表5. マルチモーダル正規画像分布における異常検出精度（ROCAUC %） # 表5.

Dataset	Geometric	DN2
CIFAR10	61.7	**71.7**
CIFAR100	57.3	**71.0**

各オブジェクトのバウンディングボックスは各軸に少なくとも120ピクセルある。それを256×256ピクセルにリサイズする。以前のデータセットと同じプロトコルに従う。画像は高解像度であるため、自己教師付きベースラインとしてHendrycks（Hendrycks et al. 結果をTab. 4. DN2がMHRotを大幅に上回っていることがわかる。これは、特徴抽出器の性能が一般的に強いことと、RotNetタイプの手法で強く使われる回転事前分布がないことの両方によるものである。画像はセンタリングされていることに注意してください。

WBC（Zheng et al., 2018）： 困難な実世界データでの性能をさらに調査するために、我々は、白血球の異なるカテゴリの高解像度顕微鏡画像からなるWBC画像データセットで実験を行った。このデータには好ましい方向がない。さらにこのデータセットは非常に小さく、1クラスあたり数十枚の画像しかない。我々は、中国のJiangxi Telecom Science Corporationから入手したデータセット1を使用し、それぞれ20枚以上の画像を含む4つの異なるクラスに分割する。各クラスの最初の80%の画像を学習セットに、最後の20%をテストセットに設定する。結果を表4に示す。4. 予想通り、DN2はMHRotを大差で上回り、実データへの適用性の高さを示している。

4.2. マルチモーダル異常検知

実際には正規分布には複数のクラスが含まれるため，単峰性の異常検出は現実的でないと主張されている（Ahmed & Courville（2019）など）。我々は、両方の設定が実際に発生すると考えているが、異常とみなされる1つのクラス（例えば、「Cat」を除くすべてのCIFAR10クラスが正常である）を除いて、すべてのクラスが正常であると指定されるシナリオの結果も提示する。我々は正常クラスを構成する異なるクラスのクラスラベルを提供せず、むしろそれらを1つのマルチモーダルなクラスとみなしていることに注意してください。これは、多くの異なるラベルのないタイプのデータからなる複雑な正常クラスを持つ現実的なケースをシミュレートしていると考えている。

CIFAR10とCIFAR100において、DN2とGeometricを比較した。全てのクラスにおける平均ROCAUCを表5に示す。5. DN2はGeometricよりも大幅に高い性能を達成しています。

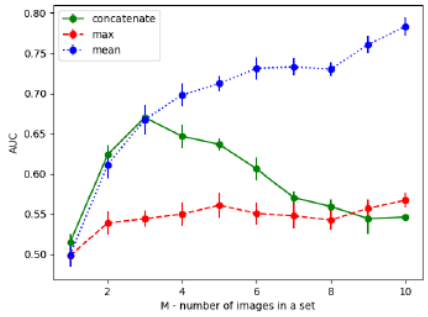

図4. 図4.グループ毎の画像数と検出ROCAUCの比較。平均プーリングを用いたグループ異常検出は、3枚以上の画像を持つグループに対して、単純な特徴連結よりも優れている。

これはGeometricが異常データに対してネットワークを汎化させないことを要求しているためであると考えられる。しかし、学習データが十分に変化すると、ネットワークは未見のクラスでも汎化できるようになり、この手法が有効でなくなる。これは特にCIFAR100で顕著である。

4.3. 小さな訓練データセットからの汎化

正常データセットでの学習を利用しないDN2の利点の1つは、非常に小さなデータセットからの汎化能力である。これは自己教師あり学習ベースの手法では不可能であり、通常のテスト画像に汎化するのに十分な一般的特徴を学習しない。CIFAR10におけるDN2とGeometricの比較を図5に示す。学習画像数と平均ROCAUCをプロットした。DN2は10枚の画像からでも非常に正確に異常を検出できるのに対し、Geometricは学習画像の枚数が減るにつれて急速に悪化することがわかる。同様のプロットをFashionMNISTについても図5に示す。Geometricは画像数が少ないと数値的な問題に悩まされるため示していない。DN2はまたもや非常に少ない画像から強力な性能を達成した。

4.4. 教師なし異常検知

学習セットが純粋な正常画像から構成されるのではなく、ラベル付けされていない正常画像と異常画像が混在する設定もある。その代わりに、異常画像は正常画像のごく一部であると仮定する。学習セット中の異常画像の割合に対するDN2の性能を図5に示す。訓練セットの不純物の割合が存在するにつれて、性能はやや低下する。性能を向上させるために、我々はクリーニングステージを提案し、トレーニングセット内の最も離れたkN Nを持つトレーニングセット画像の50%を除去する。その後、通常通りDN2を実行する。性能は図5にも示されている。我々のクリーニング手順により、DN2が大幅に改善されたことがわかる。

Readable

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 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.

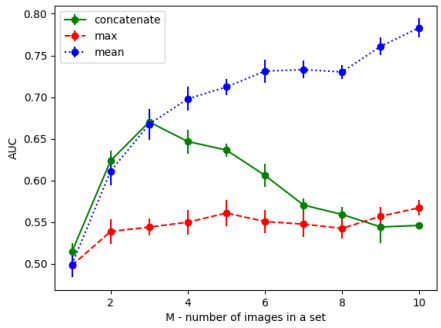

Figure 4. Number of images per group vs. detection ROCAUC. Group anomaly detection with mean pooling is better than simple feature concatenation for groups with more than 3 images.

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

Readable