

# GEOMETRIC CONFORMAL OUTLIER SYNTHESIS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Deep neural networks for image classification often exhibit overconfidence on out-of-distribution (OOD) samples. To address this, we introduce Geometric Conformal Outlier Synthesis (GCOS), a training-time regularization framework aimed at improving OOD robustness during inference. GCOS addresses a limitation of prior synthesis methods by generating virtual outliers in the hidden feature space that respect the learned manifold structure of in-distribution (ID) data. The synthesis proceeds in two stages: (i) PCA on training features identifies geometrically-informed, off-manifold directions; (ii) a Conformally-Inspired Shell, defined by the empirical quantiles of a nonconformity score from a calibration set, adaptively controls the synthesis magnitude to produce boundary samples. The shell ensures that generated outliers are neither trivially detectable nor indistinguishable from in-distribution data, facilitating smoother learning of robust features. This is combined with a contrastive regularization objective that promotes separability of ID and OOD samples in a chosen score space, such as Mahalanobis or energy-based. Experiments show that GCOS improves OOD detection relative to baselines using the standard energy-based inference approach. As an exploratory extension, the framework naturally transitions to conformal OOD inference, which translates uncertainty scores into statistically valid p-values and enables thresholds with formal error guarantees, providing a pathway toward more predictable and reliable OOD detection.

## 1 INTRODUCTION

Test-time out-of-distribution (OOD) detection is an important component of any machine learning model. For classification tasks, the ability to identify an input as an outlier rather than as an instance of one of the training set classes is crucial for robustly handling novel or unexpected inputs that inevitably arise outside the curated training environment. Classification models trained without explicit regularization for outliers can develop overconfident decision boundaries, where instances far from all classes may still be confidently assigned to the nearest class (Du et al., 2022; Vernekay et al., 2019).

To illustrate how uncertainty-aware objectives shape representations in practice, we begin by discussing Virtual Outlier Synthesis (VOS) (Du et al., 2022), which leads to more robust decision boundaries. VOS addresses the challenge of data imbalance in anomaly detection, particularly when anomalous examples are scarce or entirely absent during training. The core idea behind VOS is to artificially generate synthetic outliers, enabling the model to better differentiate between in-distribution (ID) and OOD patterns. Unlike traditional data augmentation, which typically relies on perturbations of existing samples or generation of OOD data in the data space (i.e. the space of input images), VOS models and samples from regions in the feature space that are statistically unlikely under the distribution of ID data. This approach helps simulate plausible yet diverse abnormal examples.

Among existing methods, our work is similar to VOS in its emphasis on OOD-aware representation learning. However, a fundamental limitation of the VOS-based approach is the assumption that outliers can be modeled as samples drawn from a simple distribution (e.g., Gaussian) outside the support of the normal data (Siddiqi et al., 2023). This simplification may fail to capture the complex and often non-Gaussian nature of real-world anomalies, which can exhibit structured or domain-specific characteristics (Tao et al., 2023b). Consequently, the synthesized outliers may not accurately reflect the true anomaly space, potentially leading to poor generalization. The effectiveness of VOS

054 also depends on the geometry of the learned feature space. If the latent space does not adequately  
 055 separate normal and abnormal regions, synthetic outliers may overlap with normal samples or fall  
 056 into irrelevant regions, reducing their utility.

057 One of our key contributions is to replace the reliance on pre-defined parametric distributions (such  
 058 as class-conditional Gaussians) for outlier modeling. Instead, we propose a geometric synthesis  
 059 framework that generates outliers by probing low-variance subspaces of the learned feature mani-  
 060 fold, as identified by Principal Component Analysis in the feature space. To control the difficulty  
 061 of the synthetic outliers, we aim to generate meaningful examples for regularization that are neither  
 062 too close to real data embeddings (making them inseparable), nor trivially obvious outliers that are  
 063 too easy for the model to identify as OOD.

064 Another key limitation in existing OOD literature is the heavy focus on far-OOD benchmarks, where  
 065 test data is semantically distant from the training domain (e.g., evaluating an animal classifier on  
 066 industrial objects). While useful, these benchmarks overlook what we argue is the more critical  
 067 challenge for robust AI: near-OOD detection, where models must separate fine-grained categories  
 068 within the same super-class (e.g., unseen animal breeds). Such cases are more likely to trigger  
 069 catastrophic failures in practice due to high feature-space similarity. Accordingly, alongside standard  
 070 far-OOD benchmarks, our work places strong emphasis on evaluating methods against near-OOD  
 071 tasks, where samples come from the same domain as in-distribution classes but remain unseen during  
 072 training.

073 Conformal prediction (CP) Vovk et al. addresses this challenge from a complementary perspective.  
 074 CP provides a model-agnostic framework for quantifying uncertainty with formal statistical guar-  
 075 antees. It can be applied at inference time to any base predictor (Shafer & Vovk, 2008; Saunders  
 076 et al., 1999; Vovk et al., 2017). Specifically, CP outputs a set or interval that contains the true la-  
 077 bel with guaranteed coverage. While OOD detection and CP originate from distinct lines of work,  
 078 both ultimately aim to assess when a model’s predictions should not be trusted. This motivates our  
 079 exploration of whether CP can, **during training**, strengthen a model’s robustness and its ability  
 080 to internally flag outliers. We further investigate whether combining CP **during inference** offers  
 081 a promising approach for more predictable and reliable OOD detection by translating uncertainty  
 082 scores into statistically valid p-values with formal error guarantees.

083 The contributions of this paper can be summarized as follows:

- 084 1. Introduce a novel geometrically-driven outlier synthesis approach based on a conformal  
 085 heuristic.
- 086 2. Propose a loss function incorporating nonconformity scores.
- 087 3. Explore an alternative to energy-based inference as future work: conformal hypothesis  
 088 testing for OOD detection.

089 Introducing our conformal heuristic into loss regularization via outlier synthesis (contributions 1 and  
 090 2) improves the model’s ability to filter out OOD points using energy-based inference. Contribution  
 091 3 further aims to elevate machine learning models from merely accurate tools to provably reliable  
 092 systems, by providing a formal statistical framework that governs their behavior under uncertainty.  
 093 We discuss how this approach can be applied and, in doing so, open a new avenue for future research,  
 094 with promising preliminary results on the Colored MNIST and Retinopathy datasets.

## 096 2 CONTEXT

### 097 2.1 OUTLIER EXPOSURE

098 VOS regularization is based on the principle of Outlier Exposure (Hendrycks et al., 2018). During  
 099 training, a set of virtual outliers,  $\mathcal{Z}_{ood}$ , are generated (e.g., by sampling from the tails of class-  
 100 conditional Gaussian distributions fitted to in-distribution features). The OOD detection capability  
 101 is learned by forcing a distinction between the energy scores (LeCun et al., 2007; Ngiam et al., 2011;  
 102 Grathwohl et al., 2019) of ID and virtual outlier samples.

$$103 E(\mathbf{z}; \theta) = -\log \sum_{k=1}^K \exp(f_k(\mathbf{z}; \theta)) \quad (1)$$

Following the formulation in the VOS paper, the energy score for a feature vector  $\mathbf{z}$  with model parameters  $\theta$  is defined as the negative log partition function (1) where  $f_k(\mathbf{z}; \theta)$  is the  $k$ -th logit from the classifier for  $K$  classes. A low energy score corresponds to a confident, in-distribution-like prediction. The model then learns a function  $\phi$  that maps this energy score to a new logit, which determines the probability of the input being in-distribution. The uncertainty regularization loss,  $\mathcal{L}_{\text{uncertainty}}$ , is the binary cross-entropy loss for this task, where ID samples have a target label of 1 and virtual outliers have a target label of 0.

$$\mathcal{L}_{\text{uncertainty}} = \mathbb{E}_{\mathbf{z}_{\text{in}} \sim \mathcal{Z}_{\text{in}}} \left[ -\log \frac{1}{1 + \exp(-\phi(E(\mathbf{z}_{\text{in}}; \theta)))} \right] + \mathbb{E}_{\mathbf{z}_{\text{ood}} \sim \mathcal{Z}_{\text{ood}}} \left[ -\log \frac{\exp(-\phi(E(\mathbf{z}_{\text{ood}}; \theta)))}{1 + \exp(-\phi(E(\mathbf{z}_{\text{ood}}; \theta)))} \right] \quad (2)$$

Here,  $\mathcal{Z}_{\text{in}}$  is the distribution of features of the in-domain data and  $\mathcal{Z}_{\text{ood}}$  is the distribution of synthesized virtual outliers. The first term pushes the probability of ID samples being recognized as ID towards 1, while the second term pushes the probability of virtual outliers being recognized as ID towards 0.

## 2.2 ENERGY-BASED INFERENCE

A common and widely adopted approach for OOD detection is the energy-based OOD detection score (Liu et al., 2020), which uses the value in equation (1) directly as the outlier signal, bypassing any auxiliary heads. Higher energy indicates lower model confidence and thus a greater likelihood of OOD. Conceptually, this plays the same role as a probability score in binary classification, providing a per-sample measure of “outlierness.” To ensure comparability with prior work and consistency with the VOS evaluation protocol, we report results in Section 5 using the general energy score as the OOD detection score.

While energy-based scores provide a simple heuristic, their thresholds are tuned on validation data and lack formal guarantees on novel inputs. In Section 6, we explore an alternative approach that converts raw scores into statistically valid p-values, enabling thresholds with formally controlled error rates.

## 2.3 CONFORMAL PREDICTION

Conformal prediction provides a distribution-free framework for constructing prediction sets with guaranteed coverage under the assumption that the data points are exchangeable. Let  $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$  denote a calibration set drawn from the same distribution as the test points. For a new input  $x_{n+1}$ , a nonconformity score function  $\mathcal{S}(x, \hat{y})$  measures how unusual a candidate label  $\hat{y}$  is relative to the calibration data. The conformal prediction set for  $x_{n+1}$  is then defined as  $C(x_{n+1}) = \{\hat{y} : \mathcal{S}(x_{n+1}, \hat{y}) \leq Q_{1-\alpha}\}$ , where  $Q_{1-\alpha}$  is the  $(1 - \alpha)$ -quantile of the nonconformity scores on the calibration set. Under the exchangeability assumption, this guarantees that the prediction set covers the true label with probability at least  $1 - \alpha$ . The choice of nonconformity score is flexible and can be adapted to different types of models and tasks. For classification, it is often derived from model probabilities or logits; for regression, it can be based on residuals or prediction intervals.

