PROBE: PROTECTED OBSERVATION-SUBSPACE ENERGY FOR FINE-GRAINED OPEN-SET RECOGNITION ON FROZEN CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained open-set recognition must reject unseen categories that lie close to known classes while tolerating observation-induced variation in the features. In existing post-hoc scores, feature magnitude varies with prediction confidence and image conditions, and within-class covariance alone does not distinguish observation-induced variation from directions that separate known-class means. We introduce Protected Observation-Subspace Energy (PROBE), which combines class-specific isotropic energies on unit-normalized features with an additive covariance adjustment constructed from paired feature displacements under mild image corruptions, in which a between-class penalty discourages additional tolerance along directions separating known-class means. Among post-hoc and end-to-end baselines, PROBE achieves the highest mean Hard AUROC and Hard open-set classification rate (OSCR) on three Semantic Shift Benchmark (SSB) datasets and the highest mean AUROC and OSCR on three crop-level unmanned aerial vehicle (UAV) protocols constructed here. With the covariance spectrum held fixed, directions fitted with the between-class penalty yield higher mean AUROC than directions fitted without it on every dataset. Fitting uses only known-class samples of a frozen classifier and has a closed-form solution, and inference scores a single test image without corrupting it, retraining the classifier, or changing its label predictions.

1 INTRODUCTION

Open-set recognition (OSR) must reject unknown categories while preserving accurate known-class predictions (Vaze et al., 2022; Lang et al., 2024). The fine-grained case is difficult because unknown categories occupy neighborhoods of known-class features: distinguishing attributes are often local, so semantic shift need not produce a large residual. Observation-induced variation also displaces known-class features, so expanding class support uniformly to tolerate this variation can admit nearby unknowns. This raises the question of which feature directions should receive additional tolerance without weakening rejection of semantically similar unknowns.

Post-hoc rejectors reuse a trained classifier and decouple rejection from representation learning, but existing scores do not answer this question. Confidence scores such as the maximum softmax probability (MSP) and the maximum logit are strong baselines (Vaze et al., 2022), and Gaussian scoring models known-class feature distributions explicitly (Lee et al., 2018). However, feature magnitude varies with prediction confidence and image conditions, while within-class covariance alone does not explicitly distinguish observation-induced variation from directions separating class means. Comparing an image with a mildly corrupted copy exposes observation-induced directions, but these may overlap with directions separating known-class means; tolerance along this overlap may also make it harder to reject nearby unknowns.

PROBE addresses both issues with two statistics fitted in closed form to a frozen classifier: a class-conditional energy (M1) and a protected observation subspace (M2). M1 fits an isotropic Gaussian energy to each class on unit-normalized features, so the score depends on feature direction rather than magnitude. M2 estimates a shared observation subspace from the paired displacements induced by mild corruption probes while penalizing overlap with the scatter of known-class means

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Figure 1), treating these directions as a proxy for those that separate nearby unknowns. The subspace and its spectrum define an additive covariance adjustment, and the resulting energy is the rejection score. Probes are used only to fit M2; inference scores the normalized feature of one image without corrupting the test image, and the frozen head supplies the label.

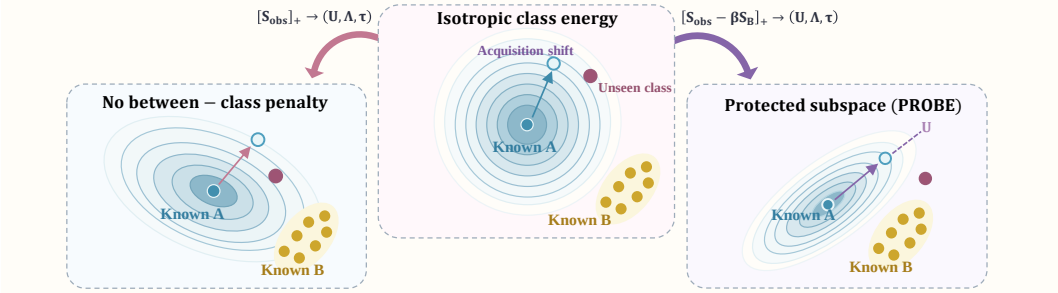

Figure 1: Schematic of the class support of known class A under observation-induced variation. Center: M1 alone has no preferred direction. Left: M2 fitted without the between-class penalty can elongate toward known class B and admit an unseen class. Right: PROBE penalizes directions separating known-class means when fitting the subspace, discouraging additional tolerance along them.

The main contributions of this paper are summarized as follows:

- We propose PROBE, a post-hoc rejection energy for frozen classifiers that combines class-specific isotropic energies on unit-normalized features with a between-class-protected additive covariance adjustment. It is fitted in closed form from known-class samples, scores a single image at inference, and leaves the frozen head's predictions unchanged.
- We test between-class protection through a penalty ablation, occupancy measurements on validation probes, and a direction control with a fixed spectrum and isotropic tail. With the spectrum held fixed, protected directions yield higher mean AUROC than unprotected directions on every dataset, and the mean gains from this adjustment extend to shared and class-specific Ledoit–Wolf covariances.
- We construct three crop-level UAV protocols. With one shared configuration, PROBE achieves the highest mean AUROC and OSCR on them and the highest mean Hard AUROC and Hard OSCR on three fine-grained benchmarks among post-hoc and end-to-end baselines.

2 Related Work

Open-set recognition. OpenMax introduced an explicit unknown probability by applying extreme-value modeling to activation responses (Bendale & Boult, 2016). Other approaches generated counterfactual pseudo-unknowns (Neal et al., 2018), represented open space with reciprocal points (Chen et al., 2020; 2022), or learned boundaries using data and classifier placeholders (Zhou et al., 2021). OpenAUC instead introduced an AUC-oriented training objective (Wang et al., 2022b). Vaze et al. (2022) showed that strong closed-set classifiers provide competitive OSR baselines and introduced the SSB. Fine-grained OSR further challenges rejection through localized differences between known and unknown categories (Lang et al., 2024). PROBE operates post hoc, fitting a rejection energy to a frozen encoder and classification head.

Feature-space geometry. Feature-space methods model deep representations beyond classification confidence. Mahalanobis scoring uses class-conditional Gaussians (Lee et al., 2018), CGDL constrains class-conditional latent distributions (Sun et al., 2020), and CSSR models class-specific reconstruction manifolds (Huang et al., 2023). Other methods represent known features as unions of low-dimensional subspaces (Zaeemzadeh et al., 2021) or use non-parametric neighbor distances (Sun et al., 2022). ViM takes a principal subspace of the training features and fuses the residual

outside it with the logits (Wang et al., 2022a). Tangent distance discounts displacements along local transformation directions when comparing prototypes (Hastie et al., 1994). PROBE instead fits its subspace to corruption displacements while penalizing directions that separate known-class means, and uses it inside class-conditional Gaussian energies as added variance, so displacements within the subspace receive a smaller quadratic penalty instead of being scored as a residual.

Invariant representation learning. Invariant representation learning reduces sensitivity to factors such as style and imaging conditions. Disentanglement methods separate variation factors (Hu et al., 2018) or learn distinct semantic and domain-specific representations (Zhang et al., 2022). Other work formulates augmentation as invariance-constrained learning (Hounie et al., 2023) or learns sample-specific local invariances (Miao et al., 2023). Domain-generalization methods separate domain information in prompts (Cheng et al., 2024) or suppress domain-sensitive features (Wen et al., 2025). Whereas these approaches train or adapt representations, PROBE uses mild ImageNet-C corruptions (Hendrycks & Dietterich, 2019) to fit additional tolerance into the rejection energy while leaving the encoder unchanged.

