WDiscOOD: Out-of-Distribution Detection via Whitened Linear Discriminant Analysis

Yiye Chen Yunzhi Lin Ruinian Xu Patricio A. Vela
Georgia Institute of Technology
{yychen2019, yunzhi.lin, rnx94, pvela}@gatech.edu

Abstract

Deep neural networks are susceptible to generating overconfident yet erroneous predictions when presented with data beyond known concepts. This challenge underscores the importance of detecting out-of-distribution (OOD) samples in the open world. In this work, we propose a novel feature-space OOD detection score based on class-specific and class-agnostic information. Specifically, the approach utilizes Whitened Linear Discriminant Analysis to project features into two subspaces - the discriminative and residual subspaces - for which the in-distribution (ID) classes are maximally separated and closely clustered, respectively. The OOD score is then determined by combining the deviation from the input data to the ID pattern in both subspaces. The efficacy of our method, named WDiscOOD, is verified on the large-scale ImageNet-1k benchmark, with six OOD datasets that cover a variety of distribution shifts. WDiscOOD demonstrates superior performance on deep classifiers with diverse backbone architectures, including CNN and vision transformer. Furthermore, we also show that WDiscOOD more effectively detects novel concepts in representation spaces trained with contrastive objectives, including supervised contrastive loss and multi-modality contrastive loss.

1. Introduction

Deep learning models are typically designed with a *closed-world* assumption [46], where test data is assumed to be drawn from the same distribution as the training data. Without any built-in mechanism to distinguish novel concepts, deep neural networks are prone to generate incorrect answers with high confidence when presented with *out-of-distribution (OOD)* data. Such a misleading decision could result in catastrophic consequences in applications, which makes the deep neural network hard to deploy in the *open world*. To address this issue, significant research efforts have been devoted to the problem of OOD detection, which aims to establish a robust method for identifying when the testing data is "unknown."

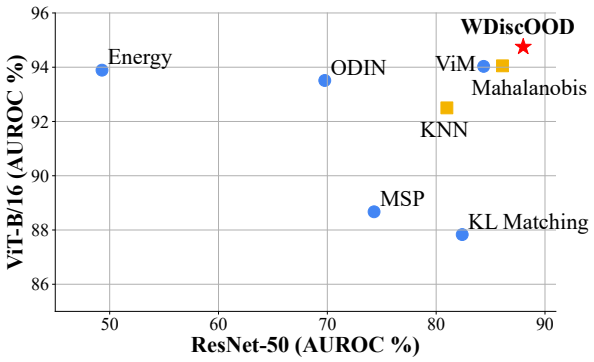

Figure 1: **Performance on two OOD detection benchmarks.** We propose a novel OOD detection method based on *Whitened Linear Discriminant Analysis*, wihch is highlighted with red star. It outperforms baselines with both ResNet-50 (x-axis) and ViT-B/16 (y-axis) model on ImageNet-1k benchmark. Blue dots and yellow square denote classifier-based and feature-based methods, respectively. The AUROCs are averaged over six OOD datasets, for which the detailed results are tabulated in Tab. 1.

Specifically, OOD detection requires establishing a scoring function to separate ID and OOD data. Designed for deep visual classifiers, the majority of works examine activation patterns from the classification layer. For example, a popular baseline is to leverage the maximum posterior probability output of a softmax classifier to indicate ID-ness [21], assuming that the network is more confident with its decision for ID data. The idea is enhanced by neural network calibration techniques [17, 32, 24], which aim to align network confidence with actual likelihood. The technique has been shown to improve OOD detection performance. Other classifier-based scores explore unnormalized posterior probabilities known as logits [20, 33], and norms of gradients backpropagated from the classification layer [25, 29].

Recently, ViM [44] argues that both class-dependent and class-agnostic information can potentially facilitate the OOD detection task. Based on the idea, it designs a scoring function by mapping the feature-space principle component residual to the logit space. Despite the great per-

formance achieved, ViM still requires supervision to train the classification layer. Meanwhile, advances in pretraining large-scale visual encoders using contrastive learning techniques [3, 27] have occured in recent years. These methods no longer rely on jointly training deep visual encoders with task heads, but instead formulate objectives directly in the feature space to produce high-quality visual representations. The learnt encoder can be applied to downstream tasks with little or no fine-tuning required [36, 39], suggesting a new paradign for addressing computer vision problems. While it is non-trivial to adopt classifier-based OOD detection methods directly to visual encoders, applying feature space methods is straightforward. For example, SSD [38] applies Mahalanobis [31] distance in the contrastive feature space, and demonstrates superior performance compared to a typical classification visual encoder.

In this work, we aim to reason about both class-specific and class-agnostic information *solely within the feature space*. To achieve this, we utilize Whitened Linear Discriminant Analysis (WLDA) to project visual features into two subspaces: a discriminative subspace and a residual subspace. The former contains compact class-discriminative signals, while the latter constrains shared information. We refer to these subspaces as Whitened Discriminative Subspace (WD) and Whitened Discriminative Residual Subspace (WDR), respectively. Given the compactness of the WD space, we detect anomalies by measuring the distance to the nearest class center. Conversely, in the WDR space where in-distribution (ID) classes are entangled, we examine the distance to the centroid of all training data as an OOD indicator. The final proposed scoring function unifies the information from both subspaces by computing a weighted sum of the scores in each space.

While subspace techniques for OOD detection have been explored previously [5], our approach differs significantly. Assuming ID data lies on a low-dimensional manifold, existing approaches measure the residual magnitude as an indicator for OOD-ness [34]. In contrast, we assume the residual space captures rich class-shared information. Furthermore, while past literature examines the residual to principles [44] or classifier weights [5], we explore the remaining information to discriminative components. This design enables us to jointly reason with both discriminative and residual information without relying on task heads, which is more applicable to stand-alone visual encoders.

