Contrastive Training for Improved Out-of-Distribution Detection

Jim Winkens[1], **Rudy Bunel**[2], **Abhijit Guha Roy**[1], **Robert Stanforth**[2], **Vivek Natarajan**[1]
Joseph R. Ledsam[2], **Patricia MacWilliams**[1], **Pushmeet Kohli**[2], **Alan Karthikesalingam**[1],
Simon Kohl[2], **Taylan Cemgil**[2], **S. M. Ali Eslami**[2] **and Olaf Ronneberger**[2]

Google Health[1], DeepMind[2]
{jimwinkens, olafr}@google.com

Abstract

Reliable detection of out-of-distribution (OOD) inputs is increasingly understood to be a precondition for deployment of machine learning systems. This paper proposes and investigates the use of contrastive training to boost OOD detection performance. Unlike leading methods for OOD detection, our approach does not require access to examples labeled explicitly as OOD, which can be difficult to collect in practice. We show in extensive experiments that contrastive training significantly helps OOD detection performance on a number of common benchmarks. By introducing and employing the *Confusion Log Probability* (CLP) score, which quantifies the difficulty of the OOD detection task by capturing the similarity of inlier and outlier datasets, we show that our method especially improves performance in the 'near OOD' classes – a particularly challenging setting for previous methods.

1 Introduction

A well-trained deep neural network f that obtains high accuracy on its test set can still make arbitrarily bad predictions when exposed to inputs drawn from an unfamiliar distribution [28, 29]. This poses a significant obstacle for real world deployment, where it is typically either prohibitively expensive or outright impossible to ensure that the network is only ever exposed to data from the training distribution.

In safety-critical applications, e.g. in medical diagnosis, it would be preferable to detect inputs unfamiliar to the trained network for separate processing (for instance by a human expert), than to make potentially inaccurate predictions using the machine learning system. This problem is known as out-of-distribution (OOD) detection, open set recognition, or anomaly detection by different research communities.

Out-of-distribution detection can be performed by approximating a probability density $p(\mathbf{x})$ of training inputs $\mathbf{x}$, and detecting test-time OOD inputs using a threshold γ: if $p(\mathbf{x}) < \gamma$ then $\mathbf{x}$ is considered OOD. The surprising finding of

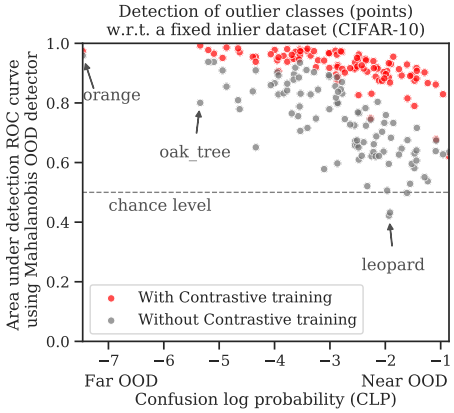


Figure 1: Each point represents performance at detecting one of the classes in CIFAR-100 as outliers with respect to a network trained on CIFAR-10. The classes are sorted by increasing similarity to the inlier classes as given by the class-wise confusion log probability (CLP), see Eq. 5. Contrastive training improves OOD detection results across the board, particularly in the near OOD regime, where the outlier and inlier classes are highly similar. CIFAR-10 contains classes similar to *leopard* (e.g. *dog, cat*) but none similar to *oak tree* or *orange*.

Preprint. Under review.

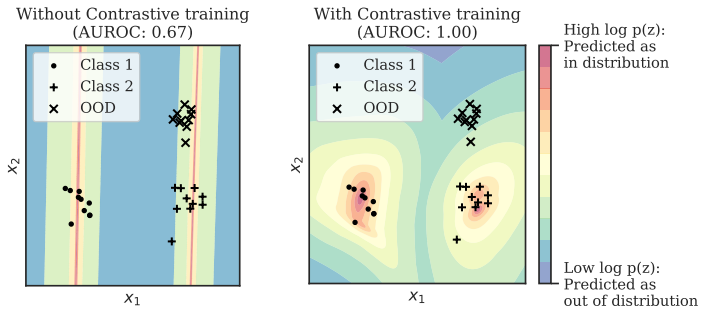


Figure 2: We show $s(\mathbf{z}) = \log p(\mathbf{z})$ of a network trained to distinguish between two classes, where $\mathbf{z}$ is its penultimate activation. Left: Supervised training of $\mathbf{z}$ alone may discard input dimensions that are unnecessary for classification but necessary for OOD detection. Right: Contrastive training makes $\mathbf{z}$ also be sensitive to the x_2 dimension.

Nalisnick et al. [25] was that even powerful neural generative models trained to estimate $p(\mathbf{x})$ (e.g. on CIFAR-10 images) can perform poorly at OOD detection, often assigning higher probabilities to out-of-distribution test examples (Street View House Numbers) than to in-distribution test examples.

Modern OOD detection techniques [12, 13, 18, 19] instead assign a scalar score $s(\mathbf{z})$ (e.g. via an approximated probability density) to activations $\mathbf{z}$ in an intermediate feature space of a discriminatively trained classifier f, and use that to detect OOD inputs. The success of these approaches highly depends on the quality of the intermediate feature space defined by f. If the feature space is not sufficiently rich, the network may be blind to properties of the image that turn out to be necessary for detection of OOD inputs. Consider, for instance, the case of visual inputs. Variation in captured images is either due to semantic differences of the objects (e.g. pose, shape, texture), or due to differences in the imaging process (e.g. lighting, camera position). Depending on the application, an unfamiliar variation of either type could lead to an input being deemed out-of-distribution. We therefore desire intermediate feature spaces defined by f that capture as many semantic properties as possible, whilst also remaining sensitive to properties of the imaging process.

Supervised learning produces semantic representations, but only to the extent that those representations discriminate between classes labeled in the dataset. The network f is not incentivized to learn features (semantic or otherwise) beyond the bare minimum necessary to classify. That is why current state-of-the-art approaches to OOD detection *enrich* the intermediate feature space beyond what would ordinarily be learned via only supervised learning on the inlier dataset, for instance by exploiting examples from an outlier dataset (Outlier Exposure, OE, [12]), or with self-supervised losses (Rotation Prediction, RP, [13]). To do this, however, OE requires access to examples labeled explicitly as OOD (which can be difficult to collect in practice), and RP relies on the assumption that an unrelated auxiliary task will produce beneficial representations, which may not always hold.

The key idea of this paper is to encourage f to learn as many high-level, task-agnostic, semantic features as possible from the in-distribution dataset, so as to enable it to detect any kind of out-of-distribution input at test time. We note recently introduced *contrastive* training techniques such as SimCLR [3] as performant and well-motivated approaches to this end. Using a set of class-preserving transformations, SimCLR introduces a loss that pulls transformed versions of the same image closer to each other, whilst pushing all other images away. This incentivizes the model to learn features that discriminate between all dataset images, even if they belong to the same class. When combined with supervised training, f learns features that are both rich and semantically discriminative.

Figure 2 demonstrates this idea on a toy example, where we aim to classify points in a 2-dimensional input space (x_1, x_2). The two classes can be distinguished by the first dimension x_1 alone. With only supervised training, f has no incentive to be sensitive to the x_2 dimension, making OOD detection using $\mathbf{z}$ impossible. Contrastive training, however, shapes $\mathbf{z}$ to remain sensitive to both dimensions and makes OOD detection possible. We show in extensive experiments that this approach scales to high-dimensional problems and consistently improves performance.

An additional difficulty of OOD detection lies in the evaluation of methods. Quantitative evaluation requires specification of an 'outlier' dataset at test time, which is itself a subjective design choice, as the notion of 'out-of-distribution' is task dependent. We therefore distinguish between *near OOD* regimes where inlier and outlier distributions are meaningfully similar, and *far OOD* regimes where the two are unrelated. Near OOD is encountered more often in practice, e.g. a system that detects medical pathologies will often encounter patients with atypical combinations of pathologies

(near OOD) and will have to be reliable nonetheless. A completely broken sensor (far OOD) is less prevalent by comparison. We therefore advocate for quantification of the 'similarity' of inlier and outlier distributions used in evaluations, and propose a metric for this which we use in our experiments. To summarize, the key contributions of the paper are as follows:

- We propose a new approach for OOD detection that incorporates contrastive training. Our approach avoids explicit inlier and outlier density modelling in the input space, can readily be incorporated into existing training setups and is simple to adapt to different datasets.
- We show that the approach consistently improves OOD detection across a wide spectrum of benchmarks, outperforming competitive methods such as Outlier Exposure (OE, [12]). Unlike OE, our method does *not* require access to data from the outlier distribution during training or tuning.
- We introduce 'Confusion Log Probability' (CLP) as a metric to evaluate OOD detection methods, which measures the similarity of inlier / outlier dataset pairs. Using this metric, we show that the proposed method improves OOD detection in both near and far OOD settings, but especially in near OOD settings. See Figure 1 for an overview.

