# Beyond Clean Training Data: A Versatile and Model-Agnostic Framework for Out-of-Distribution Detection with Contaminated Training Data

Supplementary Material

## Table of Contents

## Appendix A: Detailed discussion on data contamination

In the area of AD, there exist some works studying the issue of contaminated training data [18, 22, 35]. Among these limited works, the IAD method [18] particularly focuses on the intricacies of the weight function design to be used as a model-agnostic AD with contaminated training data. To manage and adjust the weight for each sample throughout the rounds, IAD employs a weight function, based on the sigmoid function, that correlates the anomaly scores from the current round with the weight for the subsequent found. Despite its improvement and some similarities to our method, the IAD method unfortunately grapples with multiple limitations. First, the design of the weight function is strictly (or fatedly) reliant on the median statistics of the anomaly scores, (implicitly) presupposing a fixed contamination rate wherein exactly half of the dataset's samples are contaminated, although the actual proportion of contamination could be any value (normally unknown and possibly small) within the range [0, 0.5) in practice. The assumption restricts the method's flexibility and applicability across varied scenarios that exhibit diverse contamination ratios, leading to eventual performance degradation because such design can inadvertently suppress the feature learning of genuine samples during training. Also, in the IAD, the transition of the weight function from its initial state ($t = 0$) to subsequent training rounds ($t = 1, 2, \cdots$) is executed hastily, potentially leading to an excessive counterbalancing effect that might impede the assimilation of features from anomaly samples. In addition, the IAD framework incorporates a termination criterion based on a thresholding technique that evaluates the rank of the matrix composed of the anomaly scores for all samples within the dataset, which can be complex. Furthermore, it is worth clarifying that the IAD approach remains incompatible with recently proposed OOD detection models that incorporate synthetic negative samples, a burgeoning and critically important area of research within this domain, meaning that it cannot be model-agnostic for OOD detection. Last but not least, the IAD does not have any mechanism to estimate the contamination rate within the training dataset.

In general, directly applying AD methods to OOD detection is ineffective because they pursue different goals, under different assumptions about data, and adopt different techniques evaluated by different metrics. AD's main goal is to identify rare, unusual instances in a known distribution, while OOD detection aims to distinguish samples that do not belong to the training distribution. This difference in goals leads to different assumptions about the data: AD assumes that anomalies are rare instances within a single, known distribution, whereas OOD detection assumes those samples from unknown (or previously unseen) distributions. These differences then affect the performance metrics: AD mostly adopts precision and recall for rare events within a dataset, while OOD detection typically uses AUROC. In terms of methodologies, AD techniques focus on identifying low-density regions within the training distribution; however, OOD detection employs techniques such as contrastive learning, negative sampling, etc, to accurately refine the boundary between ID and OOD data. AD methods typically lack the capability to handle samples outside the training distribution. Overall, due to those differences, AD methods are often less effective for OOD detection. Furthermore, model-agnostic AD methods within AD contexts are often not model-agnostic for OOD detection.

In addition, our investigation also partly intersects with the domain of "deep learning with noisy labels." However, the "label noise" is different from the "data contamination" considered in this work. "label noise" exists within the ID classes of the ID dataset [1, 10, 16, 44], which typically results in mislabeled samples in the datasets [36]. Within this sphere, strategies such as regularization [37, 46], deployment of robust loss functions [17, 48], and sample selection mechanisms [10, 38] are prevalently adopted to mitigate the effects of label noise during training.

Compared with "deep learning with noisy labels", our task is more practical in real-world applications. In ours, "data contamination" refers to OOD samples included in the training dataset, which is a common issue when training datasets are constructed from real-world noisy, heterogeneous sources. For example, if ImageNet-1k is entirely used as the training dataset, its OOD portion can be determined *only when* an OOD dataset is specifically defined. This issue was extensively studied [4], but from the perspective of testing (not training as ours): With ImageNet-1k being training dataset and Places being test OOD dataset—a common setup—the percentage of ID samples present in the OOD dataset is as high as 59.5% (Table 1 of [4]), resulting in an FPR score increase up to 20% (Figure 3 of [4]), also providing similar trends for other OOD datasets. [4] claims that (i) high ID-OOD contamination is common and (ii) its huge negative impact. While the solution of [4] was to manually clean up datasets, our approach is (arguably) more practical: mitigate ID-OOD contamination through intelligent training.

Another category of partly related research works is Vision-Language Models (VLMs). VLMs like CLIP have demonstrated strong performance in image classification tasks, thanks to the huge amount of data that is used in their pre-training stage [27]. As a consequence, it might seem that this kind of pre-trained VLMs is the best choice for OOD detection on the contaminated dataset, as they do not need to be trained on the contaminated ID dataset, thus preventing performance degradation from data contamination.

However, VLM-based methods still face major limitations for OOD detection on the contaminated dataset. Let's take CLIP, the most popular VLM for OOD detection, as an example. First of all, CLIP-based methods are extremely resource-intensive, which are potentially unsuitable for on-device AI, AIoT, or tiny AI, which is one of the most popular scenarios for OOD detection in real-world applications. In contrast, our framework seamlessly integrates with models of any size. In addition, though CLIP achieves satisfactory *overall* accuracy, its performance is highly inconsistent across categories, with some classes exhibiting even 0% accuracy [30]–serious concerns for its suitability for risk-sensitive applications, e.g., healthcare. By contrast, our framework integrates effortlessly with any detectors optimized for specific applications. Besides, CLIP-based detection relies on a predefined set of explicit ID class labels, while such details of the dataset are often unavailable in real-world OOD detection, e.g., medical images labeled as "normal". In contrast, our framework performs seamlessly with any training dataset without requiring such detailed knowledge.

## Appendix B: Details of weight function design for our methods

In our proposed approach, we introduce a novel weight function that correlates the normalized[1] OOD score $s_i^{(t)}$ with the weight $w_i^{(t)}$ and pioneers an OOD percentage estimation mechanism $\hat{P}_{\text{OOD}}^{(t)}$, which is the first of this field. This innovative framework facilitates the iterative segregation of OOD samples within a contaminated dataset. Details on the idea of the weight function design are illustrated in the following subsections.

### B.1: Weight function design for our method for OOD detection models that do not use negative samples

The conceptual architecture of our proposed weight function for OOD detection models that do not use negative samples is illustrated in Figure 3. Drawing inspiration from the unit step function, depicted in Figure 3(a), our design delineates discrete intervals for ID and OOD samples. To refine the weight, we have adopted a modified unit step function that ensures a bifurcated treatment of ID and OOD samples via constant functions—specifically, $1$ for ID and $0$ for OOD, as evidenced in Figure 3(b).

Nevertheless, the mere alignment of ID and OOD sample ratios within the intervals of the weight function does not guarantee enhanced OOD detection fidelity. The precision of sample categorization is augmented when OOD scores are markedly delineated from the ID/OOD interface. The proximity of OOD scores to this demarcation attenuates the reliability of the predictions. In anticipation of such classification inaccuracies, we have devised a nuanced weight function, embodied by the orange trajectory in Figure 3(c). This function tempers the influence of scores proximal to the boundary, thereby mitigating misclassification risks while preserving the equilibrium of ID and OOD sample proportions.

The application of the nuanced weight function is most efficacious during the terminal epoch of training ($1 < t <$ END), leveraging an estimated ID ratio $\hat{P}_{\text{ID}}^{(t)}$ derived from the cumulative insights accrued throughout the training continuum. Initial epochs employ a uniform function (dark blue in Figure 3(d)) to impartially consider all samples, mirroring the inherent unsupervised learning paradigm at the inception of training.

