

Table 1. Anomaly Detection Accuracy on Cifar10 (ROCAUC %)

	OC-SVM	Deep SVDD	GEOM	GOAD	MHRot	DN2
0	70.6	61.7 \pm 1.3	74.7 \pm 0.4	77.2 \pm 0.6	77.5	93.9
1	51.3	65.9 \pm 0.7	95.7 \pm 0.0	96.7 \pm 0.2	96.9	97.7
2	69.1	50.8 \pm 0.3	78.1 \pm 0.4	83.3 \pm 1.4	87.3	85.5
3	52.4	59.1 \pm 0.4	72.4 \pm 0.5	77.7 \pm 0.7	80.9	85.5
4	77.3	60.9 \pm 0.3	87.8 \pm 0.2	87.8 \pm 0.7	92.7	93.6
5	51.2	65.7 \pm 0.8	87.8 \pm 0.1	87.8 \pm 0.6	90.2	91.3
6	74.1	67.7 \pm 0.8	83.4 \pm 0.5	90.0 \pm 0.6	90.9	94.3
7	52.6	67.3 \pm 0.3	95.5 \pm 0.1	96.1 \pm 0.3	96.5	93.6
8	70.9	75.9 \pm 0.4	93.3 \pm 0.0	93.8 \pm 0.9	95.2	95.1
9	50.6	73.1 \pm 0.4	91.3 \pm 0.1	92.0 \pm 0.6	93.3	95.3
Avg	62.0	64.8	86.0	88.2	90.1	92.5

Table 2. Anomaly Detection Accuracy on Fashion MNIST and CIFAR10 (ROCAUC %)

	OC-SVM	GEOM	GOAD	DN2
FashionMNIST	92.8	93.5	94.1	94.4
CIFAR100	62.6	78.7	-	89.3

We conduct experiments against state-of-the-art methods, deep-SVDD (Ruff et al., 2018) which combines OCSVM with deep feature learning. Geometric (Golan & El-Yaniv, 2018), GOAD (Bergman & Hoshen, 2020), Multi-Head RotNet (MHRot) (Hendrycks et al., 2019). The latter three all use variations of RotNet.

For all methods except DN2, we reported the results from the original papers if available. In the case of Geometric (Golan & El-Yaniv, 2018) and the multi-head RotNet (MHRot) (Hendrycks et al., 2019), for datasets that were not reported by the authors, we run the Geometric code-release for low-resolution experiments, and MHRot for high-resolution experiments (as no code was released for the low-resolution experiments).

Cifar10: This is the most common dataset for evaluating unimodal anomaly detection. CIFAR10 contains 32×32 color images from 10 object classes. Each class has 5000 training images and 1000 test images. The results are presented in Tab. 1, note that the performance of DN2 is deterministic for a given train and test set (no variation between runs). We can observe that OC-SVM and Deep-SVDD are the weakest performers. This is because both the raw pixels as well as features learned by Deep-SVDD are not discriminative enough for the distance to the center of the normal distribution to be successful. Geometric and later approaches GOAD and MHRot perform fairly well but do not exceed 90% ROCAUC. DN2 significantly outperforms all other methods.

In this paper, we choose to evaluate the performance of without finetuning between the dataset and simulated anomalies (which improves performance on all methods including DN2). Outlier Exposure is one technique for such finetuning. Although it does not achieve the top performance by itself, it reported improvements when combined with MHRot to achieve an average ROCAUC of 95.8% on CIFAR10. This and other ensembling methods can also improve the performance of DN2 but are out-of-scope of this paper.

Fashion MNIST: We evaluate Geometric, GOAD and DN2 on the Fashion MNIST dataset consisting of 6000 training images per class and a test set of 1000 images per class. We present a comparison of DN2 vs. OCSVM, Deep SVDD, Geometric and GOAD. We can see that DN2 outperforms all other methods, despite the data being visually quite different from Imagenet from which the features were extracted.

CIFAR100: We evaluate Geometric, GOAD and DN2 on the CIFAR100 dataset. CIFAR100 has 100 fine-grained classes with 500 train images each or 20 coarse-grained classes with 2500 train images each. Following previous papers, we use the coarse-grained version. The protocol is the same as CIFAR10. We present a comparison of DN2 vs. OCSVM, Deep SVDD, Geometric and GOAD. The results are inline with those obtained for CIFAR10.

Comparisons against MHRot:

We present a further comparison between DN2 and MHRot (Hendrycks et al., 2019) on several commonly-used datasets. The experiments give further evidence for the generality of DN2, in datasets where RotNet-based methods are not restricted by low-resolution, or by image invariance to rotations.

We compute the ROCAUC score on each of the first 20 categories (all categories if there are less than 20), by alphabetical order, designated as normal for training. The standard train and test splits are used. All test images from