表3. 花、鳥、猫対犬におけるMHRotとDN2の比較（平均クラスROC AUC %）

Dataset	MHRot	DN2
Oxford Flowers	65.9	**93.9**
UCSD Birds 200	64.4	**95.2**
CatsVsDogs	88.5	**97.5**

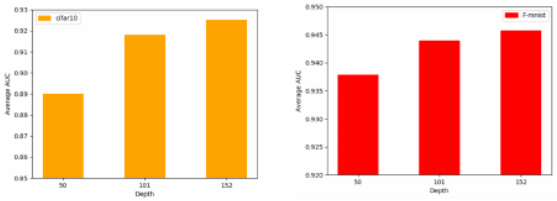


図1. ネットワークの深さ（ResNetの層数）は、Cifar10とFashionMNISTの結果を向上させた。

すべてのクラスが推論に使用され、適切なクラスが正常とされ、残りはすべて異常とされる。表示を簡潔にするため、テストされたクラスの平均ROCAUCスコアが報告されている。

102 Category Flowers (Nilsback & Zisserman, 2008)： このデータセットは102カテゴリの花からなり、それぞれ10枚の学習画像から構成される。テストセットは1クラスあたり30枚から200枚以上の画像で構成される。

Caltech-UCSD Birds 200 (Wah et al., 2011)： このデータセットは、鳥類の200のカテゴリーからなる。クラスは通常55から60の画像を含み、trainとtestで均等に分割される。

CatsVsDogs (Elson et al., 2007)： このデータセットは、犬と猫の2つのカテゴリから構成され、それぞれ10,000枚のトレーニング画像がある。テストセットは各クラス2,500枚の画像から構成される。各画像には、様々なシーンで様々な角度から撮影された犬か猫が含まれている。データはASIRRAデータセットから抽出され、各クラスを最初の10,000画像をトレーニング、最後の2,500画像をテストとして分割した。

結果を表3に示す。3. DN2は全てのデータセットにおいてMHRotを大きく上回る。

ネットワークの深さの効果：

Imagenetのような大規模データセットで訓練された深いネットワークは、浅いネットワークよりも汎化する特徴を学習する。異なる深さのネットワークからの特徴を使用した場合のDN2の性能を調査した。具体的には、50層、101層、152層のResNetの平均ROCAUCを図1にプロットする。DN2はすべてのネットワークでうまく機能するが、ネットワークの深さが深くなるほど性能は向上する。

近傍探索数の効果

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

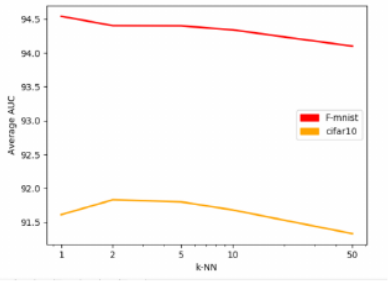
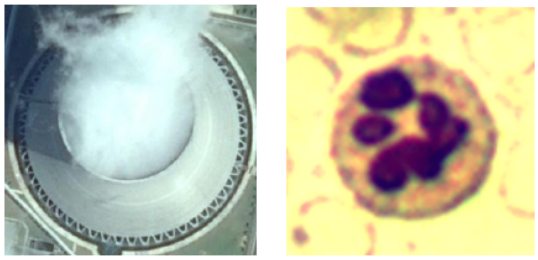


図2. 最適なKの数は約2である。



左）DIORデータセットからの煙突画像 （右）WBCデータセットからの画像。

kNNで使用されるボル 異なる最近傍数に対するCIFAR10とFashionMNISTの平均ROCAUCの比較を図2に示す。差は特に大きくないが、通常2近傍が最良である。

データ不変性の効果

幾何学的変換の予測に依存する手法、例えば（Golan & El-Yaniv, 2018; Hendrycks et al., 2019; Bergman & Hoshen, 2020）は、画像が（回転予測のための）所定の方向と（並進予測のための）中心を持つという強力なデータ事前分布を用いる。この仮定は実際の画像ではしばしば誤りである。この仮定を満たさない2つの興味深いケースは、空撮画像と顕微鏡画像で、これらは好ましい方向を持たないため、回転予測は効果がない。

DIOR (Li et al., 2020)： 航空画像データセット。画像は登録されているが、好ましい方向はない。このデータセットは、120 × 120以上の解像度を持つ50枚以上の画像を持つ19のオブジェクト・カテゴリから構成される（1クラスあたりの画像数の中央値は578枚）。バウンディング

表4. DIORとWBCの異常検出精度（ROCAUC）

Dataset	MHRot	DN2
DIOR	83.2	**92.2**
WBC	60.5	**82.9**

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

Dataset	MHRot	DN2
Oxford Flowers	65.9	**93.9**
UCSD Birds 200	64.4	**95.2**
CatsVsDogs	88.5	**97.5**

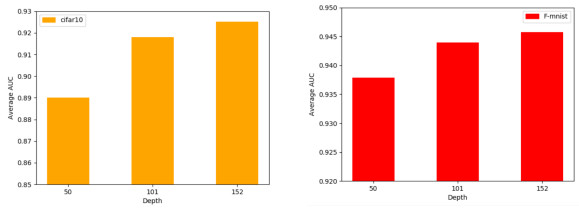


Figure 1. Network depth (number of ResNet layers) improves both Cifar10 and FashionMNIST results.

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 neighbors 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.

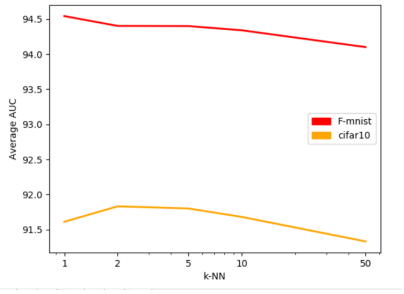


Figure 2. Number of neighbors vs ROCAUC, the optimal number of K is around 2.

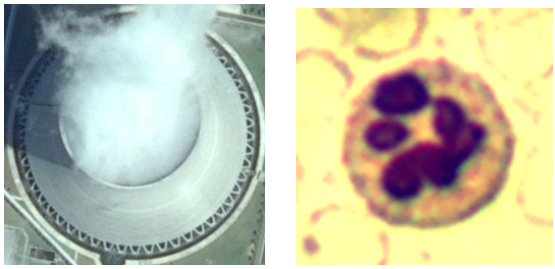


Figure 3. (left) A chimney image from the DIOR dataset (right) An image from the WBC Dataset.

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 120×120 (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	**92.2**
WBC	60.5	**82.9**