2 Related Work

Representations from classification networks. The vast majority of OOD detection methods use scores derived from models trained only with multi-class supervision. Hendrycks & Gimpel [11] propose using the maximum softmax probability (MSP) to score OOD samples, which was further improved in ODIN [19] by using temperature scaling and input pre-processing. Lee et al. [18] utilize standard Gaussian density estimates of the network's class-conditional intermediate activations. Sastry & Oore [31] showed improvements on far OOD detection by using Gram matrices from multiple feature maps.

Alternative training strategies. Beyond proposing better scoring functions, another area of research is adapting the training strategy to improve the quality of scores . MSP scoring was improved by using strategies like confidence loss [17], auxiliary objectives [6, 13, 23], margin loss [34] and outlier exposure [12]. A multi-head network architecture was used by Shalev et al. [33] to improve the intermediate feature map for OOD detection. Similarly, an approach for novelty detection based on metric learning was presented by Masana et al. [22]. Most of the above approaches make the assumption of access to OOD samples during the training process to enhance performance.

Bayesian approaches. Under the Bayesian paradigm, Blundell et al. [2], Malinin & Gales [21], Chen et al. [4] showed that model uncertainty estimates can be produced by learning distributions over network weights and Gal & Ghahramani [7] proposed Monte-Carlo dropout sampling for it.

Generative and hybrid models. An intuitive strategy for detecting OOD samples is to train a generative model from which one can compute the likelihood as an OOD score. An ensemble approach was adapted by Choi et al. [5] and likelihood ratios were estimated as an OOD scoring metric in [30]. While generative models are a promising avenue for OOD detection, applying them directly to the image space has not achieved state of the art results, even when combined with a classification network [26]. Concurrent to our work, Zhang et al. [36] show in a surprising empirical finding that if a residual flow network is attached to the penultimate layer of a classification network, and the networks are trained together, the $p(\mathbf{z})$ learned by the flow network is able to detect near OOD samples much better than baselines in the open set recognition field.

3 Proposed Method

As shown in Figure 2, training using only supervised classification losses may not produce the required features for identifying OOD samples. Contrastive training [1, 3, 9, 10] provides a remedy to this problem, by learning a representation capable of distinguishing between all individual training samples, while incorporating existing prior knowledge about identity-preserving transformations.

For image classification, camera parameter and illumination are obvious variations. Both can be approximated by translating, scaling and rotating the image, as well as applying brightness and contrast transformations. Intuitively, the contrastive loss moves augmented versions of the same image closer together in the feature space whilst pushing all other image pairs apart [3].

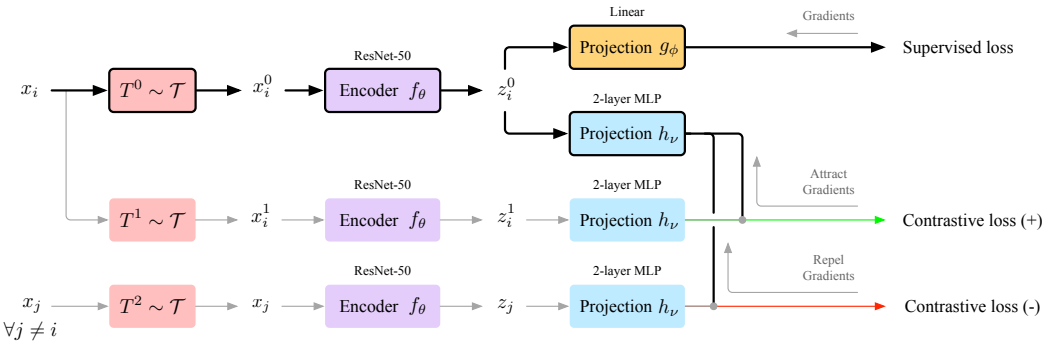


Figure 3: Schematic description of the multitask approach. $\mathbf{x}_i, \mathbf{x}_j$: training images. T: image transformation (cropping, brightness, etc.). f_θ: encoder network. $\mathbf{z}$: image represented in latent space. g_ϕ: projection to k classes. h_ν: projection to lower-dimensional embedding space.

Architecture. Our proposed architecture (Figure 3) for contrastive training is based on SimCLR [3], chosen for its simplicity. It is composed of an encoder network f_θ, followed by two projection heads g_ϕ and h_ν. g_ϕ maps to the class predictions and h_ν maps to the low dimensional embedding over which we define the contrastive loss. The desired feature space is learned on in-distribution training samples only. For a batch of images $\{\mathbf{x}_i\}_{i=1\ldots N}$, we define $\mathbf{x}_i^0 = T^0(\mathbf{x}_i)$ and $\mathbf{x}_i^1 = T^1(\mathbf{x}_i)$, where T^0 and T^1 denote two explicit transformations (such as crop-resize or color distortions) selected randomly from a set of transformations $\mathcal{T}$. The representations of the transformed images are given as $\mathbf{z}_i^0 = f_\theta(\mathbf{x}_i^0)$ and $\mathbf{z}_i^1 = f_\theta(\mathbf{x}_i^1)$ where f_θ is an encoder, a deep network with parameters θ.

The representation is learned by solving a relational classification task using cosine similarity in some embedding space. The cosine similarity between any two vectors $\mathbf{u}$, $\mathbf{w}$ is given by $\mathrm{sim}(\mathbf{u}, \mathbf{w}) = \mathbf{u}^\top \mathbf{w}/(\|\mathbf{u}\|\|\mathbf{w}\|)$. A 2-layer MLP h_ν maps a sample from the representation space to a lower-dimensional embedding space $\{\hat{\mathbf{z}}_i\}$, i.e. $\hat{\mathbf{z}}_i^0 = h_\nu(\mathbf{z}_i^0)$, and $\hat{\mathbf{z}}_i^1 = h_\nu(\mathbf{z}_i^1)$.

Objective. In this embedding space, the loss function aims to maximize the cosine similarity of sample pairs originating from the same image, i.e. $\mathrm{sim}(\hat{\mathbf{z}}_i^0, \hat{\mathbf{z}}_i^1) \to 1$, whilst minimizing all other pairs originating from two different images, i.e. $\mathrm{sim}(\hat{\mathbf{z}}_i^a, \hat{\mathbf{z}}_j^b) \to 0$, with $j \in \{1, \ldots, N\} \setminus i$ and $a, b \in \{0, 1\}$. The contrastive loss for sample i is then defined as:

$$L_{\text{con},i} = \sum_{a \in \{0,1\}} -\log \frac{\exp\left(\mathrm{sim}(\hat{\mathbf{z}}_i^a, \hat{\mathbf{z}}_i^{1-a})/\tau\right)}{\sum_{j \in \{1,\ldots,N\}} \exp\left(\mathrm{sim}(\hat{\mathbf{z}}_i^a, \hat{\mathbf{z}}_j^{1-a})/\tau\right) + \sum_{j \in \{1,\ldots,N\} \setminus i} \exp\left(\mathrm{sim}(\hat{\mathbf{z}}_i^a, \hat{\mathbf{z}}_j^a)/\tau\right)}, \quad (1)$$

where τ is a positive temperature parameter. The projection head g_ϕ consists of a linear transformation mapping the representation space to k logits for the k in-distribution classes. The logits are trained with a standard softmax cross-entropy loss L_{class}.

Training is performed in two stages. The first stage consists of using solely the contrastive loss L_{con} for a large number of epochs to help learn a good representation. In the second stage, we optimize the combined loss $L_{\text{con}} + \lambda L_{\text{class}}$ for a smaller number of epochs, where λ is the supervised loss weight.

