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

*Yiye Chen, Yunzhi Lin, Ruinian Xu, Patricio A. Vela*

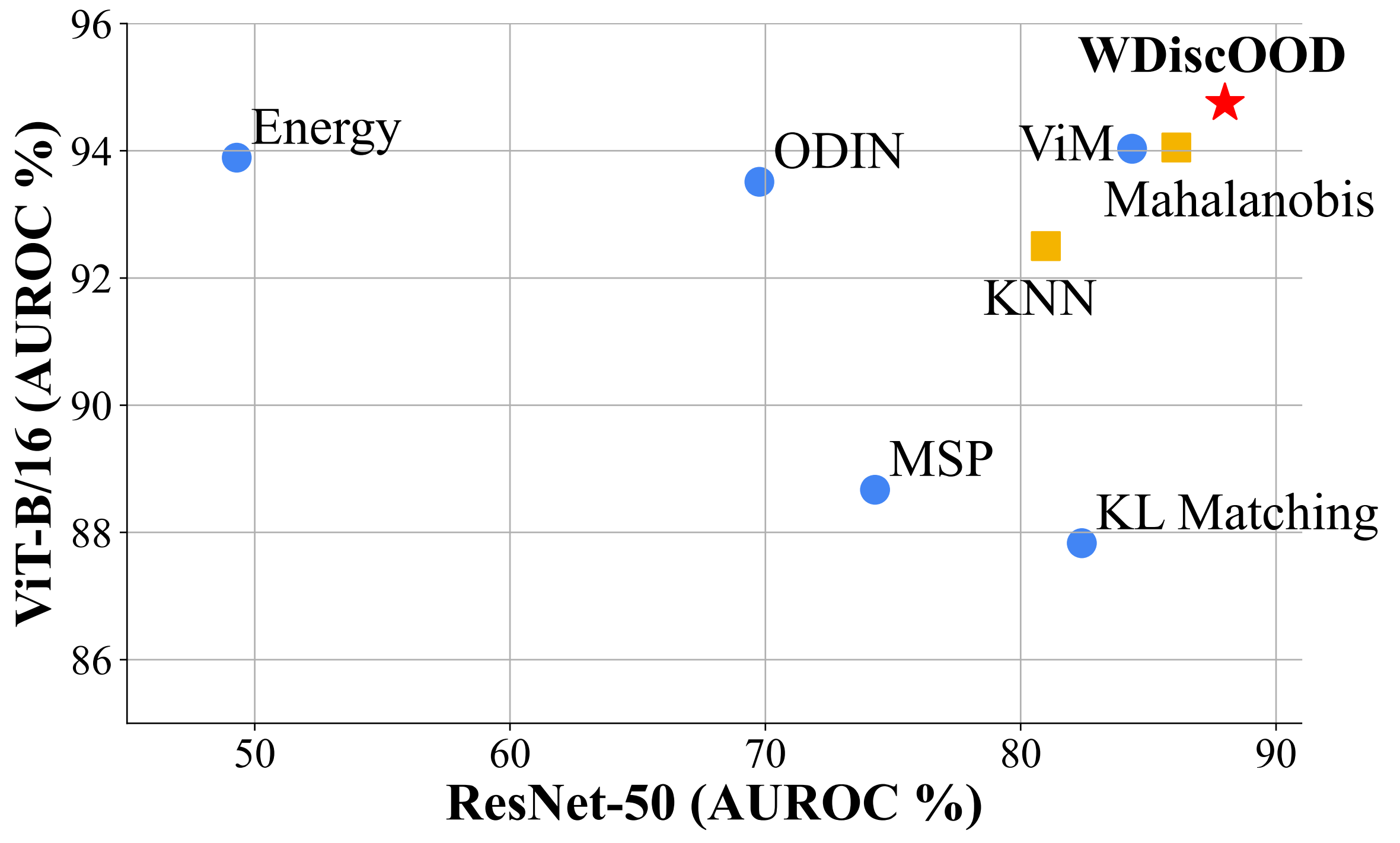
Institute for Robotics and Intelligent Machine , Georgia Tech

