Detecting Out of Distribution Samples using Class Centered Residual Energy in the Discarded PCA Subspace with Weight Alignment

Shreen Gul[1] **Mohamed Elmahallawy**[2] **Ardhendu Tripathy**[3] **Sanjay Madria**[1]

[1]Missouri University of Science and Technology, Rolla, MO 65401, USA, {sgchr,madrias}@mst.edu
[2]Washington State University, Richland, WA 99354, USA, mohamed.elmahallawy@wsu.edu
[3]ardhendutr@gmail.com

Abstract

Principal component analysis (PCA) has recently been adopted for out-of-distribution (OOD) detection, yet most existing methods focus on dominant, high-variance directions of the feature space. This energy-retaining perspective overlooks a crucial fact: discriminative signals separating in-distribution (ID) and OOD samples often lie in low-variance residual components that are typically discarded. Furthermore, prior PCA-based scores rely solely on feature geometry, limiting robustness under distributional shift. In this work, we revisit PCA for OOD detection from an *uncertainty-aware perspective*. Specifically, we propose a residual-aware OOD detection framework that explicitly *models the discarded subspace of a shared within-class* PCA basis. By constructing a residual variation score, we capture deviations that are ignored by the leading principal components. To further enhance reliability, we introduce a *complementary classifier-weight alignment score* that measures the consistency between the selected class-centered residual and the corresponding classifier weight vector. This dual-signal design enables reliable detection even when OOD samples exhibit high energy in the dominant subspaces. Extensive experiments across convolutional and transformer backbones show consistent improvements over strong baselines in AUROC and FPR at 95% TPR. This shows that modeling residual structure and integrating geometric and classifier-aware signals leads to more principled and robust OOD detection. The code to reproduce the results is available at https://github.com/sgchr273/CREWA.git.

1 INTRODUCTION

Deep neural networks can output highly confident predictions even when they are wrong Gul et al. [2026]. This overconfidence is largely a consequence of standard training objectives that *optimize accuracy on the training distribution rather than calibrated uncertainty* Azizmalayeri et al. [2024]. When test inputs deviate from the training data, these models can fail silently, making reliable out-of-distribution (OOD) detection essential in safety-critical applications such as medical imaging, autonomous driving, and crack or fraud detection Yang et al. [2024]. A reported incident in which a driverless Tesla crashed into a $3.5 million jet highlights the potential consequences of deploying high-confidence models *without principled uncertainty handling* Adversa AI [2025].

Similar concerns arise in active learning, where uncertainty estimates guide sample acquisition and can become unreliable under distribution shift, motivating stronger post hoc uncertainty and OOD scoring methods Gul et al. [2024b,a]. Most post hoc OOD detectors compute an abnormality score from a fixed, pretrained network using either internal representations or output statistics. Broadly, these approaches fall into three categories. *(i) Feature-space methods* Lee et al. [2018], Wang et al. [2022], Sun et al. [2022, 2021] leverage the empirical observation that OOD samples tend to deviate from in-distribution (ID) clusters in the penultimate feature space. *(ii) Logit-based methods* Djurisic et al. [2022], Basart et al. [2022], Wei et al. [2022] assign scores based on summary statistics of the predictive distribution. *(iii) Gradient-based techniques* Behpour et al. [2023], Huang et al. [2021] characterize distributional shifts through gradient norms or by quantifying gradient components orthogonal to an ID subspace. While effective, these approaches often rely on a single view of abnormality (feature geometry, output statistics, or sensitivity), which can limit robustness under complex distribution shifts.

In a complementary line of work, Principal Component Analysis (PCA) has been widely adopted for OOD detection

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1790–1803.

by modeling the low-dimensional structure of ID features and scoring test inputs using reconstruction error or residual energy. Guan et al. [2023] revisit PCA-based reconstruction and show that dominant ID variance directions may hinder separability, motivating regularized objectives. Fang et al. [2024] extend this idea via kernel PCA to capture nonlinear structure and compute reconstruction error in the mapped space. NECO Ammar et al. [2024] leverages neural collapse geometry by estimating a Simplex ETF subspace from class means and measuring how much feature energy lies within this subspace.

Despite these advances, many PCA-based detectors Guan et al. [2023], Fang et al. [2024] share two recurring limitations. First, they prioritize principal components that explain the largest ID variance, even though *discriminative OOD evidence often resides in the discarded, low-variance directions as we discover in this work*. As a result, an OOD sample that projects strongly onto the retained ID subspace may be *incorrectly* accepted as ID. Second, most methods rely exclusively on feature-space geometry and neglect complementary information encoded in the classifier parameters, *overlooking an additional source of uncertainty signal.*

In this work, we address the above two limitations by (i) explicitly quantifying energy in the complementary (*discarded*) subspace of a within-class PCA model, and (ii) augmenting this geometric signal with a parameter-space cue derived from the classifier head. Intuitively, ID samples should concentrate most of their mass within the retained class-conditional subspace and exhibit minimal energy in discarded directions. In contrast, challenging OOD samples that appear ID-like in the dominant subspace can still be exposed through elevated residual energy and weak alignment with class anchors encoded by the classifier weights.

Importantly, our method operates in a strictly post hoc regime: it leaves model parameters unchanged, requires no auxiliary OOD data, and relies solely on penultimate-layer representations. Concretely, we compute per-class feature means, center training features within each class, and fit a single PCA basis to the pooled class-centered residuals, corresponding to estimating a tied within-class covariance structure. At inference time, we compute class-centered residuals with respect to each ID class mean, evaluate their complement residual energies using the shared PCA basis, and select the class that yields the minimum residual energy for the final score. We further incorporate a complementary alignment penalty that quantifies the mismatch between the feature vector and the corresponding classifier weight vector. The resulting score captures both deviation from the closest class-conditioned affine structure in feature space and inconsistency with the corresponding classifier head geometry. In summary, we propose:

1. **Residual-aware OOD scoring.** We show that discriminative OOD signals often reside in low-variance directions, and propose a principled framework that scores energy in the discarded subspace of a pooled within-class PCA model, capturing discrepancies missed by dominant principal components.
2. **Feature-parameter coupling for uncertainty estimation.** We propose a novel dual-signal scoring mechanism that augments residual subspace energy with a classifier-weight alignment penalty, integrating feature-space geometry with parameter-space class anchors to improve robustness against ID-like OOD samples.
3. **Strictly post hoc and training-free design.** Our method leaves network parameters unchanged, requires no auxiliary OOD data, and operates solely on penultimate-layer representations, making it simple, efficient, and broadly applicable to pretrained models.
4. **Theoretical and empirical validation.** We provide theoretical insight into why residual directions amplify distributional shifts and demonstrate consistent improvements over strong baselines across convolutional (ResNet) and transformer (ViT) backbones on multiple benchmarks, including ImageNet-1K and CIFAR-10, achieving superior AUROC and FPR at 95% TPR.

2 RELATED WORK

PCA and Residual-based Methods. Our approach builds on PCA and residual-based scoring, but it differs from existing methods Wang et al. [2022], Lee et al. [2018], Ren et al. [2021] in this category. For example, VIM Wang et al. [2022] measures residual magnitude outside a principal subspace but constructs this subspace in a class-agnostic manner and converts the residual into a virtual OOD logit combined with classifier outputs. In contrast, our pooled within-class PCA captures within-class nuisance variation in the discarded directions and explicitly integrates a classifier-weight alignment term. Similarly, Mahalanobis Lee et al. [2018] assumes tied covariance but can overweight directions shared by both ID and OOD features Ren et al. [2021], whereas our formulation emphasizes energy leakage into the complementary subspace, where ID samples typically have minimal mass and OOD samples exhibit substantially higher residual energy. DECA Li et al. [2025] also leverages principal versus complementary decompositions, but it is built on the global feature covariance and scores samples using weighted reconstruction error over the full spectrum, whereas our method uses pooled within class covariance with nearest centroid mean subtraction. Moreover, DECA fuses seven hand engineered WPC and NPC curve statistics, while our detector is a single theoretically grounded scalar whose advantage follows directly from the Subspace SNR Lemma.