Conformal prediction in the context of OOD detection can be formulated as hypothesis testing, where the nonconformity score functions as a test statistic. For a candidate test point, CP computes a p-value by comparing its score to the calibration scores from the training data, which is analogous to accepting or rejecting a null hypothesis (Vovk et al., 2014; Barber et al., 2023; Tibshirani et al., 2019). Similar to classical hypothesis testing, controlling Type I error (coverage guarantee) often comes at the cost of low power, i.e., the ability to correctly reject false hypotheses. Some OOD points may not appear “strange” enough to be flagged, particularly if the nonconformity score computed from the model’s internal representations is weak. The nonconformity score derived from the model’s final prediction  $\hat{y}$  may not fully capture uncertainty. A single prediction must simultaneously serve as an effective uncertainty measure and as an accurate estimate of the true label or class probability, which can be a limiting constraint. In contrast, deriving the score from logits or the model’s hidden states provides more flexibility and often produces a more informative nonconformity measure.

## 3 CONFORMAL SHELL SYNTHESIS

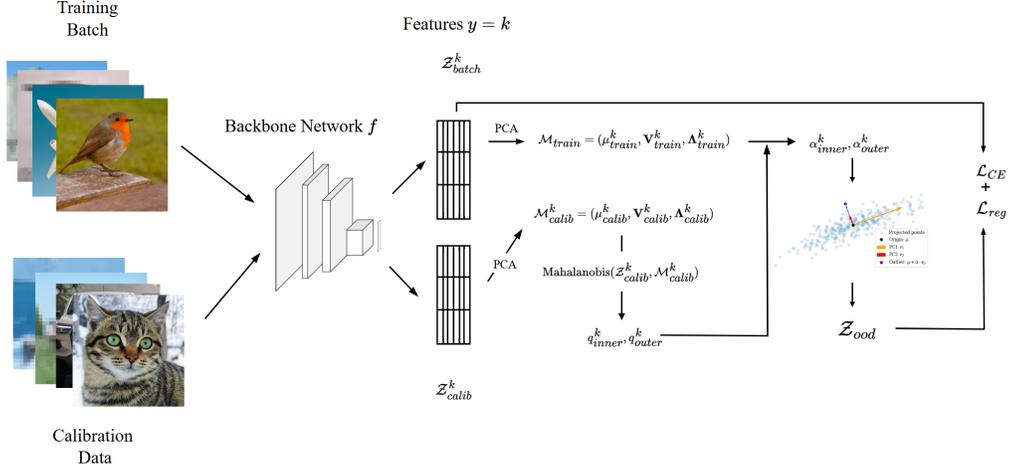


Figure 1: **GCOS Training Procedure Schematic.** Illustration of the data flow in our online synthesis and regularization method. Epoch-level calibration on  $\mathcal{D}_{calib}$  produces PCA models  $\mathcal{M}_{calib}$  and Mahalanobis quantiles  $q$ . During batch-level training, features are used to update a queue that generates proposer PCA models  $\mathcal{M}_{train}$  and identifies off-manifold directions  $v$ . Outliers  $\mathbf{z}_{ood}$  are synthesized to match the target quantiles  $q$ , as evaluated by  $\mathcal{M}_{calib}$ . The final regularization loss,  $\mathcal{L}_{reg}$ , is a contrastive objective computed on the energy scores of in-distribution batch features and the synthesized outliers  $\mathbf{z}_{ood}$  and added to cross-entropy loss,  $\mathcal{L}_{CE}$ , from the main classification task.

We introduce a geometrically-driven synthesis of virtual outliers via Principal Component Analysis (Hotelling, 1933). PCA is applied to the hidden representations  $\mathcal{Z}_{in}$  obtained from the backbone network immediately preceding the classification layer. Once the eigenvectors and eigenvalues,  $\mathbf{V}_{train}$  and  $\mathbf{\Lambda}_{train}$ , are computed, they are split into “large” and “small” components. The first  $K$  principal components that account for at least  $\eta$  (e.g. 90%) of the variance are marked as “large,” while the remaining  $\mathbf{V}_{small, train}^k$  are considered “small.” These small PCs are the main building blocks for outlier synthesis: they correspond to directions with the smallest variability in the data, meaning that moving along them produces unlikely points while still remaining near the data centroid  $\mu$ . Similar to VOS, we use a rolling per-class buffer to compute the covariance matrix  $\Sigma$  and the mean.

Let  $v$  be the direction of a chosen small PC (see the scatter plot in figure 1). We can then generate an OOD feature as  $\mathbf{z}_{ood}(\alpha) = \mu + \alpha v$ , where  $\alpha$  is a scalar. The problem of outlier synthesis is thus reduced to sampling a single scalar. Choosing  $\alpha$  requires care: very small values produce synthetic points too close to real features to be separable, while very large values produce trivially easy outliers that are not useful for generalization.

To address this, we apply a conformal prediction-inspired heuristic to determine appropriate values for  $\alpha$ . This heuristic evaluates how “strange” a point is relative to a calibration set set aside at the beginning of training, ensuring that **generated outliers avoid being overly simple, unrealistically extreme, or, at the other extreme, indistinguishably close to ID data**. For a given choice of the score (uncertainty) function  $\mathcal{S}$ , we compute the quantiles of the scores of the features derived from the calibration data. These quantiles then serve as thresholds to define a conformal shell  $[\alpha_{inner}, \alpha_{outer}]$ . Intuitively, the conformal shell defines a range of deviation magnitudes,  $\alpha$ , that produce outliers of a specific “strangeness.” The boundaries of this shell,  $\alpha_{inner}$  and  $\alpha_{outer}$ , are determined by the 95th and 99th percentile nonconformity score thresholds ( $q_{95}$  and  $q_{99}$  respectively), which are derived from the calibration set. Formally,  $\alpha_{inner}$  is the minimum deviation required such that the nonconformity score of the synthesized point equals the lower threshold, i.e.,  $\mathcal{S}(\mathbf{z}_{ood}(\alpha_{inner})) = q_{95}$ . This establishes  $\alpha_{inner}$  as the precise boundary where, for any infinitesimal  $\epsilon > 0$ , the point  $\mathbf{z}_{ood}(\alpha_{inner} - \epsilon)$  would still be considered in-distribution by this threshold, while  $\mathbf{z}_{ood}(\alpha_{inner} + \epsilon)$  would not. Similarly,  $\alpha_{outer}$  is determined using the higher score threshold  $q_{99}$ . New

216 OOD features,  $\mathbf{z}_{ood}(\alpha)$ , are then generated by sampling  $\alpha$  uniformly from this “hard-negative” shell:  
 217  $\alpha \sim \mathcal{U}[\alpha_{inner}, \alpha_{outer}]$ .

218 When choosing the direction  $v$ , we consider two options. Having identified the “small” eigenvectors  
 219  $\mathbf{V}_{small, train}^k$ , we either average them to obtain a single generation direction

$$220 \quad v = \frac{1}{|\mathbf{V}_{small, train}^k|} \sum_{v_i \in \mathbf{V}_{small, train}^k} v_i \quad (3)$$

221 or we apply the outlier synthesis procedure separately for each  $v_i \in \mathbf{V}_{small, train}^k$  or its subset. We refer  
 222 to the former approach as the *average direction* method and the latter as the *per direction* method.  
 223 Refer to Appendix H for the ablation study. Additionally, moving in the opposite direction of  $v$  is  
 224 also possible, which we implement by randomly selecting the sign of the direction.

225 Since the model is updated continuously and the calibration data participates in this feedback loop,  
 226 the strict exchangeability assumption of traditional conformal prediction is violated. To address this,  
 227 we maintain two calibration sets: one for online calibration and outlier synthesis during training, and  
 228 another for final calibration, which is later used for the conformal hypothesis testing during OOD  
 229 evaluation in Section 6.

230 If the score function is monotonic in  $\alpha$  (i.e., uncertainty increases as we move farther from the  
 231 centroid),  $\alpha$  can be determined through single-variable optimization. Given the binary nature of  
 232 OOD detection, where the goal is simply to determine if the score crosses a target quantile, a binary  
 233 search can also be employed, as described in Algorithm 1. The Mahalanobis distance (Mahalanobis,  
 234 2018) in (4) satisfies this criterion.

$$235 \quad \mathcal{S}_{Mahal}(z, \mu, \{\lambda_i\}, \{v_i\}) = \sum_{i=1}^D \frac{((z - \mu)^T v_i)^2}{\lambda_i + \epsilon} \quad (4)$$

236 To directly optimize the feature space for OOD separability, we propose a novel geometric regular-  
 237 ization loss,  $\mathcal{L}_{reg}$  (contribution 2). This loss is formulated as a contrastive objective that minimizes  
 238 the nonconformity scores of in-distribution samples while maximizing the scores of synthesized  
 239 OOD samples. Let  $\mathcal{S}_{\mathcal{L}}(\mathbf{z} | \mathcal{M}_k)$  denote the nonconformity score of a feature vector  $\mathbf{z}$  with respect  
 240 to the calibration model  $\mathcal{M}_k$  for class  $k$ . The loss is defined as (5) where  $y_{id}$  is the true class label  
 241 for the ID virtual sample  $\mathbf{z}_{id}$ .

$$242 \quad \mathcal{L}_{reg} = \mathbb{E}_{\substack{\mathbf{z}_{id} \sim \mathcal{Z}_{id} \\ \mathbf{z}_{ood} \sim \mathcal{Z}_{ood}}} \left[ \max(0, \mathcal{S}_{\mathcal{L}}(\mathbf{z}_{id} | \mathcal{M}_{y_{id}}) - \min_k \mathcal{S}_{\mathcal{L}}(\mathbf{z}_{ood} | \mathcal{M}_k) + m) \right] \quad (5)$$

243 In equation (5), the positive term,  $\mathcal{S}_{\mathcal{L}}(\mathbf{z}_{id} | \mathcal{M}_{y_{id}})$ , represents the score of an in-distribution sam-  
 244 ple against its own class’s reference model, which we aim to minimize. The negative term,  
 245  $\min_k \mathcal{S}_{\mathcal{L}}(\mathbf{z}_{ood} | \mathcal{M}_k)$ , is the score of a synthesized outlier against its best-fitting (i.e., minimum  
 246 score) class model, which we aim to maximize. The nonconformity score function  $\mathcal{S}_{\mathcal{L}}(\cdot)$  can be  
 247 either the Mahalanobis distance, in which case  $\mathcal{M}_k$  is the class-specific PCA model  $(\mu_k, \mathbf{V}_k, \mathbf{\Lambda}_k)$ ,  
 248 or the energy strangeness score (6), where  $\mathcal{M}_k$  is implicitly the shared classifier head. The adaptive  
 249 margin  $m$  is set to the difference between the 95th and 50th percentiles of the positive scores in the  
 250 batch, ensuring a dynamically scaled separation between ID and OOD score distributions. Refer to  
 251 Appendix B for adaptive margin calculation.