## Background & Goal

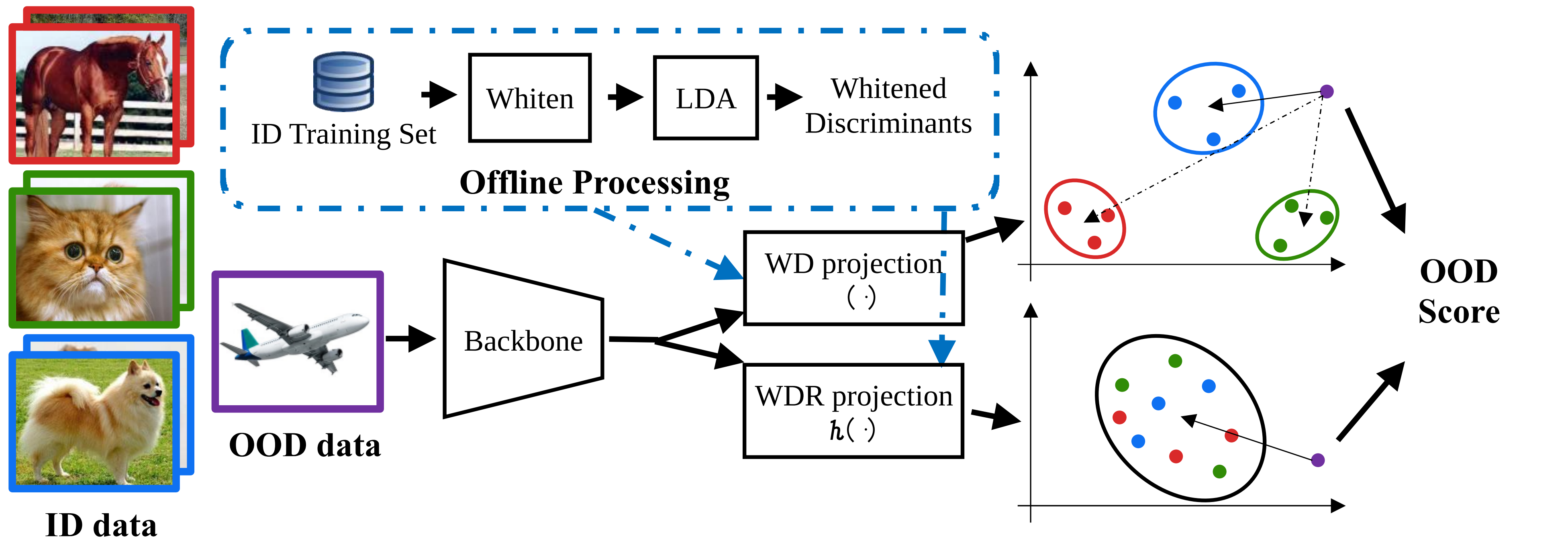
- ➢Deep vision models prone to generate incorrect predictions when given unfamiliar data (**Out-of-distribution, OOD**) vrelative to the training data (**In-distribution, ID**)
- ➢We study the OOD detection problem, where the goal is to develop a mechanism to distinguish between ID and OOD data
- ➢We aim to jointly reason about class-agnostic and class-specific information in the feature space

## Contributions

- ➢A new OOD scoring function based on Whitened Linear Discriminant Analysis (WLDA) in the feature space.
- ➢A new insight on the efficacy of the Whitened Discriminative Residual (WDR) Subspace on OOD detection.
- ➢New state-of-the-art results achieved on the large-scale ImagetNet OOD detection benchmark, under various settings including various visual classifiers (CNN & ViT) and contrastive visual encoders (SupCon & CLIP)



## Methodology



### 1. Data whitening

$$x = S_{z,w}^{-1/2}\, z$$

| | |
|---|---|
| $x$ | Whitened feature |
| $z$ | Original feature |
| $S_{z,w}^{-1/2}$ | Covariance matrix for $z$ |

### 2. Discriminative & Residual Decomposition

$$g(x) = W^T x \qquad h(x) = (I - QQ^T)x$$

| | |
|---|---|
| $g(\cdot)$ | **Whitened Discriminative (WD)** space projection |
| $h(\cdot)$ | **Whitened Discriminative Residual (WDR)** space projection |
| $W$ | Stack of top discriminants in $x$ space |
| $Q$ | Eigenvalues of $W$ |

### 3. OOD score

$$s_g(x) = -\min_c \left|\left|g(x) - \mu_c^{WD}\right|\right|_2;\ s_h(x) = -\left|\left|h(x) - \mu^{WDR}\right|\right|_2;\ \ s(x) = s_g(x) + \alpha s_h(x)$$