3 METHOD

PROBE fits class means and isotropic scales from known-class training features, together with a shared covariance adjustment constructed from paired probes (Figure 2). At inference, it evaluates the resulting class energies on a single image without applying probe transformations and uses their minimum for rejection.

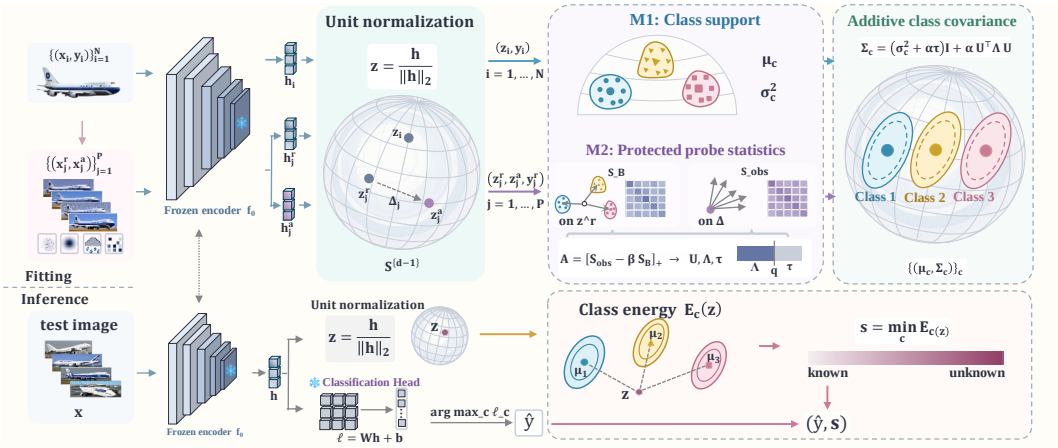

Figure 2: PROBE fitting and inference. M1 estimates per-class means and isotropic scales; M2 fits a shared protected observation subspace with its spectrum and isotropic tail. Together they define the additive class covariance Σ_c. The frozen head supplies $\hat{y}$, and the rejection score is $s = \min_c E_c(z)$.

3.1 PROBLEM FORMULATION AND PAIRED OBSERVATION PROBES

Let $\mathcal{D}_{\mathrm{K}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the known-class training set, where $y_i \in \mathcal{C}_{\mathrm{K}}$. We train a feature backbone f and a linear classification head on $\mathcal{D}_{\mathrm{K}}$ with cross-entropy and then freeze both. The frozen head produces logits $\ell(x)$ and the closed-set prediction $\hat{y} = \arg\max_{c \in \mathcal{C}_{\mathrm{K}}} \ell_c(x)$. The rejector operates only on unit-normalized column features $z = \mathrm{normalize}(f(x)) \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$.

An observation probe pairs a reference training image with a copy transformed by a single mild corruption. Let $\{(z_j^r, z_j^a, y_j^r)\}_{j=1}^{P}$ denote the resulting normalized feature pairs, where P counts pairs rather than distinct images. We use all known-class training images without additional class balancing and generate three corrupted copies per image, targeting $P = 3N$ reference–corruption pairs. Each copy uses one operator sampled uniformly with replacement from 19 *imagecorruptions* operators at severity 1 (Hendrycks & Dietterich, 2019). M1 statistics use the N normalized training features, and the two scatter matrices of M2 use the P pairs. We assume these probes preserve

the class label, although they may attenuate discriminative attributes. Appendix A.1 details probe generation and preprocessing.

3.2 M1: Class-Conditional Energy on Normalized Features

M1 models class support on unit-normalized features, making the score invariant to positive feature rescaling. A separate isotropic scale for each class accounts for differences in feature dispersion. For the normalized training features $z_i = \mathrm{normalize}(f(x_i))$, we compute

$$\mu_c = \frac{1}{n_c} \sum_{y_i=c} z_i, \qquad \sigma_c^2 = \frac{1}{n_c\, d} \sum_{y_i=c} \|z_i - \mu_c\|_2^2, \tag{1}$$

where n_c is the number of training features in class c. These parameters define an isotropic Gaussian energy in the ambient feature space, evaluated at unit-normalized features:

$$E_c^{(0)}(z) = \frac{\|z - \mu_c\|_2^2}{\sigma_c^2} + \tfrac{1}{2}\, d \log \sigma_c^2. \tag{2}$$

M1 alone uses the minimum class energy, $s(z) = \min_{c \in \mathcal{C}_\mathrm{K}} E_c^{(0)}(z)$. The distance term scales residuals by class dispersion, while the log-variance term accounts for the volume of class support. The log-volume weight is fixed at $\frac{1}{2}$, yielding, up to an additive constant, an ambient Gaussian negative log-density with covariance $\sigma_c^2 I/2$ (Appendix B.1). For unit features, Eq. 1 gives $\sigma_c^2 = (1 - \|\mu_c\|_2^2)/d$. All variances are floored at 10^{-8}.

3.3 M2: Protected Observation Subspace from Paired Displacements

Increasing isotropic variance reduces the quadratic penalty equally in every direction, providing no directional preference for accommodating observation-induced variation. M2 uses paired observation probes to select directions for additional variance, allowing greater tolerance along those directions than in the complementary subspace. A between-class penalty discourages the selection of directions that separate known-class means.

Observation scatter. The paired displacement $\Delta_j = z_j^a - z_j^r$ measures the effect of one probe. Let $\bar{\Delta}$ be the mean over the P pairs and $\mathrm{TN}(X) = X/\operatorname{tr}(X)$ denote trace normalization. The observation scatter is

$$S_\mathrm{obs} = \mathrm{TN}\left(\frac{1}{P-1} \sum_{j=1}^{P} (\Delta_j - \bar{\Delta})(\Delta_j - \bar{\Delta})^\top \right). \tag{3}$$

Trace normalization places observation and between-class scatter on a common scale, while centering captures variation around the mean probe displacement.

Between-class protection. We measure between-class scatter using the reference features of the probe pairs. Let m_c count reference pairs in class c, with mean $\tilde{\mu}_c$, and let $\tilde{\mu}$ be their overall mean. Then

$$S_\mathrm{B} = \mathrm{TN}\left(\frac{1}{P} \sum_{c \in \mathcal{C}_\mathrm{K}} m_c (\tilde{\mu}_c - \tilde{\mu})(\tilde{\mu}_c - \tilde{\mu})^\top \right). \tag{4}$$

We seek $U \in \mathbb{R}^{q \times d}$ with orthonormal rows that maximizes $\operatorname{tr}(U S_\mathrm{obs} U^\top) - \beta \operatorname{tr}(U S_\mathrm{B} U^\top)$, capturing observation scatter while penalizing overlap with between-class variation. The leading eigenvectors of $S_\mathrm{obs} - \beta S_\mathrm{B}$ attain this maximum. For covariance construction, we retain its positive spectral component:

$$A = \left[S_\mathrm{obs} - \beta S_\mathrm{B} \right]_+ = \sum_{k=1}^{d} \lambda_k u_k u_k^\top, \qquad U = [u_1, \ldots, u_q]^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q), \tag{5}$$

where $[\cdot]_+$ clips negative eigenvalues to zero and the eigenvalues are sorted in decreasing order. The requested rank q_req is capped by $d - 1$ and by the number of positive eigenvalues; q denotes the effective rank. The mean τ of the remaining $d - q$ eigenvalues of A sets an isotropic tail, which Eq. 6 adds in every direction, including the selected subspace (Appendix B.2). M2 is controlled by q_req, β and the covariance scale α introduced next.