Embedding-based and Distance-based Methods. CIDER (Compactness and DispErsion Regularized learning) learns hyperspherical embeddings by increasing inter-class angular

margins while promoting intra-class compactness Ming et al. [2022]. While effective for ID separation, the default OOD scoring relies on KNN cosine distances in the embedding space, requiring storage of a large ID feature bank and nearest-neighbor search at inference, increasing memory usage and latency.

Neural Collapse-based Methods. Recent OOD detection methods have leveraged geometric regularities of deep representations, particularly those associated with neural collapse. For instance, the Neural Collapse Inspired (NCI) detector scores test features based on their proximity to classifier weight vectors combined with feature norms Liu and Qin [2025]. While effective, these approaches rely on global class statistics, which can be fragile when different classes exhibit heterogeneous covariance or multimodal structure.

Wu et al. [2024] introduce a separation objective that encourages OOD features to lie orthogonal to the ID subspace induced by neural collapse. However, such approaches may degrade when neural collapse conditions are not well satisfied, for example, under class imbalance or limited training data. Moreover, they typically assume overparameterized feature spaces, an assumption that may not hold for compact architectures.

3 METHODOLOGY

Complement Residual Energy with Weight Alignment (CREWA) is a fully post-hoc OOD detection method. CREWA requires only ID training features and the pre-trained classifier's final-layer weights; it does not use any OOD data and does not retrain the backbone. Our key insight is that ID features concentrate their variation in a low-dimensional within-class subspace, while the orthogonal complement exhibits minimal ID variance. Consequently, residual energy measured in these low-variance directions provides a strong abnormality signal: ID samples produce near-zero complement energy, whereas OOD samples more frequently accumulate energy in this subspace.

Concretely, CREWA first fits a shared PCA basis using pooled within-class residuals from the ID training data, obtained by subtracting the corresponding class mean from each training feature. This decomposes the feature space into a retained subspace, which captures dominant ID variation, and an orthogonal complementary subspace. At inference time, CREWA does not rely solely on the classifier's top-1 prediction. Instead, it forms a class-centered residual for each ID class by subtracting each class mean from the test representation, computes the complement residual energy for each residual, and selects the class that yields the minimum residual energy. The selected class is then used for the classifier-weight alignment term. Unlike methods that focus on dominant principal components, CREWA scores test inputs using their energy in the *discarded*, low-variance directions, where *ID features exhibit minimal variation* but *OOD samples often accumulate residual signal*.

To further improve robustness—especially for near-OOD inputs that may resemble ID ('ID-like") samples along principal directions—we augment the complement energy with a weight-alignment penalty. This term measures the misalignment between the selected class-centered residual and the corresponding classifier weight vector, capturing inconsistencies not reflected in variance structure alone. Fig. 1 summarizes the CREWA framework, which proceeds as follows:

Step 1: Extract the penultimate feature representation $h(x)$ from the trained classifier and compute the corresponding logits as $g(x) = Wh(x)$.

Step 2: Using the ID training features, construct pooled within-class residuals and fit a shared PCA basis to obtain the retained subspace $V_{\|}$ and its orthogonal complement $V_{\perp}$.

Step 3: For each class c, compute the class-centered residual representation

$$z_c(x) = h(x) - \mu_c.$$

Step 4: Measure the complement-subspace residual energy for each class as

$$s_{\perp,c}(x) = \|V_{\perp}^{\top} z_c(x)\|_2^2.$$

Step 5: Select the class that yields the minimum complement residual energy:

$$c^*(x) = \arg\min_c s_{\perp,c}(x).$$

Step 6: Compute the classifier-weight alignment penalty using the selected class:

$$a(x) = 1 - \cos(z_{c^*}(x), w_{c^*}).$$

Step 7: Define the final CREWA score by combining complement residual energy with the alignment penalty:

$$s_{\mathrm{CREWA}}(x) = s_{\perp,c^*}(x) + \beta a(x).$$

In the following, we detail the problem formulation, PCA construction, residual and alignment components, and the final augmented score. We conclude with a theoretical result showing that, under mild conditions, complement-subspace energy achieves a higher signal-to-noise ratio than energy in the retained subspace.

3.1 PROBLEM STATEMENT

Let $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^N$ denote an ID training set, where each $x_i \in \mathcal{X}$ and $y_i \in \{1, \ldots, C\}$. We assume that $\mathcal{D}_{\mathrm{tr}}$ is

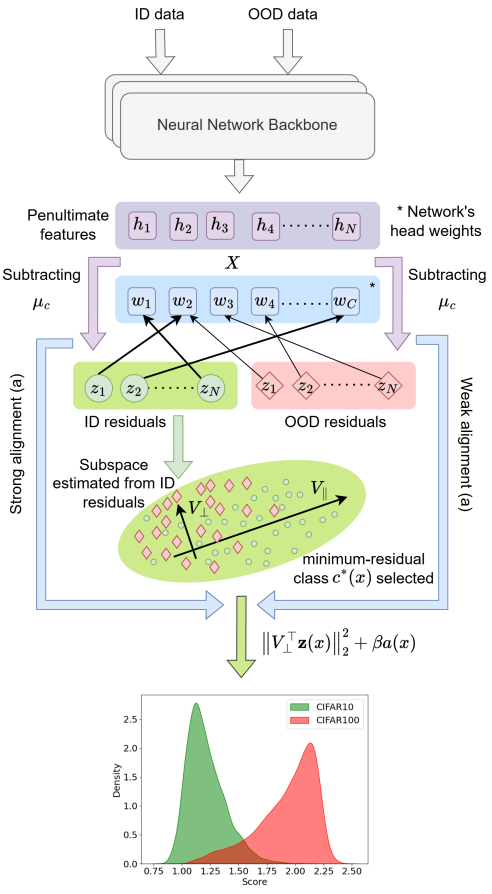


Figure 1: Overview of our framework CREWA. We compute class-centered residuals in feature space, score complement-subspace energy, and augment it with a weight-alignment penalty to separate ID and OOD samples.

drawn i.i.d. from an unknown joint distribution $\mathbb{P}^{\text{in}}_{\mathcal{X}\mathcal{Y}}$ over $\mathcal{X} \times \{1, \dots, C\}$, and write the induced marginal over inputs as $\mathbb{P}^{\text{in}}_{\mathcal{X}}$. Also, we consider a pretrained classifier consisting of a feature extractor $f : \mathcal{X} \to \mathbb{R}^D$ and a linear classification head $W \in \mathbb{R}^{C \times D}$ (bias $\mathbf{b}$ optional). For any input $x \in \mathcal{X}$, we define

$$\mathbf{h}(x) = f(x) \in \mathbb{R}^D, \qquad g(x) = W\mathbf{h}(x) \in \mathbb{R}^C,$$

and for training samples $\mathbf{h}_i = \mathbf{h}(x_i)$.