As shown in Fig. 1, WDiscOOD achieves superior performance on large-scale ImageNet-1k benchmark compared to a wide range of baselines, under both classic CNN and recent Visual Transformer (ViT) architectures. In addition, our method surpasses other feature-space approaches in distinguishing novel concepts from stand-alone contrastive encoders, involving Supervised Contrastive (SupCon) model [27] and Contrastive Language-Image Pre-Training (CLIP) model [36]. What's more, we discover that the Whitened Discriminative Residual Space is more effective in identifying anomalous responses compared to other subspace techniques or even a subset of classifier-based scoring methods, verifying the importance of the class-agnostic feature component for the task.

In summary, our contribution involves:

- A new OOD detection score, based on WLDA, that jointly considers class-discriminative and class-agnostic information solely within the feature space.
- A new insight into the effectiveness of the Whitened Discriminative Residual Subspace space, which captures the shared information stripped of discriminative signals, for detecting anomalies in OOD samples.
- New state-of-the-art results achieved by our method on the large-scale ImageNet OOD detection benchmark, for various visual classifiers (CNN and ViT) as well as contrastive visual encoders (SupCon and CLIP).

2. Related Work

Scoring Functions for Pretrained Models One line of work explores scoring functions to distinguish inliers and outliers, which is fundamental to the OOD detection task. Multiple designs have been proposed for pretrained deep visual classifiers [1, 28, 37, 16, 21, 33, 20, 25, 29, 40, 44, 34, 31, 41]. The most straightforward design is Maximum Softmax Probability (MSP) score, which considers the network confidence as an uncertainty measurement. ODIN [32] improves MSP's performance via two calibration techniques - input preprocessing and temperature scaling. Logit-space methods involve max logit [20] and energy score [33]. The latter is further enhanced by feature rectification [33]. Huang et al. [25] and Lee et al. [29] investigate gradient space, which demonstrates effectiveness in revealing ID and OOD distinction. Those methods achieve great performance on several OOD detection benchmarks, but are tailored to the classification task and not applicable to pretrained visual encoders since they are dependent on classifier outputs. On the other hand, methods based on feature space analysis, such as Mahalanobis [31] and KNN [41], not only exhibit exceptional OOD detection performance for classification models, but also showcase applicability to stand-alone visual encoders, including SupCon [27, 38] and CLIP [36, 14]. Subspace analysis [44, 34, 5] measures the residual information from low-dimensional manifold [34] or column space of classifier weights. However, to achieve state-of-the-art performance, the result needs to be converted to classifier output space to factor in class-dependent information [44]. In this work, we combine the residual and discriminative information solely in the feature space via Linear Discriminant Analysis, which makes our method applicable to any visual encorder.

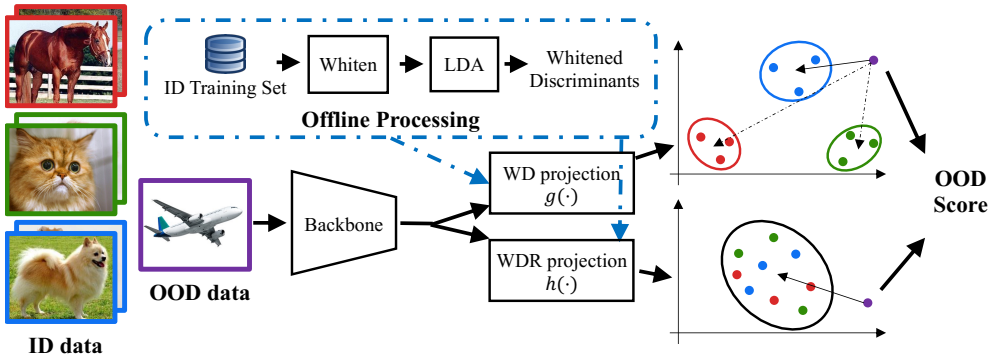

Figure 2: **Overview of WDiscOOD method.** We detect an OOD sample by projecting the feature into *Whitened Discriminative Subspace (**WD**)* and *Whitened Discriminative Residual Subspace (**WDR**)*, where ID classes are maximally separated and closely clustered, respectively. The projection functions are obtained from offline Whitened Linear Discriminant Analysis on the ID dataset. OOD samples are presumed to be far from the class clusters in WD space and the entire dataset clutter in WDR space. Therefore, we formulate the OOD score as a combination of the distance to the nearest class centroid in the WD space and to the entire dataset centroid in the WDR space.

Model Modifications & Training Constraints Another branch of OOD detection methods trains the network to respond differently to ID and OOD data, by either modifying the network structure [24] or adding specialized training regularization [26, 48, 47]. Along these lines, the most direct approach is to encourage the network to give distinguishable predictions for outlier data, such as a uniform posterior probability [22], lower energy [33], or confidence estimation from dedicated branch [7]. to do so requires an auxiliary OOD training set, commonly known as Outlier Exposure (OE) [22]. Several methods [8, 22] assume the availability of unknown data from outlier datasets, which can hardly cover all potention distribution shifts in the open world, and is not always feasible. Lee et al. [30] utilizes generative models, such as Generative Adversarial Networks (GAN) [15] for outlier data synthesis. However, generating high fidelity imagery induces optimization difficulties. To avoid these obstacles, VOS [11] synthesize outliers in the feature space, which can adapt to the ID feature geometry during training. Although effective, these methods require network re-training to endow the model with the ability to reject OOD data, which can be expensive and intractable. Especially for models trained on large-scale datasets. On the other hand, our OOD scoring function design can be directly applied on any pre-trained visual model.

3. Methods

The core of OOD detection lies in creating a scalar function $s(\cdot)$ that assigns distinguishable scores to ID and OOD data. This allows the query data to be classified as either ID or OOD by applying a threshold to the score. In this paper, we aim to assign higher scores for ID data. Our approach involves disentangling class-discriminative and class-general information from the features of a deep network's penultimate layer using Whitened Linear Discriminant Analysis. We then jointly reason with both types of information to effectively identify OOD data. We first summarize the LDA method in Sec.3.1, then explain the proposed WDiscOOD score based on LDA in Sec. 3.2 .