## Results – Classifiers (ResNet-50 & ViT)

| 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** |

| 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** |

WDiscOOD achieves superior results compared to a large set of baselines for ImageNet classifiers with various backbones including ResNet-50 and Vision Transformer (ViT)

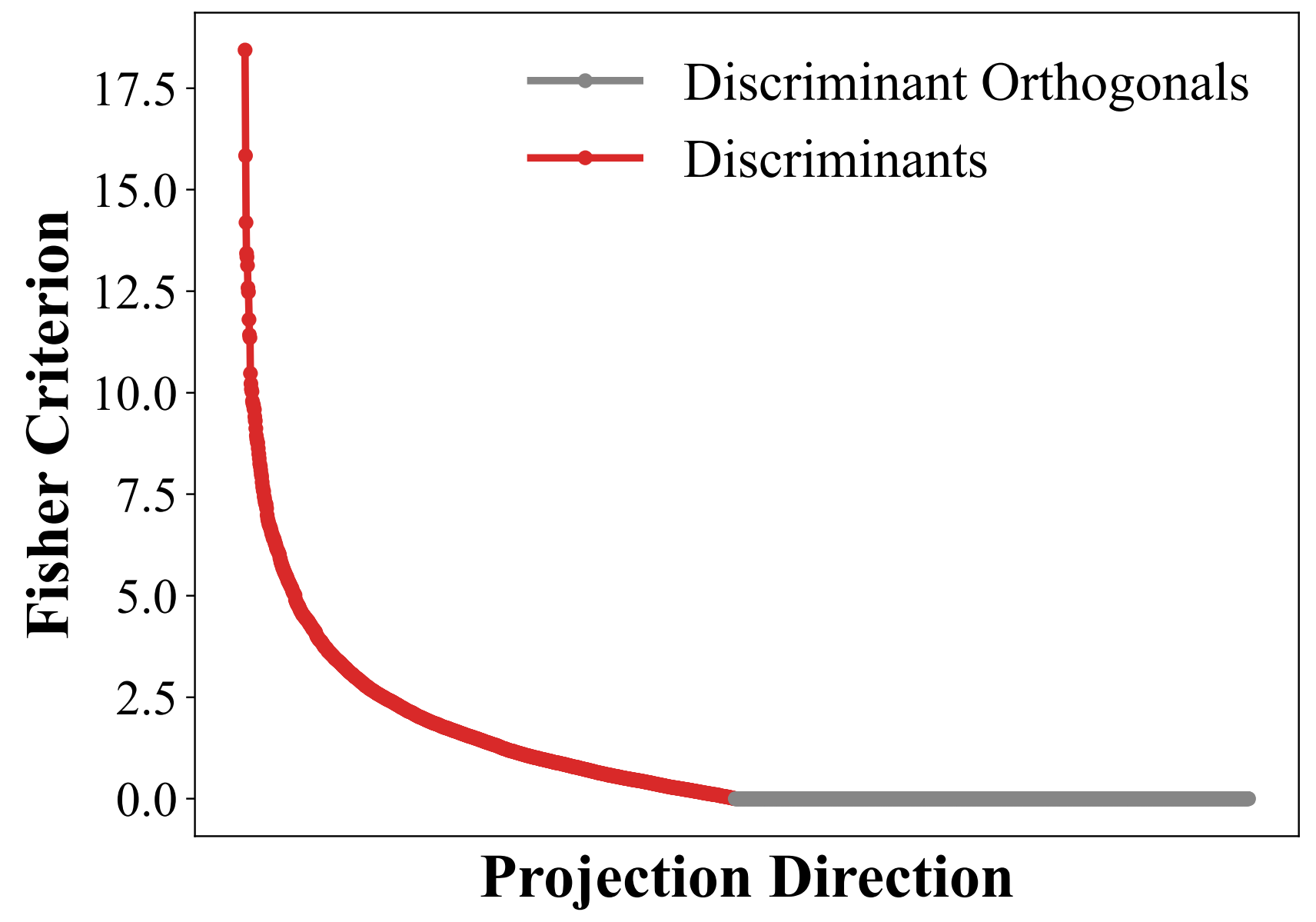
## Results – Contrastive Models (SupCon & CLIP)

| 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** |

- WDiscOOD is applicable for contrastive models as it is a feature space method that does not rely on any task head.
- It outperforms other feature space methods for SupCon and CLIP model on the ImageNet dataset.