Another pivotal element of our methodology is the progression from the initial uniform state to the ultimate nuanced weight function during the intermediate epochs ($1 < t <$ END). We advocate for a phased evolution from a uniform to a more customized weight curve, as signified by the color gradient transitioning from blue to light blue in Figure 3(d). A corpus of research [1, 10, 16, 29, 34, 44] suggests that deep learning models are predisposed to initially prioritize elementary, discernible features, subsequently advancing to more intricate features as training progresses. Given the prevalence of ID samples within contaminated datasets, these typically constitute the focal point of early training, obviating the necessity for immediate recalibrations of the weight function to counteract the influence of OOD samples, as posited by the IAD approach. Therefore, a phased transition aligns more congruently with the model's incremental refinement in learning and reliability during the training phase for OOD prediction.

### B.2: Weight function design for our method for OOD detection models that use negative samples

The logic flow of our proposed weight function for our method for OOD detection models that use negative samples is illustrated in Figure 4. Conventional OOD detection paradigms predominantly leverage synthetic negative samples, fabricated from ID data, to augment detection capabilities, premised on the notion of an unpolluted ID training subset. In stark contrast, our methodology acknowledges the prevalent issue of dataset contamination, wherein authentic OOD samples are intrinsically embedded within the training corpus. By capitalizing on these genuine OOD instances as negative exemplars, our method circumvents the synthesis of artificial negatives, thereby bolstering the veracity and efficacy of the training regimen. This strategic adjustment

[1] If the model does not inherently produce normalized OOD scores, a normalization process would be applied:

$$s_i^{(t)} = \min\left(\max\left(\frac{\hat{s}_i^{(t)} - T_{\text{low}}^{(t)}(s)}{T_{\text{high}}^{(t)}(s) - T_{\text{low}}^{(t)}(s)}\right), 1\right)$$

where $s_i^{(t)}$ and $\hat{s}_i^{(t)}$ denote the post-normalized and pre-normalized OOD scores, respectively, and $T_{\text{low}}^{(t)}(s)$ and $T_{\text{high}}^{(t)}(s)$ denote the lower and higher OOD score thresholds in this normalization.

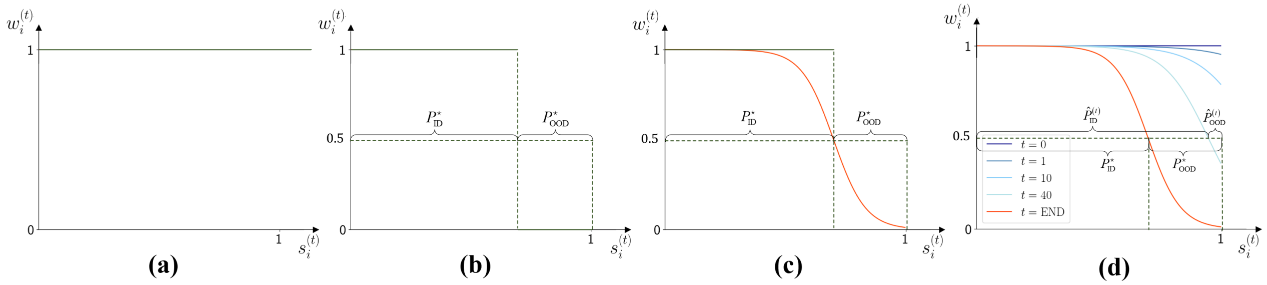


Figure 3. Design of the weight function in our proposed method for OOD detection models that do not use negative samples. (a) The unit step function, which inspires our method. (b) The ideal weight function for this method, a reverse and shift variant of the unit step function. It bifurcates all samples into two groups based on the percentage of ID and OOD samples, assigning disparate weights to them: 1 for ID samples and 0 for OOD samples. (c) The realistic weight function (in red), a curve function that maintains the same midpoint as the ideal weight function. (d) Transition of our proposed weight function from the beginning ($t = 0$) to the end ($t = \mathrm{END}$) of training. At the beginning, the weights of all samples are uniformly initiated at 1 (dark blue line). The progressively brighter blue curves illustrate the transition of the weight function throughout the training phases ($1 < t < \mathrm{END}$), with the intensification in brightness correlating with the progression in training rounds. The curve in red illustrates the weight function's ultimate configuration at the end of training ($t = \mathrm{END}$).

not only equips the models with the acumen to accurately discern authentic OOD samples during the training phase but also markedly enhances their discriminative prowess between ID and OOD instances, culminating in an optimized detection performance.

Within the IAD framework [18], the conventional weight function constrains the weight $w^{(t)}$ to be strictly non-negative (e.g., $[0, 1]$), a constraint that is inherently aligned with the treatment of positive samples. We introduce a pioneering weight function, reminiscent of the unit Sign function, which extends the weight domain to encompass the range $[-1, 1]$, as illustrated in Figures 4(a) and 4(b). This innovation permits the assignment of negative weights to OOD samples while preserving positive weights for ID samples, thereby expanding the functional ambit and augmenting the adaptability of the model. This enhanced weight function facilitates the seamless incorporation of negative weights into the loss computations of OOD models for negative samples.

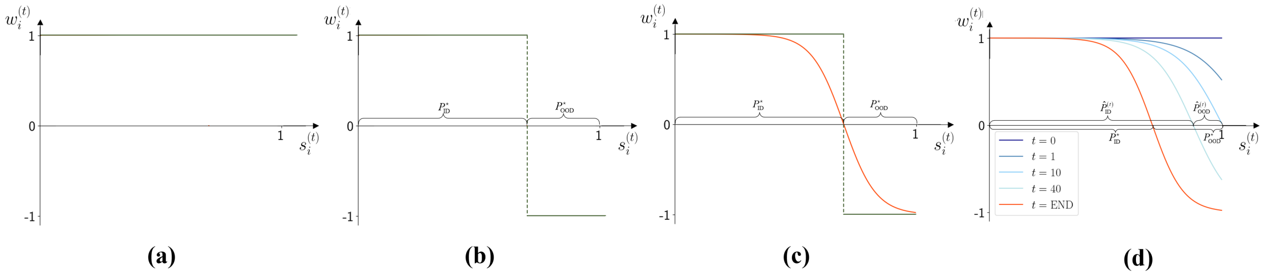
In alignment with the theoretically ideal weight function postulated in this method, we introduce a weight function when the backbone OOD detection model could produce reliable OOD scores, as depicted in Figure 4(c). This function is characterized by a curvilinear contour proximal to the demarcation between ID and OOD samples, designed to mitigate classification inaccuracies when OOD scores approximate this critical juncture, thus reinforcing the precision of predictions.

However, the backbone OOD detection model is only likely to produce reliable OOD scores at the later stage of training, which indicates that the transition mechanism by $\tau^{(t)}$ remains imperative. Thus, the final weight function for backbone OOD detection models that utilize negative samples is also equipped with the transition mechanism by $\tau^{(t)}$, as depicted in Figure 4(d).

By instituting this proposed weight function, we also empower backbone OOD detection models to efficaciously leverage real OOD samples as negative instances, a feat traditionally exclusive to the realm of synthetic negatives. This innovation not only broadens the scope of our methodology in the context of contaminated datasets but also substantially elevates the precision and operational efficiency of OOD detection models.