3.4 Observation-Aware Covariance and Open-Set Decision

M2 modifies the class covariance through an additive observation term with scale $\alpha > 0$:

$$\Sigma_c = (\sigma_c^2 + \alpha\tau)\, I + \alpha\, U^\top \Lambda U, \qquad s_c^2 = \sigma_c^2 + \alpha\tau. \tag{6}$$

The covariance combines isotropic tail variance with additional variance along the selected observation directions (Appendix B.2). Because Λ and τ are derived from the spectrum of the penalized, trace-normalized matrix A, this term is a constructed tolerance rather than an estimate of the displacement covariance, and α sets its scale relative to σ_c^2. Orthonormal rows of U give a closed-form energy without a $d \times d$ inverse. Writing $r = z - \mu_c$ and $b_k = u_k^\top r$, we obtain

$$E_c(z) = \sum_{k=1}^{q} \frac{b_k^2}{s_c^2 + \alpha\lambda_k} + \frac{\|r\|_2^2 - \sum_{k=1}^{q} b_k^2}{s_c^2} + \tfrac{1}{2}\Big(d \log s_c^2 + \sum_{k=1}^{q} \log\Big(1 + \frac{\alpha\lambda_k}{s_c^2}\Big)\Big). \tag{7}$$

With $\alpha = 0$, Eq. 7 reduces to the M1 energy in Eq. 2. For a residual $r = tu_k$, the quadratic term is $t^2/(s_c^2 + \alpha\lambda_k)$, whereas a residual orthogonal to the selected subspace contributes $\|r\|_2^2/s_c^2$. At equal residual length, the selected directions therefore receive a smaller quadratic penalty, while a positive isotropic tail also reduces penalties outside the subspace. For a fixed fitted spectrum, the log-determinant term increases with α, balancing residual tolerance against the volume of class support.

Fitting and inference. Fitting uses only known-class samples to compute the class statistics and the two scatter matrices, followed by one $d \times d$ eigendecomposition yielding U, Λ and τ. No rejection-loss optimization, calibration, or logit fusion is required. With $U\mu_c$ precomputed and Uz reused across classes, scoring costs $O(qd + |\mathcal{C}_{\mathrm{K}}|\, d)$ per feature and requires no $d \times d$ matrix inverse. Accepted samples retain the frozen head's prediction $\hat{y}$.

Covariance variants. For the covariance controls in Section 4.5, we replace the isotropic class term with shared or class-specific Ledoit–Wolf estimates fitted to unit-normalized, class-centered training features. The shared estimate pools residuals across classes, whereas the class-specific estimates are fitted separately. Each empirical covariance S is regularized as $\widehat{\Sigma} = (1-\rho)S + \rho\, \mathrm{tr}(S)I/d + 10^{-5}I$, with ρ estimated from the corresponding training residuals (Ledoit & Wolf, 2004). Both variants retain the quadratic-plus-log-determinant energy and, when enabled, the same additive M2 covariance term.

4 Experiments

The experiments address three questions: whether PROBE rejects unknowns better than post-hoc and end-to-end baselines on fine-grained and UAV benchmarks (Sections 4.2 and 4.3), whether this advantage holds across frozen backbones (Section 4.4), and how M1 and M2 contribute to its performance, including the effect of between-class protection (Section 4.5).

4.1 Experimental Setup

Datasets and protocols. FGVC-Aircraft (Maji et al., 2013), Stanford Cars (Krause et al., 2013) and Caltech-UCSD Birds (CUB) (Wah et al., 2011) follow the SSB protocol (Vaze et al., 2022). Unknown classes are partitioned into Hard and Easy subsets by semantic proximity, and images are evaluated at 448×448. We evaluate Drone-type-Set (DTS) (AlDosari et al., 2024), the Sunny subset of VisioDECT (Ajakwe et al., 2026), and ICG-Drone-3c (Xiao et al., 2024) using target crops. The classes are drone models, and each fold holds out one model as the unknown class. Their leave-one-class-out protocols have seven, six, and three folds, respectively. Known-class partitions target a 70/10/20 train/validation/test ratio, using whole recording groups when at least three are available per class and contiguous temporal blocks otherwise. Partitions remain fixed across training seeds, and the held-out class is excluded from classifier training and rejector fitting in each fold. Appendix A.2 details crop construction, sample selection, and recording-based partition rules.

Implementation details. We train cross-entropy (CE) classifiers and refer to them as Shared-CE, because all post-hoc methods share the same frozen classifier within each dataset–backbone setting and, for UAV evaluation, within each leave-one-class-out fold. All four backbones use the checkpoint with the highest known-class validation accuracy within their specified training schedule. The primary backbone is a Places365-pretrained ResNet-50 with $d = 2048$ (Zhou et al., 2018; He et al., 2016). Appendix A.3 reports initialization sources and training settings for all four backbones.

We use one shared configuration, $\alpha = 0.3$, $\beta = 0.5$, and $q_{\text{req}} = 64$, for all primary and additional-backbone evaluations. Appendix C reports sensitivity analyses. Main tables report means across three classifier training seeds; seed identifiers and sample standard deviations (SD) are provided in the appendices.

Baselines. Post-hoc baselines share the frozen Shared-CE classifier with PROBE: MSP (Hendrycks & Gimpel, 2017), maximum logit score (MLS) (Vaze et al., 2022), Energy (Liu et al., 2020), Mahalanobis (Lee et al., 2018), ViM (Wang et al., 2022a), and KNN (Sun et al., 2022). End-to-end baselines retrain the backbone from the same Places365 initialization with their official code: ARPL+CS (Chen et al., 2022), CSSR (Huang et al., 2023), OpenAUC (Wang et al., 2022b), and SupCon (Khosla et al., 2020), NegCosSch-CE (Kundu et al., 2026), and SpHOR (Bahavan et al., 2026), the last three scored with MLS. All methods use the same partitions and evaluation protocol.

Metrics. Known-class classification is evaluated with accuracy, and OSR with the area under the receiver operating characteristic curve (AUROC) and OSCR; the latter measures correct known-class prediction jointly with unknown rejection (Dhamija et al., 2018). All metrics are reported as percentages; AUROC and OSCR summarize performance across thresholds. SSB results distinguish Hard and Easy unknowns, with Hard AUROC as the primary measure; UAV evaluation also prioritizes AUROC.

4.2 Fine-Grained Open-Set Recognition

Table 1: Fine-grained open-set recognition. Mean over three seeds (%). Hard/Easy AUROC and OSCR. Post-hoc methods share the classifier; end-to-end methods use separately trained classifiers. Best AUROC and OSCR in each column are bolded; Acc is shared by all post-hoc methods and is not bolded. SD in Appendix D.