3.1. Multiclass Linear Discriminant Analysis

The objective of Linear Discriminant Analysis (LDA) [13, 12] is to find a set of projection directions, termed *discriminants*, along which multi-class data are maximally separated. Specifically, given a training dataset $\{(\boldsymbol{x}_i \in \mathbb{R}^D, c_i)\}_i^N$ of D-dimensional features from C classes ($c_i \in \{1,2,\cdots,C\}$), the objective of LDA is to find the direction $\boldsymbol{w}$ such that the *Fisher discriminant criterion* [2], defined as the ratio of inter-class variance over intra-class variance, is maximized:

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}, \tag{1}$$

where $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are between-class scatter matrix and within-class scatter matrix. Denote the cardinality for class c as N_c, then the scatter matrices are formulated as:

$$\boldsymbol{S}_w = \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_{c_i})(\boldsymbol{x}_i - \boldsymbol{\mu}_{c_i})^T \text{ and} \tag{2}$$

$$\boldsymbol{S}_b = \sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T, \tag{3}$$

where $\boldsymbol{\mu}$ is the center of the entire training dataset.

The optimization program defined in Eq. 1 is resolved by solving the generalized eigenvalue problem:

$$\boldsymbol{S}_b \boldsymbol{w} = \lambda \boldsymbol{S}_w \boldsymbol{w}, \tag{4}$$

where the generalized eigenvalue is equal to the Fisher criterion for the corresponding eigenvector. Intuitively, a larger

Fisher criterion suggests that ID classes are better separated. Therefore, projecting original features along top discriminants, known as the *Foley-Sammon Transform (FST)*, maps the original feature to a low-dimensional subspace where ID classes are compactly clustered. As described in the following section, we utilize FST to disentangle discriminative and residual information from visual feature space.

3.2. WDiscOOD: Whitened Linear Discriminant Analysis for OOD detection

For a given extracted image feature $\boldsymbol{z}$, Fig. 2 depicts the proposed WDiscOOD score to involve three steps: (1) offline data whitening with the statistics derived from the training dataset; (2) discriminative and residual space mapping with discriminants and discriminant orthogonal complement set, which are estimated based on offline LDA; (3) a weighted sum of distances to ID cluster(s) in both spaces as the final WDiscOOD score.

Data Whitening We begin by whitening the feature prior to conducting LDA. The application of whitened LDA has been extensively studied and implemented across various domains [35, 18]. According to Hariharan et al. [18], data whitening eliminates the correlation between feature elements and enhances the feature's capability to encode data similarity. Furthermore, it improves the numerical conditioning of LDA, particularly in scenarios with small sample sets. In this paper, whitening is applied to the within-class covariance matrix of Eq. 2. Suppose the eigenvalue decomposition for the covariance matrix on the feature space: $\boldsymbol{S}_{\boldsymbol{z},w} = \boldsymbol{V}_w \boldsymbol{\Lambda}_w \boldsymbol{V}_w^T$, where $\boldsymbol{V}_w \in \mathbb{R}^{D\times r}$, $\boldsymbol{\Lambda}_w \in \mathbb{R}^{r\times r}$, r denotes matrix rank, then the feature is whitened by:

$$\boldsymbol{x} = \boldsymbol{S}_{\boldsymbol{z},w}^{-1/2}\boldsymbol{z} = \boldsymbol{V}_w \boldsymbol{\Lambda}_w^{-1/2} \boldsymbol{V}_w^T \boldsymbol{z} \tag{5}$$

Discriminative and Residual Space Mapping To isolate the class-specific and class-agnostic components of the whitened visual feature, we leverage LDA to project the data onto two distinct subspaces. The first subspace, referred to as the *Whitened Discriminative Subspace (WD)*, encodes the discriminative information of the data by compactly grouping samples from the same class and separating different classes. The second subspace, referred to as the *Whitened Discriminative Residual Subspace (WDR)*, captures the residual of the discriminative information by clustering together samples from all classes.

Following [9], we solve LDA by adding a multiple of identity matrix to the within-class scatter matrix $\boldsymbol{S}_w + \rho \boldsymbol{I}$. It stablizes small eigenvalues and ensures a sufficiently well-conditioned scatter matrix. It also converts the generalized eigenvalue problem in Eq. 4 to an eigenvalue problem:

$$(\boldsymbol{S}_w + \rho \boldsymbol{I})^{-1} \boldsymbol{S}_b \boldsymbol{w} = \lambda \boldsymbol{w} \tag{6}$$

With discriminants solved from LDA, project the features into WD via Foley-Sammon Transformation. To be specific, construct the projection matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{N_D}] \in \mathbb{R}^{D\times N_D}$ by stacking the top N_D discriminants corresponding to the largest Fisher criterion. The WD projection is defined to be

$$g(\boldsymbol{x}) = \boldsymbol{W}^T \boldsymbol{x}. \tag{7}$$

To capture the class-agnostic information, project the data onto the subspace spanned by the orthogonal complements of the top-N_D discriminants, e.g., to the WDR space. The orthogonal complements correspond to projection directions with lower Fisher criterion values, which implies that the separation between classes is less significant compared to intra-class variance. Thus, WDR space captures shared information among the ID classes, which can be used to formulate additional constraints for ID data. Formally, suppose the eigendecomposition: $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{Q}\boldsymbol{\Lambda}_{\boldsymbol{W}}\boldsymbol{Q}^T$, then the WDR projection of a query feature is

$$h(\boldsymbol{x}) = (\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{x}. \tag{8}$$

WDiscOOD Score The WDiscOOD score combines ID data constraints in both WD and WDR spaces. In the discriminative subspace, where discrepency between known classes is maximized, all ID data is in close proximity to some class cluster. Formulate the WD space OOD detection score as the distance to the nearest class center $\boldsymbol{\mu}_c^{WD}$:

$$s_g(\boldsymbol{x}) = -\min_c \|g(\boldsymbol{x}) - \boldsymbol{\mu}_c^{WD}\|_2 \tag{9}$$

In the WDR space where inter-class discrepency is minimized, the data from ID classes tends to scatter around a shared centroid. Measure the OOD score in WDR space as the distance to the center of all ID training data $\boldsymbol{\mu}^{WDR}$:

$$s_h(\boldsymbol{x}) = -\|h(\boldsymbol{x}) - \boldsymbol{\mu}^{WDR}\|_2 \tag{10}$$

The design is distinct from residual norm score from prior work [44, 34], which assumes minimal information left from ID data that is irrelevant for the classification task. Instead, we assume that the residual space contains abundant shared information, motivating us to formulate the distance-based score in Eq. 10.

