

Table 5. Anomaly Detection Accuracy on Multimodal Normal Image Distributions (ROCAUC %)

Dataset	Geometric	DN2
CIFAR10	61.7	71.7
CIFAR100	57.3	71.0

boxes provided with the data, and take each object with a bounding box of at least 120 pixels in each axis. We resize it to 256×256 pixels. We follow the same protocol as in the earlier datasets. As the images are of high-resolution, we use the public code release of Hendrycks (Hendrycks et al., 2018) as a self-supervised baseline. The results are summarized in Tab. 4. We can see that DN2 significantly outperforms MHRot. This is due both to the generally stronger performance of the feature extractor as well as the lack of rotational prior that is strongly used by RotNet-type methods. Note that the images are centered, a prior used by the MHRot translation heads.

WBC (Zheng et al., 2018): To further investigate the performance on difficult real world data, we performed an experiment on the WBC Image Dataset, which consists of high-resolution microscope images of different categories of white blood cells. The data do not have a preferred orientation. Additionally the dataset is very small, only a few tens of images per-class. We use Dataset 1 that was obtained from Jiangxi Telecom Science Corporation, China, and split it to the 4 different classes that contain more than 20 images each. We set the first 80% images in each class to the train set, and the last 20% to the test set. The results are presented in Tab. 4. As expected, DN2 outperforms MHRot by a significant margin showing its greater applicability to real world data.

4.2. Multimodal Anomaly Detection

It has been argued (e.g. Ahmed & Courville (2019)) that unimodal anomaly detection is less realistic as in practice, normal distributions contain multiple classes. While we believe that both settings occur in practice, we also present results on the scenario where all classes are designated as normal apart from a single class that is taken as anomalous (e.g. all CIFAR10 classes are normal apart from "Cat"). Note that we do not provide the class labels of the different classes that compose the normal class, rather we consider them to be a single multimodal class. We believe this simulates the realistic case of having a complex normal class consisting of many different unlabelled types of data.

We compared DN2 against Geometric on CIFAR10 and CIFAR100 on this setting. We provide the average ROCAUC across all the classes in Tab. 5. DN2 achieves significantly stronger performance than Geometric. We believe this is

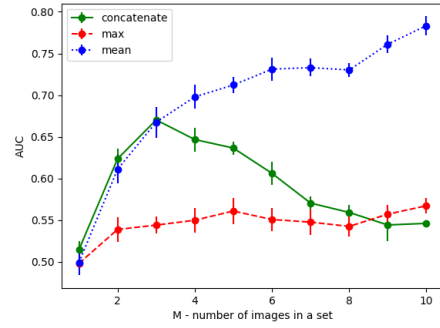


Figure 4. Number of images per group vs. detection ROCAUC. Group anomaly detection with mean pooling is better than simple feature concatenation for groups with more than 3 images.

occurs as Geometric requires the network not to generalize on the anomalous data. However, once the training data is sufficiently varied the network can generalize even on unseen classes, making the method less effective. This is particularly evident on CIFAR100.

4.3. Generalization from Small Training Datasets

One of the advantage of DN2, which does not utilize learning on the normal dataset is its ability to generalize from very small datasets. This is not possible with self-supervised learning-based methods, which do not learn general enough features to generalize to normal test images. A comparison between DN2 and Geometric on CIFAR10 is presented in Fig. 5. We plotted the number of training images vs. average ROCAUC. We can see that DN2 can detect anomalies very accurately even from 10 images, while Geometric deteriorates quickly with decreasing number of training images. We also present a similar plot for FashionMNIST in Fig. 5. Geometric is not shown as it suffered from numerical issues for small numbers of images. DN2 again achieved strong performance from very few images.

4.4. Unsupervised Anomaly Detection

There are settings where the training set does not consist of purely normal images, but rather a mixture of unlabelled normal and anomalous images. Instead we assume that anomalous images are only a small fraction of the number of the normal images. The performance of DN2 as function of the percentage of anomalies in the training set is presented in Fig. 5. The performance is somewhat degraded as the percentage of training set impurities exist. To improve the performance, we proposed a cleaning stage, which removes 50% of the training set images that have the most distant kNN inside the training set. We then run DN2 as usual. The performance is also presented in Fig. 5. Our cleaning procedure is clearly shown to significantly improve