Method	FGVC-Aircraft					Stanford Cars					CUB				
	Acc	Hard		Easy		Acc	Hard		Easy		Acc	Hard		Easy	
		AUROC	OSCR	AUROC	OSCR		AUROC	OSCR	AUROC	OSCR		AUROC	OSCR	AUROC	OSCR
Post-hoc on Shared-CE															
MSP	89.81	80.52	76.90	86.39	81.96	96.80	84.59	83.28	93.76	91.73	84.46	77.02	69.99	83.29	74.62
MLS	89.81	80.82	76.32	88.47	82.92	96.80	84.74	83.32	93.97	91.86	84.46	75.16	68.11	83.35	74.42
KNN	89.81	80.44	76.81	86.78	82.32	96.80	85.75	84.53	93.48	91.69	84.46	76.03	69.71	82.55	74.82
Energy	89.81	80.82	76.23	88.50	82.85	96.80	84.60	83.15	93.78	91.66	84.46	74.70	67.57	83.05	74.00
Mahalanobis	89.81	73.77	70.29	84.29	79.59	96.80	81.21	80.09	90.09	88.60	84.46	72.57	65.90	79.35	71.24
ViM	89.81	66.94	63.73	77.36	73.31	96.80	79.30	78.12	86.51	85.09	84.46	69.23	62.72	78.89	70.56
End-to-end															
ARPL+CS	71.28	68.95	55.10	76.52	60.10	96.75	84.35	82.86	94.19	91.93	83.19	76.35	68.70	84.60	74.57
CSSR	69.72	60.20	45.48	64.61	48.62	96.03	76.06	74.24	86.59	84.17	80.94	64.94	55.36	69.98	59.35
OpenAUC	88.56	81.01	76.35	89.48	83.27	95.95	84.32	82.58	94.22	91.53	82.28	77.40	68.80	86.21	75.01
SupCon + MLS	89.21	78.61	74.22	87.12	81.47	96.54	83.18	81.69	92.94	90.85	83.07	73.71	66.07	**86.44**	75.79
NegCosSch-CE + MLS	90.53	83.19	75.82	89.44	81.20	97.26	85.46	83.25	**94.57**	**92.10**	85.88	77.47	66.86	86.11	74.25
SpHOR + MLS	90.05	81.13	76.86	88.92	83.60	95.74	79.83	78.36	92.99	90.61	82.98	76.94	69.56	84.53	74.98
PROBE	89.81	**84.00**	**78.70**	**89.95**	**83.62**	96.80	**86.21**	**84.94**	93.43	91.62	84.46	**77.63**	**70.65**	86.09	**76.85**

PROBE achieves the highest mean Hard AUROC and Hard OSCR on the three SSB datasets (Table 1). Hard AUROC margins over the runners-up are 0.81, 0.46, and 0.16 percentage points on Aircraft, Cars, and CUB. On the Easy metrics PROBE leads on Aircraft and on CUB OSCR and is within about one point of the best score elsewhere. PROBE exceeds the raw-feature Mahalanobis baseline by 5.0–10.2 percentage points of Hard AUROC; Section 4.5 separates the contributions of M1 and M2.

4.3 Open-Set UAV Recognition

The selected UAV data contain 2,621 target crops for DTS, 6,912 for the Sunny subset of VisioDECT, and 718 for ICG-Drone-3c. Counts refer to individual annotated boxes; sample selection, crop construction, and recording-based partitions are detailed in Appendix A.2.

PROBE achieves the highest mean AUROC and OSCR on the three UAV datasets (Table 2). AUROC margins over the runners-up are 2.43, 0.38, and 3.83 percentage points on DTS, VisioDECT, and ICG-Drone-3c, respectively.

Table 2: Open-set UAV recognition under leave-one-class-out protocols. Mean over three seeds (%). Post-hoc methods share the frozen classifier within each dataset; Acc is shared by all post-hoc methods. Best AUROC and best OSCR in each column in bold; Acc is not bolded. SD in Appendix D.

Method	DTS			VisioDECT			ICG-Drone-3c		
	Acc	AUROC	OSCR	Acc	AUROC	OSCR	Acc	AUROC	OSCR
Post-hoc on Shared-CE									
MSP	90.98	77.89	74.49	95.03	76.59	74.70	85.73	65.79	60.65
MLS	90.98	73.65	70.12	95.03	74.61	72.30	85.73	66.01	60.81
KNN	90.98	81.21	77.87	95.03	92.76	90.40	85.73	72.05	66.59
Energy	90.98	73.16	69.54	95.03	74.36	72.00	85.73	66.02	60.81
Mahalanobis	90.98	78.07	75.03	95.03	91.29	89.38	85.73	71.56	66.42
ViM	90.98	76.00	73.09	95.03	92.94	90.24	85.73	71.31	66.08
End-to-end									
ARPL+CS	89.37	76.95	73.60	93.43	84.37	81.30	76.00	58.82	48.82
CSSR	80.03	60.87	52.64	89.34	72.06	67.36	69.86	59.36	40.19
OpenAUC	89.41	81.05	78.40	92.10	78.59	75.18	74.82	66.76	53.59
SupCon + MLS	93.58	81.58	79.13	93.87	91.75	90.40	84.02	64.62	56.99
NegCosSch-CE + MLS	90.27	81.95	78.20	94.54	88.55	86.75	78.41	64.72	56.41
SpHOR + MLS	91.59	82.18	77.73	93.37	93.23	89.32	86.54	65.98	58.40
PROBE	90.98	**84.61**	**81.09**	95.03	**93.61**	**90.51**	85.73	**75.88**	**68.97**

4.4 Generalization across Backbones

Table 3: Hard AUROC across backbones on the SSB datasets. Mean over three seeds (%). Methods within each dataset–backbone column share the same frozen classifier. Best in each column in bold. D-161: DenseNet-161. SD in Appendix D.

Method	FGVC-Aircraft			Stanford Cars			CUB		
	D-161	ConvNeXt-T	ViT-B/16	D-161	ConvNeXt-T	ViT-B/16	D-161	ConvNeXt-T	ViT-B/16
MSP	65.03	70.40	65.64	79.23	83.51	79.04	72.25	81.48	78.42
MLS	66.70	71.83	65.92	78.92	82.64	79.48	69.17	81.03	77.05
KNN	66.99	72.11	63.63	76.77	84.54	77.63	68.90	81.70	75.12
Energy	66.81	72.27	65.88	78.59	82.28	79.71	68.25	80.40	76.39
Mahalanobis	65.17	72.78	62.20	74.21	84.30	73.73	68.31	81.71	73.47
ViM	59.57	72.23	63.51	65.50	83.47	74.25	59.96	80.28	73.31
PROBE	**67.56**	**74.76**	**66.72**	**80.02**	**85.42**	**80.34**	**73.04**	**82.39**	**78.72**

PROBE has the highest mean Hard AUROC among post-hoc scores in all nine additional dataset–backbone settings (Table 3). Gains over Mahalanobis range from 4.5 to 6.6 percentage points on ViT-B/16 and from 0.7 to 2.0 percentage points on ConvNeXt-T. Margins over the strongest baseline in each setting range from 0.30 to 1.98 points.

4.5 Ablation and Mechanism Analysis

We now ask where PROBE's gains come from. Relative to raw-feature Mahalanobis, M1 improves mean AUROC on all six datasets; adding M2 gives a further gain of 1.3–6.0 percentage points (Table 4). The relative contributions vary across datasets: M1 accounts for most of the gain on Aircraft and CUB, whereas M2 contributes more than M1 on DTS and VisioDECT. The random-subspace variant with a refitted spectrum improves mean AUROC over M1 on four datasets and reduces it on two, while remaining 0.4–5.7 percentage points below PROBE across all six datasets.

Table 4: Configuration comparison: mean Hard AUROC (SSB) or leave-one-class-out AUROC (UAV) over three seeds (%). Mahalanobis uses raw features and tied covariance. The random-subspace row averages the scores from three random rank-q orthonormal bases, each with Λ and τ refitted from observation scatter. SDs: Appendix D. ICG: ICG-Drone-3c.