To incorporate information from both spaces, the WDiscOOD score is defined as the weighted sum of both scores:

$$s(\boldsymbol{x}) = s_g(\boldsymbol{x}) + \alpha s_h(\boldsymbol{x}) \tag{11}$$

4. Experiments

This section compares the described approach with state-of-the-art methods on a large-scale OOD detection benchmark to demonstrate the effectiveness of the core concept.

Evaluation is done for visual classifiers and stand-alone constrative visual encoders. Additionally, a comprehensive ablation study provides further insights into WDiscOOD.

ID and OOD datasets Following recent work [26], testing involves a large-scale OOD detection task with ImageNet-1k [6] as the ID dataset. Six test OOD datasets are applied for evaluation, including: *SUN* [45], *Places* [49], *iNaturalist* [43], *Textures* [4], *ImageNet-O* [23], and *OpenImage-O* [44]. The first three (*SUN*, *Places*, *iNaturalist*) use the subset curated by [26] with non-overlapping categories *w.r.t* the ID dataset. Note that this evaluation is more comprehensive than previous OOD detection literature investigating ImageNet benchmark [44, 41], in the sense that they only adopt a subset of the above OOD datasets. Utilizing a variety of OOD data sources encompasses a wider range of distributional shift patterns, thus enabling a more comprehensive evaluation of OOD detection methods.

Evaluation Metrics Evaluation uses two commonly adopted metrics that quantify a scoring function's ability to distinguish ID and OOD data. The first is *Area under the Receiver Operating Characteristic Curve (AUROC)*, a threshold-free metric that measures the area under the plot of the true positive rate (TPR) against the false positive rate (FPR) under varying classification thresholds. The AUROC metric is advantageous as it is invariant to the ratio of positive sample number to that of negative sample, making it suitable for evaluating OOD detection task, where the number of ID and OOD samples is imbalanced. Higher value indicates better performance. The second is *False positive rate at* 95% *true positive rate (FPR95)* (smaller is better).

Models Settings and Hyperparameters Testing is done with classification models for ImageNet-1k [6] having various backbones. The first is a ResNet-50 backbone [19], the most widely applied convolutional neural network. The second is a Vision Transformer (ViT), a transformer-based vision model that processes an input image as a sequence of patches. Following [41], we adopt the officially released ViT-B/16 architecture pretrained on ImageNet-21k and finetuned for classification on ImageNet-1k.

For stand-alone visual encoders, we test with Supervised Contrastive (SupCon) [27] and Contrastive Language-Image Pre-Training (CLIP) [36] models. The former is optimized to encourage smilarity between the embeddings of samples from the same class, while maximizing the distance between them for different classes. The latter is a multi-modality representation learning method, trained to pull together the features for matched image-text pairs, and push them away for non-matching pairs. Both encoders adopt a ResNet-50 backbone with officially released weights. Different from [14], we discard the language encoder from CLIP and perform OOD detection only in the visual feature space. This removes the assumption of the availability of a textural ID class name or description, which is not always feasible in real-world application. While both models utilize a projection head to a low-dimensional embedding space to formulate the training objectives, we leverage the backbone penultimate layer feature for better OOD detection performance following [41].

Per ViM [44], we adopt different hyparameter settings according to feature dimension. For ResNet-50 encoder with $D = 2048$ dimensional features, we set the number of discriminants as $N_D = 1000$ and the score weight in Eq. 11 as $\alpha = 5$. On the other hand, we adopt $N_D = 500$ and $\alpha = 1$ for ViT feature space of $D = 768$ dimensionality. When performing Linear Discriminant Analysis, we sample $N = 200,000$ training images that are evenly distributed among all ID classes for the estimation of statisticcal quantities, such as means and scatter matrices.

Baseline Methods Comparison uses nine baselines that derive scores from pretrained models without requiring network modification or finetuning. Seven logit/probability-space methods are included: *MSP* [21], *Energy* [33], *ODIN* [32], *MaxLogit* [20], *KLMatch* [20], *ReAct* [40], and *ViM* [44]. For ReAct, we use Energy+ReAct with truncation percentile $p = 99$. Two feature-space baselines are included: *Mahalanobis* [31] and *KNN* [41]. For the Mahalanobis method, we follow SSD [38] to directly apply the scoring function on the final layer feature without input-precossing nor a multi-layer feature ensemble technique.

4.1. OOD Detection Performance on Classifiers

The first experiment presents the OOD detection comparison with classification models based on ResNet-50 and ViT backbones. Tab. 1a and Tab. 1b provide the quantitative results, consisting of outcomes for each OOD dataset and average performance across them. The best AUROC and FPR95 are in bold, with second and third rank underlined. On average across all OOD datasets, WDiscOOD exhibits superior performance compared to the baselines for both ViT and ResNet-50 models. This validates the efficacy of our OOD score function design.