Density estimation. We detect OOD samples using a method analogous to Lee et al. [18], by fitting Gaussian distributions to the activations on the training data, which we shape in two significant ways. First, the contrastive loss encourages the network to encode all features capable of distinguishing between samples rather than only those necessary to discriminate between classes. Second, to simplify the distribution of the activation that is fitted, label smoothing is added to the cross-entropy loss L_{class}, so as to prevent the network from spreading out the activations in an attempt to drive the logits of the correct class to infinity. This encourages tight clustering of the activations of each class, as demonstrated by Müller et al. [24].

To take advantage of this last property, our density estimation is performed class-wise, over the activations $\mathbf{z}$ at the penultimate layer. For each class c, we estimate an n-dimensional multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, with n the dimension of $\mathbf{z}$. For the OOD score $s(\mathbf{x})$, the highest density is taken over all the class-conditional Gaussian components. A high score $s(\mathbf{x})$ indicates that the representation of a test sample in the embedding space lies close to the typical set for one of the

classes. Conversely, a low score signifies that the test sample has a representation that is far from all training set examples and is therefore likely to represent an OOD example.

$$s(\mathbf{x}) = \max_c \left[-(f_\theta(\mathbf{x}) - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (f_\theta(\mathbf{x}) - \boldsymbol{\mu}_c) - \log\left((2\pi)^n \det \boldsymbol{\Sigma}_c\right) \right], \tag{2}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are obtained using the standard estimators. Unlike Lee et al. [18], we do not perform ensembling over detectors based on different layers, since determining the optimal linear combination would require access to labelled OOD data. We also do not perform input preprocessing [19]. Instead, we rely on having a richer representation over which to define our distribution.

4 Confusion Log Probability (CLP) as a Measure of Dataset Distance

Real world test samples may vary strongly, from falling exactly within the training distribution to falling far out of the training distribution. As a consequence, a robust model should exhibit strong OOD detection performance across the entirety of the spectrum. Current benchmarks however report only a single performance statistic for each in-distribution and OOD dataset pair, such as the area under the receiver operating characteristic curve. This type of evaluation metric can therefore not resolve the performance in detecting near versus far OOD.

Different proposals toward discerning the difficulty of OOD tasks have been made in the literature. The openness score [32], used in the Open Set Recognition community, gives a difficulty measure based on the number of classes in the training set compared to the number of classes in the test set. This measure, however, ignores the visual similarity that can exist between classes, and therefore the relative difficulty of detecting different unknown classes. Another example is the usage of the maximum mean discrepancy with an L^2 distance kernel on the image space [19]. However, applying the L^2 distance directly in the image space can only identify nearly identical images. Therefore this metric cannot completely assess the spectrum of OOD cases which we seek.

We propose the **confusion log probability (CLP)** as a measure that is indicative of the difficulty of an OOD detection task. CLP is based on the probability with which a classifier confuses outliers with inliers, that has access to outlier samples during training. Given two labelled datasets $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$, with corresponding sets of classes $\mathcal{C}_{\text{in}}$ and $\mathcal{C}_{\text{out}}$, we train an ensemble of N_e classifiers $\{\hat{p}^j\}_{j=1}^{N_e}$ on the joint dataset $\mathcal{D} = \mathcal{D}_{\text{in}} \cup \mathcal{D}_{\text{out}}$ using the extended label set $\mathcal{C} = \mathcal{C}_{\text{in}} \cup \mathcal{C}_{\text{out}}$. Once the ensemble is trained, we compute an estimate for the confusion matrix between classes on held-out test data. The expected probability of a test sample $\mathbf{x}$ to be predicted as class k is given by:

$$c_k(\mathbf{x}) = \frac{1}{N_e} \sum_{j=1}^{N_e} \hat{p}^j(\hat{y} = k|\mathbf{x}). \tag{3}$$

Therefore the confusion of a set of test OOD samples $\mathcal{D}_{\text{test}}$ with the inlier classes $\mathcal{C}_{\text{in}}$, i.e. the confusion log probability (CLP) of $\mathcal{D}_{\text{test}}$, becomes:

$$\text{CLP}_{\mathcal{C}_{\text{in}}}(\mathcal{D}_{\text{test}}) = \log\left(\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{\mathbf{x} \in \mathcal{D}_{\text{test}}} \sum_{k \in \mathcal{C}_{\text{in}}} c_k(\mathbf{x}) \right). \tag{4}$$

A low CLP indicates that test samples are far OOD and a high CLP indicates that they are near OOD. Note that summing over the probability for all inlier classes $\mathcal{C}_{\text{in}}$ effectively evaluates the binary classification problem of inlier versus outlier classes on outlier samples. As such CLP is asymmetric.

We can compute a class-wise CLP with $\mathcal{D}_{\text{test}}$ being the test samples of this specific class, or a dataset CLP by using all samples of the test dataset as $\mathcal{D}_{\text{test}}$. The class-wise CLP is used in Fig. 1, where we show how OOD detection performance varies as a function of the distance between the CIFAR-100 classes and the inlier CIFAR-10 dataset. By using classifiers to estimate class confusions (5 ResNet-34 models in our work), we ground the measure in a notion of visual similarity rather than semantic or image space similarity. The choice of an ensemble of independently trained classifiers is motivated by their well-calibrated predictions over single classifiers [16]. Further implementation details of the ensemble training setup to compute CLP and further qualitative analysis using dendrograms can be found in Appendix D.

Table 1: Out-of-distribution detection performance (AUROC).

Method	**Near OOD** $\mathcal{D}_{\text{in}}$ = CIFAR-100 $\mathcal{D}_{\text{out}}$ = CIFAR-10 CLP= [-4.5 to -2.6]	**Near & Far OOD** $\mathcal{D}_{\text{in}}$ = CIFAR-10 $\mathcal{D}_{\text{out}}$ = CIFAR-100 CLP=[-7.4 to -0.8]	**Far OOD** $\mathcal{D}_{\text{in}}$ = CIFAR-10 $\mathcal{D}_{\text{out}}$ = SVHN CLP=[-12.1 to -7.6]	**Average**
Softmax probs. [11]	77.1	86.4	89.9	84.5
ODIN [19]*	77.2	85.8	96.7	86.6
Mahalanobis [18]*	77.5	88.2	99.1	88.3
Residual flows [38]*	77.1	89.4	99.1	88.5
Outlier exposure [12]†	75.7†	**93.3**†	98.4†	89.1†
Rotation pred. [13]‡	-	90.9	98.9	-
Gram matrix [31]	67.9	79.0	**99.5**	82.1
Ours	**78.3**	92.9	**99.5**	**90.2**

* Uses data explicitly labeled as out-of-distribution for tuning
† Uses data explicitly labeled as out-of-distribution for training ‡ Uses additional data for pretraining

5 Experiments

Near to far OOD spectrum benchmark. We study out-of-distribution detection on the following in-distribution dataset ($\mathcal{D}_{\text{in}}$) and out-of-distribution dataset ($\mathcal{D}_{\text{out}}$) pairs which we believe represent the most challenging pairs of common OOD detection benchmarks [11]: We use the CIFAR-10 and CIFAR-100 [15] datasets, as well as the Street View House Numbers (SVHN) dataset [27]. Note that the CIFAR-10 and CIFAR-100 classes are mutually exclusive. The distance of the dataset pairs is given by the min-max bounds of the class-wise CLP.

Evaluation metrics. As is common in the OOD detection literature, we use the area under the receiver operating characteristic curve (AUROC), with in-distribution and out-of-distribution being the labels, as the metric for OOD detection performance. Note that this is a threshold-independent and calibration-free evaluation metric. Results with additional metrics can be found in Appendix E. In addition to AUROC, we also quantify the performance on a single OOD sample using the *OOD rank*. For estimating this, we compare the score of the OOD sample $s(\mathbf{x})$ to the scores of the inlier test dataset. The OOD rank is given by the percentage of inlier test examples that have a lower $s(\mathbf{x})$ than the OOD sample. The higher the rank, the more the sample is deemed to be OOD.

Setup. We adopt a wide ResNet-50 [8] as the encoder f_θ, whose last layer has a fixed-dimensional output (6144-D). This output vector $\mathbf{z}$ is the representation on which we compute the OOD score $s(\mathbf{x})$ for a given test sample $\mathbf{x}$. $\mathbf{z}$ is followed by the supervised head g_ϕ with label smoothing, and the contrastive head h_ν with a 128-D 2-layer MLP projection with batch normalization and ReLU. For fair comparison, we use the same architecture, the same transformation function $\mathcal{T}$ and scoring function for the baselines without contrastive training reported in Table 6. Further training setup details are provided in Appendix B.