Configuration	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
Raw-feature Mahalanobis	73.77	81.21	72.57	78.07	91.29	71.56
M1	82.13	83.90	76.29	78.62	92.03	73.89
M1 + random subspace	81.93	85.12	77.25	78.92	89.95	74.06
PROBE	**84.00**	**86.21**	**77.63**	**84.61**	**93.61**	**75.88**

Removing the between-class penalty reduces mean AUROC most strongly on Aircraft and DTS, by 13.30 and 9.89 percentage points (Table 15). On both datasets, this variant falls below M1 (Tables 4 and 13).

With PROBE's spectrum held fixed, replacing protected directions with unprotected ones lowers mean AUROC by about ten points on Aircraft and DTS and by 0.88–1.39 points on the other four datasets (Table 5). For each class, PROBE and this control have identical covariance eigenvalues, traces, and log-determinants, isolating the effect of subspace directions.

Table 5: Fixed-spectrum direction control: U_p replaced by the leading eigenvectors U_0 of S_{obs}, retaining PROBE's spectrum, tail, and rank. Mean AUROC over three seeds (%); paired differences are mean±SD in percentage points. PROBE means: Table 4; control SDs: Appendix D. ICG: ICG-Drone-3c.

Configuration	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
Unprotected directions U_0	73.07	85.11	76.75	74.56	92.22	74.70
PROBE $- U_0$ (paired)	10.93 ± 1.77	1.10 ± 0.94	0.88 ± 0.60	10.05 ± 2.99	1.39 ± 1.16	1.18 ± 0.18

The isotropic same-trace control distributes the added variance uniformly across all feature dimensions, preserving its trace. PROBE exceeds this control by 1.34–5.54 points of paired mean AUROC (Table 15; construction in Appendix B.3).

Occupancy on validation probes. We evaluate the fitted subspaces on new known-class validation probes generated with a random seed disjoint from fitting. Occupancy measures the fraction of validation scatter captured by U: $R_{\text{obs}} = \text{tr}(US_{\text{obs}}U^\top)$ and $R_{\text{B}} = \text{tr}(US_{\text{B}}U^\top)$. A random rank-$q$ subspace has expected occupancy q/d (about 0.03 for ResNet-50).

For operators available during fitting (Table 6, upper block), protected mean R_{B} ranges from 0.13–0.43 across the six datasets, versus 0.68–0.99 without protection; protected mean R_{obs} ranges from 0.52–0.88. Cars shows a smaller mean reduction than Aircraft and DTS, from 0.69 ± 0.20 to 0.41 ± 0.21 (mean±SD). These known-class probe measurements are consistent with retaining observation variation while reducing overlap with directions separating known-class means.

Table 6: Validation-probe occupancy, mean over three seeds. Upper block: fitting operators with the fitted basis. Lower block: dataset-specific held-out operators after refitting. SDs: Table 16. ICG: ICG-Drone-3c.

Dataset	PROBE		$\beta = 0$	
	R_{obs}	R_{B}	R_{obs}	R_{B}
Operators available during fitting				
Aircraft	0.79	0.43	0.99	0.99
Cars	0.52	0.41	0.62	0.69
CUB	0.58	0.38	0.73	0.76
DTS	0.88	0.27	0.99	0.98
VisioDECT	0.86	0.13	0.89	0.68
ICG	0.77	0.28	0.84	0.71
Held-out operators on Aircraft and DTS				
Aircraft	0.69	0.41	0.95	0.99
DTS	0.86	0.31	1.00	0.99

We also evaluate held-out operators on Aircraft and DTS only (Table 6, lower block). Aircraft holds out impulse noise and brightness; DTS additionally holds out glass blur and JPEG compression. We refit each model's U and spectrum with these operators excluded from training probes, then measure occupancy on their known-class validation probes. This pattern also holds for the held-out operators on these two datasets.

Beyond isotropic class covariance. Adding M2 increases mean AUROC with both shared and class-specific Ledoit–Wolf covariances across all six datasets, extending its mean gains beyond isotropic class models (Table 7).

Table 7: Covariance controls on unit-normalized features: mean AUROC over three seeds (%). Variants share frozen classifiers; base covariance and M2 vary. Isotropic rows without/with M2 are M1/PROBE. Best in bold; SDs: Table 21. ICG: ICG-Drone-3c.

Base covariance	M2	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
Class-specific isotropic	—	82.13	83.90	76.29	78.62	92.03	73.89
Shared Ledoit–Wolf	—	79.94	84.26	77.28	81.49	91.19	70.53
Class-specific Ledoit–Wolf	—	80.58	85.10	76.83	82.33	91.03	69.28
Class-specific isotropic	✓	**84.00**	**86.21**	**77.63**	**84.61**	**93.61**	**75.88**
Shared Ledoit–Wolf	✓	80.91	84.74	77.40	82.53	92.69	72.31
Class-specific Ledoit–Wolf	✓	81.58	85.81	76.88	83.17	92.46	70.91

5 Conclusion

PROBE combines class-specific isotropic energies on unit-normalized features with an additive covariance adjustment constructed from paired observation probes for open-set rejection on frozen classifiers. The class energy models known-class support independently of feature magnitude, while the between-class penalty guides the selected observation directions away from directions separating known-class means.

Experiments on fine-grained benchmarks and crop-level UAV protocols support this combination. Fixed-spectrum direction controls show higher mean AUROC with protected directions across the six datasets, while validation-probe occupancy measurements characterize their reduced overlap with between-class scatter. PROBE provides a closed-form rejector fitted only from known-class samples, with single-image inference that requires no probe transformations and preserves the frozen head's predictions.

AI use statement

We used generative AI tools for literature retrieval and discovery, including identifying potentially relevant prior work. We checked suggested sources against the original publications before citing them. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Section 3 specifies PROBE's fitting and scoring rules. The appendices detail probe generation, UAV partitions, backbone training, checkpoint selection, training seeds, and the fixed-spectrum direction control. They also report seed variability, validation-probe occupancy measurements, and sensitivity analyses. All fitting and scoring ran on one NVIDIA RTX 6000 Ada Generation GPU per job.

References

Simeon Okechukwu Ajakwe, Vivian Ukamaka Ihekoronye, Golam Mohtasin, Rubina Akter, Jae Min Lee, and Dong Seong Kim. VisioDECT: A robust dataset for aerial and scenario based multi-drone detection, identification, and neutralization. *Data in Brief*, 65:112448, 2026. doi: 10.1016/j.dib.2026.112448.

Khloud AlDosari, AIbtisam Osman, Omar Elharrouss, Somaya Al-Maadeed, and Mohamed Zied Chaari. Drone-type-Set: Drone types detection benchmark for drone detection and tracking. In *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1–7. IEEE, 2024.

Thiru Thillai Nadarasar Bahavan, Sachith Seneviratne, and Saman Halgamuge. SpHOR: A representation learning perspective on open-set recognition for identifying unknown classes in deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6901–6910, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Bahavan_SpHOR_A_Representation_Learning_Perspective_on_Open-set_Recognition_for_Identifying_CVPRF_2026_paper.html.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1563–1572, 2016.

Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 507–522. Springer, 2020.

Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2022.

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23595–23604. IEEE, 2024.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18613–18624, 2020.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Trevor Hastie, Patrice Simard, and Eduard Säckinger. Learning prototype models for tangent distance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 7, 1994.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Ignacio Hounie, Luiz F. O. Chamon, and Alejandro Ribeiro. Automatic data augmentation via invariance-constrained learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 13410–13433. PMLR, 2023.

Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3399–3407, 2018.

Hongzhi Huang, Yu Wang, Qinghua Hu, and Ming-Ming Cheng. Class-specific semantic reconstruction for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4214–4228, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18661–18673, 2020.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 554–561, 2013.