At deployment, the model may receive an input x that is *not drawn* from $\mathbb{P}^{\text{in}}_{\mathcal{X}}$; such inputs are considered OOD. We define a real-valued scoring function $s : \mathcal{X} \to \mathbb{R}$ and an OOD decision rule $\delta : \mathcal{X} \to \{\text{ID}, \text{OOD}\}$ based on thresholding:

$$\delta(x) = \begin{cases} \text{OOD}, & \text{if } s(x) \geq \tau, \\ \text{ID}, & \text{if } s(x) < \tau, \end{cases}$$

where $\tau \in \mathbb{R}$ denotes the decision threshold.

3.2 CLASS MEANS AND CLASS-CENTERED RESIDUALS

Our OOD detector, CREWA, first computes class-mean feature vectors as:

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i:\, y_i = c} \mathbf{h}_i, \qquad c \in \{1, \dots, C\}, \tag{1}$$

where $N_c = |\{i : y_i = c\}|$. Using these means, we form class-centered residuals as:

$$\mathbf{z}_i = \mathbf{h}_i - \boldsymbol{\mu}_{y_i}, \qquad i = 1, \dots, N. \tag{2}$$

Stacking residuals row-wise yields $Z \in \mathbb{R}^{N \times D}$ whose ith row is $\mathbf{z}_i^\top$. Let $\boldsymbol{\mu}_G = \frac{1}{N} \sum_{i=1}^N \mathbf{h}_i$ denotes the global mean. The within-class scatter matrix can then be expressed as:

$$S_W = \sum_{c=1}^{C} \sum_{i:\, y_i = c} (\mathbf{h}_i - \boldsymbol{\mu}_c)(\mathbf{h}_i - \boldsymbol{\mu}_c)^\top.$$

The between-class scatter matrix can be expressed as:

$$S_B = \sum_{c=1}^{C} N_c \, (\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^\top.$$

Fitting PCA on class-centered residuals isolates *within-class* variation, avoiding domination by the between-class scatter S_B that can arise when PCA is applied to globally centered pooled features $\mathbf{h}_i - \boldsymbol{\mu}_G$. This formulation therefore emphasizes nuisance directions that are orthogonal to class-mean structure.

3.3 POOLED WITHIN-CLASS PCA AND RETAINED/COMPLEMENT SUBSPACES

We perform singular value decomposition (SVD) on Z:

$$Z = U \Sigma V^\top,$$

where $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $r = \text{rank}(Z)$ denotes the rank of the matrix. The matrix $V = [\mathbf{v}_1, \dots, \mathbf{v}_r] \in \mathbb{R}^{D \times r}$ contains orthonormal principal directions.

A common approach chooses the smallest k that satisfies an energy criterion. Defining

$$E(k) = \frac{\sum_{j=1}^k \sigma_j^2}{\sum_{j=1}^r \sigma_j^2},$$

one may pick k as the smallest integer such that $E(k) \geq \rho$ for a target $\rho \in (0, 1)$. However, when, in practice, the representation spectrum is relatively flat, this criterion can retain nearly all directions, resulting in a weakly expressive complement subspace. To address this limitation, we instead

determine $k^\star$ (the number of retained principal directions) based on the stable rank of Z:

$$\mathrm{srank}(Z) = \frac{\|\Sigma\|_F^2}{\|\Sigma\|_2^2} = \frac{\sum_{j=1}^{r} \sigma_j^2}{\max_j \sigma_j^2}. \tag{3}$$

We then set

$$k^\star = \max(1,\ \min(\lfloor \alpha \cdot \mathrm{srank}(Z) \rfloor\ ,\ r - d_{\min})), \tag{4}$$

where $\alpha = 0.5$ and $d_{\min} = 64$. These values are chosen empirically and discussed further in Section 4.1. When $r \geq d_{\min}+1$, this ensures a complement dimension $p = r-k^\star \geq d_{\min}$, corresponding to the perpendicular subspace used for residual scoring.

The resulting decomposition yields a retained basis as:

$$V_{\|} = [\mathbf{v}_1, \ldots, \mathbf{v}_k^\star] \in \mathbb{R}^{D \times k},$$

and a complementary basis as:

$$V_{\perp} = [\mathbf{v}_{k^\star+1}, \ldots, \mathbf{v}_r] \in \mathbb{R}^{D \times p},$$

where the complement dimension is $p = r - k^\star$.

3.4 COMPLEMENT PROJECTION AND RESIDUAL ENERGY SCORE

For a test input x, we compute the feature representation $h(x) = f(x)$. Instead of conditioning only on the classifier's top-1 prediction, we evaluate the class-centered residual with respect to each ID class mean:

$$z_c(x) = h(x) - \mu_c, \qquad c \in \{1, \ldots, C\}.$$

Using the shared complement basis $V_\perp$, we compute the class-conditional complement residual energy

$$s_{\perp,c}(x) = \left\| V_\perp^\top z_c(x) \right\|_2^2.$$

CREWA then selects the class that gives the minimum residual energy:

$$c^*(x) = \arg \min_{c \in \{1,\ldots,C\}} s_{\perp,c}(x).$$

The residual energy used by the detector is therefore

$$s_\perp(x) = s_{\perp,c^*}(x).$$

Larger $s_\perp(x)$ indicates that even the closest class-conditioned residual has elevated energy in directions that exhibit low variance under the pooled within-class residual model of ID training features.

3.5 CLASSIFIER-WEIGHT ALIGNMENT PENALTY

Residual energy alone may be insufficient for near-OOD inputs that lie predominantly within the retained residual subspace. To improve sensitivity in such cases, we incorporate a complementary signal derived from classifier geometry.

Let $w_c^\top$ denote the c-th row of the linear classification head W, where $w_c \in \mathbb{R}^D$. After selecting the minimum-residual class $c^*(x)$, we use the corresponding class-centered residual

$$u(x) = z_{c^*}(x)$$

for cosine alignment. We define the normalized directions as

$$\tilde{u}(x) = \frac{u(x)}{\|u(x)\|_2 + \eta}, \qquad \tilde{w}_{c^*} = \frac{w_{c^*}}{\|w_{c^*}\|_2 + \eta},$$

where $\eta > 0$ ensures numerical stability. The cosine alignment and alignment penalty are

$$\cos(x) = \tilde{u}(x)^\top \tilde{w}_{c^*}, \qquad a(x) = 1 - \cos(x).$$

Larger $a(x)$ indicates weaker alignment between the selected class-centered residual and the corresponding classifier weight direction.

3.6 AUGMENTED OOD SCORE AND CALIBRATION

The final CREWA score combines the minimum class-conditioned complement residual energy and the alignment penalty:

$$s_{\mathrm{CREWA}}(x) = s_{\perp,c^*}(x) + \beta a(x),$$

where $\beta > 0$ balances the two terms and $c^*(x)$ is the minimum-residual class selected in Section. (3.4). The decision rule

$$\delta(x) = \mathbb{I}\{s_{\mathrm{CREWA}}(x) \geq \tau\}$$

is applied with a threshold τ chosen using ID validation data or to meet a desired TPR. Theoretical analysis on why the complement subspace can yield a stronger discriminative signal is in Appendix B.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate CREWA on two widely used ID benchmarks: CIFAR-10 Krizhevsky and Hinton [2009] and ImageNet-1K Deng et al. [2009]. For CIFAR-10, we follow the OpenOOD benchmark Zhang et al. [2023] and evaluate against six OOD datasets. Specifically, CIFAR-100 Krizhevsky and Hinton [2009] and TinyImageNet-200 Le and Yang [2015] are treated as *near-OOD* datasets,

Table 1: Subspace configuration for each backbone with $\alpha = 0.5$ fixed across architectures.

