 / 76.00 / 74.85 / 57.05 / 1.66	62.61 / 9.43 / 37.50 / 3.64
	average	83.92 / 57.12 / 47.60 / 76.36 / **21.39** / 71.67 / 70.70 / 61.61 / **5.41**	58.47 / 15.01 / 43.59 / 7.35
CIFAR-100 (DenseNet)	iSUN	83.85 / 62.10 / 17.70 / 82.45 / 3.99 / 40.35 / 18.40 / 16.20 / 5.61	89.67 / 3.21 / 76.09 / 0.92
	LSUN(R)	82.75 / 58.55 / 15.90 / 77.45 / 2.38 / 34.45 / 16.40 / 14.40 / 4.36	89.83 / 1.98 / 73.39 / 0.57
	LSUN(C)	73.40 / 42.25 / 81.40 / 47.05 / 34.70 / 45.80 / 35.10 / 35.10 / 8.57	53.88 / 26.35 / 33.94 / 16.89
	ImageNet(R)	81.90 / 56.95 / 17.10 / 76.85 / 4.15 / 39.70 / 13.55 / 11.95 / 3.09	89.25 / 3.85 / 75.50 / 1.19
	ImageNet(C)	75.45 / 46.75 / 36.85 / 58.65 / 10.92 / 33.30 / 22.75 / 21.90 / 3.65	80.14 / 8.09 / 60.05 / 3.40
	Guassian	100.0 / 100.0 / 0.00 / 100.0 / 0.00 / 100.0 / 0.00 / 0.00 / 0.00	100.0 / 0.00 / 100.0 / 0.00
	Places365	76.15 / 51.80 / 83.05 / 58.20 / 45.56 / 53.00 / 36.35 / 36.35 / 15.21	62.32 / 41.32 / 46.77 / 32.74
	Textures	75.15 / 56.50 / 59.70 / 58.85 / 36.38 / 47.25 / 25.90 / 25.90 / 14.07	69.91 / 31.87 / 48.95 / 23.20
	iNaturalist	70.30 / 51.80 / 75.35 / 48.90 / 48.61 / 43.05 / 45.25 / 44.85 / 21.27	74.11 / 48.27 / 64.23 / 44.94
	SVHN	73.75 / 44.30 / 53.65 / 65.95 / 10.52 / 48.55 / 20.50 / 20.50 / 3.08	75.38 / 11.63 / 69.06 / 9.72
	average	79.27 / 57.10 / 44.07 / 67.44 / **19.72** / 48.55 / 23.42 / 22.71 / **7.89**	78.45 / 17.66 / 64.80 / 13.36

the experimental results of FPR95%, AUROC and AUPR values on ImageNet. Due to the lack of large-scale high-resolution outlier exposure datasets, which should theoretically be much larger than the ID dataset, none of the outlier-exposure OOD detectors had been evaluated on the ImageNet benchmark in literature. Therefore, only OOD detectors without requiring outlier exposure datasets are evaluated. It can be observed that the proposed method also achieves the highest averaged detection accuracy on the ImageNet benchmark. It is noteworthy that our method performs relatively poorly on the iNaturalists and Places365 compared to other OOD datasets, as is also the case for other traditional methods. It can be because the ImageNet includes semantically overlapping objects as the iNaturalists and Places365, as is shown in Appendix A.9, even though we have used the conceptually disjoint subsets of Textures and iNaturalists provided by the previous study [1] for model evaluation.

Ablation studies We conduct ablation experiments on CIFAR-100 using the WideResNet network. Table 2 verifies the effectiveness of each component, i.e., density consistency regularization and contrastive distribution regularization. We remark that the baseline model without regularization corresponds theoretically to the Energy model with no outlier exposure and temperature coefficient of $T = 1$. Table 3 verifies the robustness of the proposed method to hyperparameters, including batch size, regularization weight γ and constant r. The model performance is quite consistent when increasing or decreasing the default values by 100% and 50%, respectively. Please note that unlike many previous approaches involving hyperparameter tuning [4, 8–10, 17, 32], our method uses fixed

Table 2: Contribution of the proposed regularization terms to OOD detection performance. DCR: density consistency regularization. CDR: contrastive distribution regularization. The result is the average value across all OOD datasets.

in-dist (model)	DCR	CDR	mean AUROC (↑)	mean AUPR (↑)	mean FPR95% (↓)
CIFAR-100 (WideResNet)	-	-	80.29	95.19	76.73
	✓	-	92.71	98.40	37.29
	✓	✓	98.85	99.75	5.41

Table 3: Ablation studies on CIFAR-100 (ID dataset) with WideResNet demonstrating the proposed method's robustness to hyperparameters. The result is the average value across all OOD datasets.