Results. Table 1 shows our main results. We observe that our proposed method improves OOD detection across the spectrum of our proposed benchmark. Specifically, on the near OOD dataset pair with CIFAR-100 as $\mathcal{D}_{\text{in}}$ and CIFAR-10 as $\mathcal{D}_{\text{out}}$, we obtain an AUROC of 78.3 which is, to the best of our knowledge, a new state-of-the-art result, outperforming [18]. With CIFAR-10 as $\mathcal{D}_{\text{in}}$ and CIFAR-100 as $\mathcal{D}_{\text{out}}$, containing both test samples that are near as well as far OOD, we obtain an AUROC of 92.9, which is more than two points better than the next best method not using labeled out-of-distribution data during training. Figure 4a shows contrastive training helps to differentiate OOD classes that are highly similar to inlier classes, where the baseline results is worse than random performance. In the far OOD setting (Figure 4b), contrastive training is also a significant component for further separation between $\mathbf{z}$ of dissimilar images and inlier dataset images. The performance gap is the largest for the high CLP regime (+18 AUROC points for $-5 \leq \text{CLP} < 0$), and remains for the whole spectrum (Figure 4c). On the far OOD dataset pair with CIFAR-10 as $\mathcal{D}_{\text{in}}$ and SVHN as $\mathcal{D}_{\text{out}}$, we obtain an AUROC of 99.5 which is on par with the current state-of-the-art [31]. Results on additional dataset pairs are reported in Appendix E.

Table 2: **Ablation study of objective function.** The baseline model only performs supervised training and we evaluate the impact of incorporating label smoothing (LS) and contrastive training (CT). We report AUROC, as well as the standard deviation of the OOD rank between 5 runs.

Training strategy		AUROC			OOD rank Average standard deviation		
LS	CT	$\mathcal{D}_{\text{in}}$ = CIFAR-100 $\mathcal{D}_{\text{out}}$ = CIFAR-10	CIFAR-10 CIFAR-100	CIFAR-10 SVHN	CIFAR-100 CIFAR-10	CIFAR-10 CIFAR-100	CIFAR-10 SVHN
x	x	63.9 ± 0.3	81.1 ± 0.2	96.7 ± 0.2	23.3	20.7	6.7
✓	x	74.1 ± 0.4	90.8 ± 0.1	99.2 ± 0.1	20.3	12.8	2.0
x	✓	72.1 ± 0.4	90.9 ± 0.2	98.8 ± 0.2	23.7	12.5	2.4
✓	✓	78.3 ± 0.2	92.9 ± 0.2	99.5 ± 0.1	19.3	10.6	1.7

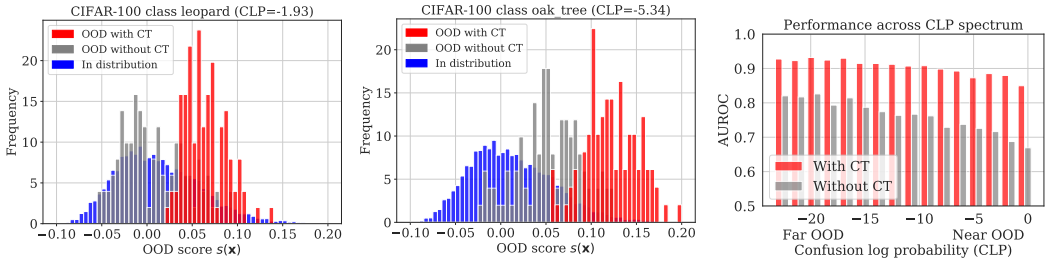
When considering the average performance on all three dataset pairs as representative of the entire near to far OOD spectrum, our method obtains an AUROC of 90.2 outperforming the previous state-of-the-art method [12]. It is worth pointing out that, unlike others, our approach *does not* assume access to additional data or labels from an outlier distribution during training or tuning.

Impact of activation space shaping. We perform an ablation study to investigate the impact of label smoothing and contrastive training, with results reported in Table 6. We report results at the task level, but also investigate the variation on a per-sample basis, by looking at the OOD rank variation between several randomly initialized runs. This allows us to quantify whether the OOD samples considered in distribution always remain the same or whether the errors vary.

We observe that while using contrastive training or label smoothing alone leads to improvements over the supervised baseline, significantly better results are obtained when the two of them are combined together to shape the activation space. We hypothesize that scoring with standard Gaussian density estimation benefits significantly from tighter class clusters obtained via label smoothing, and it is required for the combination of contrastive training and the scoring function as seen in Eq. 2 to work effectively. Furthermore, our feature shaping strategies allow for a significant reduction in variation of the OOD rank of samples between runs. We hypothesize that for the supervised baseline, only features required for class discrimination will be created while the presence of task-agnostic features useful for near OOD detection will only be present by chance due to randomness in training.

More details on additional ablation tests and discussions are available in Appendix F, where we study the impact of parameters like relative weighting between supervised and contrastive loss, temperature and model capacity on OOD detection performance.

Failure mode analysis. Figure 5 shows failure cases for the baseline and our model. We show samples from $\mathcal{D}_{\text{out}}$ = CIFAR-100 most likely to be *incorrectly* predicted as in-distribution for networks trained on $\mathcal{D}_{\text{in}}$ = CIFAR-10. The percentile rank of the OOD score w.r.t. the scores of the inlier dataset indicates the success of the OOD detection (higher is better). For most of the baseline's worst mistakes, our approach succeeds in identifying that sample as an outlier.



(a) Near OOD (b) Far OOD (c) CLP Spectrum

Figure 4: Experimental results with contrastive training (CT) under the $\mathcal{D}_{\text{in}}$ = CIFAR-10, $\mathcal{D}_{\text{out}}$ = CIFAR-100 setting, showing the histograms of OOD scores $s(\mathbf{x})$ with respect to those of a fixed inlier dataset for **(a)** a near OOD class and **(b)** a far OOD class. **(c)**: Out-of-distribution detection performance across the sample-wise CLP spectrum.

(a) Worst mistakes by baseline (BL)

(b) Worst mistakes by contrastive training (CT)

Figure 5: **Failure mode analysis.** $\mathcal{D}_{\text{in}}$ = CIFAR-10; $\mathcal{D}_{\text{out}}$ = CIFAR-100; **GT**: CIFAR-100 ground truth label. **BL**: baseline, **CT**: contrastive training. The numbers show the percentile rank of the OOD score. In brackets the class of the most activated inlier Gaussian. Green: percentile rank > 50%, indicating relative success of a method at detecting that input as an outlier.

6 Discussion and Conclusion

In this paper, we proposed a simple contrastive training-based approach to OOD detection that outperforms recent methods across a variety of settings. Furthermore we introduced a new metric to quantify the difficulty of a specific OOD task, the confusion log probability (CLP), that captures the similarity of inlier and outlier datasets. We showed with extensive experiments that representations obtained through contrastive training improve OOD detection performance beyond what is possible with purely supervised training. The representations are shaped by joint training, in which the contrastive loss pushes the representations apart, even within each class, while the supervised loss acts to cluster the representations by class. Unlike previous competitive methods our approach does not require access to data from the outlier distribution during training or tuning. Moreover, in contrast to many other representation learning approaches, the underlying SimCLR [3] has a well-motivated training objective, and can scale to large images and datasets.

Of course this additional training objective can not guarantee that all necessary features for near OOD classes are found. See Figure 5 for a failure mode analysis for the setting $\mathcal{D}_{\text{in}}$ = CIFAR-10 and $\mathcal{D}_{\text{out}}$ = CIFAR-100. Here the model saw 'automobiles' and 'trucks' during training, but it was not able to distinguish the new classes 'pickup truck' and 'bus' from the existing ones. These class-pairs are however also the most challenging for fully supervised training as CLP shows.

This work only examines improvements that arise from training a richer representation $\mathbf{z}$. With regards to scoring of $\mathbf{z}$, we employed standard Gaussian density modeling like prior methods [14]. In concurrent work, Zhang et al. [36] show that scoring with a deep flow-based network can also significantly improve performance. We expect that these improvements in density estimation are complementary to our proposed contrastive training.

We use a larger neural network than is typical in the literature. This capacity is needed to encode the additional features that SimCLR training creates. We have shown that the Mahalanobis approach [14] does not benefit from additional model capacity. For other baselines we can only speculate that the authors tried to add more capacity, and published the network that yielded best performance.