## Method Understanding & Ablation Study

➢The separation of class-agnostic and class specific information



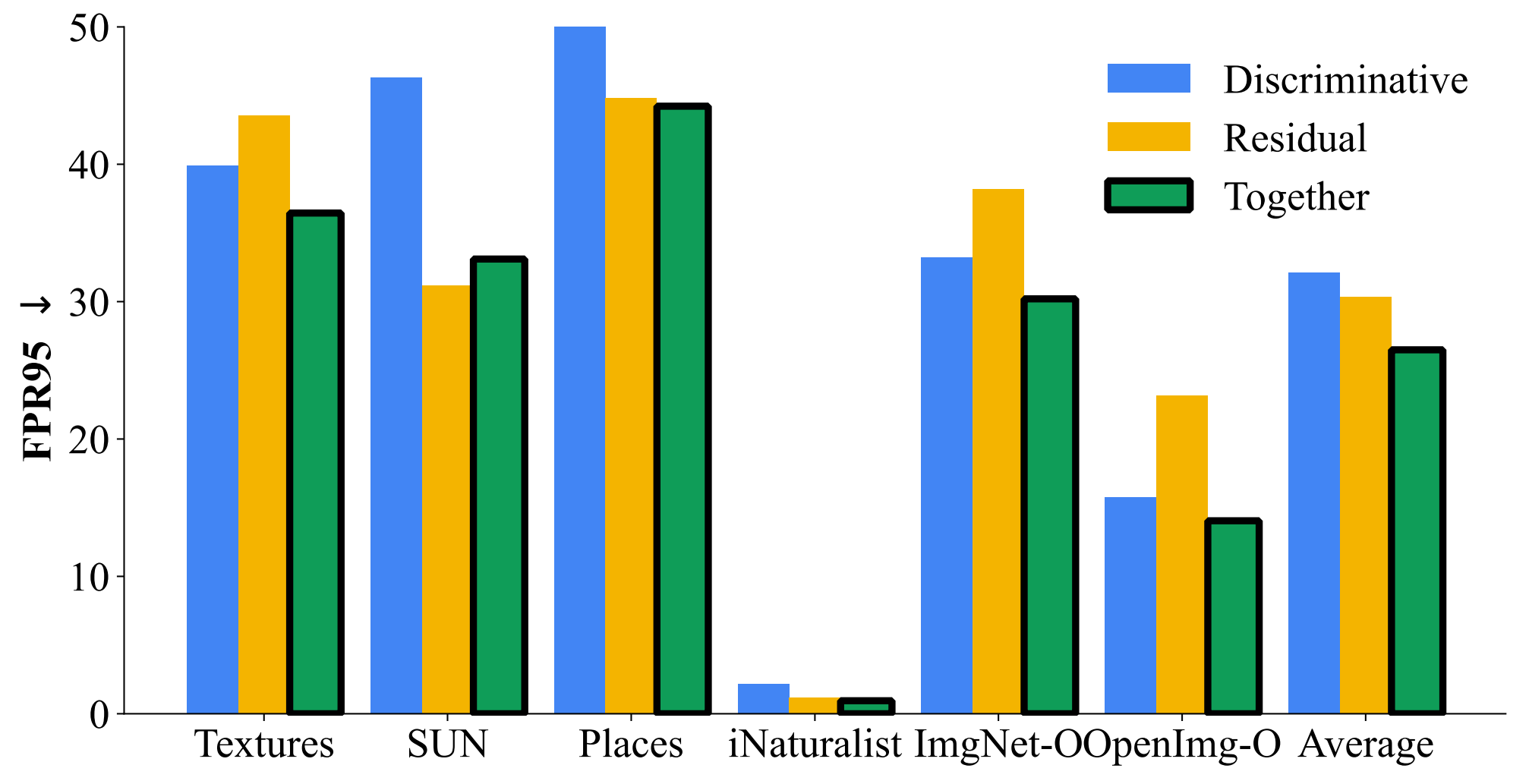
- *Fisher Criterion values are lower for discriminant orthogonals than discriminants*
- *This suggests that the feature projections in WD space are maximally separated into classes, and are closely clustered in WDR space.*

➢The importance of data whitening for OOD detection

| 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 |

*Feature whitening greatly improves the performance of feature-distance-based OOD detection, regardless of the distance type.*

➢WDR and WD spaces are complementary



- *The integrated score $s(\cdot)$ performs better than individual components $s_g(\cdot)$ and $s_h(\cdot)$.*
- *Residual space is more critical than the discriminant space.*

Evaluate under FPR95, which is the lower the better