252 Building on the general framework, we next present its instantiation in our proposed GCOS con-  
 253 figuration. First, the set of virtual outliers  $\mathcal{Z}_{ood}$  is generated using our geometric synthesis method  
 254 where the conformal shell boundaries are determined by quantiles of the Mahalanobis distance score  
 255 function. Second, for the loss in equation (5), we define the nonconformity score  $\mathcal{S}_{\mathcal{L}}(\cdot)$  to be the  
 256 Energy Strangeness Score:

$$257 \quad \mathcal{S}_{\mathcal{L}}(\mathbf{z}) = \log \sum_{i=1}^K w_i \cdot \exp(h_{\phi}(\mathbf{z})_i) \quad (6)$$

258 Since this score is class-agnostic, the  $\min_k \mathcal{S}_{\mathcal{L}}(\mathbf{z}_{ood} | \mathcal{M}_k)$  term in the loss simplifies to  $\mathcal{S}_{\mathcal{L}}(\mathbf{z}_{ood})$ .  
 259 This hybrid approach leverages the geometric properties of the feature space to propose effective  
 outlier locations, while directly optimizing the energy landscape which has shown to be a robust

indicator for OOD detection (Du et al., 2022). We also evaluated our outlier synthesis approach using the regularization loss from VOS, and examined the case where both synthesis and loss regularization scores are given by the Mahalanobis distance (i.e.,  $\mathcal{S}_{\mathcal{L}}(\cdot) = \mathcal{S}(\cdot) = \mathcal{S}_{Mahal}(\cdot)$ ), detailed in Appendix F.

#### 4 EXPERIMENTAL SETUP

Many previous works have primarily focused on far-outliers, i.e., they evaluated models on OOD datasets that are unrelated to the training data (e.g., CIFAR-10 for ID and Textures for OOD). We argue that such scenarios are not realistic in practice. Therefore, we propose evaluating on near-OOD data - images that are not included in the training set but originate from the same field or domain.

We consider four problems. The first dataset, *Colored MNIST*, assigns each digit in MNIST (LeCun, 1998) a fixed RGB color. The in-distribution split uses consistent digit-color associations, while the out-of-distribution split permutes these associations so that each digit appears with a different color. This construction yields an ID training set, an ID test set, and an OOD evaluation set, enabling controlled assessment of generalization under distribution shifts. The second dataset, *MVTec*, is based on images of defective industrial objects. From the original MVTEc dataset (Bergmann et al., 2019), we construct the in-distribution training and test sets by splitting the “good”, non-defective, samples of each class, while the OOD evaluation set consists of all defective samples from the same classes. This setup enables the evaluation of models on both normal variations and diverse anomalous patterns. The third dataset, based on *Stanford Dogs*, involves classifying multiple dog breeds (Khosla et al., 2011). In this case, the OOD data comprises breeds that are similar, but not identical, to those in the training set (e.g., golden retriever in the training set and labrador retriever as an OOD breed). Finally, the *Retinopathy* dataset (Karthik & Dane, 2019) consists of retinal fundus images curated for the classification of eye disorders. The in-distribution subset includes five diabetic retinopathy (DR) severity levels: No DR, Mild, Moderate, Severe, and Proliferative. To evaluate OOD detection, the dataset also provides fundus images with other ocular pathologies (Schwartz, 2020), such as glaucoma, age-related macular degeneration (AMD), and pathological myopia.

### 5 RESULTS

#### 5.1 BASELINE METHODS

We evaluate the trained models on the OOD detection task, where unseen ID test samples and OOD samples are mixed together, and the model is required to distinguish between them. We compare our approach against a model trained with Gaussian outlier synthesis with VOS regularization loss with an energy-based inference method, and a baseline model without regularization under energy-based inference. Additionally, we selected a representative set of baselines spanning different OOD detec-

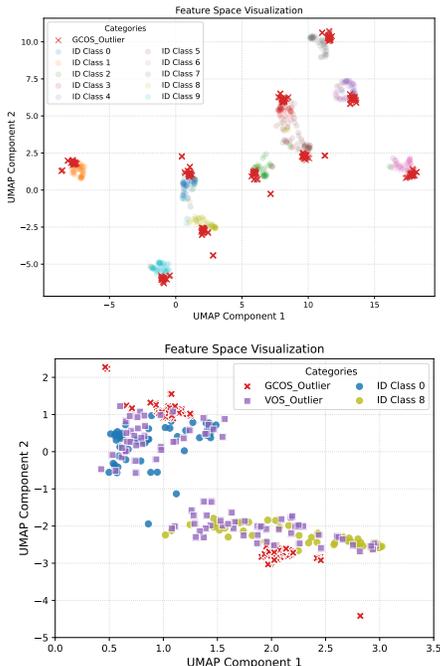


Figure 2: **UMAP Projection of Learned Features.** Top: the overall feature space, showing that classes form varying shapes and are largely well separated. Bottom: a zoomed-in view highlighting the distribution of GCOS outliers in off-manifold regions and VOS outliers near cluster edges for two clusters. The panels illustrate how GCOS generates points in challenging regions beyond the main clusters, while VOS outliers remain close to class boundaries.

Method	C-MNIST			Stanford Dogs			MVTec			Retinopathy			Avg AUROC
	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	
No Reg.	85.60	99.47	32.50	88.35	97.72	80.00	94.86	98.91	27.69	69.74	98.43	81.50	84.64
VOS	94.71	99.87	18.50	99.25	99.85	5.00	80.37	95.83	70.77	70.52	98.56	80.00	86.21
MSP	92.10	89.88	21.12	56.15	13.81	96.96	56.58	34.72	82.20	67.02	94.09	90.52	67.96
MaxLogit	88.88	85.68	27.59	58.42	14.27	93.69	55.30	39.20	90.24	77.21	96.41	81.39	69.95
ReAct	89.41	86.60	24.69	57.77	14.08	94.04	57.31	36.74	87.09	77.37	96.43	75.84	70.47
GradNorm	93.62	94.09	36.74	57.13	91.40	94.00	55.62	76.35	99.08	64.87	16.14	86.43	67.81
KL-Matching	91.96	89.70	21.41	56.39	13.82	93.11	56.29	34.55	83.62	69.00	94.48	84.28	68.41
ViM	96.54	96.80	19.26	40.50	87.23	99.00	60.80	77.54	95.09	77.08	36.95	76.82	68.73
NPOS	2.80	64.61	100.00	61.45	87.17	95.00	91.34	98.25	52.31	24.62	82.05	99.65	45.05
GCOS + $\mathcal{L}_{reg}$ (Ours)	<b>99.50</b>	<b>99.99</b>	<b>1.00</b>	<b>99.55</b>	<b>99.91</b>	<b>0.00</b>	<b>95.61</b>	<b>99.08</b>	<b>23.08</b>	<b>79.23</b>	<b>99.16</b>	<b>73.00</b>	<b>93.47</b>

Table 1: Model comparison on OOD detection for energy-based inference.

tion paradigms. Among classical scoring methods, we utilize MSP (Hendrycks & Gimpel, 2016), which relies on the maximum softmax probability as a direct confidence proxy. Similarly, MaxLogit (Hendrycks et al., 2019a) bypasses softmax normalization entirely, demonstrating that unnormalized logit magnitudes often carry more discriminative information regarding uncertainty than probability distributions. Moving to feature-space regularization, we evaluate ReAct (Sun et al., 2021), which truncates activations in the penultimate layer at a fixed threshold to suppress the abnormally high feature responses often triggered by OOD data. We also include ViM (Wang et al., 2022), which explicitly models the ID manifold by combining standard logits with a residual score derived from the projection of features onto the null space of the principal components. In the gradient-based category, we use GradNorm (Huang et al., 2021), leveraging the observation that the vector norm of gradients backpropagated from the output layer is consistently higher for in-distribution inputs. We further represent distribution matching techniques with KL-Matching (Hendrycks et al., 2019b), which detects anomalies by minimizing the Kullback-Leibler divergence between a sample’s prediction and the mean class-conditional distributions. Finally, we compare against NPOS (Tao et al., 2023a), a non-parametric synthesis framework that generates challenging virtual outliers via rejection sampling based on K-nearest neighbor distances, thereby avoiding the structural limitations of parametric Gaussian assumptions.

For evaluation, we report the area under the ROC curve (AUROC; higher is better), the area under the precision-recall curve (AUPR; higher is better), and the false positive rate at 95% true positive rate (FPR95; lower is better). The first two metrics are threshold-agnostic, whereas the latter is defined for a threshold  $\gamma$  such that 95% of in-distribution images are correctly identified as ID.

## 5.2 DISCUSSION

As shown in Table 1, our approach - conformally driven outlier exposure with geometric regularization loss under energy-based inference - outperforms competing methods in terms of AUROC and AUPR. Moreover, it consistently achieves substantially lower FPR95 than VOS and no-regularization across all datasets (23.08% vs. 27.69% and 70.77%; 1% vs. 32.5% and 18.5%; < 0.01% vs. 80% and 5%; 73% vs. 81.50% and 80%). Interestingly, VOS reduces the false positive rate compared to no-regularization on the Colored MNIST dataset; however, the trend reverses on the MVTec dataset, where the model without outlier exposure performs better than VOS. Nevertheless, our approach demonstrates superior and consistent performance across both datasets.

To more comprehensively evaluate GCOS, we include seven additional OOD detection methods spanning classical scoring approaches, post-hoc regularization techniques, gradient-based methods, distribution matching, and non-parametric outlier synthesis. GCOS achieves the best performance across all evaluation scenarios, with an average AUROC of 93.47%. Notably, GCOS outperforms even strong recent methods such as ViM (average AUROC: 68.73%), ReAct (70.47%) and NPOS (45.05%), demonstrating substantial margins. The consistent superiority across diverse datasets suggests that geometry-aware outlier synthesis with conformal calibration provides fundamental advantages over both classical scoring methods that operate on fixed representations and synthesis methods that lack geometric guidance. The particularly large gaps on datasets with complex class distributions (such as Colored MNIST) indicate that adaptive per-class calibration is crucial when class geometries are heterogeneous, while the more modest improvements on MVTec suggest these datasets may have simpler decision boundaries where standard methods already perform near-optimally.