Amit Kumar Kundu, Vaishnavi S. Patil, and Joseph Jaja. Boosting open set recognition performance through modulated representation learning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=vpBKry7kL5.

Nico Lang, Vésteinn Snæbjarnarson, Elijah Cole, Oisin Mac Aodha, Christian Igel, and Serge Belongie. From coarse to fine-grained open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17804–17814. IEEE, 2024.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

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 (NeurIPS)*, volume 31, 2018.

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

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Ning Miao, Tom Rainforth, Emile Mathieu, Yann Dubois, Yee Whye Teh, Adam Foster, and Hyunjik Kim. Learning instance-specific augmentations by capturing local invariances. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 24720–24736. PMLR, 2023.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 620–635. Springer, 2018.

Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13480–13489. IEEE, 2020.

Yiyou Sun, Yifei Ming, Xiaojin 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, pp. 20827–20840. PMLR, 2022.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930. IEEE, 2022a.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. OpenAUC: Towards AUC-oriented open-set recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 25033–25045, 2022b.

Changsong Wen, Zelin Peng, Yu Huang, Xiaokang Yang, and Wei Shen. Domain generalization in CLIP via learning with diverse text prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9559–9569. IEEE, 2025.

Jiaping Xiao, Jian Hui Chee, and Mir Feroskhan. Real-time multi-drone detection and tracking for pursuit-evasion with parameter search. *IEEE Transactions on Intelligent Vehicles*, 9(10):6027–6037, 2024.

Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9452–9461. IEEE, 2021.

Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8024–8034. IEEE, 2022.

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.

Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410. IEEE, 2021.

A Experimental Protocols

A.1 Probe Generation

We enumerate all images in the known-class training directories, with class directories ordered by name and no additional class-balanced subsampling. Each image receives three corruptions, with operators drawn independently and uniformly with replacement from Table 8. The image identifier and corruption index determine each probe's random seed, independently of the classifier training seed. All training seeds share these corrupted images, but each frozen encoder produces its own reference and corrupted feature embeddings.

RGB images are resized to 448×448 for SSB or 224×224 for UAV before applying the selected corruption at severity 1. The corrupted image is then resized to the model input dimensions, converted to a tensor, and normalized using the ImageNet mean and standard deviation. For UAV datasets, probes operate on the precomputed target crops rather than on the original full frames.

Table 8: Probe operators. The list contains all 19 *imagecorruptions* operators, comprising 15 common corruptions and four additional validation corruptions. Sampling is uniform over operators, rather than over the four families.

Family	Operators
Noise	`gaussian_noise`, `shot_noise`, `impulse_noise`, `speckle_noise`
Blur	`defocus_blur`, `glass_blur`, `motion_blur`, `zoom_blur`, `gaussian_blur`
Weather	`snow`, `frost`, `fog`, `brightness`, `spatter`
Digital	`contrast`, `elastic_transform`, `pixelate`, `jpeg_compression`, `saturate`

A.2 UAV Data Construction and Partitions

We use the officially released frames without extracting additional frames from the original videos at 30 fps. Each target box is expanded by 15% on each side. The crop is enlarged as needed to ensure a minimum short side of 64 pixels, then letterboxed to 224×224. All sample counts refer to individual annotated target boxes, not videos or unique source frames (Table 9).

Table 9: UAV sample selection and evaluation scope. Target crops are counted by individual annotated boxes.

Dataset	Selection	Target crops	Leave-one-class-out folds
DTS	Official list; demonstration annotations and exact duplicate frames removed	2621	7
VisioDECT	Sunny subset only	6912	6
ICG-Drone-3c	All QAV samples; Tello and F450 each capped at 268, prioritizing complete videos	718	3

Partitions are established once before leave-one-class-out evaluation. Classes with at least three recording groups are partitioned by whole groups, without placing a group in multiple splits. For classes with fewer than three groups, each recording is divided into contiguous 70/10/20 temporal blocks. The temporal blocks meet at adjacent boundaries, with no temporal gap between splits. Whole-group allocation can produce realized proportions that differ from the target ratios.

DTS recording groups are defined by aircraft type, resolution, and contiguous recording segment. VisioDECT and ICG-Drone-3c use the official recording-source groups. VisioDECT contains only one or two recording groups per class, so its partitions predominantly use temporal blocks from shared recordings. The five QAV recording groups in ICG-Drone-3c are assigned as whole groups.

In each fold, the held-out category supplies unknown test samples and contributes no samples to classifier training or rejector fitting. The remaining categories retain their fixed train/validation/test partitions. DTS holds out each of seven categories: Bebop, dji phantom 3, dji phantom 4 pro, Emax, Intel Aero RTF, Mambo, and YH 19HW. VisioDECT uses six class-held-out folds, and ICG-Drone-3c holds out Tello, F450, and QAV in turn.

A.3 Backbone Training and Checkpoint Selection

ResNet-50 and DenseNet-161 are initialized from the official CSAIL checkpoints pretrained for supervised scene classification on Places365, `resnet50_places365.pth.tar` and `densenet161_places365.pth.tar`, respectively (Zhou et al., 2018). Places365 initialization of the primary ResNet-50 follows Vaze et al. (2022). ConvNeXt-T and ViT-B/16 use torchvision ImageNet-1K weights with the `IMAGENET1K_V1` identifier. For ViT-B/16 at 448×448, positional embeddings are interpolated from a 14×14 grid to a 28×28 grid.

We use classifier training seeds 0, 1, and 2 for all reported three-seed evaluations. Models complete their specified training schedules, after which the checkpoint with the highest known-class validation accuracy is selected. Tables 10 and 11 report the SSB settings at 448×448. The primary ResNet-50 UAV experiments run for 120 epochs on DTS and ICG-Drone-3c and 480 epochs on VisioDECT. They retain the ResNet-50 optimizer, learning rate, batch size, and RandAugment settings reported in Table 10.

Table 10: Backbone training settings on SSB. LR denotes learning rate, WD weight decay, RRC random resized crop, and RE random erasing. RandAugment settings are (N, M) (Cubuk et al., 2020).

Backbone	Epochs	Optimization and schedule	Additional settings
ResNet-50	600	SGD; LR 10^{-3}; constant	Batch size 12; RandAugment(2,30)
DenseNet-161	200	AdamW; LR 10^{-4}; cosine	Batch size 12; Cars uses RRC and flip; Aircraft/CUB additionally use RandAugment(2,14)
ConvNeXt-T	200	AdamW; LR 5×10^{-5}; WD 10^{-8}; cosine; no warmup	Layer decay 0.7; label smoothing 0.1; head initialization scale 0.001; RE 0.25
ViT-B/16	200	SGD; LR 0.01; momentum 0.9; WD 0; cosine; 5-epoch warmup	Gradient clipping 1

Table 11: Batch sizes for the additional backbones on SSB. ViT-B/16 uses a physical batch size of 8 with four gradient-accumulation steps on Aircraft and Cars, giving an effective batch size of 32.

Backbone	Aircraft	Cars	CUB
DenseNet-161	12	12	12
ConvNeXt-T	24	24	80
ViT-B/16	8×4 accumulation	8×4 accumulation	88

B Energy Interpretation and Covariance Controls

B.1 Ambient Gaussian Energy

Up to a constant independent of z and c, the energy $r^\top \Sigma_c^{-1} r + \frac{1}{2} \log\det \Sigma_c$ is the negative log-density of an ambient Gaussian with covariance $\Sigma_c/2$. The energy is evaluated on unit-normalized features without a spherical normalization constant.