Model	Rank (r)	Stable Rank	Kept ($k^\star$)	Discarded (p)
ResNet50	2048	546	273	1775
Swin-T	768	473	236	532
ResNet18	512	24	12	500
ViT-B/16	768	26	13	755

Table 2: Representation geometry diagnostics across CNN and transformer backbones. Transformer backbones show stronger class mean–classifier weight alignment and lower within–between scatter ratios, indicating a more class-aligned representation geometry.

Backbone	$\overline{\cos}(\mu_c, w_c)$	S_W	S_B	S_W/S_B
ResNet18	0.303	0.237	0.136	1.744
ResNet50	0.569	0.510	0.411	1.243
ViT	0.630	0.447	0.448	0.997
Swin-T	0.766	0.482	0.480	1.003

while Textures Cimpoi et al. [2014], Places365 Zhou et al. [2018], SVHN Netzer et al. [2011], and MNIST LeCun et al. [1998] are considered *far-OOD* datasets.

For ImageNet-1K, we follow the evaluation protocol of NECO Ammar et al. [2024] and use five OOD datasets: Places365, SUN Xiao et al. [2010], Textures, iNaturalist Van Horn et al. [2017], and OpenImage-O Wang et al. [2022].

Metrics and Backbones. We use two widely adopted OOD detection metrics to evaluate CREWA: the false positive rate at 95% true positive rate (FPR95) and the area under the receiver operating characteristic curve (AUROC).

Experiments are conducted with ResNet, Vision Transformer (ViT), and Swin Transformer backbones. For ResNet and Swin-T, we use off-the-shelf pretrained weights provided by `torchvision`. On CIFAR-10, we fine-tune the ResNet and ViT models for ID classification. On ImageNet-1K, we directly use pretrained checkpoints of ResNet-50 and Swin-T.

Baselines. We compare CREWA against state-of-the-art post-hoc OOD detection methods. For a fair comparison, we restrict our evaluation to post-hoc approaches that require no retraining of the backbone beyond standard ID classification training. Specifically, we evaluate against *ten* baselines: Mahalanobis Lee et al. [2018], Energy Liu et al. [2020], Maximum Softmax Probability (MSP) Hendrycks and Gimpel [2017], Logit Gate Guan et al. [2023], KPCA Fang et al. [2024], GradOrth Behpour et al. [2023], NECO Ammar et al. [2024], NCI Liu and Qin [2025], VIM Wang et al. [2022], and DECA Li et al. [2025]. All methods are implemented using the official repositories. We do not include LLM-based or multimodal LLM-based OOD detectors as direct baselines because they rely on additional language-model reasoning, prompting, or multimodal inputs, whereas CREWA is designed for post-hoc scoring of fixed vision backbones using only penultimate features and classifier weights.

Hyperparameter Selection: Our method CREWA involves two tunable quantities: the dimensionality k of the pooled within-class residual PCA basis and the mixing weight β in the augmented score s_{CREWA} defined in the Section. 3.6.

A. Stable Rank Retention Factor (α). The retained dimensionality k is determined implicitly by the stable rank retention factor α, which controls how much of the within-class residual spectrum is preserved and thus balances principal structure retention with complement expressiveness. Energy-based thresholds are unreliable when the spectrum is flat, as they may retain most directions and weaken the complement subspace. We therefore select k using the stable rank criterion in Section 3.3, with fixed $\alpha = 0.5$ and $d_{\min} = 64$ to ensure a sufficiently expressive complement. Further details are provided in Section 4.2.2, and the resulting stable ranks, retained dimensions k, and complement dimensions p are reported in Table 1.

B. Calibration Coefficient (β). Since the complement residual energy and the alignment penalty may operate on different numerical scales across architectures and datasets, we calibrate β using ID data only by matching their typical magnitudes. Specifically, we define:

$$\beta^\star = \gamma \frac{\mathrm{median}(s_\perp(x))}{\mathrm{median}(a(x))}, \qquad x \sim \mathbb{P}^{\mathrm{in}}_{\mathcal{X}}, \text{ with } \gamma = 1 \text{ by default.}$$

The coefficient β is computed once on ID data and then fixed for all evaluations. Median-based calibration ensures robustness to skewness and heavy-tailed score distributions due to its insensitivity to extreme values.

4.2 RESULT ANALYSIS

4.2.1 Comparison with Baselines

In this section, we compare CREWA with state-of-the-art baselines across multiple backbones and OOD datasets to demonstrate its effectiveness under fair post-hoc settings.

Table 3 reports OOD detection results using CIFAR-10 as the in-distribution (ID) dataset. We categorize CIFAR-100 and TinyImageNet-200 as near-OOD datasets, and Textures, Places, SVHN, and MNIST as far-OOD datasets. The best-performing method is highlighted in bold, while the second- and third-best results are underlined.

For the ResNet-18 backbone (top half), NCI performs well on near OOD but degrades on far OOD. In contrast, GradOrth and Logit Gate perform strongly on far OOD yet are less competitive on near OOD. Overall, CREWA is the

Table 3: OOD detection results for two backbones on CIFAR10 as ID, evaluated on six OOD datasets. Metrics are reported as percentages. Best is bold; second and third best are underlined.

Model	Method	CIFAR100		TinyImageNet 200		Textures		Places		SVHN		MNIST		Average	
		AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓
ResNet18	Mahalanobis	86.66	50.31	87.25	45.33	99.27	3.69	94.90	23.54	**99.94**	**0.17**	99.58	1.65	94.60	20.78
	Energy	78.71	67.54	88.73	42.22	89.03	45.53	90.74	31.49	79.03	67.54	87.98	34.13	85.70	48.08
	MSP	87.71	45.10	91.73	36.58	94.48	23.28	94.59	21.72	92.70	26.75	96.46	11.34	92.94	27.46
	Logit Gate	86.63	52.73	89.59	40.35	99.35	3.40	95.89	20.03	99.90	0.36	99.38	2.96	95.12	19.97
	KPCA RFF	85.69	53.86	84.93	53.09	99.10	4.51	94.04	28.11	99.90	0.38	**99.78**	**0.82**	93.91	23.46
	GradOrth	86.93	50.83	89.31	41.36	99.36	3.18	96.06	18.91	99.91	0.30	99.44	2.77	95.17	19.56
	NECO	87.89	47.08	92.43	31.50	98.95	5.84	95.22	20.54	99.67	1.43	95.87	16.67	95.01	20.51
	NCI	90.94	**36.14**	93.13	28.99	98.63	6.89	96.86	12.52	99.44	2.31	97.86	5.98	96.14	15.47
	VIM	84.83	53.55	85.76	47.93	98.89	6.87	94.23	24.80	99.76	1.03	99.67	1.30	93.86	22.58
	DECA	76.84	73.61	74.57	72.72	98.18	11.09	88.78	46.15	99.68	1.46	99.55	1.82	89.60	34.48
	CREWA (Ours)	**91.48**	36.78	**93.58**	**26.46**	**99.48**	**2.71**	**97.27**	**12.48**	99.84	0.62	99.41	2.30	**96.84**	**13.56**
ViT B/16	Mahalanobis	97.01	15.31	97.33	**16.14**	**99.98**	**0.03**	99.14	4.03	99.62	1.78	99.86	0.71	98.82	6.33
	Energy	93.43	31.58	92.75	35.06	98.02	8.42	94.71	22.94	99.06	4.39	97.93	7.58	95.98	18.33
	MSP	97.83	10.53	96.95	19.33	99.85	0.38	98.86	5.65	99.57	1.86	99.24	2.56	98.72	6.72
	Logit Gate	96.99	15.34	96.95	17.88	99.97	0.10	**99.15**	**3.92**	99.22	3.67	99.75	0.87	98.67	6.96
	KPCA RFF	97.28	13.79	96.98	18.10	99.94	0.15	99.06	4.40	99.54	1.99	99.74	0.93	98.76	6.56
	GradOrth	96.98	15.65	97.01	18.05	99.97	**0.03**	99.13	3.98	99.46	2.52	99.87	0.49	98.74	6.79
	NECO	**98.16**	**9.14**	**97.39**	16.22	99.93	0.16	98.94	5.09	99.72	1.16	99.51	1.56	**98.94**	**5.56**
	NCI	97.32	13.95	96.57	19.86	99.88	0.34	98.97	4.98	**99.91**	**0.34**	99.44	1.41	98.68	6.81
	VIM	96.09	20.04	95.70	25.90	99.88	0.17	98.48	9.05	99.16	4.04	99.83	**0.30**	98.19	9.92
	DECA	93.96	31.90	94.92	33.66	99.90	0.15	98.62	7.60	97.85	10.28	**99.90**	0.32	97.53	13.99
	CREWA (Ours)	97.95	11.06	97.07	17.33	99.93	0.17	99.00	4.88	99.56	1.87	99.79	0.51	98.88	5.97