We decided to visualize the resulting feature space of our model using a UMAP projection (McInnes et al., 2018) in figure 2. We observe that our method separates most of the classes effectively. The shapes of the resulting “clusters” are not uniform; some appear more spherical while others are elongated. This may indicate that class-conditional Gaussian tail sampling does not fully capture the complex, low-dimensional manifold structure of learned feature representations, and that methods leveraging the geometry of the feature space are necessary. Another interesting property of the obtained representations is the behavior of GCOS outliers: they tend to cluster on the opposite sides of two nearby classes. This can be clearly seen in the bottom panel of figure 2, where VOS outliers are distributed around the cluster, including regions between classes where the decision boundary lies, whereas GCOS outlier synthesis behaves as intended. Rather than attempting to affect the main classification boundary, it instead pushes the decision boundary beyond both the cluster and the classification region closer to the data. As noted before if an outlier lies far beyond both the classification boundary between clusters and one of the clusters, models trained without OOD regularization will be confidently misclassifying such points as belonging to the adjacent class. By focusing on off-manifold directions, our method generates points in these challenging regions, thereby “flanking” the decision boundary more tightly around the data clusters.

To summarize, GCOS (contributions 1 + 2) improve OOD detection by generating outliers that respect the geometry of in-distribution data, consistently outperforming baseline methods across multiple metrics. Feature-space visualizations show that GCOS tightly enclose the decision boundary around data clusters, reducing overconfidence on challenging outliers. These results highlight the value of geometry-aware outlier synthesis for robust neural networks and point toward future integration with formal conformal guarantees.

## 6 MOVING TOWARDS OOD DETECTION WITH STATISTICAL GUARANTEES

At inference, the goal of OOD detection is to decide whether an unseen sample comes from the in-distribution or not. Traditional approaches assign a score and compare it against a threshold tuned on validation data. While often effective, such heuristic thresholding provides no formal guarantee of performance on unseen data.

**Inference via Conformal Hypothesis Testing.** OOD detection using conformal inference provides a principled way to assess how unusual a new data point is, translating raw scores into statistically meaningful p-values and enabling formal hypothesis testing. In contrast, simple energy-based thresholding relies on heuristics - a cutoff tuned on past data that may fail when conditions change. By offering thresholds with rigorous error guarantees, conformal inference ensures that decisions remain reliable even on unseen data.

Knowing the uncertainty measure in advance may help guide the model’s hidden representations to facilitate inference-time conformal prediction, either during training or post-training. Therefore, our work integrates procedures from conformal inference into the learning process of deep learning models, illustrated on a computer vision classification task with OOD detection. We believe that this idea could represent a promising new direction in research. The manual choice of the non-conformity score function is reminiscent of handcrafted filters in the early days of computer vision (Szeliski, 2022; LeCun et al., 2015), which were later largely replaced by learned convolutional layers. In a similar spirit, recent methods have explored ways to integrate uncertainty-aware mechanisms directly into the learning process (Kendall & Gal, 2017; Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Kuleshov et al., 2018). After training concludes, we run a single calibration step on a separate held-out calibration set to prepare the model for inference. This final calibration is distinct from the online calibration performed during training, which relies on the network’s temporary state at the start of each epoch to build an adaptive heuristic for guiding outlier synthesis. Further details are provided in Appendix C.

**Mixed Results for Contribution 3.** During inference, we attempt to apply a formal hypothesis testing framework based on conformal prediction to determine whether an unseen sample ID or OOD. This procedure replaces the energy-based inference in (1), where OOD detection relies on a computing energy score from the logits of the main classification head.

Given a test sample with feature vector  $\mathbf{z}^{\text{test}}$ , we first compute its nonconformity score,  $S(\mathbf{z}^{\text{test}})$ , using the score function obtained during final calibration (e.g., Energy or Mahalanobis). This score is then used to derive a p-value, which quantifies the likelihood that  $\mathbf{z}^{\text{test}}$  belongs to the ID distribution under the null hypothesis.

To account for the multi-class nature of the ID data, we calculate a p-value for the test sample with respect to each class’s reference distribution. For class  $k$ , this p-value is the fraction of samples in the final calibration set  $\mathcal{S}_{\text{calib},k}^{\text{final}}$  whose scores are greater than or equal to the test sample’s score:

$$p_k(\mathbf{z}^{\text{test}}) = \frac{1 + |\{\mathbf{s} \in \mathcal{S}_{\text{calib},k}^{\text{final}} \mid \mathbf{s} \geq S(\mathbf{z}^{\text{test}} | \mathcal{M}_k)\}|}{1 + |\mathcal{S}_{\text{calib},k}^{\text{final}}|} \quad (7)$$

The overall p-value for the test sample is the maximum across all classes,  $p_{\text{final}}(\mathbf{z}^{\text{test}}) = \max_{k \in \{1, \dots, K\}} p_k(\mathbf{z}^{\text{test}})$ , reflecting its compatibility with the most likely ID class. A sample is classified as OOD if this final p-value is less than the significance level, and as ID otherwise.

	MVTec			Colored MNIST			Stanford Dogs			Retinopathy		
	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
GCOS + $\mathcal{L}_{\text{uncertainty}}$	50.92	83.68	100.00	98.75	99.97	2.00	50.00	83.33	100.00	67.60	98.37	77.00
GCOS + $\mathcal{L}_{\text{reg}}$ (Mahalanobis)	63.67	88.07	100.00	98.93	99.98	1.50	50.00	83.33	100.00	37.58	96.27	97.00
GCOS + $\mathcal{L}_{\text{reg}}$ (Energy)	70.62	88.50	72.31	76.00	97.85	34.00	2.40	65.43	100.00	31.11	95.30	98.00

Table 2: Results using formal hypothesis testing (7) as OOD classification head.

We can observe from Table 2 that the use of conformal hypothesis testing as an OOD detection head, instead of energy-based inference discussed in section 2.2, yields mixed results. In some cases, this technique collapses to a nearly random classifier. On the simple Colored MNIST dataset, it achieves competitive scores, with an outstanding 1.5% FPR95 when applying geometric score-based regularization loss using the Mahalanobis distance. Interestingly, geometric loss with energy shows performance improvement on the MVTec dataset, reaching an AUROC of approximately 70.62%. For the Retinopathy dataset, our outlier synthesis combined with the logistic regression uncertainty loss from (2) (GCOS +  $\mathcal{L}_{\text{uncertainty}}$ ) achieves comparable AUROC performance to VOS and the model without regularization evaluated under energy-based inference and reported in Table 1, while also improving FPR95.

Despite mixed performance, these results come with statistical guarantees. At present, the technique demonstrates promising potential and represents an emerging area for further research. In the main results section, we demonstrated that incorporating conformal-inspired techniques improve the model’s accuracy on OOD detection, representing a step toward a more advanced framework. If the community succeeds in making the procedure more consistent and accurate while retaining conformal guarantees, it could establish a framework in which the predictive model is inherently equipped with uncertainty quantification - a property that is essential in domains such as medicine, where the need for reliable uncertainty estimation is known in advance, even before training the model.

## REFERENCES

- Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL <https://arxiv.org/abs/2107.07511>.
- Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.
- Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.
- Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.
- Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

- 486 Andrija Djuricic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation  
487 shaping for out-of-distribution detection, 2023. URL [https://arxiv.org/abs/2209.](https://arxiv.org/abs/2209.09858)  
488 09858.
- 489  
490 Lars Doorenbos, Raphael Sznitman, and Pablo Márquez-Neila. Non-linear outlier synthesis for  
491 out-of-distribution detection. *arXiv preprint arXiv:2411.13619*, 2024.
- 492 Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don’t know by  
493 virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.
- 494  
495 Xuefeng Du, Yiyu Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with  
496 diffusion models. *Advances in Neural Information Processing Systems*, 36:60878–60901, 2023.
- 497 Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help  
498 out-of-distribution detection? *arXiv preprint arXiv:2402.03502*, 2024.
- 499  
500 Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model  
501 uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059.  
502 PMLR, 2016.
- 503 Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi,  
504 and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like  
505 one. *arXiv preprint arXiv:1912.03263*, 2019.
- 506  
507 Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution  
508 examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- 509  
510 Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier  
511 exposure. *arXiv preprint arXiv:1812.04606*, 2018.
- 512  
513 Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi,  
514 Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings.  
515 *arXiv preprint arXiv:1911.11132*, 2019a.
- 516  
517 Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learn-  
518 ing can improve model robustness and uncertainty. *Advances in neural information processing*  
519 *systems*, 32, 2019b.
- 520  
521 Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal*  
522 *of educational psychology*, 24(6):417, 1933.
- 523  
524 Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-  
525 of-distribution image without learning from out-of-distribution data. In *Proceedings of the*  
526 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 10951–10960, 2020.
- 527  
528 Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic  
529 space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,  
530 pp. 8710–8719, 2021.
- 531  
532 Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distribu-  
533 tional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.
- 534  
535 M. Karthik and S. Dane. Aptos 2019 blindness detection. [https://www.kaggle.com/](https://www.kaggle.com/competitions/aptos2019-blindness-detection)  
536 [competitions/aptos2019-blindness-detection](https://www.kaggle.com/competitions/aptos2019-blindness-detection), 2019. Accessed: 2025-08-30.
- 537  
538 Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer  
539 vision? *Advances in neural information processing systems*, 30, 2017.
- 540  
541 Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-  
542 grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual*  
543 *categorization (FGVC)*, volume 2, 2011.
- 544  
545 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.  
546 2009.