## Appendix C: Dataset pre-processing

In this study, we conformed to the popular dataset setups in existing works of the OOD detection field. Our experiments encompass an array of datasets, which include grayscale images datasets (MNIST, EMNIST, and FMNIST), color images datasets (CIFAR-10, SVHN, GTSRB, and Celeb_A). Furthermore, we have incorporated larger-scale color image datasets such as CIFAR-100, Mini-ImageNet, Tiny-ImageNet, ImageNet-1k, iNaturalist, and OpenImage-O. Most of the datasets in grayscale and natural images were retrieved either directly from their official websites or using the built-in TensorFlow implementation and were used without further modifications. Specifically, for the EMNIST dataset, we selectively utilized the "Letters" partition. The datasets SVHN and Celeb_A were subjected to a cropping pre-processing step to ensure the centralization of the primary subjects within the resultant images. To achieve uniformity in image resolution for CIFAR-100, Tiny-ImageNet, and Mini-ImageNet, the CIFAR-100 images were upscaled to a resolution of 64 × 64 pixels from an initial 32 × 32 pixels, aligning with the resolutions of Mini-ImageNet and Tiny-ImageNet. We also adopt the popular pre-processing practice for ImageNet-1k, iNaturalist, and OpenImage-O. For experimental validation, we constituted a validation set from each dataset, representing 10% of the total image, extracted from the original training set. Comprehensive details pertaining to the datasets are delineated in Table 9. To con-



**(a)** **(b)** **(c)** **(d)**

Figure 4. Design of the weight function in our method for backbone OOD detection models that use negative samples. (a) The unit Sign function, which inspires our method. (b) The ideal weight function for this method, a reverse and shift variant of the unit Sign function. (c) The realistic weight function (in red), a curve function that maintains the same midpoint as the ideal weight function. (d) Transition of our proposed weight function from the beginning ($t = 0$) to the end ($t = \text{END}$) of training. At the beginning, the weights of all samples are uniformly initiated at 1 (dark blue line). The progressively brighter blue curves illustrate the transition of the weight function throughout the training phases ($1 < t < \text{END}$), with the intensification in brightness correlating with the progression in training rounds. The curve in red illustrates the weight function's ultimate configuration at the end of training ($t = \text{END}$).

struct the contaminated dataset, a portion of samples from the OOD dataset are selected to be merged with the samples from the ID dataset based on each test case. The labels of the OOD samples are randomly set based on the ID labels.

## Appendix D: Utilization of backbone OOD detection models with our method

### D.1: Utilization of DB with our method

For the DB [2] backbone OOD detection model, we employ a Variational Autoencoder (VAE) [19] as the neural network, consistent with the neural network selection in the original DB approach. Specifically, we design three different VAEs for smaller datasets (grayscale images datasets and color images datasets with an image size of $32 \times 32$), slightly larger datasets (color images datasets with an image size of $64 \times 64$), and large datasets (color images datasets with an image size of $224 \times 224$), as shown in Tables 10, 11, and 12, respectively. The VAE for smaller datasets, as specified in Table 10, mirrors the convolutional VAE architecture utilized in DCGAN [26], whereas the VAE described in Table 11 and 12 features enhanced depth and capacity in its encoder and decoder structures, optimized for better performance on larger datasets.

In our experiments with DB as the backbone OOD detection model, the latent space dimensions were set to 20 for the smaller VAE and 32 for the larger VAE. With regard to the selection of the number of filters (nf) and the number of channels (nc), we standardized nf to 32 and nc to 1 for all grayscale images datasets, and nf to 64 and nc to 3 for all color images datasets. Regarding the hyperparameters, we set the value of $\alpha$ to 12 to modulate the slope of the weight function. As for the parameter $k$ in the transition function $\tau^{(t)}$, we set it to be 1.2.

In the training procedure with DB as the backbone OOD detection model with our method, we adhere to the original loss function design in DB, which is the Evidence Lower Bound (ELBO) loss. Thus, $L = L_{\text{ELBO}}$ is the training loss used for Equation (1) of the main text. Inference procedures retain the original DB configuration, employing corrected log-likelihood, either being corrected analytically or algorithmically, depending on the distribution of the decoder in VAE. Consistent with the DB approach, contrast normalization is applied during both the training and inference phases.

### D.2: Utilization of LReg with our method

For the implementation of the LReg [39] backbone OOD detection model, we employ the identical neural network configurations as used in the DB, as detailed in Tables 10, 11, and 12. This approach ensures uniformity in network architecture and scale across different models, thereby minimizing performance variances that could arise from architectural discrepancies. Consistency was maintained not only in the neural network setup but also in the selection of dimensional parameters and hyperparameters, mirroring those used in the experiments with DB. The decoder in the VAE was configured to follow a categorical distribution, in accordance with the specifications set forth in the original publication.

In the training phase, we stick to the original design of LReg and use the same ELBO loss $L_{\text{ELBO}}$ as loss $L$ for Equation (1) of the main text. During the inference phase, we adopt the original Likelihood Regret metric in LReg, which is calculated based on adding a preparatory training phase of 100 epochs for the encoder, executed before each sample's inference.

### D.3: Utilization of LRat with our method

In the implementation of the LRat [28] backbone OOD detection model, we use the same neural network setup for DB, as shown in Tables 10, 11, and 12. The choice of dimension

Table 9. Details of the 13 datasets used in experiments for grayscale images datasets, color image datasets, and larger scale color images datasets, including the number of train and test samples and resolution.

| **Dataset** | **Type** | **Number of train samples** | **Number of test samples** | **Resolution** ($H \times W \times C$) |
|---|---|---|---|---|
| MNIST | Grayscale | 54000 | 10000 | $32 \times 32 \times 1$ |
| Fashion-MNIST | Grayscale | 54000 | 10000 | $32 \times 32 \times 1$ |
| EMNIST-Letters | Grayscale | 79920 | 14800 | $32 \times 32 \times 1$ |
| SVHN | Color | 65932 | 26032 | $32 \times 32 \times 3$ |
| Celeb_A | Color | 146493 | 19962 | $32 \times 32 \times 3$ |
| GTSRB | Color | 35289 | 12630 | $32 \times 32 \times 3$ |
| CIFAR-10 | Color | 45000 | 10000 | $32 \times 32 \times 3$ |
| CIFAR-100 | Larger Color | 45000 | 10000 | $64 \times 64 \times 3$ |
| Mini-ImageNet | Larger Color | 45000 | 10000 | $64 \times 64 \times 3$ |
| Tiny-ImageNet | Larger Color | 100000 | 10000 | $64 \times 64 \times 3$ |
| ImageNet-1k | Larger Color | 1281167 | 100000 | $224 \times 224 \times 3$ |
| iNaturalist | Larger Color | 579184 | 95986 | $224 \times 224 \times 3$ |
| OpenImage-O | Larger Color | - | 17632 | $224 \times 224 \times 3$ |

Table 10. Detailed structure of encoder and decoder for the VAE model on smaller scale datasets: MNIST, Fashion-MNIST, EMNIST-Letters, SVHN, Celeb_A, GTSRB, and CIFAR-10. nc: number of channels; nf: number of filters; nz: number of latent dimensions; BN: batch normalization; Conv: convolution layer; DeConv: deconvolution layer; ReLU: rectified linear unit.

| **Encoder** |
|---|
| Input image of shape $32 \times 32 \times$ nc |
| $4 \times 4$ Conv$_{\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{2\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{4\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{2\times\text{nz}}$ Stride=1 |
| **Decoder** |
| Input latent code, reshape to $1 \times 1 \times$ nz |
| $4 \times 4$ DeConv$_{4\times\text{nf}}$ Stride=1, BN, ReLU |
| $4 \times 4$ DeConv$_{2\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ DeConv$_{\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ DeConv$_{\text{nc}}$ Stride=2 |