ablation studies	parameters	mean AUROC (↑)	mean AUPR (↑)	mean FPR95% (↓)
batch size	128	98.81	99.74	5.54
	192	98.71	99.72	6.07
	256	98.85	99.75	5.41
	384	98.74	99.73	6.10
	512	98.73	99.73	6.26
regularization weight γ	0.50×10^{-3}	98.75	99.73	6.27
	0.75×10^{-3}	98.76	99.73	5.99
	1.00×10^{-3}	98.85	99.75	5.41
	1.50×10^{-3}	98.72	99.72	5.82
	2.00×10^{-3}	98.74	99.72	5.74
constant r	50	98.79	99.74	6.05
	75	98.82	99.75	5.57
	100	98.85	99.75	5.41
	150	98.85	99.75	5.31
	200	98.88	99.76	5.17

hyperparameters for OOD detection across different ID datasets, OOD datasets and networks. It further confirms the robustness of the proposed method.

5 Conclusion

In this paper, we propose a theoretical probabilistic framework for OOD detection, which derives analytical density regularization by investigating the inherent consistent and contrastive characteristics of the ID sample density to build an accurate sample density estimator. Unlike many previous studies, our method neither relies on unproven assumptions about the latent feature distribution nor requires additional outlier exposure datasets to calibrate ID density estimates or tune hyperparameters. A simple yet effective computational algorithm is also developed, which can be used for any pretrained classification network. Extensive evaluations on benchmark datasets show that the proposed approach significantly outperforms existing state-of-the-art OOD detectors. Furthermore, our method even achieves comparable or better performance than the approaches requiring external large-scale outlier exposure datasets. A limitation of our approach is that the fixed hyperparameters may not be optimal for a specific task, although it enables our fairly robust model to achieve state-of-the-art performances across benchmarks. We leave it for future work to investigate adaptively setting hyperparameters with the ID dataset. We believe our accurate sample density estimator can also guide future studies in related machine learning fields, such as active learning, open-world classification, semi-supervised learning, etc.

References

[1] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, pages 144–157, 2021.

[2] Keunseo Kim, JunCheol Shin, and Heeyoung Kim. Locally most powerful bayesian test for out-of-distribution detection using deep generative models. In *Advances in Neural Information Processing*

Systems, pages 14913–14924, 2021.

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

[4] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475, 2020.

[5] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10948–10957, 2020.

[6] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *8th International Conference on Learning Representations*, 2020.

[7] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *Advances in Neural Information Processing Systems*, pages 20685–20696, 2020.

[8] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8491–8501, 2020.

[9] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[10] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations*, 2019.

[11] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14680–14691, 2019.

[12] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9517–9525, 2019.

[13] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations*, 2018.

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations*, 2015.

[15] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations*, 2017.

[17] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations*, 2018.

[18] 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 560–574, 2018.

[19] Zhi Zhou, Lan-Zhe Guo, Zhanzhan Cheng, Yu-Feng Li, and Shiliang Pu. Step: Out-of-distribution detection in the presence of limited in-distribution labeled data. In *Advances in Neural Information Processing Systems*, pages 29168–29180, 2021.

[20] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–50, 2019.

[21] Vikash Sehwag, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *9th International Conference on Learning Representations*, 2021.

[22] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, pages 11839–11852, 2020.

[23] Ziyu Wang, Bin Dai, David Wipf, and Jun Zhu. Further analysis of outlier detection with deep generative models. In *Advances in Neural Information Processing Systems*, pages 8982–8992, 2020.

[24] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

[25] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations*, 2019.

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

[27] Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In *Advances in Neural Information Processing Systems*, pages 21038–21049, 2020.

[28] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 481–490, 2019.

[29] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[30] Alireza Shafaei, Mark Schmidt, and James J. Little. A less biased evaluation of out-of-distribution sample detectors. In *30th British Machine Vision Conference*, 2019.

[31] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, pages 20578–20589, 2020.

[32] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021.

[33] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *8th International Conference on Learning Representations*, 2020.

[34] Yann LeCun, Sumit Chopra, Raia Hadsell, Fu Jie Huang, and et al. A tutorial on energy-based learning. In *Predicting structured data*. MIT Press, 2006.

[35] Jun Shao. *Mathematical statistics*. Springer Science & Business Media, 2003.

[36] François Chollet et al. Keras. `https://keras.io`, 2015.

[37] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[39] 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.

[40] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

[42] 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.

[43] 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.

[44] 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.

[45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[46] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[47] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[48] Ilya Loshchilov and Hutter Frank. Sgdr: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations*, 2017.

[49] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488, 2010.

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section Conclusion
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] We use standard datasets freely available for academic research. They are well-known so we did not mention their licenses.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provide code for the proposed method in the supplementary material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]