- 540 Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning  
541 using calibrated regression. In *International conference on machine learning*, pp. 2796–2804.  
542 PMLR, 2018.
- 543 Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive  
544 uncertainty estimation using deep ensembles. *Advances in neural information processing systems*,  
545 30, 2017.
- 547 Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.  
548
- 549 Yann LeCun, Sumit Chopra, M Ranzato, and F-J Huang. Energy-based models in document recog-  
550 nition and computer vision. In *Ninth International Conference on Document Analysis and Recog-  
551 nition (ICDAR 2007)*, volume 1, pp. 337–341. IEEE, 2007.
- 552 Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444,  
553 2015.
- 555 Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers  
556 for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.
- 557 Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting  
558 out-of-distribution samples and adversarial attacks. *Advances in neural information processing  
559 systems*, 31, 2018.
- 561 Hengzhuang Li and Teng Zhang. Outlier synthesis via hamiltonian monte carlo for out-of-  
562 distribution detection. *arXiv preprint arXiv:2501.16718*, 2025.
- 563 Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution  
564 image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- 566 Qilin Liao, Shuo Yang, Bo Zhao, Ping Luo, and Hengshuang Zhao. Bood: Boundary-based out-of-  
567 distribution data generation. *arXiv preprint arXiv:2508.00350*, 2025.
- 568 Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution dete-  
569 ction. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475,  
570 2020.
- 572 Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian  
573 Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018.
- 574 Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and  
575 projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- 577 Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling.  
578 In *International conference on machine learning*, pp. 15650–15665. PMLR, 2022.
- 579 Jiquan Ngiam, Zhenghao Chen, Pang W Koh, and Andrew Y Ng. Learning deep energy models.  
580 In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 1105–  
581 1112, 2011.
- 583 Paul Novello, Joseba Dalmau, and Léo Andeol. Out-of-distribution detection should use conformal  
584 prediction (and vice-versa?), 2024. URL <https://arxiv.org/abs/2403.11532>.
- 585 Drew Prinster, Samuel Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for  
586 any data distribution (and how to find them), 2024. URL [https://arxiv.org/abs/2405.  
587 06627](https://arxiv.org/abs/2405.06627).
- 589 Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler.  
590 Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference  
591 on computer vision and pattern recognition*, pp. 14318–14328, 2022.
- 592 Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram  
593 matrices. In *International conference on machine learning*, pp. 8491–8501. PMLR, 2020.

- 594 Craig Saunders, Alex Gammerman, and Volodya Vovk. Transduction with confidence and credibil-  
595 ity. 1999.
- 596
- 597 Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component anal-  
598 ysis. In *International conference on artificial neural networks*, pp. 583–588. Springer, 1997.
- 599
- 600 K. Schwartz. Ocular disease recognition (odir-5k). [https://www.kaggle.com/datasets/  
601 andrewmvd/ocular-disease-recognition-odir5k](https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k), 2020. Accessed: 2025-08-30.
- 602 Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning  
603 Research*, 9(3), 2008.
- 604 Shafaq Siddiqi, Faiza Qureshi, Stefanie Lindstaedt, and Roman Kern. Detecting outliers in non-iid  
605 data: A systematic literature review. *IEEE Access*, 11:70333–70352, 2023.
- 606
- 607 Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activa-  
608 tions. *Advances in neural information processing systems*, 34:144–157, 2021.
- 609
- 610 Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest  
611 neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.
- 612 Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.
- 613
- 614 Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive  
615 learning on distributionally shifted instances. *Advances in neural information processing systems*,  
616 33:11839–11852, 2020.
- 617 Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. *arXiv  
618 preprint arXiv:2303.02966*, 2023a.
- 619
- 620 Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis, 2023b.  
621 URL <https://arxiv.org/abs/2303.02966>.
- 622 Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal pre-  
623 diction under covariate shift. *Advances in neural information processing systems*, 32, 2019.
- 624
- 625 Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal  
626 prediction under covariate shift, 2020. URL <https://arxiv.org/abs/1904.06019>.
- 627 Sachin Vernekar, Ashish Gaurav, Taylor Denouden, Buu Phan, Vahdat Abdelzad, Rick Salay, and  
628 Krzysztof Czarnecki. Analysis of confident-classifiers for out-of-distribution detection. *arXiv  
629 preprint arXiv:1904.12220*, 2019.
- 630
- 631 Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*.  
632 Springer.
- 633
- 634 Vladimir Vovk, Ivan Petej, and Valentina Fedorova. From conformal to probabilistic prediction. In  
635 *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 221–  
636 230. Springer, 2014.
- 637
- 638 Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distribu-  
639 tions based on conformal prediction. In *Conformal and probabilistic prediction and applications*,  
640 pp. 82–102. PMLR, 2017.
- 641
- 642 Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-  
643 logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern  
644 recognition*, pp. 4921–4930, 2022.
- 645
- 646 Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment  
647 distributions for out-of-distribution detection. *Advances in neural information processing systems*,  
36:73274–73286, 2023.
- 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.

648 Puning Yang, Jian Liang, Jie Cao, and Ran He. Auto: Adaptive outlier optimization for online  
649 test-time ood detection. *arXiv preprint arXiv:2303.12267*, 2023.  
650

651 Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun:  
652 Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv  
653 preprint arXiv:1506.03365*, 2015.

654 Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10  
655 million image database for scene recognition. *IEEE transactions on pattern analysis and machine  
656 intelligence*, 40(6):1452–1464, 2017.  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701

## APPENDIX

## A RELATED WORK

**Out-of-distribution detection** has been widely studied across several related tasks, including anomaly detection, novelty detection, and open-set recognition. Yang et al. (2024) provide a unified framework that situates these tasks as special cases, reviewing classification-, density-, and distance-based approaches and highlighting challenges such as calibration and realistic outlier modeling.

A common strategy for improving OOD detection is to leverage auxiliary outlier data. Hendrycks et al. (2018) introduced Outlier Exposure, showing that training on auxiliary outliers improves detection of unseen anomalies. Wang et al. (2023) extend this idea by crafting a set of worst-case OOD distributions around auxiliary data, theoretically reducing the discrepancy to unseen OOD samples. Li & Zhang (2025) propose synthesizing virtual outliers directly from in-distribution data using Hamiltonian Monte Carlo, generating diverse OOD samples without requiring an external dataset. Ming et al. (2022) also focus on informative outlier selection, using posterior sampling to learn a compact decision boundary between ID and OOD data.

Tao et al. (2023a) introduced Non-Parametric Outlier Synthesis (NPOS), which relaxes the restrictive Gaussian assumption of prior works by employing a non-parametric framework to synthesize outliers via rejection sampling. Addressing the challenge of high-dimensional generation, Du et al. (2023) proposed Dream-OOD, utilizing diffusion models to generate photo-realistic outliers in pixel space by sampling from low-likelihood regions of a text-conditioned latent space. Similarly, Liao et al. (2025) developed BOOD, a framework that specifically targets the decision boundary by perturbing in-distribution features to cross into OOD regions before decoding them into images using diffusion models. Doorenbos et al. (2024) presented NCIS to further enhance synthetic outlier quality, operating directly within a diffusion model’s embedding space and employing a conditional volume-preserving network to model complex class manifolds. Finally, in the specific context of industrial anomaly detection, Roth et al. (2022) proposed PatchCore, which uses a memory bank of nominal patch-level features to achieve state-of-the-art performance on benchmarks like MVTEC AD without requiring outlier supervision.

Other methods focus on feature-space metrics or model activations. Lee et al. (2018) use Mahalanobis distance to define a confidence score, while Sun et al. (2022) employ non-parametric nearest-neighbor distances for flexible, assumption-free detection. Hsu et al. (2020) decompose confidence scoring and modify input preprocessing to improve detection without relying on OOD data. Sun et al. (2021) show that controlling internal activations can mitigate overconfidence on OOD samples. Several works address scalability and structured label spaces. Huang & Li (2021) propose MOS scoring, grouping labels to simplify decision boundaries and reduce false positives. Hendrycks & Gimpel (2016) establish the maximum softmax probability as a competitive baseline, and Du et al. (2024) demonstrate that unlabeled data can be leveraged to identify candidate outliers and train OOD classifiers with theoretical guarantees. Test-time adaptation has been explored as a dynamic approach to OOD detection. Yang et al. (2023) propose AUTO, which selectively updates the model using streaming pseudo-labeled in/out samples to improve robustness without requiring auxiliary OOD data.

**Conformal prediction** (Vovk et al.) has emerged as a practical and widely applicable framework for uncertainty quantification. Angelopoulos & Bates (2022) provide an accessible overview of the methodology, emphasizing its distribution-free validity guarantees, ease of use, and adaptability to diverse domains such as computer vision, natural language processing, and reinforcement learning. Tibshirani et al. (2020) extend the approach to handle covariate shift by introducing a weighted version of conformal prediction that incorporates likelihood ratios between training and test distributions, thereby enabling valid prediction intervals even when data distributions differ. The interplay between conformal prediction and OOD detection has been explored, but to a limited extent. Novello et al. (2024) show that conformal methods can provide conservative corrections to OOD evaluation metrics, while OOD scores can improve conformal prediction sets when used as nonconformity measures. Finally, the theoretical scope of conformal prediction has been expanded substantially, as Prinster et al. (2024) prove that validity guarantees hold under any data distribution, including sequential feedback settings, and propose tractable algorithms for scenarios such as black-box optimization and active learning.

## B MODEL TRAINING

Algorithm 3 summarizes the end-to-end training procedure of our proposed GCOS framework. The process consists of two phases embedded in the standard training loop: an epoch-level online calibration step and a batch-level synthesis and regularization step.

**Epoch-Level Online Calibration.** The purpose of this step is to establish a stable “Judge” model that guides synthesis throughout the epoch. At the beginning of each epoch (after a warm-up period  $E_{start}$ ), the model is evaluated on the held-out calibration set  $\mathcal{D}_{calib}$  to obtain feature representations  $\mathcal{Z}_{calib}$  for all in-distribution classes. For each class, a PCA model  $\mathcal{M}_{calib}^k$  is fit to its features, and the resulting Mahalanobis score distribution  $S_{calib}^{Mahal,k}$  is computed. From this distribution, the quantiles  $q_{inner}^k$  and  $q_{outer}^k$  are extracted and stored as synthesis boundaries for the upcoming training iterations. For correctness of the PCA results, the data should be normalized across all dimensions so that dimensions with larger magnitudes do not bias the resulting components. We opt for the linear version of PCA (Hotelling, 1933) rather than kernel PCA (Schölkopf et al., 1997) because we work in the feature space immediately before the final classification layer, where the data should already be linearly separable by the main classification head.

**Batch-Level Training.** Each batch begins with a standard forward pass and cross-entropy loss computation. The extracted features  $Z_{id}$  are stored in a running feature buffer  $\mathcal{Q}_{train}$ , which serves as a “Proposer” by maintaining an estimate of the recent feature distribution. When the buffer is filled and the warm-up phase is complete, the GCOS synthesis and regularization procedure is applied. A Proposer PCA model  $\mathcal{M}_{train}$  is constructed from  $\mathcal{Q}_{train}$ , and synthetic outliers are generated on a per-class basis. Specifically, for each class, we identify the low-variance principal components  $\mathbf{V}_{small,train}^k$ . Along these directions, the `FindBoundaryAlpha` subroutine determines the  $\alpha$  values that place a candidate feature on the pre-computed quantile shells  $q_{inner}^k$  and  $q_{outer}^k$ . Outliers  $\mathbf{z}_{ood}$  are then sampled from these shells as perturbations of the class mean  $\mu_{train}^k$ .