Table 11. Detailed structure of encoder and decoder for the VAE model on bigger scale datasets: CIFAR-100, Mini-ImageNet, and Tiny-ImageNet. Other conventions are the same as Table 10.

| **Encoder** |
|---|
| Input image of shape $64 \times 64 \times$ nc |
| $4 \times 4$ Conv$_{\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{2\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{4\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{8\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{16\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{2\times\text{nz}}$ Stride=1 |
| **Decoder** |
| Input latent code, reshape to $1 \times 1 \times$ nz |
| $4 \times 4$ DeConv$_{16\times\text{nf}}$ Stride=1, BN, ReLU |
| $4 \times 4$ DeConv$_{8\times\text{nf}}$ Stride=1, BN, ReLU |
| $4 \times 4$ DeConv$_{4\times\text{nf}}$ Stride=1, BN, ReLU |
| $4 \times 4$ DeConv$_{2\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ DeConv$_{\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ DeConv$_{\text{nc}}$ Stride=2 |

parameters in the neural network and hyperparameters in our method also remains the same as the values in the experiments with DB as the backbone OOD detection model. For the distribution of the decoder in VAE, we set it to categorical distribution, which aligns with the choices delineated in the original research.

Regarding the training procedure with LRat as the backbone OOD detection model with our method, we stick to the original design of LRat and use the same ELBO loss $L_{\text{ELBO}}$ as loss $L$ for Equation (1) of the main text for the standard model. As for the background model in LRat, we adhere to the prescribed noise corruption technique for VAE configurations. Herein, the mutation parameter is designated as 0.3 for grayscale images and 0.1 for color images. Additionally, both models incorporate a substantial weight decay parameter as 100. In the inference procedure, we use the original Likelihood Ratio metric in LRat, which is calculated based on the ratio of log-likelihood between the standard model and the background model.

Table 12. Detailed structure of encoder and decoder for the VAE model on larger-scale datasets: ImageNet-1k, iNaturalist, and OpenImages-O. The design of the encoder and decoder are based on ResNet-18. Other conventions are the same as Table 10.

| **Encoder** |
|---|
| Input image of shape $224 \times 224 \times$ nc |
| $7 \times 7$ Conv$_{\text{nf}}$ Stride=2, BN, ReLU, MaxPool $3 \times 3$ |
| ResNet-18 Block 1: $3 \times 3$ Conv$_{\text{nf}}$, Stride=1, Repeated |
| ResNet-18 Block 2: $3 \times 3$ Conv$_{2\times\text{nf}}$, Stride=2, Downsampling |
| ResNet-18 Block 3: $3 \times 3$ Conv$_{4\times\text{nf}}$, Stride=2, Downsampling |
| ResNet-18 Block 4: $3 \times 3$ Conv$_{8\times\text{nf}}$, Stride=2, Downsampling |
| $4 \times 4$ Conv$_{2\times\text{nz}}$, Stride=1 |
| **Decoder** |
| Input latent code, reshape to $1 \times 1 \times$ nz |
| $4 \times 4$ DeConv$_{8\times\text{nf}}$, Stride=1, BN, ReLU |
| $4 \times 4$ DeConv$_{4\times\text{nf}}$, Stride=2, BN, ReLU |
| $4 \times 4$ DeConv$_{2\times\text{nf}}$, Stride=2, BN, ReLU |
| $4 \times 4$ DeConv$_{\text{nf}}$, Stride=2, BN, ReLU |
| $7 \times 7$ DeConv$_{\text{nc}}$, Stride=2 |

## D.4: Utilization of WAIC with our method

For the implementation of the WAIC [7] backbone OOD detection model, we replicate the neural network configuration for DB, as shown in Tables 10, 11, and 12. The values of dimension parameters in the neural network and hyperparameters in our method are also the same as in the experiments with DB as the backbone OOD detection model. For the distribution of the decoder in VAE, we set it to continuous Bernoulli distribution, which aligns with the choices delineated in the original research.

During the training phase with WAIC as the backbone OOD detection model with our method, we stick to the original design of WAIC and use the same ELBO loss $L_{\text{ELBO}}$ as loss $L$ for Equation (1) of the main text. In the inference procedure, we use the original inference metric in WAIC, which is formulated as $\mathbb{E}_\theta\left[\log p_\theta(x)\right] - \text{Var}_\theta\left[\log p_\theta(x)\right]$, where $\theta$ is the parameter of the decoder in VAE. The mathematical expectation $\mathbb{E}_\theta$ and the variance $\text{Var}_\theta$ are computed across an ensemble of six VAEs.

## D.5: Utilization of G-ODIN with our method

Unlike the backbone OOD detection models in Appendices E.1 to E.4, G-ODIN [14] leverages a discriminative classifier. In the original G-ODIN work, the authors use ResNet-50 as the neural network. However, to facilitate a more equitable comparison with the VAE-based backbone OOD detection models, we opted for discriminative classifiers with a parameter count commensurate to those of the VAEs detailed in Tables 10, 11, and 12. The structural details of the classifiers for both smaller and larger datasets are presented in Tables 13, 14, and 15, respectively.

Table 13. Detailed structure of classifier model on smaller scale datasets: MNIST, Fashion-MNIST, EMNIST-Letters, SVHN, Celeb_A, GTSRB, and CIFAR-10. FC: fully connected layer. cls: number of classes. Other conventions are the same as Table 10.

| **Classifier** |
|---|
| Input image of shape $32 \times 32 \times$ nc |
| $4 \times 4$ Conv$_{\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{2\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{4\times\text{nf}}$ Stride=2, BN, ReLU |
| Flatten output to 1D |
| FC-128, ReLU |
| FC-cls, Softmax |

Table 14. Detailed structure of classifier model on larger scale datasets: CIFAR-100, Mini-ImageNet, and Tiny-ImageNet. AAP: adaptive average pooling. Other conventions are the same as Table 13.

| **Classifier** |
|---|
| Input image of shape $64 \times 64 \times$ nc |
| $4 \times 4$ Conv$_{\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{2\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{4\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{8\times\text{nf}}$ Stride=2, BN, ReLU |
| $4 \times 4$ Conv$_{16\times\text{nf}}$ Stride=2, BN, ReLU |
| AAP to $1 \times 1$ |
| FC-512, ReLU |
| FC-256, ReLU |
| FC-cls, Softmax |

In our implementation with G-ODIN as the backbone OOD detection model, the configuration for the number of filters (nf) and the number of channels (nc) was standardized; we set nf to 32 and nc to 1 for all grayscale image datasets, and nf to 64 and nc to 3 for all color image datasets. Regarding other hyperparameters, we set the value of $\alpha$ to 12 to modulate the slope of the weight function, and $k$ to be 1.2 in the transition function $\tau^{(t)}$.

In the training procedure with G-ODIN as the backbone OOD detection model with our method, we adhere to the original loss function design in G-ODIN, which is the standard Cross-Entropy loss for discriminative classifier. Thus, $L = L_{\text{CE}}$ is the training loss used for Equation (1). Furthermore, temperature scaling—a pivotal component of the original G-ODIN framework—was incorporated into our training phase. Also, the preprocessing of inputs entails a comprehensive search operation across the entire validation set, a process meticulously aligned with the procedural integrity of the original G-ODIN study. For the inference procedure, we employ the original inference metric configuration in G-ODIN, which is calculated by $\max_i h_i$, where $h_i$ represents the output of the softmax function of the classifier.