B.2 Additive Observation Term and Tail

The additive observation term adds an isotropic term $\alpha\tau I$ across all d dimensions and additional variances $\alpha\lambda_k$ along the selected directions. Unlike probabilistic PCA, which uses $\lambda_k - \tau$ along retained directions, the leading eigenvalues are not reduced by τ, so the added covariance has trace $\alpha(\sum_{k\le q} \lambda_k + d\tau)$ rather than $\alpha \operatorname{tr}(A)$; the construction is therefore additive, not a trace-preserving approximation of A.

B.3 Isotropic Same-Trace Control

For a fitted PROBE model, the exact control is

$$q_{\text{iso}} = 0, \qquad \tau_{\text{iso}} = \tau + \frac{1}{d}\sum_{k=1}^{q} \lambda_k, \qquad \Sigma_c^{\text{iso}} = \left(\sigma_c^2 + \alpha\tau + \frac{\alpha}{d}\sum_{k=1}^{q} \lambda_k\right) I. \tag{8}$$

Here q, λ_k, and τ are taken from the original fitted PROBE model. The classifier, M1 statistics, α, and energy rule remain unchanged. With $q_{\text{iso}} = 0$, the energy takes the same scalar-variance form as M1, but with an enlarged variance. Its covariance spectrum and log-determinant generally differ from PROBE's. Results are reported in Table 15.

B.4 Fixed-Spectrum Direction Control

For each fitted PROBE model, we retain its Λ_p, τ_p, α, effective rank q, and M1 statistics. We replace U_p with U_0, whose rows are the leading q eigenvectors of S_{obs}, ordered by decreasing eigenvalue. The control covariance is

$$\Sigma_c^{\text{ctrl}} = (\sigma_c^2 + \alpha\tau_p) I + \alpha U_0^\top \Lambda_p U_0. \tag{9}$$

The control uses the same classifier, test samples, and energy rule as PROBE. The protected spectrum and tail come from PROBE fitted with $(\alpha, \beta, q_{\text{req}}) = (0.3, 0.5, 64)$; U_0 is computed without the between-class penalty.

C Hyperparameter Settings and Sensitivity

We vary one parameter at a time around the shared configuration $(\alpha, \beta, q_{\text{req}}) = (0.3, 0.5, 64)$ while fixing the other two. These sweeps assess sensitivity on the reported benchmarks.

All tested positive values of α improve mean AUROC over M1 on all six datasets, although the best-performing value varies by dataset. Between-class protection has its largest effect on Aircraft and DTS; increasing its weight further does not uniformly improve performance. Across the tested ranks, the shared rank-64 configuration remains within 0.24 percentage points of the highest mean AUROC on each dataset.

Table 12: Observation-strength sensitivity. Mean AUROC over three training seeds (%): Hard AUROC on SSB and leave-one-class-out AUROC on UAV. $\beta = 0.5$ and $q_{\mathrm{req}} = 64$ are fixed. Bold marks the configuration used in the primary experiments. ICG denotes ICG-Drone-3c.

α	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
0	82.13	83.90	76.29	78.62	92.03	73.89
0.1	83.94	86.05	77.72	84.14	93.48	75.78
0.2	84.02	86.16	77.65	84.36	93.63	75.78
0.3	**84.00**	**86.21**	**77.63**	**84.61**	**93.61**	**75.88**
0.5	83.98	86.00	77.66	84.45	93.72	75.82
1.0	83.89	86.22	77.58	84.40	93.77	75.90

Table 13: Between-class penalty sensitivity. Mean AUROC over three training seeds (%): Hard AUROC on SSB and leave-one-class-out AUROC on UAV. $\alpha = 0.3$ and $q_{\mathrm{req}} = 64$ are fixed. Bold marks the configuration used in the primary experiments. $\beta = 0$ removes the between-class penalty. ICG denotes ICG-Drone-3c.

β	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
0	70.70	84.52	71.98	74.72	90.84	70.80
0.1	77.68	85.15	77.53	81.49	93.73	76.12
0.25	80.21	85.12	77.61	84.14	93.59	75.71
0.5	**84.00**	**86.21**	**77.63**	**84.61**	**93.61**	**75.88**
1.0	84.04	85.87	77.37	84.02	93.36	75.19
2.0	83.81	86.22	77.23	83.52	93.12	75.18

Table 14: Subspace-rank sensitivity. Mean AUROC over three training seeds (%): Hard AUROC on SSB and leave-one-class-out AUROC on UAV. $\alpha = 0.3$ and $\beta = 0.5$ are fixed. Bold marks the configuration used in the primary experiments. The effective rank equals q_{req} in every reported setting. ICG denotes ICG-Drone-3c.

q_{req}	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
64	**84.00**	**86.21**	**77.63**	**84.61**	**93.61**	**75.88**
128	84.04	86.22	77.72	84.60	93.46	75.44
256	84.09	86.03	77.71	84.68	93.46	75.37
512	84.07	86.04	77.67	84.80	93.25	75.38
1024	84.12	86.05	77.63	84.85	93.18	75.39

D Supplementary Results and Variability

D.1 Statistical Reporting

We report the mean and sample SD across the three training seeds (Appendix A.3). Dataset partitions and fitting probes remain fixed across these runs (Appendix A.1). For UAV datasets, fold scores are macro-averaged within each seed before calculating statistics across training seeds. These SDs describe training-seed variability under fixed partitions and probes, rather than uncertainty across alternative recording splits or probe draws. For the SD-only tables, the corresponding means are in the main tables. Paired differences are computed within each seed before calculating their mean and sample SD. Differences are expressed in percentage points and calculated before rounding the displayed scores.

D.2 Supplementary Controls

Table 15: Supplementary controls. The isotropic same-trace construction is defined in Appendix B.3. Values are paired PROBE-minus-control AUROC differences, mean ± sample SD in percentage points. SSB uses Hard AUROC. ICG denotes ICG-Drone-3c.

Control	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
$\beta = 0$	13.30 ± 2.65	1.69 ± 1.58	5.65 ± 3.86	9.89 ± 3.20	2.77 ± 1.72	5.08 ± 1.14
Isotropic same-trace expansion	2.00 ± 0.70	1.34 ± 0.12	1.34 ± 0.66	5.54 ± 1.73	1.58 ± 1.46	1.91 ± 1.52

D.3 Validation-Probe Occupancy Variability

Section 4.5 describes both validation-probe settings and reports their mean occupancies in Table 6. Table 16 gives the corresponding sample SDs across the three training seeds.

Table 16: Validation-probe occupancy, sample SD over three seeds. Means and probe settings are in Table 6. ICG: ICG-Drone-3c.

Dataset	PROBE		$\beta = 0$	
	R_{obs}	R_{B}	R_{obs}	R_{B}
Operators available during fitting				
Aircraft	±0.06	±0.10	±0.00	±0.00
Cars	±0.15	±0.21	±0.17	±0.20
CUB	±0.11	±0.14	±0.22	±0.22
DTS	±0.03	±0.00	±0.00	±0.01
VisioDECT	±0.08	±0.02	±0.08	±0.24
ICG	±0.08	±0.01	±0.09	±0.07
Held-out operators on Aircraft and DTS				
Aircraft	±0.14	±0.10	±0.03	±0.01
DTS	±0.02	±0.03	±0.00	±0.01

D.4 Performance and Ablation Variability

Table 17: Fine-grained open-set recognition, sample SD over three seeds (%). Means in Table 1.