After the synthesis step, the regularization loss is computed. In GCOS-Hybrid (Mahalanobis score for synthesis, energy-based score for regularisation loss), this involves calculating energy-based strangeness scores for both the real ID features ( $S_{pos}$ ) and the synthesized outliers ( $S_{neg}$ ). An adaptive margin, determined from the batch statistics of  $S_{pos}$ , is used to construct the contrastive regularization loss  $\mathcal{L}_{reg}$ . The final training loss combines the cross-entropy objective with the weighted regularization term,  $\lambda\mathcal{L}_{reg}$ , which is minimized through backpropagation to update the model.

**Algorithm 1** FindBoundaryAlpha

---

**Require:** Start point  $\mu_{train}$ , direction  $\mathbf{v}_j$ , target score  $q_{target}$ , score function  $\mathcal{S}$ , search range  $\alpha_{max}$ , steps  $N_{steps}$

**Ensure:** Boundary deviation  $\alpha_{boundary}$

- 1:  $s_0 \leftarrow \mathcal{S}(\mu_{train})$
- 2: **if**  $s_0 \geq q_{target}$  **then**
- 3:     **return** 0
- 4: **end if**
- 5:  $s_{max} \leftarrow \mathcal{S}(\mu_{train} + \alpha_{max} \cdot \mathbf{v}_j)$
- 6: **if**  $s_{max} < q_{target}$  **then**
- 7:     **return**  $\alpha_{max}$
- 8: **end if**
- 9:  $\alpha_{low} \leftarrow 0, \alpha_{high} \leftarrow \alpha_{max}$
- 10: **for**  $i = 1$  to  $N_{steps}$  **do**
- 11:      $\alpha_{mid} \leftarrow (\alpha_{low} + \alpha_{high})/2$
- 12:      $\mathbf{z}_{cand} \leftarrow \mu_{train} + \alpha_{mid} \cdot \mathbf{v}_j$
- 13:      $s_{cand} \leftarrow \mathcal{S}(\mathbf{z}_{cand})$
- 14:     **if**  $s_{cand} < q_{target}$  **then**
- 15:          $\alpha_{low} \leftarrow \alpha_{mid}$
- 16:     **else**
- 17:          $\alpha_{high} \leftarrow \alpha_{mid}$
- 18:     **end if**
- 19: **end for**
- 20: **return**  $\alpha_{high}$

---

**Algorithm 2** AdaptiveMargin

---

**Require:** Set of positive scores  $\mathcal{S}_{pos}$

**Require:** Low percentile  $p_{low}$ , high percentile  $p_{high}$

**Require:** Default margin  $m_{default}$

**Ensure:** Margin  $m$

- 1: **if**  $|\mathcal{S}_{pos}| > 1$  **then**
- 2:      $q_{low} \leftarrow p_{low}/100.0$
- 3:      $q_{high} \leftarrow p_{high}/100.0$
- 4:      $S_{typical} \leftarrow \text{Quantile}(\mathcal{S}_{pos}, q_{low})$
- 5:      $S_{boundary} \leftarrow \text{Quantile}(\mathcal{S}_{pos}, q_{high})$
- 6:      $m \leftarrow \max(0, S_{boundary} - S_{typical})$
- 7: **else**
- 8:      $m \leftarrow m_{default}$
- 9: **end if**
- 10: **return**  $m$

---

## C FINAL POST-HOC CALIBRATION FOR INFERENCE

After training is complete, we perform a final, one-time calibration to prepare the model for inference. It is crucial to distinguish this final step from the “online” calibration performed during training. The online calibration uses the network’s transient state at the start of each epoch to create an adaptive heuristic for guiding outlier synthesis.

In contrast, this final calibration operates on the fixed, fully-trained model and on another held-out calibration set. We perform a full forward pass of the calibration data,  $\mathcal{D}_{calib}$ , through this final model to generate a permanent reference distribution of nonconformity scores, denoted  $\mathcal{S}_{calib}^{final}$ . Crucially, this final distribution is distinct from the ephemeral, epoch-dependent distributions,  $\mathcal{S}_{calib}^{(e)}$ , used during training (i.e.,  $\mathcal{S}_{calib}^{final} \neq \mathcal{S}_{calib}^{(e)}$  for any epoch  $e$  before training completion). Because the model is now a fixed function, this final calibration step satisfies the exchangeability assumption of conformal prediction. This allows  $\mathcal{S}_{calib}^{final}$  to be used for generating statistically valid p-values for OOD hypothesis testing on unseen data, a formal guarantee the online heuristic cannot provide.

## D DATASETS

Examples of images used for model training and evaluation are presented in figure 3. For the training set, we employ a set of augmentations to increase sample diversity and enhance model generalization. The standard pipeline consists of the following sequential operations:

1. **Resizing and Cropping:** Images are first resized so that their shorter edge is 256 pixels. A patch of  $224 \times 224$  pixels is then randomly cropped from the image. For datasets with substantial object size variation, such as Stanford Dogs, a more aggressive RandomResizedCrop with a scale range from 0.3 to 1.0 of the original image area is applied.

**Algorithm 3** GCOS Training Procedure

---

```

864 Require: Training data  $\mathcal{D}_{train}$ , Calibration data  $\mathcal{D}_{calib}$ , Network  $(f_\theta, h_\phi)$  (backbone  $f$ , classifier head  $h$ ),
865  $E$  epochs, start epoch  $E_{start}$ , loss weight  $\lambda$ , shell percentiles  $(p_{inner}, p_{outer})$ 
866 1: Initialize network parameters  $\theta, \phi$ 
867 2: Initialize feature queues  $\mathcal{Q}_{train} \leftarrow \emptyset$ 
868 3: for epoch  $e = 1$  to  $E$  do
869 4:   if  $e \geq E_{start}$  and  $\mathcal{D}_{calib}$  is available then
870 5:     # Epoch-Level Online Calibration (The "Judge")
871 6:     Set network to eval mode
872 7:     for each class  $k$  do
873 8:       Extract features  $Z_{calib}^k$  from  $\mathcal{D}_{calib}$  using  $f_\theta$ 
874 9:       Compute  $\mathcal{M}_{calib}^k = (\mu_{calib}^k, \mathbf{V}_{calib}^k, \Lambda_{calib}^k) \leftarrow \text{PCA}(Z_{calib}^k)$ 
875 10:      Compute Mahalanobis scores  $S_{calib}^{Mahal,k} \leftarrow \text{Mahalanobis}(Z_{calib}^k, \mathcal{M}_{calib}^k)$ 
876 11:      Compute thresholds  $q_{inner}^k, q_{outer}^k \leftarrow \text{Quantiles}(S_{calib}^{Mahal,k}, p_{inner}, p_{outer})$ 
877 12:    end for
878 13:    Set network to train mode
879 14:  end if
880 15:  for batch  $(X, Y) \in \mathcal{D}_{train}$  do
881 16:     $Z_{id} \leftarrow f_\theta(X); L \leftarrow h_\phi(Z_{id})$ 
882 17:     $\mathcal{L}_{CE} \leftarrow \text{CrossEntropy}(L, Y)$ 
883 18:    Update  $\mathcal{Q}_{train}$  with  $Z_{id}$ 
884 19:     $\mathcal{L}_{reg} \leftarrow 0$ 
885 20:  if  $\mathcal{Q}_{train}$  is full and  $e \geq E_{start}$  then
886 21:    # GCOS Synthesis (Geometric Proposer)
887 22:    Compute  $\mathcal{M}_{train} = (\mu_{train}, \mathbf{V}_{train}, \Lambda_{train})$  from  $\mathcal{Q}_{train}$ 
888 23:     $\mathcal{Z}_{ood} \leftarrow \emptyset$ 
889 24:    for each class  $k$  do
890 25:      Identify small components  $\mathbf{V}_{small,train}^k$  from  $\mathbf{V}_{train}^k$  based on variance explained threshold  $\eta$ 
891 26:      for direction  $\mathbf{v}_j \in \mathbf{V}_{small,train}^k$  do
892 27:        # Algorithm 1.
893 28:         $\alpha_{inner} \leftarrow \text{FindBoundaryAlpha}(\mu_{train}, \mathbf{v}_j, q_{inner}^k, \mathcal{S})$ 
894 29:         $\alpha_{outer} \leftarrow \text{FindBoundaryAlpha}(\mu_{train}, \mathbf{v}_j, q_{outer}^k, \mathcal{S})$ 
895 30:        Sample  $\alpha \sim U[\alpha_{inner}, \alpha_{outer}]$ 
896 31:         $\mathbf{z}_{ood} \leftarrow \mu_{train}^k + \text{sign} \cdot \alpha \cdot \mathbf{v}_j$ 
897 32:         $\mathcal{Z}_{ood} \leftarrow \mathcal{Z}_{ood} \cup \{\mathbf{z}_{ood}\}$ 
898 33:      end for
899 34:    end for
900 35:    # Regularization Loss (Energy-Based)
901 36:    Compute Energy strangeness scores:  $S_{pos} \leftarrow \text{logsumexp}(h_\phi(Z_{id})), S_{neg} \leftarrow \text{logsumexp}(h_\phi(\mathcal{Z}_{ood}))$ 
902 37:     $m \leftarrow \text{AdaptiveMargin}(S_{pos})$ 
903 38:     $\mathcal{L}_{reg} \leftarrow \text{Mean}(\text{ReLU}(S_{pos} - S_{neg} + m))$ 
904 39:  end if
905 40:   $\mathcal{L}_{total} \leftarrow \mathcal{L}_{CE} + \lambda \mathcal{L}_{reg}$ 
906 41:  Update  $\theta, \phi$  using gradients of  $\mathcal{L}_{total}$ 
907 42: end for
908 43: end for
909 Ensure: Trained network  $(f_\theta, h_\phi)$ 

```

---

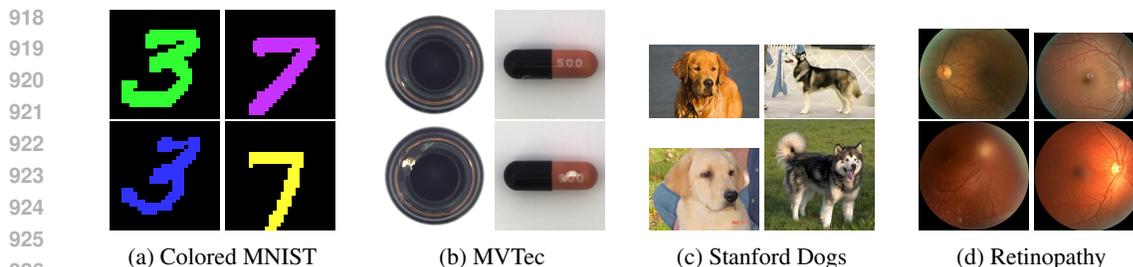
2. **Geometric Transformations:** Random rotations (up to  $\pm 15$  degrees) and random horizontal flips are applied to encourage rotational and reflectional invariance.
3. **Photometric Transformations:** For datasets where color variations are informative, such as Stanford Dogs, `ColorJitter` is applied to randomly modify brightness, contrast, and saturation. Additionally, a small probability of converting the image to grayscale is included for the Stanford Dogs dataset.
4. **Tensor Conversion and Normalization:** Finally, the augmented image is normalized using the standard ImageNet per-channel mean and standard deviation.