Table 15. Detailed structure of classifier model for OOD detection on large-scale datasets: ImageNet-1k, iNaturalist, and OpenImages-O. The design of this classifier is based on ResNet-34. Other conventions are the same as Table 13.

| **Classifier** |
|---|
| Input image of shape $224 \times 224 \times \text{nc}$ |
| $7 \times 7$ $\text{Conv}_{\text{nf}}$, Stride=2, BN, ReLU, MaxPool $3 \times 3$ |
| ResNet-34 Block 1: $3 \times 3$ $\text{Conv}_{\text{nf}}$, Stride=1, Repeated 3 times |
| ResNet-34 Block 2: $3 \times 3$ $\text{Conv}_{2\times\text{nf}}$, Stride=2, Downsampling, Repeated 4 times |
| ResNet-34 Block 3: $3 \times 3$ $\text{Conv}_{4\times\text{nf}}$, Stride=2, Downsampling, Repeated 6 times |
| ResNet-34 Block 4: $3 \times 3$ $\text{Conv}_{8\times\text{nf}}$, Stride=2, Downsampling, Repeated 3 times |
| AAP to $1 \times 1$ |
| FC-1024, ReLU |
| FC-512, ReLU |
| FC-cls, Softmax |

### D.6: Utilization of Energy with our method

Similar to G-ODIN [14], Energy [24] also leverages a discriminative classifier. Thus, we follow the same procedure as for G-ODIN for the classifier design to facilitate a more equitable comparison with the VAE-based backbone OOD detection models. The structural details of the classifiers for both smaller and larger datasets are presented in Tables 13, 14, and 15, respectively. In addition, the configuration (e.g., the number of filters (nf) and the number of channels (NC)) of the neural network is set to be the same as for G-ODIN, and the choice of other hyperparameters is also the same as in our experiments with G-ODIN as the backbone OOD detection model.

In the training procedure with Energy as the backbone OOD detection model with our method, we adhere to the original loss function design in Energy, which is a combination of the standard Cross-Entropy loss for discriminative classifier and a regularization loss defined in terms of energy. For the inference procedure, we employ the original inference metric configuration in Energy, which is their proposed energy score.

### D.7: Utilization of ReAct with our method

As for ReAct [31], we use a similar setup as for G-ODIN and Energy, which also uses a discriminative classifier. The structural details of the classifiers for both smaller and larger datasets are shown in Tables 13, 14, and 15, respectively. Also, the configuration (e.g., the number of filters (nf) and the number of channels (NC)) of the neural network is set to be the same as for G-ODIN and Energy, and the choice of other hyperparameters is also the same as in our experiments with them as the backbone OOD detection model.

In the training procedure with ReAct as the backbone OOD detection model with our method, we adhere to the original loss function design in ReAct, which is the standard Cross-Entropy loss for the discriminative classifier. Thus, $L = L_{\text{CE}}$ is the training loss used for Equation (1). Also, the ReAct operation is performed on the classifier after the training procedure. For the inference procedure, we employ the original inference metric configuration in ReAct, which is the softmax score from the classifier.

### D.8: Utilization of CSI with our method

For the CSI [33] backbone OOD detection model, we use the same neural network setup for DB, as shown in Tables 10, 11, and 12. The choice of dimension parameters in the neural network in our method is also the same as in the experiments with DB as the backbone OOD detection model. As for the hyperparameters, we set the value of $\alpha$ to 18 to modulate the slope of the weight function, and the parameter $k$ in the transition function $\tau^{(t)}$ to be 1.5.

In the training phase with CSI as the backbone OOD detection model with our method, we choose to keep only the loss function for ID samples: $L_{\text{ID}} \geq 0$ and drop the loss function for OOD samples: $L_{\text{OOD}} \leq 0$ in the original CSI study, as we have proposed a weight function (see Equation (10) of the main text) to expand the range of $w_i^{(t)}$ to $[-1, 1]$. Thus, $L = L_{\text{ID}} = L_{\text{ELBO}}$ is the training loss used for Equation (1) of the main text. As for the inference configuration, we retain the original inference metric in the CSI research, which is formulated by $\log p_\theta(x) + \log p_\theta(\tilde{x})$, where $\tilde{x} = t\left(\tilde{x}|x\right)(x)$ and $t\left(\tilde{x}|x\right)$ is a randomly selected data augmentation technique presented in SimCLR [6].

### D.9: Utilization of CnC with our method

For the implementation of CnC [11] backbone OOD detection model, we replicate the same neural network setup for DB, as shown in Tables 10, 11, and 12. Additionally, the choice of dimension parameters and hyperparameters is aligned with the experiments using CSI as the backbone OOD detection model.

During the training procedure with CnC as the backbone OOD detection model with our method, we choose to also keep only the loss function for ID samples: $L_{\text{ID}} \geq 0$ and drop the loss function for OOD samples: $L_{\text{OOD}} \leq 0$ in the original CnC study, which is similar to the setup for CSI. Thus, $L = L_{\text{ID}} = L_{\text{ELBO}}$ is the training loss used for Equation (1) of the main text. Furthermore, we incorporate the Mix-up data augmentation technique from the original CnC study. For the inference phase, we use the standard log-likelihood estimation $\log p_\theta(x)$ as the inference metrics for CnC.

### D.10: Utilization of VOS with our method

In the implementation of the VOS [9] backbone OOD detection model, we adhere to the same neural network setup for DB, as shown in Tables 10, 11, and 12. Furthermore, the dimension parameters of the neural network and the hyperparameters are selected to align with those used in the experiments involving CSI as the backbone OOD detection model.

In the training phase with VOS as the backbone OOD detection model with our method, we elect to also keep only the loss function for ID samples: $L_{\text{ID}} \geq 0$ and drop the loss function for OOD samples: $L_{\text{OOD}} \leq 0$ in the original VOS study, which is similar to the setup for CSI. Thus, $L = L_{\text{ID}} = L_{\text{ELBO}}$ is the training loss used for Equation (1) of the main text. For the inference configuration, we adopt the design of the original inference metric in VOS research, which is expressed as $\frac{e^{\log p_\theta(x)}}{1+e^{\log p_\theta(x)}}$.

## Appendix E: Supplementary results

### E.1: Extended comparison results in OOD detection performance

A comprehensive comparison of OOD detection performance is shown in Table 16 (for OOD detection models that do not use negative samples) and Table 17 (for OOD detection models that use negative samples), including multiple test cases with varying OOD percentages on the raw backbone model, backbone model with the IAD method, and backbone model equipped with our proposed method. The results indicate that conventional backbone OOD models experience significant performance declines when confronted with a contaminated dataset, even at a very small $P^\star_{\text{OOD}} = 1\%$. As detailed in Table 16, our method demonstrates superior performance in comparison to the IAD approach across a variety of datasets, OOD detection models, and OOD percentages $P^\star_{\text{OOD}}$ in our method for OOD detection models that do not use negative samples. Furthermore, the experimental outcomes presented in Table 17 affirm that our method consistently outperforms the IAD method across all three backbone OOD detection models under different test conditions in our method for OOD detection models that use negative samples.

For test cases varying ID and OOD datasets, we also evaluated our proposed method on MNIST, EMNIST, and FMNIST, for OOD detection models that do not use negative samples (Table 18) and use negative samples (Table 19). Our approach consistently demonstrates better performance compared with the IAD method across different backbone OOD detection models, which is similar to the observation on harder test cases with CIFAR-10 and CIFAR-100 datasets as in Table 2.

In addition, we also evaluate our proposed framework on the original OOD detection task ($P^\star_{\text{OOD}} = 0\%$) without data contamination in Tables 20 and 21. It could be observed that our proposed method maintains a very similar performance of OOD detection on the original OOD detection task for both the OOD detection models that do and do not use negative samples in almost all test cases. Also, the performance stability is consistently better than the IAD approach in all test cases.

