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.”

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 occurred 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 paradigm 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 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 encoder.
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 potential 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] synthesizes 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 \(\{(x_i \in \mathbb{R}^D, c_i)\}_i\) of \(D\)-dimensional features from \(C\) classes \(\{c_i \in \{1, 2, \ldots, C\}\}\), the objective of LDA is to find the direction \(w\) such that the Fisher discriminant criterion [2], defined as the ratio of inter-class variance over intra-class variance, is maximized:

\[
\max_w J(w) = \frac{w^T S_b w}{w^T S_w w},
\]

where \(S_b\) and \(S_w\) 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:

\[
S_w = \sum_{i=1}^{N} (x_i - \mu_c)(x_i - \mu_c)^T \quad \text{and} \quad S_b = \sum_{c=1}^{C} N_c(\mu_c - \mu)(\mu_c - \mu)^T,
\]

where \(\mu\) is the center of the entire training dataset. The optimization program defined in Eq. 1 is resolved by solving the generalized eigenvalue problem:

\[
S_b w = \lambda S_w w,
\]

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 $z$, 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: $S_{zw} = V_u \Lambda_u V_u^T$, where $V_u \in \mathbb{R}^{D \times r}$, $\Lambda_u \in \mathbb{R}^{r \times r}$, $r$ denotes matrix rank, then the feature is whitened by:

$$x = S_{zw}^{-1/2} z = V_u \Lambda_u^{-1/2} V_u^T z$$

(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 $S_{uw} + \rho I$. It stabilizes small eigenvalues and ensures a sufficiently well-conditioned scatter matrix. It also converts the generalized eigenvalue problem in Eq. 4 to an eigenvalue problem:

$$(S_{uw} + \rho I)^{-1} S_{uw} = \lambda w$$

(6)

With discriminants solved from LDA, project the features into WD via Foley-Sammon Transformation. To be specific, construct the projection matrix $W = \{w_1, w_2, \cdots, w_{N_D}\} \in \mathbb{R}^{D \times N_D}$ by stacking the top $N_D$ discriminants corresponding to the largest Fisher criterion. The WD projection is defined to be

$$g(x) = W^T x.$$  

(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: $WW^T = QA_WQ^T$, then the WDR projection of a query feature is

$$h(x) = (I - QQ^T)x.$$  

(8)

WDiscOOD Score The WDiscOOD score combines ID data constraints in both WD and WDR spaces. In the discriminative subspace, where discrepancy 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 $\mu_{cWD}$:

$$s_D(x) = -\min_c \|g(x) - \mu_{cWD}\|_2$$

(9)

In the WDR space where inter-class discrepancy 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 $\mu_{WD}$:

$$s_R(x) = -\|h(x) - \mu_{WD}\|_2$$

(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:

$$s(x) = s_D(x) + \alpha s_R(x)$$

(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 constrictive 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 similarity between the embeddings of samples from the same class, while maximizing the distance between them for different classes. The latter is a multimodality 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 [31].

Per ViM [44], we adopt different hyperparameter 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 statistical quantities, such as means and scatter matrices.

**Baseline Methods** Comparison uses nine baselines that derive scores from pre-trained models without requiring network modification or fine-tuning. 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-precessing 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 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
4.2. Results on Contrastive Visual Encoders

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 Sup-Con [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.

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] 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.

### 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.
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.

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.

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.

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.

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 insufficient 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 WDLA has similar performance as Euclidean, but requires additional scatter matrix estimates in subspaces. Therefore our design is justified.

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 Nq = 1000 and test with different scales from the set: α ∈ {0.01, 0.1, 0.5, 1, 2, 5, 10, 100}.
Fig. 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.

Fig. 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.

Fig. 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.

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**