For simpler datasets, such as CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), a similar but smaller-scale pipeline is used, consisting of random horizontal flips and random crops on the original  $32 \times 32$  images, followed by normalization using CIFAR-specific statistics.

Table 3: Number of ID and OOD images in each dataset.

Dataset	ID <sub>train</sub>	ID <sub>test</sub>	OOD	Total
Colored MNIST	60,000	10,000	10,000	80,000
MVTec	1,319	326	635	2,280
Stanford Dogs	1,394	100	856	2,350
Retinopathy	28,086	7,022	865	35,973

The specific class compositions for the ID and OOD splits used in our experiments are detailed below.



927 Figure 3: Example images from four datasets. First row in each box: in-distribution; second row:  
928 outliers.

929  
930  
931 **Colored MNIST**

- 932
- 933 • **ID Classes:** Digits 0–9, each correlated with a specific color.
  - 934 • **OOD Data:** Test set digits with novel color-digit pairings.

935  
936 **Stanford Dogs**

- 937
- 938 • **ID Classes (10):** Beagle, Boxer, Border Terrier, Dingo, German Shepherd, Giant Schnauzer, Golden Retriever, Old English Sheepdog, Siberian Husky, Standard Schnauzer.
  - 939 • **OOD Classes (5):** Airedale, Labrador Retriever, Malamute, Malinois, Standard Poodle.

940  
941 **MVTec AD**

- 942
- 943 • **ID Classes (7 defect-free categories):** Bottle, Cable, Capsule, Metal Nut, Screw, Transistor, Zipper.
  - 944 • **OOD Data (Anomalies):** Test set images of the same categories containing various defects.

945  
946  
947 **Retinopathy**

- 948
- 949 • **ID Classes (5 DR severity levels):** No DR, Mild, Moderate, Severe, Proliferative.
  - 950 • **OOD Data (Other Pathologies):** Fundus images depicting other diseases (e.g., Glaucoma, AMD, Pathological Myopia).
- 951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

## E OTHER EVALUATIONS

### E.1 NEAR-ODD DATASETS

We also evaluate our approaches on datasets that are commonly used in outlier exposure studies. Many previous works train classification models on datasets such as CIFAR-10 or CIFAR-100 and evaluate OOD detection on classical benchmarks like LSUN (Yu et al., 2015), Places365 (Zhou et al., 2017), and Textures (Cimpoi et al., 2014). A key limitation of this evaluation setting is that these datasets are primarily far-ODD: the ID data consist of categories such as cats, cars, and birds, while the OOD sets contain images of ground, bedrooms, towers, and similar content. Although this task is not necessarily easier, it is less representative of practical scenarios. In contrast, the main part of this paper emphasizes near-ODD classes, which are more relevant for safety-critical AI systems.

Nevertheless, we also evaluated our methods on these widely used far-ODD datasets and compared GCOS not only with VOS but also with other methods from the literature. As shown in Tables 4, 5, and Tables 6, 7, our approaches achieve competitive performance relative to existing methods.

	CIFAR-100			LSUN-C			Places			Textures			Mean		
	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95
Gaussian OOD + $\mathcal{L}_{\text{uncertainty}}$	<b>99.61</b>	<b>87.08</b>	<b>49.0</b>	<b>99.97</b>	98.56	<b>6.5</b>	99.66	89.1	<b>39.5</b>	99.68	88.34	<b>49.0</b>	99.73	90.77	<b>36.0</b>
GCOS OOD + $\mathcal{L}_{\text{uncertainty}}$	99.48	83.63	56.5	99.97	<b>98.74</b>	7.0	99.6	88.36	42.5	99.48	82.39	60.5	99.64	88.28	41.63
GCOS OOD + $\mathcal{L}_{\text{reg}}$	99.54	82.48	72.5	99.96	98.12	8.5	<b>99.82</b>	<b>92.3</b>	41.5	<b>99.76</b>	<b>90.19</b>	50.5	<b>99.77</b>	<b>90.77</b>	43.25

Table 4: Comparison of out-of-distribution detection performance trained on CIFAR-10.

	CIFAR-10			LSUN-C			Places			Textures			Mean		
	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95	AUPR	AUROC	FPR95
Gaussian OOD + $\mathcal{L}_{\text{uncertainty}}$	<b>99.35</b>	<b>76.47</b>	<b>80.0</b>	99.92	96.54	21.5	99.32	75.51	84.0	<b>99.3</b>	74.53	84.5	99.47	80.76	67.5
GCOS OOD + $\mathcal{L}_{\text{uncertainty}}$	99.33	76.13	83.5	<b>99.94</b>	<b>97.1</b>	<b>17.0</b>	<b>99.37</b>	<b>77.18</b>	<b>81.0</b>	99.23	72.96	87.5	<b>99.47</b>	<b>80.84</b>	<b>67.25</b>
GCOS OOD + $\mathcal{L}_{\text{reg}}$	99.16	72.53	89.5	99.79	91.63	38.0	99.16	72.95	84.5	99.3	<b>76.79</b>	<b>78.0</b>	99.35	78.47	72.5

Table 5: Comparison of out-of-distribution detection performance trained on CIFAR-100.

	Textures		Places365			Textures		Places365	
	AUROC	FPR95	AUROC	FPR95		AUROC	FPR95	AUROC	FPR95
Free Energy (Liu et al., 2020)	85.35	52.46	90.02	40.11	Free Energy (Liu et al., 2020)	76.35	79.63	75.65	80.18
ASH (Djurisic et al., 2023)	86.07	50.90	89.79	40.89	ASH (Djurisic et al., 2023)	83.59	63.69	74.87	79.70
VOS (Du et al., 2022)	88.34	49.00	89.10	39.50	VOS (Du et al., 2022)	74.53	84.50	75.51	84.00
GCOS + $\mathcal{L}_{\text{reg}}$	90.19	50.50	92.3	41.50	GCOS + $\mathcal{L}_{\text{reg}}$	76.79	78.00	72.95	84.50

Table 6: Results on CIFAR-10.

Table 7: Results on CIFAR-100.

### E.2 MULTI-SEED VALIDATION

We also evaluate GCOS performance across multiple random seeds ( $n=5$  per dataset) showing mean  $\pm$  standard deviation. Low variance across all metrics demonstrates hyperparameter robustness and result reproducibility; AUPR remains highly stable across all seeds. This consistency suggests that the method is not dependent on initialization lucky streaks.

Dataset	Mean AUROC	Mean AUPR	Mean FPR95	Std(AUROC)	Std(AUPR)
Colored MNIST	92.90%	99.81%	23.50%	$\pm 3.01\%$	$\pm 0.10\%$
Stanford Dogs	97.39%	99.51%	24.00%	$\pm 2.11\%$	$\pm 0.39\%$
Retinopathy	76.60%	99.05%	80.00%	$\pm 0.55\%$	$\pm 0.02\%$
MVTec	98.71%	99.66%	3.08%	$\pm 0.29\%$	$\pm 0.12\%$

Table 8: Multi-Seed Validation.

## E.3 OBJECT DETECTION

Method	FPR95 ↓	AUROC ↑	mAP (ID) ↑
OOD: MS-COCO / OpenImages			
MSP (Hendrycks & Gimpel, 2016)	70.99 / 73.13	83.45 / 81.91	48.7
ODIN (Liang et al., 2017)	59.82 / 63.14	82.20 / 82.59	48.7
Mahalanobis (Lee et al., 2018)	96.46 / 96.27	59.25 / 57.42	48.7
Energy score (Liu et al., 2020)	56.89 / 58.69	83.69 / 82.98	48.7
Gram matrices (Sastry & Oore, 2020)	62.75 / 67.42	79.88 / 77.62	48.7
Generalized ODIN (Hsu et al., 2020)	59.57 / 70.28	83.12 / 79.23	48.1
CSI (Tack et al., 2020)	59.91 / 57.41	81.83 / 82.95	48.1
GAN-synthesis (Lee et al., 2017)	60.93 / 59.97	83.67 / 82.67	48.5
VOS-ResNet50	47.53±2.9 / 51.33±1.6	88.70±1.2 / 85.23±0.6	48.9±0.2
VOS-RegX4.0	47.77±1.1 / 48.33±1.6	89.00±0.4 / 87.59±0.2	51.6±0.1
<b>GCOS (Ours)</b>	<b>51.28±&lt;0.01 / 60.69±&lt;0.01</b>	<b>87.72±&lt;0 / 86.57±&lt;0.01</b>	<b>48.8±&lt;0.01</b>

Table 9: OOD detection comparison on PASCAL-VOC (In-Distribution) with MS-COCO and OpenImages as OOD datasets.

We also investigate how the GCOS method performs in object detection settings. As shown in Table 9, GCOS represents a significant advancement over traditional OOD detection methods, achieving 4-6 percentage point improvements in AUROC on both COCO and OpenImages OOD datasets. GCOS achieves competitive performance with the state-of-the-art VOS method, particularly excelling in AUROC on OpenImages (86.57 vs 85.23 for VOS-ResNet50) while maintaining comparable in-distribution mAP (48.8 vs 48.9), demonstrating that the method does not sacrifice primary task performance. The method successfully demonstrates that conformal prediction-based outlier synthesis can effectively improve OOD detection while maintaining strong in-distribution performance. GCOS shows robust performance across both OOD datasets, with standard deviation less than 0.01 across three independent training runs, indicating high reproducibility and stability.

In conclusion, the key contribution of GCOS is a principled, geometry-based approach to OOD detection that outperforms heuristic-based methods and competes with the best synthesis-based approaches across both classification and object detection scenarios.