Consistency across OOD distributions. Tajwar et al [42] shows on small-scale OOD benchmarks that existing OOD detectors do not have a consistent performance across OOD data sources. Consistency is important as the OOD data distribution is unpredictable in the open world. WDiscOOD consistently achieves top performance across all OOD datasets. As an illustration, it achieves top-3 performance in 11 out of 12 metrics (2 - AUROC and FPR95 - for each of the 6 OOD datasets) when used with the ViT-B/16 model, with 6 of them being the highest compared to the

Method	Textures		SUN		Places		iNaturalist		ImgNet-O		OpenImg-O		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Classifier-dependent methods														
MSP [21]	72.98	74.92	70.98	78.75	73.43	76.65	60.90	84.40	95.65	53.13	69.73	81.17	73.94	74.84
Energy [33]	95.74	48.60	97.93	50.12	97.77	48.90	98.12	50.86	92.80	48.23	95.41	52.33	96.30	49.84
ODIN [32]	75.94	69.33	75.51	74.05	77.54	71.28	68.60	79.88	94.95	51.19	73.98	76.15	77.75	70.31
MaxLogit [20]	75.92	69.33	75.51	74.05	77.55	71.28	68.57	79.88	94.95	51.19	73.97	76.15	77.74	70.31
KLMatch [20]	57.57	86.09	70.36	82.91	74.04	80.65	46.83	90.81	89.75	68.86	58.21	88.31	66.13	82.94
ReAct [40]	98.05	34.51	99.66	23.68	99.80	22.86	100.00	23.13	99.40	37.31	99.86	23.86	99.46	27.56
ViM [44]	25.18	92.63	69.22	81.39	74.90	76.40	30.02	93.38	76.15	77.08	46.70	88.60	53.70	84.91
Feature space methods														
Maha [31]	31.17	91.62	66.29	84.31	70.27	81.45	25.64	95.38	81.45	75.65	**44.36**	**91.41**	53.20	86.64
KNN [41]	**23.26**	**93.11**	88.59	74.01	89.00	71.07	74.60	85.83	**71.05**	**81.15**	70.29	84.01	69.47	81.53
WDiscOOD	29.20	91.90	**56.83**	**86.74**	**64.40**	**83.13**	**22.39**	**95.59**	81.60	75.52	44.67	90.51	**49.85**	**87.23**

(a) **ResNet-50 [19]**.

Method	Textures		SUN		Places		iNaturalist		ImgNet-O		OpenImg-O		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Classifier-dependent methods														
MSP [21]	52.43	85.42	53.22	86.93	57.75	85.72	13.66	97.00	51.75	85.81	31.99	92.48	43.47	88.89
Energy [33]	36.13	91.25	34.44	93.28	**42.80**	**90.98**	5.60	98.94	30.30	93.36	16.06	96.87	27.56	94.11
ODIN [32]	38.57	90.86	37.45	92.81	44.68	90.66	6.03	98.81	33.50	92.69	17.83	96.54	29.68	93.73
MaxLogit [20]	38.56	90.86	37.45	92.81	44.68	90.66	6.03	98.81	33.50	92.69	17.83	96.54	29.68	93.73
KLMatch [20]	51.22	85.12	56.04	85.45	61.08	83.86	13.68	96.32	49.90	85.62	31.38	91.93	43.88	88.05
ReAct [40]	**36.35**	91.17	34.55	93.22	43.32	90.83	5.61	98.94	30.30	93.40	16.01	96.88	27.69	94.07
ViM [44]	38.67	91.38	**32.47**	**93.41**	44.23	89.86	1.40	99.68	31.80	94.05	16.61	97.10	27.53	94.25
Feature space methods														
Maha [31]	36.61	91.67	35.37	92.89	46.08	89.55	0.96	99.78	30.45	94.22	**13.85**	**97.50**	27.22	94.27
KNN [41]	38.28	90.74	46.08	90.73	54.50	87.54	6.75	98.70	38.95	92.53	20.59	96.12	34.19	92.72
WDiscOOD	36.58	**91.79**	32.62	93.34	43.74	89.91	**0.89**	**99.81**	**30.15**	**94.36**	14.30	97.44	**26.38**	**94.44**

(b) **ViT-B/16 [10]**.

Table 1: **Results on ResNet-50 [19] and ViT [10] classification models**. We test all methods on six OOD datasets and compute the average performance. Both metrics AUROC and FPR95 are in percentage. We highlight the best performance in bold, and underline the 2nd and 3rd ones. Our method consistently outperforms all classifier-dependent or feature-space baselines under both network architectures in terms of average performance.

baseline methods. Close baselines are ViM and Maha (both achieve 7 top-3 including 2 top-1). but their top-1 results are achieved for one dataset, whereas WDiscOOD does so for three. Similar outcomes occur for ResNet-50. Examining both discriminative and residual information promotes the detection of a broader range of OOD patterns.

Robustness across encoder architectures. Another property of WDiscOOD is its model-agnostic nature. Its consistently high performance across network types demonstrates an ability to handle the diverse feature manifolds induced by distinct network architectures.

4.2. Results on Contrastive Visual Encoders

The WDiscOOD implementation is applied to the SupCon [27] and CLIP [36] models to verify its applicability on modern visual encoders. Since classifier-dependent baselines are not applicable to the stand-alone visual feature extractors, baseline comparison involves only the feature-space baselines - Mahalanobis [31] and KNN [41]. Tab. 2 collects the results. WDiscOOD surpasses the baselines for both models, suggesting that it is more effective in identifying OOD patterns across the feature spaces induced by diverse representation learning objectives.

Method	SupCon [27]		CLIP [36]	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Mahalanobis	46.95	89.78	78.00	75.31
KNN	42.51	90.35	82.59	67.22
WDiscOOD	**40.10**	**90.89**	**77.57**	**75.74**

Table 2: **Results on SupCon [27] and CLIP [36] visual encoders.** Our method surpasses other feature-space methods on these representation learning models. This table reports only average performance, while detailed results for each dataset can be found in the Supplementary material.

Method	ResNet-50		ViT	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
PR	56.72	84.01	34.59	92.44
WDR	**53.74**	**86.56**	**30.35**	**93.72**

Table 3: **Comparison between subspace techniques.** Our proposed WDR space is more effective than the Principle Residual (PR) space [44] for the OOD detection task. Furthermore, its AUROC on ResNet-50 outperforms 8 out of 9 baselines, which demonstrates that WDR is informative for OOD detection.

