1. Appendix

In this section we will further explain the datasets we used, present the gaussian nuture of the pre-trained features and demonstrate our method's robustness.

1.1. Datasets Details

CIFAR consists of two well known datasets, Cifar10 and Cifar100, that are used for various tasks including semantic anomaly detection [5]. Each dataset contains $60,000$ 32×32 color natural images, split into $50,000$ training images and $10,000$ test images. Cifar10 is composed of 10 equal-sized classes, whereas cifar100 has 100 equal-sized fine-grained classes or 20 equal-sized coarse-grained classes. Following the previous papers, we use the coarse-grained classes notation.

Fashion MNIST consists of $60,000$ train samples and $10,000$ examples test samples [8]. Each example is a 28×28 grayscale image labeled with one of 10 different categories.

Cats Vs Dogs is a dataset of images of cats and dogs. The training set contains $10,000$ images of cats and $10,000$ images of dogs, while the test set contains $2,500$ dog images and $2,500$ cat images. There is either a dog or a cat in every image, appearing in a variety of poses and scenes. Following previous work [4, 6], we split each class to the first $10,000$ images for training and the last 2,500 for testing.

Dior contains aerial images with 19 object categories. Following previous papers [4, 6], we used the bounding boxes provided with the data, and we took objects with at least 120 pixels in each axis as well as only classes with more than 50 images. This preprocessing phase led to 19 classes, with an average training size of 649 images. The sample sizes in each class are not equal, as the lowest sample size in the training set is 116 and the highest is 1890.

Blood Cells [7] contains 320×240 augmented color images of four different cell types. The training set contains approximately $2,500$ images for each blood cell type, whereas the test set contains approximately 620 images for each type of blood cell.

Covid19 [1] is a dataset of Chest X-ray images of Covid19, Pneumonia and normal patients.We ignore the Pneumonia patients' scans and have used just the Covid19 and normal scans. Covid19 patients' chest X-rays have been divided into 460 images in the training set and 116 images in the test set. The chest X-ray images of normal patients have been divided into $1,266$ images for the training set and 317 images for the test set. Normal patients' scans are obviously considered normal, while Covid19 patients' scans are considered anomalous.

View Recognition [2] is an image dataset of natural scenes around the world. This dataset is composed of six different classes such as images of forest and streets. The training set contains approximately $2,300$ images for each class, while the test set contains approximatly 500 images for each class.

Weather Recognition [3] is a multi-class dataset of weather images designed for image classification. There are four types of outdoor weather images in this dataset, including shine and rain. The training set consists of approximately 225 images per class, while the test set contains approximately 55 images per class.

Concrete Crack Classification [9] contains 227×227 color concrete images with and without cracks. There are $16,000$ images per class in the training set and $4,000$ per class images in the test set. Images of concrete without cracks are considered normal, while images of concrete with cracks are considered anomalous.

1.2. Gaussian nature of data

In this section, we presents the fine-tuned feature empirical distribution, that explains why a Gaussian is used to model this data. Figure 1 shows the Teacher-Student fine-tuned features of the last ViT block, using class 0 samples as the normal training set. As one can observe, the fine-tuned features follow a distribution close to Gaussian, which motivate us to use a Gaussian to model the data. We observed similar empirical distributions using different ViT blocks and other normal classes.

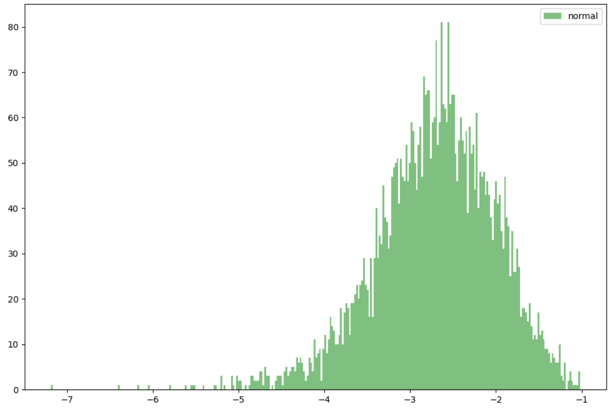

Figure 1. **Gaussian nature of data:** We show here the Teacher-Student fine-tuned features from the last ViT block using class 0 samples as the normal training set. This distribution can easily be fitted with a Gaussian model, explaining the good results we get using this module. The behaviour of other fine-tuned features of other classes is similar.

1.3. Transformaly Robust Results

To further demonstrate our robust results, we repeated our unimodal experiments three times. In each of the three trials, we calculate the average AUROC score across all

Dataset	CSI	DN2	PANDA	MSAD	Ours
CIFAR10	94.3	92.5	96.2	97.2	**98.34**(±0.018)
CIFAR100	89.6	94.1	94.1	96.4	**97.60**(±0.184)
FMNIST	-	94.5	**95.6**	94.21	94.37(±0.041)
CatsVsDogs	86.3	96.0	97.3	99.3	**99.47**(±0.037)
DIOR	78.5	92.2	94.3	97.2	**98.33**(±0.177)

Table 1. **AUROC scores of the unimodal setting:** In each trial we calculate the mean AUROC score across all classes of the datasets. We repeat this process for three trials reporting its means and standard deviations. Other benchmarks's AUROC scores are copied from the original table.

possible class choices. Table 1 shows the mean and standard deviation scores of our method, calculated over these three trails. Transformaly achieved similar results, still outperforming other methods on all datasets, except for FMNIST.

References

[1] Chest X-ray (Covid-19 & Pneumonia). 1

[2] Intel Image Classification | Kaggle. 1

[3] Gbeminiyi Ajayi. Multi-class Weather Dataset for Image Classification. 1, Sept. 2018. Publisher: Mendeley Data. 1

[4] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020. 1

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[6] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021. 1

[7] shenggan. BCCD Dataset, Oct. 2021. original-date: 2017-12-07T11:54:25Z. 1

[8] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 1

[9] Çağlar Fırat Özgenel. Concrete Crack Images for Classification. 2, July 2019. Publisher: Mendeley Data. 1