Another advantage of our setup is the ability to extract useful information from completely unlabelled images that can come from arbitrary distributions. Other approaches that learn a density directly from the training set (e.g. Zhang et al. [36]) rely on the fact that all training examples come only from the 'in-distribution' set. Especially in real-world applications, for instance medical imaging, a large set of unlabelled images is easily available from routine imaging. Removing outlier images from this set would be an expensive manual labelling task.

All in all, this work demonstrates that the challenges in OOD detection (especially for the near OOD setting) are closely related to those in unsupervised representation learning. We believe that viewing the OOD detection problem from this perspective opens up new avenues for progress.

Broader Impact

Deep neural networks have demonstrated superhuman performance across a wide range of applications, with the promise of significant positive impact on the world. Despite this, our ability to safely deploy models in real world settings is still limited. One such obstacle is the inability of existing models to accurately withhold prediction for an input that is meaningfully different from typical examples encountered during training. In contrast to human experts, deep neural network based systems tend to struggle to account for the uncertainties for taking an appropriate decision, e.g. refer a difficult case for a second opinion. Moreover, models can fail in unexpected ways, making errors difficult to identify in practice. In domains with the greatest potential for societal impact, such as medical imaging and self-driving cars, appropriate recognition of OOD inputs is essential to avoid catastrophic errors that may cause harm.

This work proposes a new scalable approach to OOD detection based on recent advancements in representation learning, and takes a step towards the ultimate goal of enabling safe real-world deployment of machine learning models in safety-critical domains. Additionally, this work also defines and recognizes the importance of evaluation of OOD detection performance in a spectrum of regimes, in particular near and far OOD settings. The experimental results have been reported on standard benchmark datasets for considerations of reproducible research, but both the OOD detection method and the evaluation approach are addressing potential issues encountered in safety critical domains. For example, in medical imaging, pathologies, poor-quality images and other artefacts that have not previously been encountered in model training are both common and important. Progress in methods for near OOD detection is therefore particularly relevant to addressing the often subtle but significant challenges of operating safely in a real world environment.

References

[1] Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, 2019.

[2] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

[3] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

[4] Chen, W., Shen, Y., Jin, H., and Wang, W. A variational dirichlet framework for out-of-distribution detection. *arXiv preprint arXiv:1811.07308*, 2018.

[5] Choi, H., Jang, E., and Alemi, A. A. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

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

[7] Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

[8] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[9] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[10] Hénaff, O. J., Srinivas, A., Fauw, J. D., Razavi, A., Doersch, C., Eslami, S. M. A., and van den Oord, A. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, 2020.

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

[12] Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[13] Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pp. 15637–15648, 2019.

[14] Kamoi, R. and Kobayashi, K. Why is the mahalanobis distance effective for anomaly detection? *arXiv preprint arXiv:2003.00402*, 2020.

[15] Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

[16] Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

[17] Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[18] Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

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

[20] Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[21] Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

[22] Masana, M., Ruiz, I., Serrat, J., van de Weijer, J., and Lopez, A. M. Metric learning for novelty and anomaly detection. In *British Machine Vision Conference*, 2018.

[23] Mohseni, S., Pitale, M., Yadawa, J., and Wang, Z. Self-supervised learning for generalizable out-of-distribution detection. In *Association for the Advancement of Artificial Intelligence*, 2020.

[24] Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4696–4705, 2019.

[25] Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

[26] Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Hybrid models with deep and invertible features. In *International Conference on Machine Learning*, 2019.

[27] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems*, 2011.

[28] Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

[29] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

[30] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pp. 14680–14691, 2019.

[31] Sastry, C. S. and Oore, S. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, 2020.

[32] Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., and Boult, T. E. Toward open set recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012.

[33] Shalev, G., Adi, Y., and Keshet, J. Out-of-distribution detection using multiple semantic label representations. In *Advances in Neural Information Processing Systems*, pp. 7375–7385, 2018.

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

[35] You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[36] Zhang, H., Li, A., Guo, J., and Guo, Y. Hybrid models for open set recognition. *arXiv preprint arXiv:2003.12506*, 2020.

[37] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[38] Zisselman, E. and Tamar, A. Deep residual flow for novelty detection. *arXiv preprint arXiv:2001.05419*, 2020.

A Detailed Performance Across CLP Spectrum

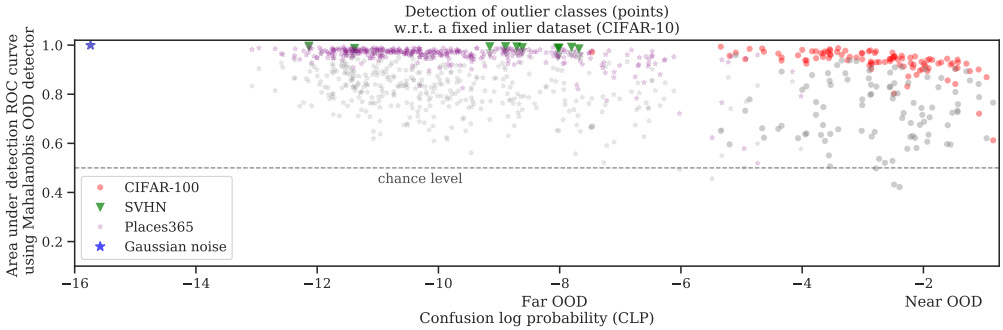


Figure 6: Each point represents performance at detecting one of the classes in CIFAR-100, SVHN, Places365 and Gaussian noise as outliers with respect to a network trained on CIFAR-10. Colored and gray markers correspond to performance from our best model with contrastive training, and our best model without contrastive training respectively. The classes are sorted by increasing similarity to the inlier classes as given by the class-wise confusion log probability (CLP). Contrastive training improves OOD detection results across the board, particularly in the near OOD regime, where the outlier and inlier classes are highly similar.

In Fig. 6, we show the class-wise OOD detection performance for the full CLP spectrum with and without contrastive training with $\mathcal{D}_{\text{in}}$ being CIFAR-10 and $\mathcal{D}_{\text{out}}$ being the combination of CIFAR-100, SVHN, Places365 and Gaussian noise test samples. Details about the training process for CLP computation are given in Sec. C.

B Implementation Details

We use a Resnet-50 with a $3\times$ width multiplier for all experiments. We first pretrain the model with a batch size of 2048 for 1000 epochs using only the contrastive loss and then finetune with a joint supervised and contrastive loss for 100 epochs in case of CIFAR-10 and 200 epochs for CIFAR-100. We use a supervised loss multiplier of $\lambda = 100$ during the finetuning stage. The models are trained using the LARS [35] optimizer with momentum 0.9 and weight decay 1×10^{-6}. Furthermore, we use an initial learning rate of 1.0, with a linear warmup for the first 30 epochs followed by a cosine decay schedule without restarts following [20] for both stages. For label smoothing, we use $\alpha = 0.01$ for CIFAR-10 as $\mathcal{D}_{\text{in}}$ and $\alpha = 0.1$ for CIFAR-100 as $\mathcal{D}_{\text{in}}$.

The data augmentation operation T as described in Section 3 follows [3], which is a sequence of random cropping followed by a random left-right flip and random color distortion.

C CLP Training Details

For computing CLP scores, we trained an ensemble of five ResNet-34 model instances. In order to accurately estimate class confusions, we would ideally like to train classifiers on a dataset that is representative of all possible images. As an approximation, we independently train each of the model instances on the union of five datasets: CIFAR-10, SVHN, CIFAR-100, Places365 and independent Gaussian noise. The entire collection has 486 classes. We used a training batch size of 1024, and ensured that examples from each of the 486 classes are uniformly sampled in a batch. We used SGD with momentum of 0.9 for training the models. The models were trained for 500 epochs with a cosine decay learning rate schedule initialized at 0.2. A weight decay parameter of 10^{-6} was used for regularization. Each of the model instances only differ in the random initializations of the weights.

Once all the model instances are trained, we use them to compute CLP between any given dataset pair. As a specific example, let us consider the case where the inlier dataset $\mathcal{D}_{\text{in}}$ is CIFAR-10 and the outlier dataset $\mathcal{D}_{\text{out}}$ is Places365, with 10 and 365 classes respectively. To calculate the CLP score of Places365 (with respect to CIFAR-10), we compute, for each of the 5 model instances, the

softmax output for all the test examples in Places365. The outputs of the model ensemble instances are averaged to have a 1×486 vector output, where 486 is the total number of classes (in the union of all datasets). We compute the total probability of the 10 outputs corresponding to CIFAR-10 classes as $\mathcal{D}_{\text{in}}$ and report the log of this probability as CLP as an estimate of confusing a Places365 example with a CIFAR-10 example.