4.3. Understanding WDiscOOD

Discriminant Residual v.s. Principle Residual To confirm the efficacy of WDR space for OOD detection, it is compared to other subspace residual designs. ViM [44] bases their method on the residual of principle space, with empirical evidence on its effectiveness over the classifier null space method [5]. Comparative evaluation of this approach against the residual score in Eq. 10 on classification models is provided in Tab. 3.

The WDR space scoring function outperforms the previous subspace technique in terms of AUROC and FPR95 for both ResNet-50 and ViT architectures. Additionally, *without* the discriminative information, it performs competitively compared to SOTA in Tab. 1. In particular, the WDR space score achieves the third highest AUROC on ResNet-50 classifier, behind the integrated WDiscOOD and Mahalanobis scores. This placing supports the claim that WDR space is highly informative for OOD detection.

Effect of Feature Whitening We evaluate WDiscOOD with or without feature whitening on the ResNet-50 and ViT classification models. To isolate the effect of feature decorrelation on LDA, we also ablate on Euclidean distance vs Mahalanobis distance (whitening + Euclidean) in Eq. 9 and Eq. 10. Tab. 4 collects the results, which indicate that feature whitening improves (FR95) performance by a large margin. Replacing Euclidean distance with Mahalanobis after LDA slightly improves the results without whitening, but is insufficent to fill the gap. This suggests that feature whitening facilitates isolation of class-specific and residual information for LDA, which is crucial for our method. Mahalanobis *after* WLDA has similar performance as Euclidean, but requires additional scatter matrix estimates in subspaces. Therefore our design is justified.

Config		ResNet-50		ViT	
Whiten†	Dist	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
✗	Maha	53.65	86.20	29.81	93.47
✗	Eucl	74.56	81.17	32.21	93.52
✓	Maha	49.85	87.23	26.60	94.40
✓	Eucl	49.86	87.23	26.49	94.41

Table 4: **Ablation on Data whitening**. †: The option of data whitening before DLA. Whitening the data prior to DLA improves over no whitening (✗+ Eucl) or whitening after DLA (✗+ Maha). Furthermore, with prior data whitening, Euclidean (✓+Eucl) performs similar as Mahalanobis (✓+Maha) while requiring less computation, which justifies our design.

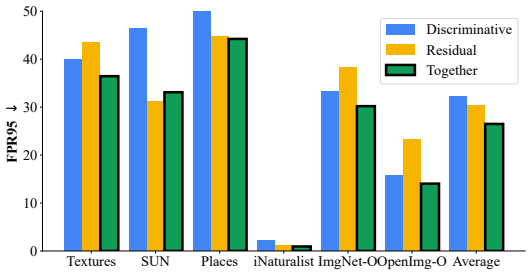

Figure 3: **The individual performance in WD and WDR space.** Our results demonstrate that two subspaces are effective in detecting OOD data from distinct distributions. Moreover, the integrated score (black border) yields a superior performance compared to each individual, evidencing that our method unifies class-discriminative and class-agnostic information.

Contribution by WD and WDR Spaces To understand the role played by each subspace, we break down the scoring function and evaluate the OOD detection *solely* within WD or WDR, with Eq. 9 or Eq. 10 as OOD scores, separately. The FPR95 performance is illustrated in Fig. 3. Our results demonstrate that two subspaces are effective in detecting different OOD patterns. Specifically, WD outperforms WDR on three OOD datasets (SUN, Places, INaturalist), while underperforming on the others (Textures, ImageNet-O, OpenImage-O). Additionally, the integrated score achieves better performance than each individual space on average, indicating that our method effectively leverages the complementary information provided by both subspaces. Unlike ViM [44], we combine the class-specific and class-agnostic information for OOD detection without relying on task heads, which improves its general use.

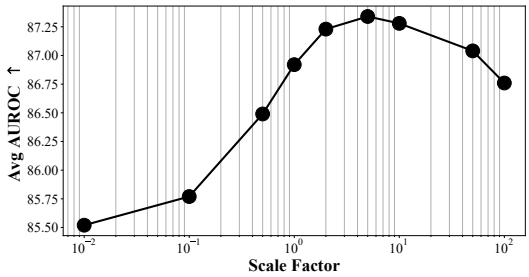

Figure 4: **Ablate on scoring scale factor**. The integrated performance is better than individual subspace over a wide range of scales, indicating that our simple linear combination is effective in factoring in information from both subspaces for OOD detection.

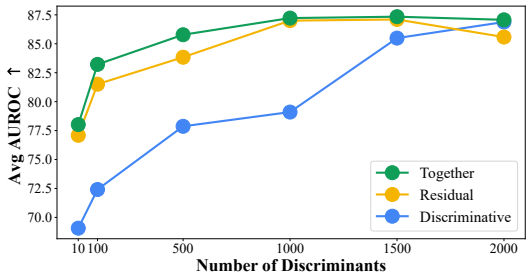

Figure 5: **Ablate on Discriminant Number N_D.** Our method achieves consistently high performance when the discriminant number is sufficiently large, enabling complete disentanglement of discriminative and residual information.

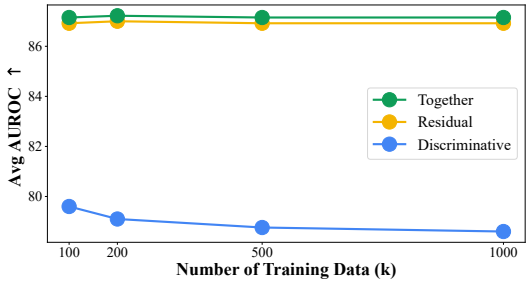

Figure 6: **Ablate on Training Data Number.** Our method achieves near-optimal performance using only 200K training data out of over 1000K, indicating that it is not excessively data-hungry.