Table 4: OOD detection on ImageNet-1K (ID). Metrics are reported as percentages. Average is the unweighted mean over the five OOD datasets (Textures, iNaturalist, SUN, Places, OpenImage-O). Best is in bold; second and third best are underlined.

Model	Method	Textures		iNaturalist		SUN		Places		OpenImage-O		Average	
		AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓	AUROC↑	FPR↓
ResNet50	Mahalanobis	93.47	31.01	92.16	26.44	75.45	58.82	73.10	65.59	**80.41**	**59.57**	82.92	48.29
	Energy	53.92	91.03	74.66	73.03	72.31	78.97	73.65	76.49	51.53	89.56	65.21	81.82
	MSP	83.95	62.45	88.20	41.40	81.44	57.54	78.34	66.91	66.59	74.75	79.70	60.61
	Logit Gate	85.99	45.26	63.52	71.81	50.84	85.79	54.58	85.17	73.01	65.55	65.59	70.72
	KPCA RFF	91.15	47.77	79.95	50.84	71.39	65.49	69.89	68.28	80.16	61.70	78.51	58.82
	GradOrth	87.63	42.57	80.63	52.37	60.06	76.26	56.98	78.96	75.11	68.48	72.08	63.73
	NECO	62.91	75.48	68.07	64.88	67.64	73.22	70.73	68.79	56.00	83.74	65.07	73.22
	NCI	85.54	60.90	89.42	37.71	**84.12**	**51.81**	**80.74**	63.24	67.20	74.37	81.40	57.61
	VIM	89.96	44.99	78.55	59.45	64.22	75.98	59.85	78.21	77.34	67.33	73.98	65.19
	DECA	91.53	39.09	73.71	64.11	57.79	84.52	51.57	87.97	69.41	80.12	68.80	71.16
	CREWA (Ours)	**93.75**	**29.86**	**95.14**	**21.55**	79.48	59.06	79.24	**60.10**	75.79	65.45	**84.68**	**47.20**
SwinT	Mahalanobis	91.08	39.98	97.28	10.88	**84.53**	51.39	81.37	62.44	78.07	**57.42**	**86.47**	**44.42**
	Energy	64.28	85.94	54.28	90.84	51.16	92.38	52.39	91.75	55.17	91.21	55.46	90.42
	MSP	72.88	71.44	90.41	33.37	83.09	56.91	81.51	60.67	54.80	93.83	76.54	63.24
	Logit Gate	90.10	45.86	80.41	60.28	60.37	89.85	52.29	94.70	75.35	69.87	71.70	72.11
	KPCA RFF	86.02	45.79	84.40	40.05	75.53	62.20	72.99	68.71	76.96	59.58	79.18	55.27
	GradOrth	87.71	43.42	87.10	42.27	67.88	75.77	63.27	81.48	73.59	67.55	75.91	62.10
	NECO	81.41	60.19	92.35	28.54	78.94	61.82	75.31	70.14	61.24	92.17	77.85	62.57
	NCI	80.75	53.68	91.06	29.55	83.32	**49.34**	**82.00**	**55.22**	71.14	67.48	81.65	51.05
	VIM	**94.35**	**28.28**	90.88	37.96	75.13	74.32	69.08	85.18	**79.07**	62.48	81.70	57.64
	DECA	89.75	44.68	77.32	71.98	75.05	78.98	67.53	89.97	74.98	73.02	76.93	71.73
	CREWA (Ours)	90.46	39.94	**97.64**	**9.68**	84.29	53.23	81.43	62.57	77.06	59.24	86.18	44.93

most consistent across both regimes, achieving an average AUROC of 96.84% and an average FPR of 13.56%.

The bottom half shows the ViT-B/16 backbone. Here, CREWA ranks second overall, trailing NECO by a small margin (0.06% AUROC and only a modest gap in FPR). While VIM also exploits a complementary subspace signal, CREWA improves FPR by 3.95% on average, suggesting that the weight alignment term provides additional, non redundant information beyond residual energy alone.

Table 4 summarizes results on ID=ImageNet-1K across five OOD datasets. With ResNet-50, CREWA achieves the best overall performance, with average AUROC of 84.68% and average FPR of 47.20%, narrowly outperforming Mahalanobis (82.92% AUROC, 48.29% FPR). With Swin-T, CREWA achieves the second best overall FPR, falling just behind Mahalanobis by 0.51% FPR and 0.29% AUROC, while outperforming VIM, NCI, NECO, and DECA by a clear margin in both AUROC and FPR.

The reduced margin observed with transformer backbones may be attributed to their more class-aligned representation geometry. As shown in Table 2, ViT and Swin-T exhibit stronger average class mean–classifier weight alignment, $(\frac{1}{C}\sum_{c=1}^{C}\cos(\mu_c, w_c))$, with values of 0.630 and 0.766, respectively, compared with 0.303 for ResNet18 and 0.569 for ResNet50. They also yield lower normalized within–between scatter ratios, approximately 1.0 for both ViT and Swin-T, compared with 1.744 for ResNet18 and 1.243 for ResNet50. These trends indicate that transformer features are already more tightly organized around classifier-aligned class structure, which can reduce the additional separation gained from CREWA's residual subspace decomposition.

4.2.2 Ablation Studies

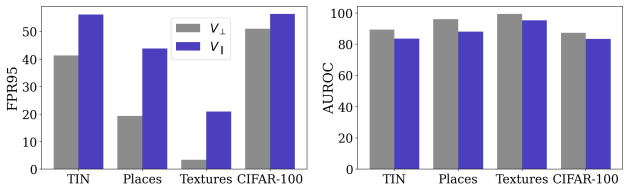
We perform ablation studies of CREWA on the CIFAR-10 dataset using ResNet-18 as the backbone. We evaluate FPR and AUROC across four OOD datasets: TinyImageNet-200 (TIN), CIFAR-100, Textures, and Places.

Ablation on Retained versus Complement PCA Directions. This experiment investigates whether the OOD score should be computed using the retained directions $V_{\|}$ or the complementary directions $V_{\perp}$, while keeping the same gating and centering procedure from Section 3.4. Specifically, we compare the main complement score $s_{\perp}(x)$ from Section 3.4 to the retained-subspace energy as:

$$s_{\|}(x) = \left\| V_{\|}^{\top} \mathbf{z}(x) \right\|_2^2,$$

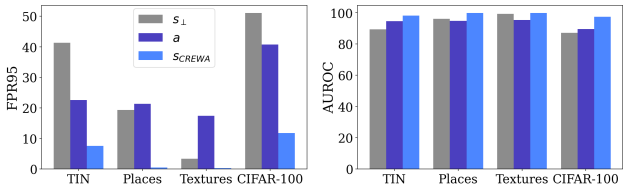
where $\mathbf{z}(x) = \mathbf{h}(x) - \boldsymbol{\mu}_{c^*(x)}$ and $c^*(x)$ is the minimum-residual class.