Method	FGVC-Aircraft					Stanford Cars					CUB				
	Acc	Hard		Easy		Acc	Hard		Easy		Acc	Hard		Easy	
		AUROC	OSCR	AUROC	OSCR		AUROC	OSCR	AUROC	OSCR		AUROC	OSCR	AUROC	OSCR
Post-hoc on Shared-CE															
MSP	±0.33	±0.38	±0.23	±1.22	±1.17	±0.13	±0.77	±0.70	±0.18	±0.20	±0.63	±0.51	±0.48	±0.80	±0.61
MLS	±0.33	±0.19	±0.19	±1.63	±1.57	±0.13	±0.77	±0.73	±0.09	±0.06	±0.63	±0.83	±0.65	±0.77	±0.57
KNN	±0.33	±0.55	±0.31	±0.53	±0.65	±0.13	±0.04	±0.09	±0.19	±0.23	±0.63	±0.54	±0.48	±0.85	±0.72
Energy	±0.33	±0.20	±0.22	±1.65	±1.59	±0.13	±0.76	±0.73	±0.10	±0.05	±0.63	±0.83	±0.65	±0.75	±0.55
Mahalanobis	±0.33	±0.73	±0.47	±1.35	±0.90	±0.13	±0.29	±0.32	±0.11	±0.13	±0.63	±0.58	±0.59	±0.79	±0.84
ViM	±0.33	±0.98	±0.81	±0.65	±0.49	±0.13	±0.35	±0.32	±0.32	±0.24	±0.63	±0.06	±0.06	±1.05	±0.93
End-to-end															
ARPL+CS	±0.26	±1.06	±0.66	±1.12	±0.75	±0.18	±0.48	±0.38	±0.09	±0.26	±0.53	±0.09	±0.12	±0.36	±0.56
CSSR	±0.43	±2.02	±1.31	±1.06	±0.73	±0.15	±1.01	±0.96	±0.73	±0.65	±0.44	±0.53	±0.61	±0.70	±0.73
OpenAUC	±0.74	±0.68	±0.71	±0.77	±0.63	±0.19	±0.46	±0.49	±0.38	±0.43	±0.88	±0.30	±0.90	±0.06	±0.77
SupCon + MLS	±1.31	±2.70	±3.26	±0.86	±1.78	±0.07	±0.38	±0.43	±0.13	±0.17	±0.47	±0.30	±0.38	±0.40	±0.10
NegCosSch-CE + MLS	±0.42	±0.19	±0.38	±0.61	±0.74	±0.13	±0.51	±0.59	±0.34	±0.31	±0.30	±0.93	±0.64	±0.63	±0.34
SpHOR + MLS	±0.34	±0.76	±0.81	±0.39	±0.20	±0.21	±0.49	±0.51	±0.27	±0.33	±0.40	±0.29	±0.44	±0.38	±0.59
PROBE	±0.33	±0.05	±0.23	±0.76	±0.36	±0.13	±0.28	±0.25	±0.36	±0.41	±0.63	±0.76	±0.87	±1.11	±0.63

Table 18: Open-set UAV recognition, sample SD over three seeds (%). Means in Table 2.

Method	DTS			VisioDECT			ICG-Drone-3c		
	Acc	AUROC	OSCR	Acc	AUROC	OSCR	Acc	AUROC	OSCR
Post-hoc on Shared-CE									
MSP	±1.44	±2.09	±2.41	±0.25	±6.05	±5.97	±1.81	±4.21	±3.54
MLS	±1.44	±3.07	±3.13	±0.25	±7.42	±7.28	±1.81	±3.77	±3.20
KNN	±1.44	±1.44	±2.02	±0.25	±1.66	±1.23	±1.81	±0.61	±0.77
Energy	±1.44	±3.15	±3.17	±0.25	±7.59	±7.44	±1.81	±3.72	±3.17
Mahalanobis	±1.44	±1.24	±1.96	±0.25	±1.01	±0.71	±1.81	±1.45	±1.09
ViM	±1.44	±1.41	±1.54	±0.25	±1.04	±1.14	±1.81	±1.39	±0.95
End-to-end									
ARPL+CS	±0.64	±5.33	±4.73	±1.36	±8.54	±7.01	±2.35	±11.59	±8.45
CSSR	±1.75	±0.91	±2.56	±2.18	±5.20	±4.36	±3.25	±3.09	±1.30
OpenAUC	±2.58	±1.01	±1.09	±0.70	±3.97	±3.44	±3.82	±4.08	±1.87
SupCon + MLS	±0.97	±2.40	±2.78	±0.55	±3.13	±5.22	±1.31	±1.54	±1.95
NegCosSch-CE + MLS	±1.40	±1.43	±1.81	±0.14	±1.30	±1.68	±0.78	±3.07	±2.35
SpHOR + MLS	±1.10	±1.06	±1.42	±0.20	±0.63	±0.57	±1.59	±4.04	±3.12
PROBE	±1.44	±1.34	±1.96	±0.25	±1.04	±0.32	±1.81	±0.50	±0.24

Table 19: Hard AUROC across backbones, sample SD over three seeds (%). Means in Table 3.

Method	FGVC-Aircraft			Stanford Cars			CUB		
	D-161	ConvNeXt-T	ViT-B/16	D-161	ConvNeXt-T	ViT-B/16	D-161	ConvNeXt-T	ViT-B/16
MSP	±2.33	±0.21	±1.18	±1.07	±0.57	±0.14	±0.15	±0.20	±0.23
MLS	±2.93	±0.07	±1.01	±1.08	±0.43	±0.55	±0.49	±0.14	±0.06
KNN	±2.85	±0.64	±0.88	±0.55	±0.87	±0.75	±0.43	±0.67	±0.45
Energy	±2.72	±0.16	±0.94	±1.16	±0.36	±0.59	±0.54	±0.70	±0.08
Mahalanobis	±1.39	±0.97	±1.18	±0.70	±0.76	±0.43	±1.01	±0.77	±0.23
ViM	±0.93	±0.88	±1.02	±1.00	±0.77	±0.45	±1.65	±0.99	±0.37
PROBE	±3.77	±0.19	±1.17	±0.39	±0.21	±0.65	±0.20	±0.76	±0.29

Table 20: Configuration comparison and direction control, sample SD over three seeds (%). Means in Tables 4 and 5. ICG: ICG-Drone-3c.

Configuration	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
Raw-feature Mahalanobis	±0.73	±0.29	±0.58	±1.24	±1.01	±1.45
M1	±0.63	±0.33	±0.10	±1.28	±1.62	±1.15
M1 + random subspace	±0.60	±0.54	±0.56	±1.18	±1.29	±1.45
PROBE	±0.05	±0.28	±0.76	±1.34	±1.04	±0.50
Unprotected directions U_0	±1.73	±0.84	±1.32	±2.99	±0.49	±0.44

Table 21: Covariance controls, sample SD over three seeds (%). Means in Table 7. ICG: ICG-Drone-3c.

Base covariance	M2	Aircraft	Cars	CUB	DTS	VisioDECT	ICG
Class-specific isotropic	—	±0.63	±0.33	±0.10	±1.28	±1.62	±1.15
Shared Ledoit–Wolf	—	±0.32	±0.60	±0.21	±2.21	±1.99	±1.57
Class-specific Ledoit–Wolf	—	±0.42	±0.52	±0.20	±1.58	±1.42	±1.52
Class-specific isotropic	✓	±0.05	±0.28	±0.76	±1.34	±1.04	±0.50
Shared Ledoit–Wolf	✓	±0.30	±0.43	±0.17	±2.66	±1.62	±3.05
Class-specific Ledoit–Wolf	✓	±0.52	±0.87	±0.54	±1.94	±1.23	±3.56