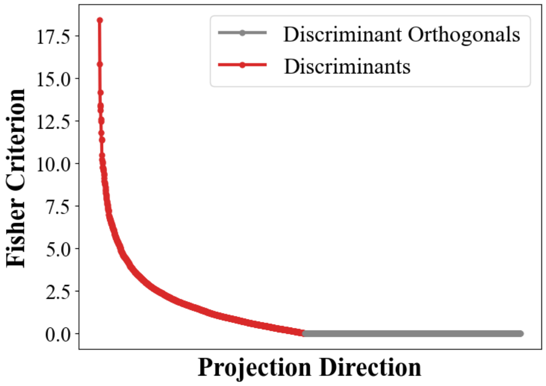

Figure 7: **Empirical Fisher Criterion (FC) values for ResNet-50 classification model.** FC values for discriminants are non-trivial whereas that for discriminant orthogonals approach zero. This verifies our assumption that WLDA disentangle discriminative and residual information from the feature space.

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.