Effect of scale factor α We ablate on the scale factor α in Eq. 11 on the ResNet-50 classification model. We fix the number of discriminants $N_D = 1000$ and test with different scales from the set: $\alpha \in \{0.01, 0.1, 0.5, 1, 2, 5, 10, 100\}$. Fig. 4 depicts the average AUROC under varied scaling factor. When the scaling factor is extreme, the score function biases towards one subspace over the other, leading to lower performance at both ends of the curve. The integrated scoring function outperforms the individual subspace scores across a wide range of scales, indicating that the proposed linear combination is effective at considering information from both subspaces.

Effect of Discriminant Number N_D We also ablate on the effect of discriminant number N_D, by testing our method with $N_D \in \{10, 100, 500, 1000, 1500, 2000\}$ on the same model. For each discriminant number, we test with varied scaling factor within the same set as above, and report the highest AUROC results. We further report individual performance from each subspace. The results are demonstrated in Fig. 5. As the discriminant number increases, *the individual performance in both WD and WDR space improves as the result of better separation between discriminative and residual information*. The trend stops when the number is sufficiently large, as the separation saturates. As a result, the intergrated performance becomes stable with increased discriminant number.

Robustness against Training Data Number N Here we explore the effects of the training data quantity N on the performance of WDiscOOD. We vary the quantity and plot the resulting average AUROC for individual subspace performance and integrated performance in Fig. 6. Near-optimal results occur with 20% of the training data (200K out of over 1000K), indicating the data efficiency of our method.

5. Conclusion

This paper presents a new OOD detection method based on Whitened Linear Discriminant Analysis (WLDA), named WDiscOOD. It jointly reasons with class-specific and class-agnostic information by disentangling the discriminative and residual information from the feature space via WLDA. Comprehensive evaluation shows that WDiscOOD establishes superior results on multiple large-scale benchmark, demonstrating robustness across model architectures as well as representation learning objectives. Analysis reveals the efficacy of Whitened Discriminative Residual Subspace (WDR) on OOD detection compared to other subspace techniques, showing the importance of understanding the behavior of OOD activation in the residual subspace.

References

[1] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1563–1572, 2016. 2

[2] Paul Bodesheim, Alexander Freytag, Erik Rodner, Michael Kemmler, and Joachim Denzler. Kernel null space methods for novelty detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3374–3381, 2013. 3

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the

wild. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 5

[5] Matthew Cook, Alina Zare, and Paul Gader. Outlier detection through null space analysis of neural networks. *arXiv preprint arXiv:2007.01263*, 2020. 2, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[7] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3

[8] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[9] Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. Deep linear discriminant analysis. *arXiv preprint arXiv:1511.04707*, 2015. 4

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[11] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022. 3

[12] J Duchene and S Leclercq. An optimal transformation for discriminant and principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):978–983, 1988. 3

[13] Donald H. Foley and John W Sammon. An optimal set of discriminant vectors. *IEEE Transactions on computers*, 100(3):281–289, 1975. 3

[14] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 2, 5

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[16] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *European Conference on Computer Vision*, pages 500–517. Springer, 2022. 2

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 1

[18] Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 459–472. Springer, 2012. 4

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5, 6

[20] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, pages 8759–8773. PMLR, 2022. 1, 2, 5, 6

[21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017. 1, 2, 5, 6

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 3

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 5

[24] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020. 1, 3

[25] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 1, 2

[26] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 3, 5

[27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2, 5, 6, 7, 12

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[29] Jinsol Lee and Ghassan AlRegib. Gradients as a measure of uncertainty in neural networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 2416–2420, 2020. 1, 2

[30] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 3

[31] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2, 5, 6, 11, 12

[32] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1, 2, 5, 6

[33] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 1, 2, 3, 5, 6

[34] Ibrahima Ndiour, Nilesh Ahuja, and Omesh Tickoo. Out-of-distribution detection with subspace techniques and probabilistic modeling of features. *arXiv preprint arXiv:2012.04250*, 2020. 2, 4

[35] Vo Dinh Minh Nhat, Sung Young Lee, and Hee Yong Youn. Whitened lda for face recognition. In *ACM international conference on Image and video retrieval*, pages 234–241, 2007. 4

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5, 6, 7, 12

[37] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *2020 International Joint Conference on Neural Networks*, pages 1–9. IEEE, 2020. 2

[38] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021. 2, 5, 11

[39] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022. 2

[40] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 2, 5, 6, 11

[41] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *International Conference on Machine Learning*, 2022. 2, 5, 6, 11, 12

[42] Fahim Tajwar, Ananya Kumar, Sang Michael Xie, and Percy Liang. No true state-of-the-art? ood detection methods are inconsistent across datasets. *arXiv preprint arXiv:2109.05554*, 2021. 5

[43] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 5

[44] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 1, 2, 4, 5, 6, 7, 11

[45] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 5

[46] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[47] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *IEEE/CVF international conference on computer vision*, pages 9518–9526, 2019. 3

[48] Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9452–9461, 2021. 3

[49] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 5

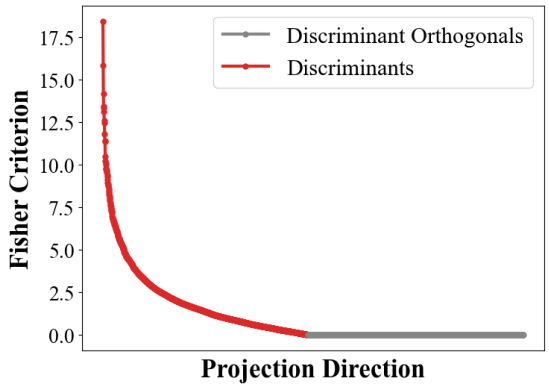

Figure 7: **Empirical Fisher Criterion (FC) values for ResNet-50 classification model.** FC values for discriminants are non-trivial whereas that for discriminant orthogonals approach zero. This verifies our assumption that WLDA disentangle discriminative and residual information from the feature space.