Besides, we also compared our proposed method with a popular method in the domain of "deep learning with noisy labels": Co-teaching [10]. The comparison results between our proposed method and Co-teaching are shown in Tables 22 and 23. It could be observed that our proposed method is able to consistently surpass the Co-teaching method in all test cases, for both OOD detection methods that do and do not use negative samples. Also, our method only requires half of the runtime memory compared with Co-teaching, as it needs two networks operating in parallel.

### E.2: Performance of backbones OOD detection models on clean ID datasets

Conventional OOD detection on clean ID datasets could be regarded as a special and simplified variant of OOD detection on contaminated datasets, where the OOD percentage $P^\star_{\text{OOD}}$ of the training set is $0\%$. In order to provide a better view and comparison with the performance of OOD detection on contaminated datasets, we also present the conventional OOD detection performance of all backbone OOD detection models on uncontaminated (clean) ID datasets, shown in Table 24.

### E.3: Impact of OOD percentage $P^\star_{\text{OOD}}$

To rigorously evaluate the influence of the OOD percentage $P^\star_{\text{OOD}}$ on the efficacy of our method, we executed a sequence of empirical analyses wherein $P^\star_{\text{OOD}}$ was systematically varied from $1\%$ to $10\%$. These analyses were performed using the most effective backbone OOD detection models identified within the frameworks of our method for OOD detection models that do and do not use negative samples. The experimental results are shown in Table 25.

The performance comparison results reveal that our method consistently outperforms the IAD method across

Table 16. Performance comparison of our method and IAD on test cases for OOD detection models that do not use negative samples with different $P^{\star}_{\text{OOD}}$. The numbers reported are AUROC scores. “Backbone” indicates the backbone OOD detection models. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | $P^{\star}_{\text{OOD}}$ | Backbone | Methods | | |
|---|---|---|---|---|---|---|
| | | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| MNIST | EMNIST | 1% | DB | 78.1 | 85.6 | **91.4** |
| | | | LReg | 80.2 | 88.3 | **93.5** |
| | | | LRat | 76.5 | 84.2 | **90.1** |
| | | | WAIC | 74.7 | 82.6 | **87.2** |
| | | | G-ODIN | 81.4 | 89 | **93.1** |
| | | 2% | DB | 75.7 | 82.9 | **88.6** |
| | | | LReg | 78.5 | 86.1 | **90.4** |
| | | | LRat | 74.6 | 82.4 | **88.6** |
| | | | WAIC | 73.2 | 80.9 | **85.6** |
| | | | G-ODIN | 81.4 | 89 | **93.1** |
| CIFAR-10 | Celeb_A | 1% | DB | 58.6 | 60.1 | **62.2** |
| | | | LReg | 63.4 | 65.2 | **68.5** |
| | | | LRat | 58.0 | 59 | **60.9** |
| | | | WAIC | 56.9 | 57.7 | **59.1** |
| | | | G-ODIN | 59.1 | 62.5 | **65.7** |
| | | 2% | DB | 56.7 | 58.4 | **60.6** |
| | | | LReg | 62.2 | 64 | **66.9** |
| | | | LRat | 55.9 | 57.8 | **59.5** |
| | | | WAIC | 55.2 | 56.5 | **58.2** |
| | | | G-ODIN | 57.2 | 60.8 | **64.3** |

Table 17. AUROC performance comparison of our method and IAD on test cases for OOD detection models that use negative samples with different $P^{\star}_{\text{OOD}}$. The numbers reported are AUROC scores. “Backbone” indicates the backbone OOD detection models. Numbers in **bold** indicate the best performance in each test case.

| ID dataset | OOD dataset | $P^{\star}_{\text{OOD}}$ | Backbone | Methods | | |
|---|---|---|---|---|---|---|
| | | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| MNIST | EMNIST | 1% | CSI | 82.6 | 89.3 | **94.7** |
| | | | CnC | 83.4 | 89.9 | **95.2** |
| | | | VOS | 84.8 | 90.4 | **96.0** |
| | | 2% | CSI | 81.2 | 87.8 | **92.4** |
| | | | CnC | 82.0 | 87.6 | **93.8** |
| | | | VOS | 82.3 | 88.6 | **93.7** |
| CIFAR-10 | Celeb_A | 1% | CSI | 58.6 | 61.6 | **64.3** |
| | | | CnC | 59.4 | 62.0 | **64.2** |
| | | | VOS | 59.8 | 61.8 | **63.1** |
| | | 2% | CSI | 57.1 | 60.2 | **62.6** |
| | | | CnC | 56.8 | 60.5 | **62.3** |
| | | | VOS | 56.5 | 59.9 | **62.1** |

different $P^{\star}_{\text{OOD}}$ based on our methods for OOD detection models that use negative samples or not with different backbone OOD detection models. This substantial enhancement attests to the robustness of our proposed framework.

Table 18. AUROC performance comparison of our method and IAD on test cases for OOD detection models that do not use negative samples with different ID and OOD datasets. The numbers reported are AUROC scores. “Backbone” indicates the backbone OOD detection models. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | $P^{\star}_{\text{OOD}}$ | Backbone | Methods | | |
|---|---|---|---|---|---|---|
| | | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| MNIST | EMNIST | 1% | DB | 78.1 | 85.6 | **91.4** |
| | | | LReg | 80.2 | 88.3 | **93.5** |
| | | | LRat | 76.5 | 84.2 | **90.1** |
| | | | WAIC | 74.7 | 82.6 | **87.2** |
| | | | G-ODIN | 81.4 | 89 | **93.1** |
| | FMNIST | | DB | 80.7 | 89.1 | **95.4** |
| | | | LReg | 81.3 | 89.5 | **95.8** |
| | | | LRat | 80.6 | 88.9 | **95.4** |
| | | | WAIC | 81.1 | 89.0 | **94.9** |
| | | | G-ODIN | 80.2 | 88.6 | **95.1** |

Table 19. AUROC performance comparison of our method and IAD on test cases for OOD detection models that use negative samples with different ID and OOD datasets. The numbers reported are AUROC scores. “Backbone” indicates the backbone OOD detection models. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | $P^{\star}_{\text{OOD}}$ | Backbone | Methods | | |
|---|---|---|---|---|---|---|
| | | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| MNIST | EMNIST | 1% | CSI | 82.6 | 89.3 | **94.7** |
| | | | CnC | 83.4 | 89.9 | **95.2** |
| | | | VOS | 84.8 | 90.4 | **96.0** |
| | FMNIST | | CSI | 80.2 | 90.4 | **97.6** |
| | | | CnC | 81.1 | 87.9 | **98.4** |
| | | | VOS | 81.6 | 93.0 | **99.5** |

### E.4: Results on performance stability and termination criterion

The stability of performance represents a critical metric in which our methodology demonstrates superiority over the IAD approach. Within the IAD framework, the authors proffer a termination criterion predicated on a rank-based algorithm utilizing a matrix of OOD scores for all samples of the dataset, as discussed in the main text. While the authors of IAD assert that this criterion enhances the efficacy of the raw IAD method, we contend that the complexity of this criterion is unnecessary. Contrarily, our method adopts a more streamlined termination criterion based on the sequential increments of the OOD percentage estimation, $\hat{P}^{(t)}_{\text{OOD}}$, during training. This criterion emerges naturally from the underlying principle that the stability of the backbone OOD detection model is intrinsically linked to the stabilized prediction of $P^{\star}_{\text{OOD}}$.