Fig. 2 shows that using the complementary directions consistently yields higher AUROC and lower FPR95 compared to scoring in the retained directions across all evaluated



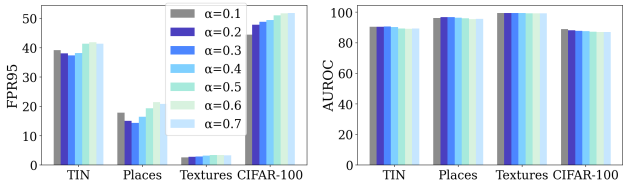
(a) FPR95 comparison. (b) AUROC comparison.

Figure 2: Sensitivity analysis of the retained ($V_{\|}$) and complementary ($V_{\perp}$) PCA subspaces across four OOD datasets.



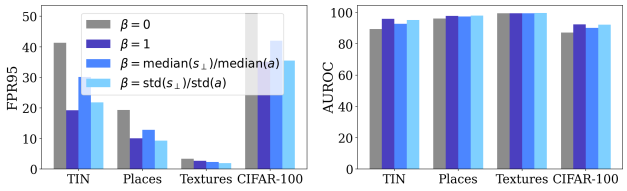
(a) FPR95 comparison. (b) AUROC comparison.

Figure 3: Sensitivity analysis of the components of the s_{CREWA} on performance across four OOD datasets.



(a) FPR95 comparison. (b) AUROC comparison.

Figure 4: Sensitivity analysis of the retention fraction α on performance across four OOD datasets.



(a) FPR95 comparison. (b) AUROC comparison.

Figure 5: Sensitivity analysis of the calibration coefficient β on performance across four OOD datasets.

OOD datasets. This behavior is expected because $V_{\|}$ captures the dominant within-class variation of ID features, and OOD examples can also project strongly onto these high-variance directions, reducing separability. In contrast, $V_{\perp}$ corresponds to low-variance directions under the pooled within-class residual model, so ID residual energy concentrates near zero while OOD inputs often exhibit elevated energy in this subspace. Consequently, $s_{\perp}(x)$ provides a more discriminative abnormality score.

Ablation on Final Score Components. We analyze the

contribution of each component in the augmented CREWA score, $s_{\mathrm{CREWA}}(x)$, by decomposing it into the complement-subspace residual energy $s_{\perp}(x)$ and the weight-alignment penalty $a(x)$. Fig. 3 presents the performance of each component individually, as well as their combination. Across four OOD benchmarks, the combined score $s_{\mathrm{CREWA}}(x)$ consistently achieves the lowest FPR95 of as low as 2.4 on Textures and the highest AUROC of as high as 99.48 on Textures, demonstrating that the two signals provide complementary information. While $a(x)$ alone can be competitive on certain datasets, its performance is not consistently strong—for example, it produces higher FPR95 on Places and Textures compared to $s_{\perp}(x)$. By contrast, incorporating $a(x)$ with $s_{\perp}(x)$ significantly reduces false positives and improves AUROC, yielding the best overall performance.

Ablation on the Subspace Retention Fraction. We evaluate the effect of the retention fraction α, which sets the number of retained within-class principal directions as $k = \lfloor \alpha \cdot \mathrm{srank}(Z) \rfloor$. As shown in Fig. 4, we sweep $\alpha \in 0.1, 0.2, \ldots, 0.7$, where smaller α values retain fewer dominant directions and leave a larger complement subspace for residual scoring.

Overall, AUROC remains relatively stable, whereas FPR95 is more sensitive to α. Larger values, particularly $\alpha \geq 0.5$, tend to increase FPR95 on datasets such as Places and CIFAR-100, suggesting that an overly small complement subspace can weaken the residual signal. The strongest performance is generally observed for moderate values, approximately $\alpha = 0.2$–0.4, while very small values do not provide consistent gains. In our experiments, we use a fixed OOD-free default value of $\alpha = 0.5$ selected from held-out ID validation data to avoid dataset-specific tuning. However, the observed sensitivity indicates that the retained/complement split can be model- and dataset-dependent. Developing an adaptive ID-only criterion, such as selecting α by controlling the upper tail of the ID score distribution, is a natural direction for future work.

Ablation on the Calibration Coefficient. In this experiment, we analyze the effect of the weighting parameter β in the augmented score $s_{\mathrm{CREWA}}(x) = s_{\perp}(x) + \beta\, a(x)$. Fig. 5 evaluates the sensitivity of performance to different choices of β. Setting $\beta = 0$ (i.e., using only the complement-subspace residual energy) consistently yields the weakest performance, producing the highest FPR95 and lowest AUROC across all OOD datasets. This confirms that the alignment penalty contributes complementary information beyond $s_{\perp}(x)$. A fixed $\beta = 1$ performs strongly overall, achieving the best FPR95 on Tiny-ImageNet and CIFAR-100; however, it is not uniformly optimal across all benchmarks. In particular, ID-based scale calibration—such as $\beta = \mathrm{median}(s_{\perp})/\mathrm{median}(a)$ or $\beta = \mathrm{std}(s_{\perp})/\mathrm{std}(a)$—improves robustness on Places and Textures, with the standard-deviation ratio achieving the lowest FPR95 on these datasets.

Overall, simple ID-only calibration strategies reduce sensitivity to the relative scale of $s_{\perp}(x)$ and $a(x)$, providing competitive and stable performance without any tuning on OOD data.

4.2.3 Complexity Analysis.

CREWA operates on the feature representation after the backbone forward pass. Since the revised inference procedure evaluates complement residual energy for each class, the dominant scoring operation is the set of projections $V_{\perp}^{\top} z_c(x)$ for $c = 1, \ldots, C$, where $V_{\perp} \in \mathbb{R}^{D \times p}$. Thus, the per-sample class-selection cost is $O(CDp)$. After the minimum-residual class is selected, the classifier-weight alignment term requires only one cosine computation and adds $O(D)$ cost. For a batch of size B, the scoring complexity is $O(BCDp)$.

5 CONCLUSION AND FUTURE WORK

In this work, we introduced CREWA, a post-hoc class-aware residual energy framework for out-of-distribution detection. CREWA models within-class variation by fitting a shared PCA basis on pooled class-centered ID residuals and scores test inputs using their energy in the discarded complementary subspace, where ID samples exhibit limited variation while OOD samples often accumulate residual signal. To reduce dependence on the classifier's top-1 prediction, CREWA evaluates class-centered residual energy with respect to each class mean and selects the minimum-residual class for scoring. The residual energy is further combined with a classifier-weight alignment term, allowing the method to integrate complementary information from both feature geometry and the classifier head. Across CIFAR-10 and ImageNet-1K benchmarks, CREWA achieves competitive performance against strong post-hoc OOD detection baselines on both convolutional and transformer backbones, while requiring no OOD data, no retraining, and only penultimate features and the classifier weights. These results show that modeling residual structure in low-variance directions, together with lightweight classifier-aware alignment, provides an effective and practical direction for post-hoc OOD detection. Future work will explore adaptive ID-only criteria for selecting the retained/complement subspace split, rather than relying on a fixed retention factor across datasets and architectures.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation (NSF), USA, under Grant No. CNS - 2219615.