The confusion matrix of the our model is shown in Fig. 7, by combining all the classes of a given dataset. The rows indicate the true labels and columns indicate the predictions. We observe that CIFAR-10 and CIFAR-100 confuses more among themselves compared to datasets: Places365, SVHN and Gaussian noise, which are far OOD to them. Also note that these datasets have almost 100% classification accuracy. Thus the relative ordering of these far OOD datasets from in-distribution CIFAR-10/100 is only as good as the state-of-the art in calibrated uncertainty.

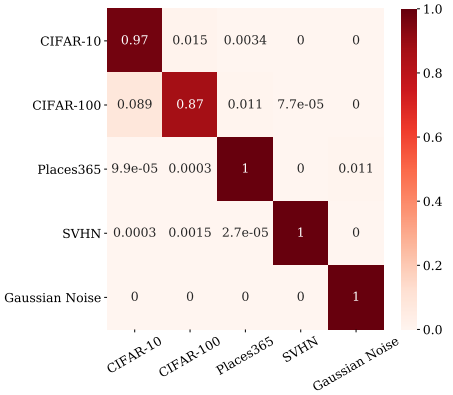


Figure 7: Confusion matrix of the classifier ensemble using the combination of the five reported datasets. Rows and columns represent the true labels and the predictions respectively.

D Qualitative Analysis of CLP

Visualization of Class Similarities

To qualitatively ascertain our CLP estimates, we use the model jointly trained on CIFAR-10, CIFAR-100, SVHN, Places365 and Gaussian noise resulting in a 486-way classifier. We compute a confusion probability for each pair of classes i, j as

$$u_{i \to j} = \frac{1}{|\mathcal{D}_{\text{test},j}|} \sum_{\mathbf{x} \in \mathcal{D}_{\text{test},j}} c_i(\mathbf{x}) \ , \tag{5}$$

using the expected probability $c_i(\mathbf{x})$ of a test sample to be predicted as class i (Eq. (3) in the main paper), and the test set for class j, $\mathcal{D}_{\text{test},j}$. We translate the probabilities into symmetric distances using

$$d(i,j) = \sqrt{-\log\left(\frac{1}{2}\left(u_{i \to j} + u_{j \to i}\right)\right)} \ . \tag{6}$$

We then use these pairwise distances to perform hierarchical agglomerative clustering (with 'average' linkage). Figure 8 shows the dendrogram. We observe that the relationships captured by this distance measure are a good representation of the visual similarity of the classes: all the SVHN classes (shown in blue) are well separated from the CIFAR-10 and CIFAR-100 classes indicating that it is a far OOD dataset. CIFAR-10 (shown in red) and CIFAR-100 classes (shown in green) on the other hand are frequently confused, indicating that the two datasets are closer to one another than to SVHN. The Places365 dataset builds a separate cluster (shown in purple). Surprisingly the CIFAR-100 classses 'orange' and 'apple' are quite dissimilar to the remaining CIFAR-100 and CIFAR-10 classes and get clustered closer to the house numbers.

The individual classes within the data sets that get clustered together match very well our impression of "visual similarity". We take this as another confirmation that our proposed CLP metric can be used to quantify the difficulty of OOD tasks.

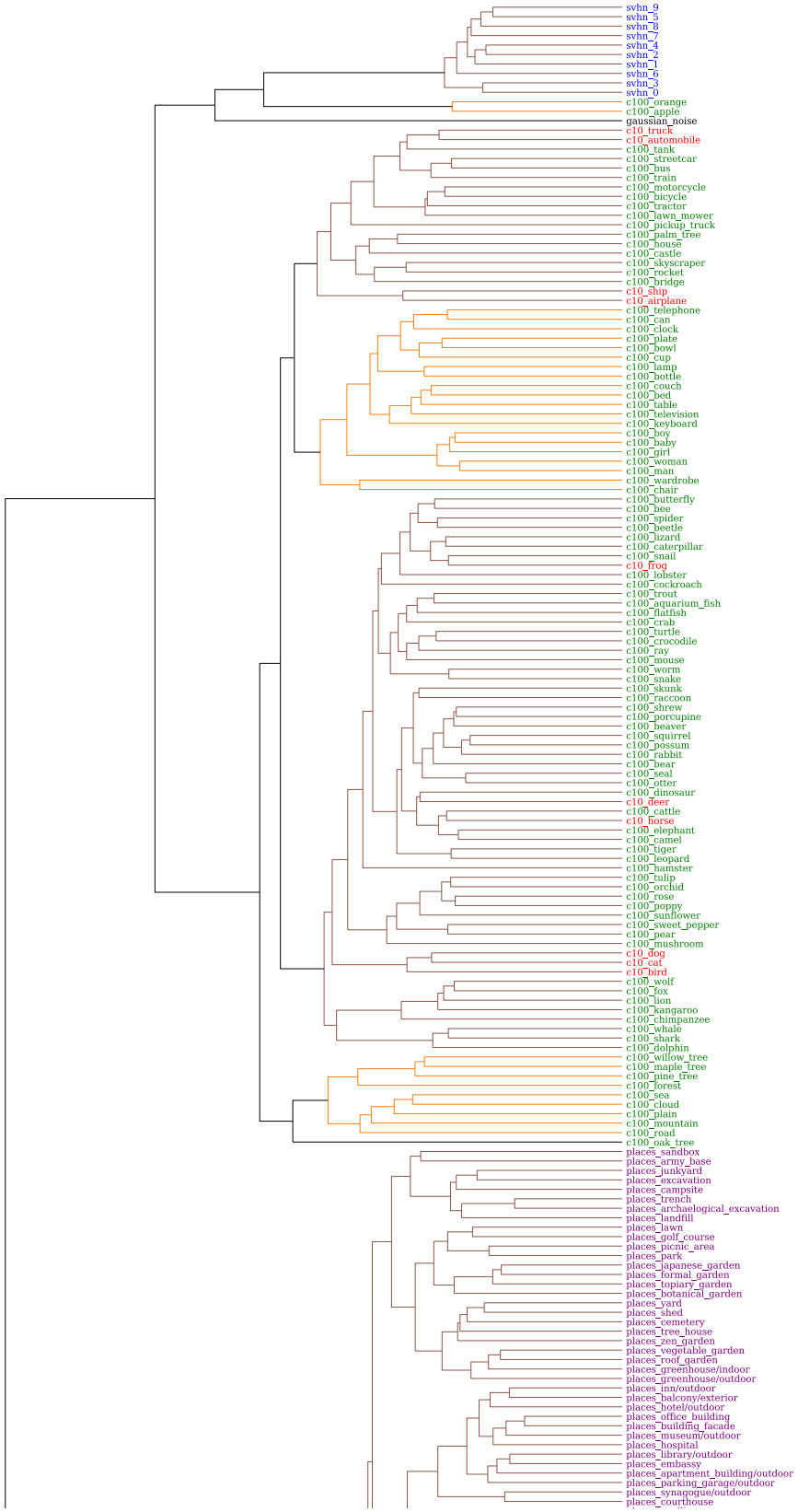
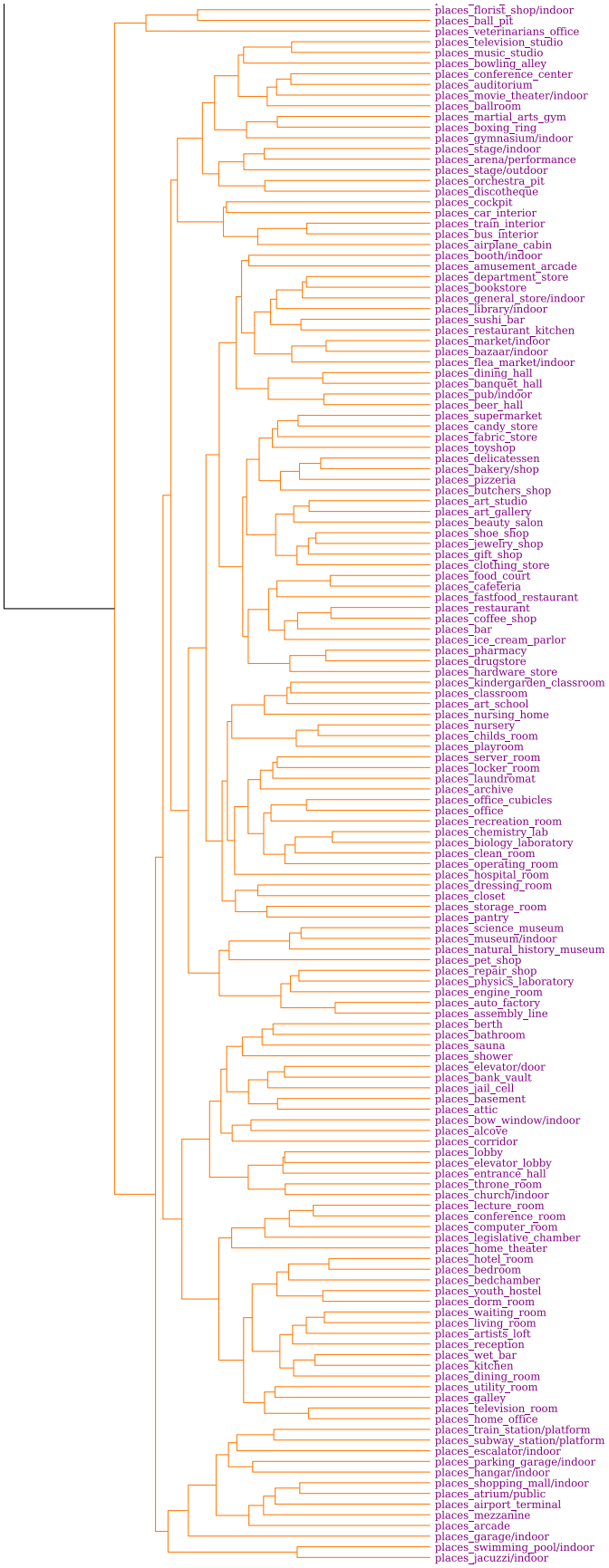