## F ABLATIONS

## F.1 SCORE FUNCTION AND LOSS

Throughout this work, we use the Mahalanobis distance to define the synthesis shell because its value reliably increases as a point moves away from a class mean along a principal component. In contrast, the Energy Strangeness score lacks this property at the beginning of training: its landscape can be complex and non-monotonic, so a larger geometric deviation  $\alpha$  does not necessarily correspond to a higher, more OOD-like energy score, making it unsuitable for our boundary-finding search algorithm and thus for synthesis.

	MVTec			Colored MNIST			Stanford Dogs			Retinopathy		
	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95	AUROC	AUPR	FPR95
GCOS + $\mathcal{L}_{\text{uncertainty}}$	93.07	98.56	33.85	93.21	99.82	25.50	<b>99.42</b>	<b>99.89</b>	<b>0.86</b>	<b>80.05</b>	<b>99.23</b>	<b>71.00</b>
GCOS + $\mathcal{L}_{\text{reg}}$ (Mahalanobis)	<b>97.56</b>	<b>99.51</b>	<b>20.00</b>	<b>95.65</b>	<b>99.91</b>	<b>25.00</b>	97.90	99.61	20.00	77.89	99.17	79.00

Table 10: **Ablation study of contributions 1 and 2.** OOD detection results on MVTec, Colored MNIST, Stanford Dogs and Retinopathy datasets. First row: contribution 1 - GCOS outliers combined with the standard energy-based separation loss  $\mathcal{L}_{\text{uncertainty}}$  in (2). Second row: contributions 1 + 2 - GCOS outliers combined with our geometric loss  $\mathcal{L}_{\text{reg}}$  in (5) using the Mahalanobis score function  $\mathcal{S}_{\mathcal{L}} = \mathcal{S}_{\text{Mahal}}$ .

It is a notable finding that while optimizing for Mahalanobis geometric loss,  $\mathcal{L}_{\text{reg}}$ (Mahalanobis), it effectively regularizes the energy-based confidence of the classifier, the reverse is not true, as the

non-monotonic nature of the energy landscape makes it an unreliable guide for our geometric synthesis search. From Table 10, we observe high scores for the GCOS outliers +  $\mathcal{L}_{reg}$  (Mahalanobis) configuration, along with a slight improvement over the GCOS outliers +  $\mathcal{L}_{uncertainty}$  uncertainty loss (2) for MVTEC and Colored MNIST. This suggests a deep and powerful connection between the geometric structure of the feature space and the confidence landscape of the classifier’s logits. This relationship may be explained as follows:

- **Geometric Compactness as a Regularizer.** Constraining feature clusters to remain geometrically compact through a Mahalanobis-based geometric loss encourages the network to place in-distribution class features on low-dimensional manifolds. This facilitates the final classifier layer’s ability to assign high-confidence logits, leading to consistently low energy scores.
- **Coupling of Geometric and Confidence-Based Strangeness.** A feature vector that is geometrically atypical - i.e., associated with a high Mahalanobis score - also tends to elicit classifier uncertainty, manifested as high energy. Since the classifier learns linear boundaries in feature space, points located far from their cluster are more likely to fall into ambiguous regions with weak logits. Consequently, optimizing geometric OOD scores based on Mahalanobis distance can effectively serve as a proxy for energy-based OOD optimization.

## F.2 CONFORMAL SHELL NECESSITY

*Why Adaptivity Matters.* Comparing the adaptive Conformal Shell against a fixed synthesis range ( $\alpha \sim U[0, 200]$ ). Without the shell’s ability to automatically find the correct feature scale, performance collapses. This demonstrates that static ranges fail to capture the nuance required for high-precision tasks.

Dataset	AUROC / Impact
	(GCOS vs Fixed Range)
C-MNIST	99.20% vs 78.00% (↓ 21.2)
CIFAR-10	90.77% vs 86.56% (↓ 4.21)

Table 11: Performance with and without shell.

## G CONFORMAL RISK CONTROL

**Note on Decision-Making with Conformal Risk Control.** In addition to the direct hypothesis test described in equation (7), the p-values can be converted into a continuous OOD score to enable explicit conformal risk control (Angelopoulos et al., 2022). We define the final OOD score for a test sample as  $S_{OOD}(\mathbf{z}^{\text{test}}) = 1 - p_{\text{final}}(\mathbf{z}^{\text{test}})$ , where higher values indicate a higher likelihood of being OOD. Using this score, a decision threshold  $\tau$  can be chosen to control a desired error rate on the in-distribution data. For instance, in VOS, the logistic regression threshold was set so that 95% of ID data is labeled as in-distribution; here, we adopt a similar logic using p-value thresholding.

Specifically, to control the False Negative Rate - the rate at which ID samples are incorrectly classified as OOD-at a level of  $\alpha_{risk}$ , we may set  $\tau$  to be the  $(1 - \alpha_{risk})$ -quantile of the OOD scores computed on the final calibration set  $\mathcal{D}_{calib}$ . A test sample is then classified as OOD if its OOD score exceeds  $\tau$ , and as ID otherwise.

This approach ensures that the FNR on future data is approximately controlled at  $\alpha_{risk}$ . While the direct p-value test enforces a fixed statistical significance level, the risk-controlled method provides flexibility by allowing threshold selection to achieve a specific desired performance on the ID data, such as the 95% True Positive Rate (corresponding to a 5% FNR) often used for computing FPR@95TPR.

## H HYPERPARAMETER SENSITIVITY

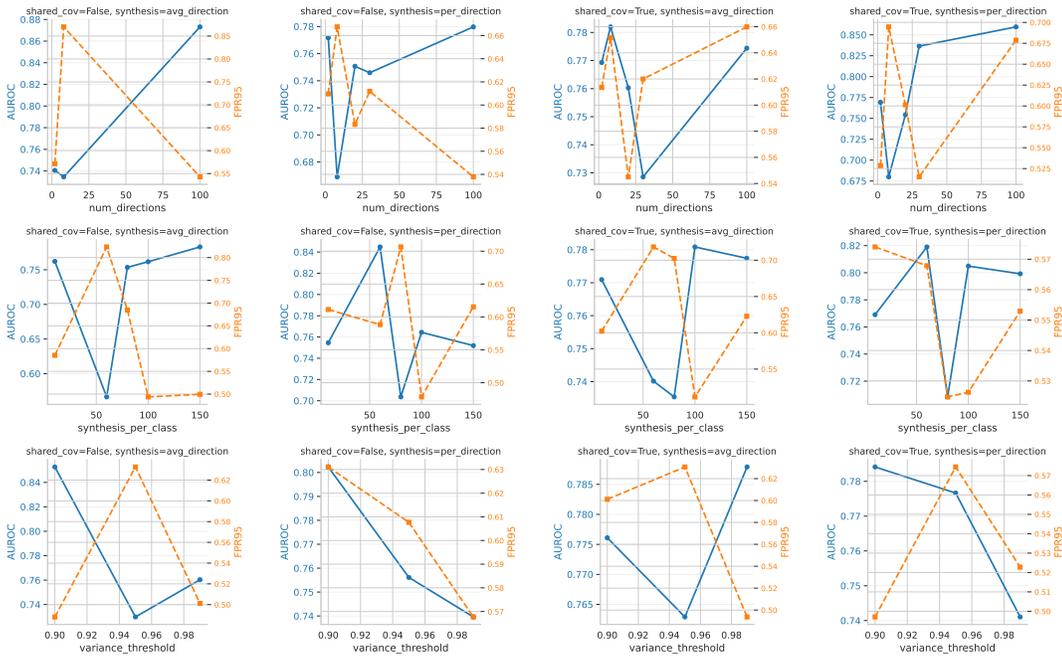


Figure 4: Hyperparameter ablation study.

Our approach introduces several hyperparameters that influence model performance across different settings, as illustrated in figure 4. The hyperparameters appear to exhibit synergies and correlated effects on performance.

The number of directions denotes the number of vectors  $\tilde{V} \subseteq \mathbf{V}_{\text{small, train}}^k$ , a random subsample of  $\mathbf{V}_{\text{small, train}}^k$ . For the *avg direction* configuration, a single outlier is generated along the direction of the average vector  $v$  as in (3), whereas *per direction* generates  $|\tilde{V}|$  OOD vectors. The effect of the number of directions (first row of the figure 4) depends on the values of other hyperparameters. For *avg direction* (columns 1 and 3), increasing *num directions* consistently improves AUROC and decreases FPR95, suggesting that averaging more diverse directions produces better and more informative outliers.

Shared Covariance controls the estimation strategy for the covariance matrix of the “Proposer” PCA models. When enabled, a single shared covariance matrix is estimated from the pooled, class-centered features of all categories in the training queue, under the assumption of a common underlying geometric structure. When disabled, a distinct per-class covariance is estimated, allowing for more specific manifold modeling at the expense of statistical robustness for classes with fewer samples. Its impact appears most notable for very small values of *num directions*; beyond this regime, its effect on performance is minimal.

The value of *synthesis per class* (second row of the figure 4) dictates the number of virtual outliers generated for each class per training iteration. It directly controls the cardinality of the synthetic OOD set,  $\mathcal{Z}_{\text{ood}}$ , used to compute the regularization loss, thereby modulating the strength of the outlier exposure signal in each optimization step. The effect of this hyperparameter depends on the values of other hyperparameters, with no clear trend observed, suggesting that the total number of outliers may be less important than their quality and diversity.

The variance threshold (third row of the figure 4),  $\eta \in (0, 1)$ , determines the partitioning of the principal components of the “Proposer” models into a high-variance signal subspace,  $\mathbf{V}_{\text{large}}$ , and a low-variance off-manifold subspace,  $\mathbf{V}_{\text{small}}$ . Components are classified according to the minimum number of principal components required to explain at least  $\eta$  proportion of the total feature variance; the remaining components form  $\mathbf{V}_{\text{small}}$  and serve as directions for outlier synthesis. A value of

1188 0.9 appears to provide the best performance, being sufficiently permissive to include more compo-  
1189 nents in  $\mathbf{V}_{\text{small, train}}^k$ , which allows for more diverse and flexible outlier generation. As the threshold  
1190 approaches 1.0, performance tends to decline, indicating that a more conservative definition of off-  
1191 manifold directions (e.g.,  $\eta = 0.99$ ) leaves too few directions in  $\mathbf{V}_{\text{small, train}}^k$  for effective synthesis.  
1192 This highlights the importance of aggressively defining off-manifold directions.  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240

---

1241 \*Portions of the text were polished for grammar and clarity with the assistance of a large language model.  
The research ideas, methods, and results are entirely the authors' own.