References

Adversa AI. Towards trusted AI week 17: Tesla crashed into a $3.5 million cirrus vision jet and others. Online, 2025. URL https://shorturl.at/NYRqr. Accessed: 2025-03-11.

M. B. Ammar, N. Belkhir, S. Popescu, A. Manzanera, and G. Franchi. NECO: Neural collapse based out-of-distribution detection. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Mohammad Azizmalayeri, Ameen Abu-Hanna, and Giovanni Cinà. Mitigating overconfidence in out-of-distribution detection by capturing extreme activations. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=nwf2mKQVhP.

Steven Basart, Mantas Mazeika, Mohammad Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

S. Behpour, T. L. Doan, X. Li, W. He, L. Gou, and L. Ren. GradOrth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In *Advances in Neural Information Processing Systems*, volume 36, pages 38206–38230, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sami Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

K. Fang, Q. Tao, K. Lv, M. He, X. Huang, and J. Yang. Kernel PCA for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 37, pages 134317–134344, 2024.

Xinyu Guan, Zhaowei Liu, W.-S. Zheng, Yuqian Zhou, and Ruixuan Wang. Revisit PCA-based technique for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19431–19439, 2023.

Shreen Gul, Mohamed Elmahallawy, Sanjay Madria, and Ardhendu Tripathy. Fishermask: Enhancing neural network labeling efficiency in image classification using fisher information. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1302–1308. IEEE, 2024a.

Shreen Gul, Mohamed Elmahallawy, Sanjay Madria, and Ardhendu Tripathy. Lplgrad: Optimizing active learning through gradient norm sample selection and auxiliary model training. In *2024 IEEE International Conference on Big Data (BigData)*, pages 900–909. IEEE, 2024b.

Shreen Gul, Mohamed Elmahallawy, Ardhendu Tripathy, and Sanjay Madria. Prototype fusion: A training-free multi-layer approach to ood detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 597–611. Springer, 2026.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

R. Huang, A. Geng, and Y. Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, volume 34, pages 677–689, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. CS231N, 2015.

Yann LeCun, Corinna Cortes, and C. J. C. Burges. The MNIST database of handwritten digit images for machine learning research. Online, 1998.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ting Li, Mengling Fan, Yan Fu, Junlin Zhou, Yuke Fang, and Duanbing Chen. Deca: Dual error curves analysis for out-of-distribution detection via low-rank decomposition. *Knowledge-Based Systems*, page 114795, 2025.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475, 2020.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022.

Yuval Netzer, Tao Wang, Adam Coates, B. Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, pages 144–157, 2021.

Yiyou Sun, Yifei Ming, Xuan Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research (PMLR)*, pages 20827–20840, 2022.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist challenge 2017 dataset. *arXiv preprint arXiv:1707.06642*, 1(2):4, 2017.

H. Wang, Z. Li, L. Feng, and W. Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4921–4930, 2022.

H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research (PMLR), pages 23631–23644, 2022.

Yingwen Wu, Ruiji Yu, Xinwen Cheng, Zhengbao He, and Xiaolin Huang. Pursuing feature separation based on neural collapse for out-of-distribution detection. *arXiv preprint arXiv:2405.17816*, 2024.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, Y. Li, Z. Liu, et al. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. arXiv, 2023.

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, 2018.

Appendix A
(Supplementary Material)

Shreen Gul[1] **Mohamed Elmahallawy**[2] **Ardhendu Tripathy**[3] **Sanjay Madria**[1]

[1]Missouri University of Science and Technology, Rolla, MO 65401, USA, {sgchr,madrias}@mst.edu
[2]Washington State University, Richland, WA 99354, USA, mohamed.elmahallawy@wsu.edu
[3]ardhendutr@gmail.com

A ASSUMPTIONS OF COMPLEMENT SCORING.

Let $\xi = V^\top z(x)$ denote the class centered residual in the PCA basis of the pooled within class residuals. The analysis assumes that ID residuals have zero mean and diagonal covariance in this basis, $\mathbb{E}[\xi \mid \mathrm{ID}] = 0$ and $\mathrm{Cov}(\xi \mid \mathrm{ID}) = \Lambda$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$. A Gaussian assumption is only used to obtain the closed form ID variance of the subspace energies; the mean shift expressions require only first and second moments. The OOD covariance is not assumed to be tied to the ID covariance.

Under this model, complement scoring is favored when the OOD perturbation in the complement coordinates is large relative to the ID variability in those coordinates. Specifically, with

$$\Delta_\parallel = \|\delta_\parallel\|_2^2 + \mathrm{tr}(\Lambda_{\mathrm{ood},\parallel} - \Lambda_\parallel), \qquad \Delta_\perp = \|\delta_\perp\|_2^2 + \mathrm{tr}(\Lambda_{\mathrm{ood},\perp} - \Lambda_\perp),$$

the complement score has larger SNR than the retained score when

$$\frac{\Delta_\perp}{\sqrt{\sum_{j=k+1}^{r} \lambda_j^2}} > \frac{\Delta_\parallel}{\sqrt{\sum_{j=1}^{k} \lambda_j^2}}.$$

This condition also identifies when the method may fail: if OOD samples mainly perturb high variance retained directions, or if the ID spectrum is nearly flat, then the retained subspace may provide comparable or stronger detection power. Thus, CREWA is most suitable when OOD deviations appear in directions that are weakly expressed under the ID within class residual distribution.

B THEORETICAL ANALYSIS OF COMPLEMENT SIGNAL STRENGTH

To motivate complement scoring, we quantify detection power via a standardized signal-to-noise ratio (SNR). For any real-valued statistic $s(x)$ define

$$\mathrm{snr}(s) \;=\; \frac{\mathbb{E}[s(x) \mid \mathrm{OOD}] - \mathbb{E}[s(x) \mid \mathrm{ID}]}{\sqrt{\mathrm{Var}(s(x) \mid \mathrm{ID}) + \varepsilon}}, \tag{5}$$

where $\varepsilon > 0$ ensures numerical stability. Below we formalize conditions under which the complement energy $s_\perp(x)$ attains a larger SNR than the retained energy $s_\parallel(x) = \|V_\parallel^\top \mathbf{z}(x)\|_2^2$.

Lemma B.1 (Subspace energy shifts and SNR). *Let $Z \in \mathbb{R}^{N \times D}$ be the pooled within-class residual matrix (Section 3.2) and $\widehat{\Sigma}_W = \frac{1}{N} Z^\top Z$. Write the eigendecomposition*

$$\begin{aligned} \widehat{\Sigma}_W &= V \Lambda V^\top, \\ \Lambda &= \mathrm{diag}(\lambda_1, \ldots, \lambda_r), \qquad \lambda_1 \geq \cdots \geq \lambda_r > 0, \end{aligned} \tag{6}$$

with $V = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$. *Fix* $k \in \{1, \ldots, r-1\}$ *and partition* $V = [V_\parallel \; V_\perp]$ *with* $V_\parallel = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ *and* $V_\perp = [\mathbf{v}_{k+1}, \ldots, \mathbf{v}_r]$.

For a test input x, *let* $z(x) = h(x) - \mu_{c^*(x)}$, *where* $c^*(x)$ *is the minimum-residual class selected in Section 3.4, and define*

$$s_\parallel(x) = \|V_\parallel^\top \mathbf{z}(x)\|_2^2, \qquad s_\perp(x) = \|V_\perp^\top \mathbf{z}(x)\|_2^2.$$

Let $\boldsymbol{\xi}(x) = V^\top \mathbf{z}(x) = [\boldsymbol{\xi}_\parallel(x)^\top, \boldsymbol{\xi}_\perp(x)^\top]^\top$.