Figure 8: **Qualitative Analysis of CLP.** Three-part dendrogram plot of the classes of CIFAR-10 (red), CIFAR-100 (green), SVHN (blue), Places365 (purple) and Gaussian noise (black) based on the expected confusion matrix combining the datasets.


places_pavilion
places_gazebo/exterior
places_beach_house
places_residential_neighborhood
places_manufactured_home
places_house
places_cottage
places_schoolhouse
places_mansion
places_oast_house
places_barn
places_hunting_lodge/outdoor
places_cabin/outdoor
places_chalet
places_water_tower
places_industrial_area
places_skyscraper
places_downtown
places_palace
places_church/outdoor
places_castle
places_tower
places_mausoleum
places_mosque/outdoor
places_temple/asia
places_pagoda
places_viaduct
places_aqueduct
places_village
places_ruin
places_kasbah
places_moat/water
places_canal/natural
places_boathouse
places_canal/urban
places_lock_chamber
places_dam
places_pier
places_harbor
places_promenade
places_bridge
places_boardwalk
places_railroad_track
places_amphitheater
places_highway
places_boat_deck
places_ski_resort
places_ice_skating_rink/outdoor
places_igloo
places_water_park
places_swimming_pool/outdoor
places_fountain
places_arch
places_plaza
places_campus
places_courtyard
places_playground
places_amusement_park
places_slum
places_construction_site
places_porch
places_patio
places_balcony/interior
places_gas_station
places_fire_station
places_motel
places_loading_dock
places_hangar/outdoor
places_street
places_crosswalk
places_parking_lot
places_driveway
places_medina
places_alley
places_fire_escape
places_ticket_booth
places_phone_booth
places_doorway/outdoor
places_barndoor
places_market/outdoor
places_bazaar/outdoor
places_carrousel
places_beer_garden
places_restaurant_patio
places_diner/outdoor
places_staircase
places_elevator_shaft
places_catacomb
places_burial_chamber
places_stable
places_kennel/outdoor
places_shopfront
places_general_store/outdoor
places_garage/outdoor
places_bus_station/indoor
places_auto_showroom
places_ice_skating_rink/indoor
places_basketball_court/indoor
places_arena/hockey
places_runway
places_heliport
places_landing_deck
places_airfield
places_raceway
places_oilrig
places_soccer_field
places_football_field
places_baseball_field
places_athletic_field/outdoor
places_stadium/soccer
places_stadium/football
places_stadium/baseball
places_volleyball_court/outdoor
places_bullring
places_arena/rodeo
places_racecourse
places_corral
places_underwater/ocean_deep
places_grotto
places_aquarium
places_mountain_path
places_desert/vegetation
places_tundra
places_valley
places_mountain
places_cliff
places_butte
places_badlands
places_rock_arch
places_canyon
places_field_road
places_desert_road
places_forest_road
places_forest_path
places_hayfield
places_vineyard
places_field/cultivated
places_wheat_field
places_corn_field
places_pasture
places_farm
places_rice_paddy
places_tree_farm
places_orchard
places_field/wild
places_rainforest
places_forest/broadleaf
places_bamboo_forest
places_hot_spring
places_creek
places_river
places_lake/natural
places_pond
places_swamp
places_marsh
places_watering_hole
places_rope_bridge
places_waterfall
places_swimming_hole
places_raft
places_fishpond
places_mountain_snowy
places_glacier
places_snowfield
places_ski_slope
places_ice_shelf
places_ice_floe
places_crevasse
places_wave
places_iceberg
places_islet
places_lagoon
places_beach
places_ocean
places_coast
places_volcano
places_sky
places_desert/sand
places_windmill
places_lighthouse
places_wind_farm


places_florist_shop/indoor
places_ball_pit
places_veterinarians_office
places_television_studio
places_music_studio
places_bowling_alley
places_conference_center
places_auditorium
places_movie_theater/indoor
places_ballroom
places_martial_arts_gym
places_boxing_ring
places_gymnasium/indoor
places_stage/indoor
places_arena/performance
places_stage/outdoor
places_orchestra_pit
places_discotheque
places_cockpit
places_car_interior
places_train_interior
places_bus_interior
places_airplane_cabin
places_booth/indoor
places_amusement_arcade
places_department_store
places_bookstore
places_general_store/indoor
places_library/indoor
places_sushi_bar
places_restaurant_kitchen
places_market/indoor
places_bazaar/indoor
places_flea_market/indoor
places_dining_hall
places_banquet_hall
places_pub/indoor
places_beer_hall
places_supermarket
places_candy_store
places_fabric_store
places_toyshop
places_delicatessen
places_bakery/shop
places_pizzeria
places_butchers_shop
places_art_studio
places_art_gallery
places_beauty_salon
places_shoe_shop
places_jewelry_shop
places_gift_shop
places_clothing_store
places_food_court
places_cafeteria
places_fastfood_restaurant
places_restaurant
places_coffee_shop
places_bar
places_ice_cream_parlor
places_pharmacy
places_drugstore
places_hardware_store
places_kindergarden_classroom
places_classroom
places_art_school
places_nursing_home
places_nursery
places_childs_room
places_playroom
places_server_room
places_locker_room
places_laundromat
places_archive
places_office_cubicles
places_office
places_recreation_room
places_chemistry_lab
places_biology_laboratory
places_clean_room
places_operating_room
places_hospital_room
places_dressing_room
places_closet
places_storage_room
places_pantry
places_science_museum
places_museum/indoor
places_natural_history_museum
places_pet_shop
places_repair_shop
places_physics_laboratory
places_engine_room
places_auto_factory
places_assembly_line
places_berth
places_bathroom
places_sauna
places_shower
places_elevator/door
places_bank_vault
places_jail_cell
places_basement
places_attic
places_bow_window/indoor
places_alcove
places_corridor
places_lobby
places_elevator_lobby
places_entrance_hall
places_throne_room
places_church/indoor
places_lecture_room
places_conference_room
places_computer_room
places_legislative_chamber
places_home_theater
places_hotel_room
places_bedroom
places_bedchamber
places_youth_hostel
places_dorm_room
places_waiting_room
places_living_room
places_artists_loft
places_reception
places_wet_bar
places_kitchen
places_dining_room
places_utility_room
places_galley
places_television_room
places_home_office
places_train_station/platform
places_subway_station/platform
places_escalator/indoor
places_parking_garage/indoor
places_hangar/indoor
places_shopping_mall/indoor
places_atrium/public
places_airport_terminal
places_mezzanine
places_arcade
places_garage/indoor
places_swimming_pool/indoor
places_jacuzzi/indoor

