

Table 3. MHRot vs. DN2 on Flowers, Birds, CatsVsDogs (Average Class ROCAUC %)

Dataset	MHRot	DN2
Oxford Flowers	65.9	<b>93.9</b>
UCSD Birds 200	64.4	<b>95.2</b>
CatsVsDogs	88.5	<b>97.5</b>

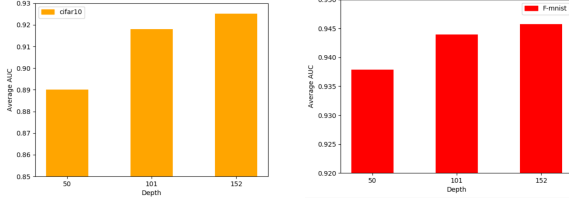


Figure 1. Network depth (number of ResNet layers) improves both Cifar10 and FashionMNIST results.

all classes are used for inference, with the appropriate class designated normal and all the rest as anomalies. For brevity of presentation, the average ROCAUC score of the tested classes is reported.

*102 Category Flowers (Nilsback & Zisserman, 2008)*: This dataset consists of 102 categories of flowers, consisting of 10 training images each. The test set consists of 30 to over 200 images per-class.

*Caltech-UCSD Birds 200 (Wah et al., 2011)*: This dataset consists of 200 categories of bird species. Classes typically contain between 55 to 60 images split evenly between train and test.

*CatsVsDogs (Elson et al., 2007)*: This dataset consists of 2 categories; dogs and cats with 10,000 training images each. The test set consist of 2,500 images for each class. Each image contains either a dog or a cat in various scenes and taken from different angles. The data was extracted from the ASIRRA dataset, we split each class to the first 10,000 images as train and the last 2,500 as test.

The results are shown in Tab. 3. DN2 significantly outperforms MHRot on all datasets.

#### Effect of network depth:

Deeper networks trained on large datasets such as Imagenet learn features that generalize better than shallow network. We investigated the performance of DN2 when using features from networks of different depths. Specifically, we plot the average ROCAUC for ResNet with 50, 101, 152 layers in Fig. 1. DN2 works well with all networks but performance is improved with greater network depth.

#### Effect of the number of neighbors:

The only free parameter in DN2 is the number of neigh-

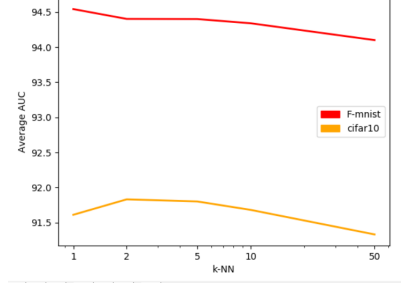


Figure 2. Number of neighbors vs ROCAUC, the optimal number of K is around 2.

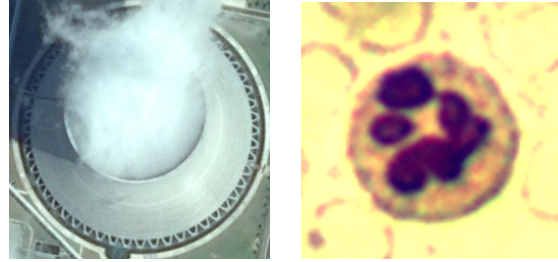


Figure 3. (left) A chimney image from the DIOR dataset (right) An image from the WBC Dataset.

bors used in kNN. We present in Fig. 2, a comparison of average CIFAR10 and FashionMNIST ROCAUC for different numbers of nearest neighbors. The differences are not particularly large, but 2 neighbors are usually best.

#### Effect of data invariance:

Methods that rely on predicting geometric transformations e.g. (Golan & El-Yaniv, 2018; Hendrycks et al., 2019; Bergman & Hoshen, 2020), use a strong data prior that images have a predetermined orientation (for rotation prediction) and centering (for translation prediction). This assumption is often false for real images. Two interesting cases not satisfying this assumption, are aerial and microscope images, as they do not have a preferred orientation, making rotation prediction ineffective.

*DIOR (Li et al., 2020)*: An aerial image dataset. The images are registered but do not have a preferred orientation. The dataset consists of 19 object categories that have more than 50 images with resolution above  $120 \times 120$  (the median number of images per-class is 578). We use the bounding

Table 4. Anomaly Detection Accuracy on DIOR and WBC (ROCAUC %)

Dataset	MHRot	DN2
DIOR	83.2	<b>92.2</b>
WBC	60.5	<b>82.9</b>