To substantiate the enhanced stability of our proposed methodology and its associated termination criterion, we provided the results for this study in Table 26. We present the data in terms of the mean AUROC scores, accompanied by the Standard Error of the Mean (SEM) for each case.

Our empirical evidence reveals that our method consistently attains superior performance, as evidenced by elevated mean AUROC scores and diminished SEM values across various test scenarios, utilizing our method for OOD detection models that do and do not use negative samples with different backbone OOD detection models and $P^{\star}_{\text{OOD}}$. Notably, the SEM values associated with our methodology closely approximate those attributed to the raw backbone OOD detection models when applied to the corresponding test cases, in stark contrast to the IAD method. This congruence underscores the efficacy of our termination criterion and attests to the reliability of a design predicated upon the OOD percentage estimation.

Table 20. AUROC score of our method and IAD for OOD detection models that do not use negative samples with different ID and OOD datasets on the original OOD detection task ($P^{\star}_{\text{OOD}} = 0\%$). Backbone OOD detection model names with square ($\square$) symbols represent the ones that do not utilize negative samples. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | Backbone | Methods | | |
|---|---|---|---|---|---|
| | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| CIFAR-100 | Tiny-ImageNet | DB$^{\square}$ | 58.1 | 57.9 | **58.1** |
| | | G-ODIN$^{\square}$ | 62.0 | 61.5 | **62.0** |
| | Mini-ImageNet | DB$^{\square}$ | 55.9 | 55.8 | **55.9** |
| | | G-ODIN$^{\square}$ | 61.4 | 60.7 | **61.3** |
| ImageNet-1k | iNaturalist | DB$^{\square}$ | 61.0 | 61.0 | **61.0** |
| | | G-ODIN$^{\square}$ | 65.3 | 65.0 | **65.3** |
| | OpenImage-O | DB$^{\square}$ | 55.2 | 53.4 | **55.2** |
| | | G-ODIN$^{\square}$ | 58.4 | 58.0 | **58.4** |

Table 21. AUROC score of our method and IAD for OOD detection models that use negative samples with different ID and OOD datasets on the original OOD detection task ($P^{\star}_{\text{OOD}} = 0\%$). Backbone OOD detection model names with triangle ($\triangle$) symbols represent the ones that utilize negative samples. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | Backbone | Methods | | |
|---|---|---|---|---|---|
| | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| CIFAR-100 | Tiny-ImageNet | VOS$^{\triangle}$ | 62.9 | 61.7 | **62.9** |
| | Mini-ImageNet | | 60.2 | 59.6 | **60.1** |
| ImageNet-1k | iNaturalist | VOS$^{\triangle}$ | 63.8 | 63.6 | **63.8** |
| | OpenImage-O | | 56.0 | 55.1 | **56.0** |

### E.5: Extended results on OOD percentage estimation performance

For OOD percentage estimation results, we also evaluate our proposed method on various other datasets. Comprehensive OOD percentage estimation evaluation results for OOD detection models that do and do not use negative samples are shown in Tables 27 and 28, respectively. It could be observed that our approach demonstrates a consistent, accurate estimation of the real OOD percentage across different test cases and backbone OOD detection models.

Evaluation of OOD percentage estimation performance is also conducted on test cases incorporating diverse ID and OOD datasets for both OOD detection models that use negative samples or not. Table 29 illustrates that our method maintains its precision in estimating $P^{\star}_{\text{OOD}}$ across an array of test cases, unaffected by the specific ID and OOD datasets combined. Analogously, Table 30 affirms the ability of our method to provide close estimations of $P^{\star}_{\text{OOD}}$ for OOD detection models that use negative samples. This consistent accuracy in estimation underscores our method's robustness and its capacity to handle a variety of unknown OOD datasets.

### E.6: Extended results on ablation study of transition $\tau^{(t)}$

Besides, we also show that this transition $\tau^{(t)}$ is also crucial for the OOD percentage estimation in the training procedure, as shown in Table 31. It could be observed that $\tau^{(t)}$ also contributes to a better OOD percentage estimation on various datasets, backbone OOD detection models, and $P^{\star}_{\text{OOD}}$.

Table 22. AUROC score of our method and Co-teaching for OOD detection models that do not use negative samples with different ID and OOD datasets with $P^{\star}_{\text{OOD}} = 1\%$). Backbone OOD detection model names with square (□) symbols represent the ones that do not utilize negative samples. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | Backbone | Methods | | |
|---|---|---|---|---|---|
| | | | Backbone only | Backbone + Co-teaching [10] | Backbone + Ours |
| CIFAR-100 | Tiny-ImageNet | DB$^\square$ | 54.8 | 56.5 | **57.2** |
| | | G-ODIN$^\square$ | 56.9 | 60.2 | **61.3** |
| | Mini-ImageNet | DB$^\square$ | 54.3 | 55.0 | **55.6** |
| | | G-ODIN$^\square$ | 53.3 | 58.2 | **60.4** |
| ImageNet-1k | iNaturalist | DB$^\square$ | 57.9 | 59.6 | **60.7** |
| | | G-ODIN$^\square$ | 60.4 | 63.5 | **64.8** |
| | OpenImage-O | DB$^\square$ | 53.6 | 54.2 | **54.3** |
| | | G-ODIN$^\square$ | 54.5 | 56.2 | **57.7** |

Table 23. AUROC score of our method and Co-teaching for OOD detection models that use negative samples with different ID and OOD datasets with $P^{\star}_{\text{OOD}} = 1\%$. Backbone OOD detection model names with triangle (△) symbols represent the ones that utilize negative samples. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | Backbone | Methods | | |
|---|---|---|---|---|---|
| | | | Backbone only | Backbone + Co-teaching [10] | Backbone + Ours |
| CIFAR-100 | Tiny-ImageNet | VOS$^\triangle$ | 57.2 | 60.9 | **62.1** |
| | Mini-ImageNet | | 57.6 | 59.1 | **59.6** |
| ImageNet-1k | iNaturalist | VOS$^\triangle$ | 61.0 | 62.5 | **63.2** |
| | OpenImage-O | | 53.8 | 54.9 | **55.2** |

Table 24. AUROC performance of all backbone OOD detection models for conventional OOD detection on uncontaminated (clean) ID datasets.

| ID Dataset | OOD dataset | Backbone | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DB | LReg | LRat | WAIC | G-ODIN | Energy | ReAct | CSI | CnC | VOS |
| MNIST | EMNIST | 100 | 100 | 100 | 96 | 100 | 100 | 100 | 100 | 99 | 100 |
| | FMNIST | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| CIFAR-10 | Celeb_A | 63 | 70 | 45 | 40 | 67 | 69 | 74 | 66 | 65 | 64 |
| | SVHN | 57 | 85 | 13 | 55 | 61 | 65 | 70 | 58 | 58 | 59 |
| | GTSRB | 66 | 60 | 49 | 26 | 65 | 72 | 79 | 70 | 66 | 68 |
| CIFAR-100 | Tiny-ImageNet | 58 | - | - | - | 62 | 68 | 74 | - | - | 63 |
| | Mini-ImageNet | 56 | - | - | - | 61 | 65 | 66 | - | - | 60 |
| ImageNet-1k | iNaturalist | 61 | - | - | - | 65 | 69 | 72 | - | - | 64 |
| | OpenImage-O | 55 | - | - | - | 58 | 62 | 66 | - | - | 56 |

Table 25. AUROC performance comparison of our method and IAD on test cases of OOD detection models that do and do not use negative samples varying $P^{\star}_{\text{OOD}}$. Numbers in **bold** indicate the best performance in each test case.