E Additional Results

As an extension to Table 1, we report additional metrics for models trained with our methods. We chose to report the whole

- **Area Under Receiver Operating Characteristic (AUROC)** corresponds to the area under the Receiver Operating Characteristic curve, which plots the True Positive Rate (TPR) as a function of the False Positive Rate (FPR). This metric has the advantage of not requiring to choose the choice of a threshold and evaluating the OOD detector in all range of operations. It can be interpreted as the probability that, when picking an out of distribution sample and an in-distribution sample, the in-distribution sample gets considered as more in-distribution that the OOD sample.
- **Area Under Precision Recall (AUPR)** is a closely related metric. It corresponds to the area under the Precision-Recall Curve, which plots Precision against Recall.
- **False Positive Rate at 95% True Positive Rate (FPR@95%TPR)** is the probability that an OOD example is correctly identified when the true positive rate (TPR) is 95%. TPR is computed as $TPR = TP/(TP + FN)$, where TP and FN denote true positive and false negative respectively.
- **Detection Accuracy (DtAcc)** is a measurement of the maximum classification accuracy that we can achieve between in-distribution and out-of-distribution examples, by choosing the optimal threshold.

We also report results on additional out of distribution datasets.

- **Gaussian noise** is a synthetic dataset composed of 32×32 pixel random images where the intensity of each pixel is sampled from a gaussian distribution with mean 0.5 and scale 0.25.
- **Places365** [37] is a dataset containing scene photographs. We rescale them from their original size down to 32×32 to match the size of our in-distribution images.

Table 3: Additional out-of-distribution detection results for our proposed method. All results are an average over five independent runs. We report results both for our method and the baseline that does not include label smoothing or contrastive training.

			Baseline / Ours			
$\mathcal{D}_{\text{in}}$	$\mathcal{D}_{\text{out}}$	CLP range	FPR@95%TPR ↓	AUROC ↑	AUPR ↑	DtAcc ↑
CIFAR-10	CIFAR-100	[-7.4 to -0.8]	67.1 / 39.9	81.3 / 92.9	81.2 / 93.7	73.8 / 85.9
	SVHN	[-12.3 to -7.6]	20.5 / 2.8	96.2 / 99.5	96.6 / 99.6	89.5 / 96.7
	Places365	[-13.1 to -4.0]	69.2 / 24.3	82.0 / 95.3	82.6 / 95.3	74.7 / 89.1
	Gaussian noise	-15.7	3.4 / 0.3	99.1 / 100	99.4 / 100	97.1 / 99.2
CIFAR-100	CIFAR-10	[-4.5 to -2.6]	93.1 / 81.8	63.9 / 78.3	68.1 / 80.2	60.8 / 72.0
	SVHN	[-6.4 to -8.6]	56.5 / 23.3	87.8 / 95.6	88.9 / 95.8	80.1 / 89.2
	Places365	[-12.1 to -4.4]	96.4 / 69.2	51.8 / 82.0	58.3 / 82.6	55.1 / 78.4
	Gaussian noise	-15.1	97.3 / 0.2	53.9 / 99.9	65.9 / 99.9	63.4 / 98.0

F Additional Ablations

Sensitivity to supervised loss multiplier. We investigate the sensitivity of our models to the supervised loss multiplier parameter λ in our objective function defined in Section 3 which controls the strength of the supervised loss during the finetuning stage. We consider $\lambda \in [0, 1, 10, 100, 1000]$ for our experiments. We observe that increasing the loss multiplier leads to particularly significant improvement on the near OOD dataset pair with $\mathcal{D}_{\text{in}}$ as CIFAR-100 and $\mathcal{D}_{\text{out}}$ as CIFAR-10 but gains are marginal beyond $\lambda = 100$. We conjecture that the large loss ratios are necessary due to the difference in scale of the supervised and contrastive loss.

Impact of increasing model capacity. Our reference model is a wide ResNet-50 with a multiplier of 3. We investigate the impact of model capacity by training models for other width multipliers,

Table 4: **Ablation study of weight between supervised loss and contrastive loss**

λ	$\mathcal{D}_{\text{in}}$ = CIFAR-100 $\mathcal{D}_{\text{out}}$ = CIFAR-10	CIFAR-100 SVHN	CIFAR-100 Places365	CIFAR-10 CIFAR-100	CIFAR-10 SVHN	CIFAR-10 Places365
0.1	72.1 ± 0.6	94.5 ± 3.6	86.8 ± 1.0	88.3 ± 0.1	99.3 ± 0.08	91.4 ± 0.4
1	77.1 ± 0.3	95.9 ± 0.3	87.3 ± 0.3	91.9 ± 0.1	99.4 ± 0.1	94.1 ± 0.3
10	77.9 ± 0.3	95.5 ± 0.6	86.1 ± 0.5	92.7 ± 0.1	99.4 ± 0.04	94.4 ± 0.2
100	78.6 ± 0.5	95.3 ± 0.4	86.2 ± 0.1	92.9 ± 0.2	99.5 ± 0.1	94.5 ± 0.5
1000	77.9 ± 0.1	95.9 ± 0.3	85.9 ± 0.4	92.9 ± 0.2	99.3 ± 0.1	95.0 ± 0.4

ranging from 1 (the default ResNet-50) to 4. We observe that decreasing the ResNet-50 width to 1 leads to a significant drop on both near and far OOD dataset pairs. On the other hand, increasing the width to 4 does not result in improvement on OOD detection performance. We hypothesize that using contrastive training mandates the need for a higher capacity model in order to capture a richer representation with general task-agnostic features, and thus to achieve optimal performance compared to a supervised only baseline.

Table 5: **Ablation study of model capacity**

Width	$\mathcal{D}_{\text{in}}$ = CIFAR-100 $\mathcal{D}_{\text{out}}$ = CIFAR-10	CIFAR-100 SVHN	CIFAR-100 Places365	CIFAR-10 CIFAR-100	CIFAR-10 SVHN	CIFAR-10 Places365
1	74.2 ± 0.4	96.5 ± 0.4	85.4 ± 0.7	89.8 ± 0.2	97.8 ± 0.3	93.9 ± 0.7
2	77.1 ± 0.6	96.4 ± 0.3	86.8 ± 0.2	92.7 ± 0.1	99.2 ± 0.2	94.6 ± 0.3
3	78.3 ± 0.3	95.4 ± 0.1	85.3 ± 0.5	92.9 ± 0.2	99.5 ± 0.1	94.7 ± 0.3
4	78.9 ± 0.2	95.4 ± 0.2	85.9 ± 0.9	92.6 ± 0.2	99.3 ± 0.04	94.3 ± 0.4

Effect of temperature parameter of the contrastive loss. Finally, we run experiments to understand the importance of the temperature parameter τ used in the contrastive loss. We consider τ values $\in [0.01, 0.1, 0.5, 1, 2]$ for our experiments. Unlike [3] which obtains optimal performance with a τ of 0.1, we find that higher temperature leads to improved performance for the OOD detection task across the board. The optimal performance is obtained with a τ of 1.

Table 6: **Ablation study of contrastive loss temperature**

τ	$\mathcal{D}_{\text{in}}$ = CIFAR-100 $\mathcal{D}_{\text{out}}$ = CIFAR-10	CIFAR-100 SVHN	CIFAR-100 Places365	CIFAR-10 CIFAR-100	CIFAR-10 SVHN	CIFAR-10 Places365
0.01	57.8 ± 3.0	52.2 ± 3.9	62.5 ± 1.9	49.5 ± 4.0	48.1 ± 14.5	54.1 ± 2.2
0.1	77.4 ± 0.3	94.4 ± 0.7	84.5 ± 1.3	91.8 ± 0.2	99.3 ± 0.03	94.8 ± 0.4
0.5	77.1 ± 0.1	96.6 ± 0.1	86.7 ± 0.2	92.9 ± 0.06	99.5 ± 0.1	94.9 ± 0.1
1	78.7 ± 0.5	95.7 ± 0.5	86.6 ± 0.5	92.9 ± 0.1	99.5 ± 0.04	94.7 ± 0.1
2	78.6 ± 0.2	96.4 ± 0.5	84.6 ± 0.3	91.9 ± 0.2	99.0 ± 0.1	93.4 ± 0.4