Assume the ID moments satisfy $\mathbb{E}[\boldsymbol{\xi} \mid \mathrm{ID}] = \mathbf{0}$ *and* $\mathrm{Cov}(\boldsymbol{\xi} \mid \mathrm{ID}) = \Lambda$. *Under OOD assume* $\mathbb{E}[\boldsymbol{\xi} \mid \mathrm{OOD}] = \boldsymbol{\delta} = [\boldsymbol{\delta}_\parallel^\top, \boldsymbol{\delta}_\perp^\top]^\top$ *and* $\mathrm{Cov}(\boldsymbol{\xi} \mid \mathrm{OOD}) = \Lambda_{\mathrm{ood}} \succeq 0$.

Then the mean shifts of the subspace energies are

$$\begin{aligned} \mathbb{E}[s_\parallel \mid \mathrm{OOD}] - \mathbb{E}[s_\parallel \mid \mathrm{ID}] &= \|\boldsymbol{\delta}_\parallel\|_2^2 + \mathrm{tr}\big(\Lambda_{\mathrm{ood},\parallel} - \Lambda_\parallel\big), \\ \mathbb{E}[s_\perp \mid \mathrm{OOD}] - \mathbb{E}[s_\perp \mid \mathrm{ID}] &= \|\boldsymbol{\delta}_\perp\|_2^2 + \mathrm{tr}\big(\Lambda_{\mathrm{ood},\perp} - \Lambda_\perp\big), \end{aligned} \tag{7}$$

where $\Lambda_\parallel = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ *and* $\Lambda_\perp = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_r)$.

If additionally $\boldsymbol{\xi} \,|\, \mathrm{ID} \sim \mathcal{N}(\mathbf{0}, \Lambda)$, *then*

$$\begin{aligned} \mathrm{Var}(s_\parallel \mid \mathrm{ID}) &= 2\sum_{j=1}^{k} \lambda_j^2, \\ \mathrm{Var}(s_\perp \mid \mathrm{ID}) &= 2\sum_{j=k+1}^{r} \lambda_j^2. \end{aligned} \tag{8}$$

and hence the SNRs take the closed forms

$$\begin{aligned} \mathrm{snr}(s_\parallel) &= \frac{\|\boldsymbol{\delta}_\parallel\|_2^2 + \mathrm{tr}(\Lambda_{\mathrm{ood},\parallel} - \Lambda_\parallel)}{\sqrt{2\sum_{j=1}^{k} \lambda_j^2} + \varepsilon}, \\ \mathrm{snr}(s_\perp) &= \frac{\|\boldsymbol{\delta}_\perp\|_2^2 + \mathrm{tr}(\Lambda_{\mathrm{ood},\perp} - \Lambda_\perp)}{\sqrt{2\sum_{j=k+1}^{r} \lambda_j^2} + \varepsilon}. \end{aligned} \tag{9}$$

Proof. Since V has orthonormal columns, $V_\parallel^\top \mathbf{z} = \boldsymbol{\xi}_\parallel$ and $V_\perp^\top \mathbf{z} = \boldsymbol{\xi}_\perp$, hence $s_\parallel = \|\boldsymbol{\xi}_\parallel\|_2^2$ and $s_\perp = \|\boldsymbol{\xi}_\perp\|_2^2$. For any random vector $\mathbf{u}$ with mean $\mathbf{m}$ and covariance Σ we have $\mathbb{E}\|\mathbf{u}\|_2^2 = \|\mathbf{m}\|_2^2 + \mathrm{tr}(\Sigma)$, which yields (7). Under the Gaussian ID assumption the coordinates are independent with variances $\{\lambda_j\}$, and $\mathrm{Var}(\xi_j^2) = 2\lambda_j^2$; summing these variances gives (8) and the SNR expressions (9). □

Corollary B.2 (Condition for $\mathrm{snr}(s_\perp) > \mathrm{snr}(s_\parallel)$)**.** *Under the assumptions of Lemma B.1, define*

$$\begin{aligned} \Delta_\parallel &= \|\boldsymbol{\delta}_\parallel\|_2^2 + \mathrm{tr}(\Lambda_{\mathrm{ood},\parallel} - \Lambda_\parallel), \\ \Delta_\perp &= \|\boldsymbol{\delta}_\perp\|_2^2 + \mathrm{tr}(\Lambda_{\mathrm{ood},\perp} - \Lambda_\perp). \end{aligned} \tag{10}$$

If

$$\frac{\Delta_\perp}{\sqrt{\sum_{j=k+1}^{r} \lambda_j^2}} > \frac{\Delta_\parallel}{\sqrt{\sum_{j=1}^{k} \lambda_j^2}},$$

then $\mathrm{snr}(s_\perp) > \mathrm{snr}(s_\parallel)$.

Proof. Immediate from the SNR expressions (9) by comparing the numerators scaled by the square-rooted ID variances (the common ε does not affect a strict inequality when it is identical for both terms). □

Lemma B.1 suggests a practical diagnostic for when complement scoring is likely to outperform retained-subspace scoring. Empirically we compute the sample SNRs of the two scores, $s_\parallel(x) = \|V_\parallel^\top \mathbf{z}(x)\|_2^2$ and $s_\perp(x) = \|V_\perp^\top \mathbf{z}(x)\|_2^2$, by estimating the ID means and variances in (5). Figure 6 summarizes this diagnostic: bars report empirical SNR values while overlaid markers indicate the AUROC obtained by thresholding each score. Consistent with the theory, the complement energy often exhibits a larger standardized mean shift and correspondingly stronger discrimination.

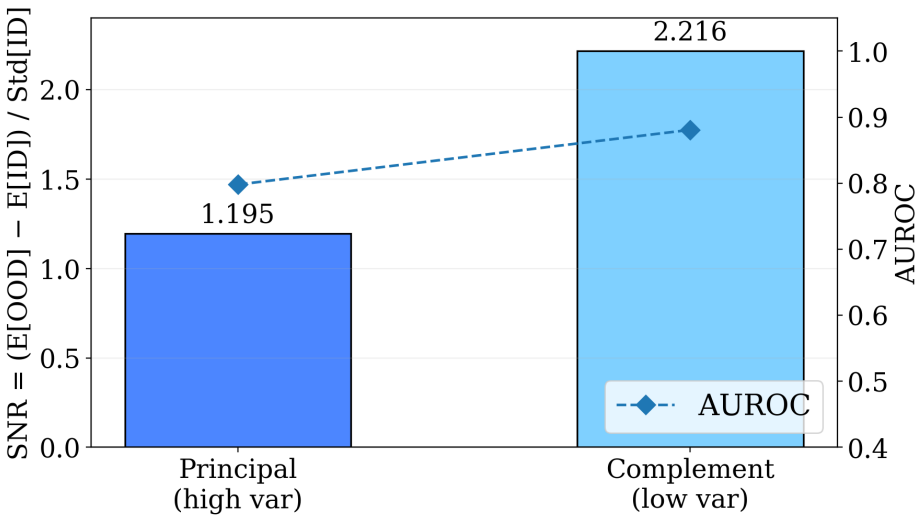


Figure 6: Subspace SNR diagnostic. Bars report $\mathrm{snr}(s_{\parallel})$ and $\mathrm{snr}(s_{\perp})$ computed from (5) using ID as the reference distribution. Diamond markers report AUROC obtained by thresholding the corresponding subspace energy score. Higher SNR and AUROC for the complement indicate that OOD deviations are more detectable in the low variance complement coordinates of the pooled within class residual model.