| ID Dataset | OOD dataset | $P^{\star}_{\text{OOD}}$ | Backbone | Methods | | |
|---|---|---|---|---|---|---|
| | | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| MNIST | EMNIST | 1% | DB | 78.1 | 85.6 | **91.4** |
| | | | VOS | 84.8 | 90.4 | **96.0** |
| | | 2% | DB | 75.7 | 82.9 | **88.6** |
| | | | VOS | 82.3 | 88.6 | **93.7** |
| | | 5% | DB | 59.2 | 76.5 | **82.4** |
| | | | VOS | 72.1 | 81.5 | **90.4** |
| | | 10% | DB | 52.1 | 65.9 | **69.8** |
| | | | VOS | 60.5 | 71.8 | **79.9** |

Table 26. Performance comparison of our method and IAD on stability based on 10 repeated simulations. The numbers reported are mean AUROC scores and Standard Error of the Mean (SEM) values.

| ID Dataset | OOD dataset | $P^{\star}_{\text{OOD}}$ | Backbone | Methods | | |
|---|---|---|---|---|---|---|
| | | | | Backbone only | Backbone + IAD [18] | Backbone + Ours |
| MNIST | EMNIST | 1% | DB | 78.1±0.13 | 85.6±0.55 | 91.4±0.21 |
| | | | VOS | 84.8±0.11 | 90.4±0.44 | 96.0±0.26 |
| CIFAR-10 | Celeb_A | | DB | 58.6±0.14 | 60.1±0.39 | 62.2±0.24 |
| | | | VOS | 59.8±0.13 | 61.8±0.46 | 63.1±0.20 |
| MNIST | EMNIST | 2% | DB | 75.7±0.21 | 82.9±1.17 | 88.6±0.38 |
| | | | VOS | 82.3±0.24 | 88.6±0.89 | 93.7±0.30 |
| CIFAR-10 | Celeb_A | | DB | 56.7±0.21 | 58.4±0.82 | 60.6±0.37 |
| | | | VOS | 56.5±0.31 | 59.9±1.03 | 62.1±0.52 |

Table 27. Normalized OOD percentage estimation error $\varepsilon$ comparison for our method and IAD for OOD detection models that do and do not use negative samples with MNIST as ID dataset and EMNIST as OOD dataset of different $P^{\star}_{\text{OOD}}$. "Any" indicates any of the backbone OOD detection models.

| Methods | Backbone | $P^{\star}_{\text{OOD}}$ | | |
|---|---|---|---|---|
| | | 0.010 | 0.020 | 0.050 |
| Backbone + IAD [18] | Any$^{\square,\triangle}$ | N/A (0.980) | N/A (0.960) | N/A (0.900) |
| Backbone + Ours | DB$^{\square}$ | 0.001 | 0.003 | 0.013 |
| | LReg$^{\square}$ | 0.001 | 0.002 | 0.006 |
| | LRat$^{\square}$ | 0.002 | 0.002 | 0.009 |
| | WAIC$^{\square}$ | 0.001 | 0.002 | 0.012 |
| | G-ODIN$^{\square}$ | 0.001 | 0.003 | 0.014 |
| Backbone + Ours | CSI$^{\triangle}$ | 0.002 | 0.003 | 0.010 |
| | CnC$^{\triangle}$ | 0.001 | 0.003 | 0.006 |
| | VOS$^{\triangle}$ | 0.001 | 0.002 | 0.010 |

Table 28. Normalized OOD percentage estimation error $\varepsilon$ comparison for our method and IAD for OOD detection models that do and do not use negative samples with CIFAR-10 as ID dataset and Celeb_A as OOD dataset of different $P^{\star}_{\text{OOD}}$. "Any" indicates any of the backbone OOD detection models.

| Methods | Backbone | $P^{\star}_{\text{OOD}}$ | | |
|---|---|---|---|---|
| | | 0.010 | 0.020 | 0.050 |
| Backbone + IAD [18] | Any$^{\square,\triangle}$ | N/A (0.980) | N/A (0.960) | N/A (0.900) |
| Backbone + Ours | DB$^{\square}$ | 0.001 | 0.002 | 0.016 |
| | LReg$^{\square}$ | 0.000 | 0.001 | 0.000 |
| | LRat$^{\square}$ | 0.000 | 0.001 | 0.006 |
| | WAIC$^{\square}$ | 0.000 | 0.001 | 0.016 |
| | G-ODIN$^{\square}$ | 0.001 | 0.001 | 0.012 |
| Backbone + Ours | CSI$^{\triangle}$ | 0.001 | 0.002 | 0.014 |
| | CnC$^{\triangle}$ | 0.001 | 0.002 | 0.004 |
| | VOS$^{\triangle}$ | 0.000 | 0.002 | 0.012 |

Table 29. Normalized OOD percentage estimation error $\varepsilon$ comparison for our method and IAD for OOD detection models that do not use negative samples with different datasets of $P^{\star}_{\text{OOD}} = 1\%$. "Any" indicates any of the backbone OOD detection models.

| ID Dataset | OOD Dataset | Method | Backbone | | |
|---|---|---|---|---|---|
| | | | LReg$^{\square}$ | LRat$^{\square}$ | WAIC$^{\square}$ |
| Any | Any | Backbone + IAD [18] | N/A (0.980) | N/A (0.980) | N/A (0.980) |
| MNIST | EMNIST | Backbone + Ours | 0.001 | 0.001 | 0.002 |
| | FMNIST | | 0.001 | 0.001 | 0.001 |
| CIFAR-10 | Celeb_A | | 0.000 | 0.000 | 0.000 |
| | SVHN | | 0.001 | 0.000 | 0.001 |

Table 30. Normalized OOD percentage estimation error $\varepsilon$ comparison for our method and IAD for OOD detection models that use negative samples with different datasets of $P^{\star}_{\text{OOD}} = 1\%$. "Any" indicates any of the backbone OOD detection models.

| ID Dataset | OOD Dataset | Method | Backbone | | |
|---|---|---|---|---|---|
| | | | CSI$^{\triangle}$ | CnC$^{\triangle}$ | VOS$^{\triangle}$ |
| Any | Any | Backbone + IAD [18] | N/A (0.980) | N/A (0.980) | N/A (0.980) |
| MNIST | EMNIST | Backbone + Ours | 0.002 | 0.001 | 0.001 |
| | FMNIST | | 0.002 | 0.001 | 0.000 |
| CIFAR-10 | Celeb_A | | 0.001 | 0.001 | 0.000 |
| | SVHN | | 0.001 | 0.002 | 0.001 |

Table 31. Normalized OOD percentage estimation error $\varepsilon$ comparison for our method and IAD for OOD detection models that use negative samples with different datasets of $P^{\star}_{\mathrm{OOD}} = 1\%$. “Any” indicates any of the backbone OOD detection models.

| ID Dataset | OOD Dataset | Method | Backbone | |
|---|---|---|---|---|
| | | | DB$^{\square}$ | VOS$^{\triangle}$ |
| Any | Any | Backbone + IAD [18] | N/A (0.980) | N/A (0.980) |
| MNIST | EMNIST | Backbone + Ours | 0.001 | 0.001 |
| | FMNIST | | 0.000 | 0.000 |
| CIFAR-10 | Celeb_A | | 0.001 | 0.000 |
| | SVHN | | 0.000 | 0.001 |
| MNIST | EMNIST | Backbone + Ours (w/o transition $\tau^{(t)}$) | 0.003 | 0.003 |
| | FMNIST | | 0.002 | 0.001 |
| CIFAR-10 | Celeb_A | | 0.002 | 0.005 |
| | SVHN | | 0.002 | 0.004 |