A. Model details

For ResNet-50 and ViT-B/16 classifiers, we adopt the feature encoder trained with a single-layer classification head on the ImageNet-1k training dataset. ViT-B/16 refers to the base model variant (layer=12, dimension $D = 768$, heads=12) with 16×16 input patch size. For the CLIP visual encoder, we adopt the ResNet-50 model trained with ViT-B/32 language encoder. We discard the language model and only use visual encoder in our experiments. The input data is cropped and resized to 224×224 for ResNet models (including ResNet-50 classification encoder, SupCon, and CLIP), and 384×384 for ViT-B/16. For both Mahalanobis [31] and WDiscOOD, we $L2$-normalize the feature for models directly trained on inner products between visual features (including SupCon model with Supervised Contrastive loss on normalized feature, and ViT with attention mechanism). We found that a normalized feature space enhances OOD detection performance when the inner product between image features is trained to encode similarity.

B. Baseline details

Mahalanobis We remove the input preprocessing and feature ensemble techniques proposed in the original paper [31] for small-scale benchmarks, as we find that they compromise the performance on large-scale benchmarks. Instead, we follow SSD [38] and apply the Mahalanobis distance directly to the penultimate layer feature. $200,000$ random training samples are used for calculating the precision matrix and class-wise centroids.

KNN For all models, we $L2$-normalize the features following the original work [41]. The KNN score is calculated on $200,000$ random training data. The nearest number is downscaled proportionally based on $k = 1000$ for the full training set.

ReAct Following the practice of [44], we use the most effective Energy+ReAct setting. We also adopt rectification percentile $p = 99$ instead of $p = 90$ from original work [40] for better performance.

Principle Residual (PR) The settings for Principle Residual (PR) baseline evaluated in Sec. 4.3 are adopted from ViM [44]. Prior to principle component estimation, we center the features based on classification layer weights and bias. 1000 principle components are used when the feature dimension is greater than 1500 (ResNet-50, SupCon,CLIP), otherwise 512 principle components are used (for ViT).

C. Empirical Fisher Criterion Value

To empirically verify our assumption that WLDA disentangles discriminative and residual information, we compare the Fisher Criterion (FC) values for discriminants and discriminant residuals from ResNet-50 classification model trained on ImageNet; see Fig. 7. The FC values for discriminants are non-trivial, indicating separation of ID features along the directions. On the other hand, the projections along discriminant orthogonal directions are non-separable, as the FC values in those subspaces are close to zero. This verifies our assumption that WLDA separate class-specific and class-agnostic information, and explains the superior performance of WDiscOOD method.

D. Detailed Results on SupCon and CLIP

Sec. 4.2 provides the average results for WDiscOOD and the feature-space baselines on SupCon and CLIP visual encoders. Here, Tab. 5 gives the AUROC and FPR95 measures on all six OOD datasets

E. OOD detection in the Embedding space

Both SupCon and CLIP formulate the contrastive loss in a low-dimensional feature space obtained from a projection head. As explained in Sec. 4, we follow KNN [41] and SSD [31] to apply all feature-space methods on the penultimate layer feature space for better performance. To further verify the claim, we test all feature-space methods, including KNN, Mahalanobis, and the proposed WDiscOOD, in the embedding spaces. For hyperparameters in WDiscOOD, we choose $N_D = 512$ and $\alpha = 1$ for CLIP embeddings with $D = 1024$ dimensions, and $N_D = 50$ and $\alpha = 1$ for SupCon model with embedding dimension as $D = 128$.

Method	Textures		SUN		Places		iNaturalist		ImgNet-O		OpenImg-O		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Maha [31]	14.80	95.62	63.09	86.76	68.93	84.20	33.41	95.06	**65.50**	83.00	35.96	94.05	46.95	89.78
KNN [41]	15.18	95.62	47.97	89.29	58.33	85.45	30.30	94.83	66.10	**83.88**	37.18	93.05	42.51	90.35
WDiscOOD	**13.94**	**95.84**	**47.87**	**89.40**	**58.21**	**86.34**	**21.49**	**96.21**	66.50	83.27	**32.59**	**94.26**	**40.10**	**90.89**

(a) **SupCon [27]**.

Method	Textures		SUN		Places		iNaturalist		ImgNet-O		OpenImg-O		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Maha [31]	54.11	89.77	**81.36**	77.45	83.87	78.21	97.74	56.41	76.50	**74.89**	**74.42**	**75.13**	78.00	75.31
KNN [41]	59.61	88.92	89.65	69.86	90.33	70.76	99.59	36.52	**75.35**	73.48	80.98	63.77	82.59	67.22
WDiscOOD	**54.10**	**89.85**	81.45	**78.33**	**81.54**	**80.14**	**96.81**	**57.69**	76.95	74.38	74.59	74.05	**77.57**	**75.74**

(b) **CLIP [36]**.

Table 5: **Results on SupCon [27] and CLIP [36] visual encoders**. We test all methods on six OOD datasets and compute the average performance. Both metrics AUROC and FPR95 are in percentage. We highlight the best performance in bold. WDiscOOD more consistently outperforms the alternatives for both encoders in terms of average FPR95 and AUROC.

Method	Space	SupCon [27]		CLIP [36]	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
Maha	Embed	54.74	86.61	97.06	67.83
KNN		55.96	86.40	96.42	61.81
WDisc		53.10	87.22	92.25	63.26
Maha	Last Layer	46.95	89.78	78.00	75.31
KNN		42.51	90.35	82.59	67.22
WDisc		**40.10**	**90.89**	**77.57**	**75.74**

Table 6: **Comparison between penultimate feature space and embedding space for SupCon [27] and CLIP [36] for all feature-space methods.** Low-dimensional embeddings are less information for OOD detection compared to penultimate layer features, suggesting potential loss of information critical for the task.

Comparision between performance in the embedding space and penultimate feature space is in Tab. 6, where all methods suffer from performance degradation in the embedding space. The results show that distance between visual features in the embedding space do not imply similarity, which is potentially caused by lost information due to limited dimensionality.