Contrastive Conformal Sets

Yahya Alkhatib[1] **Wee Peng Tay**[1]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Abstract

Contrastive learning produces coherent semantic feature embeddings by encouraging positive samples to cluster closely while separating negative samples. However, existing contrastive learning methods lack a principled construction of geometric sets in the semantic feature space with distribution-free guarantees at any user-specified coverage level. We extend conformal prediction to this setting by introducing covering sets equipped with learnable generalized hyper-ball constraints. We propose a method that constructs conformal sets guaranteeing user-specified coverage of positive samples while maximizing negative sample exclusion. We theoretically motivate volume minimization as a proxy for negative exclusion, enabling our approach to operate effectively even when negative pairs are unavailable. The positive inclusion guarantee inherits the distribution-free coverage property of conformal prediction, while negative exclusion is maximized through learned set geometry optimized on a held-out training split. Experiments on simulated and real-world image datasets demonstrate improved inclusion-exclusion trade-offs compared to standard distance-based conformal baselines.

1 INTRODUCTION

Effective machine learning fundamentally relies on learning embeddings that capture semantic relationships between data samples [Boser et al., 1992, Dinh et al., 2026, Wen et al., 2016, Schroff et al., 2015]. In high-dimensional feature spaces, this objective is formalized as placing semantically similar samples in proximity while distributing dissimilar samples to distinct regions. Contrastive learning provides a principled framework for achieving this goal by training encoders to minimize distances to similar samples (positive pairs) while maximizing distances to dissimilar samples (negative pairs) under an appropriate embedding metric [Chen et al., 2020, van den Oord et al., 2018, Zhu et al., 2020, Shi et al., 2023]. A key strength of contrastive learning is its applicability to both supervised settings, where similarity is defined by labels, and unsupervised settings, where positive and/or negative pairs are generated through data augmentation [Zhu et al., 2020]. This flexibility has established contrastive learning as a dominant paradigm for pretraining transferable representations that achieve strong performance across diverse downstream tasks [Ardeshir and Azizan, 2022]. In high-stakes domains such as healthcare and finance, there is an increasing demand for models that provide rigorous and principled quantification of uncertainty, rather than relying on heuristic measures. This challenge is particularly pronounced in high-dimensional semantic embedding spaces, where point estimates often lack guarantees of reliability.

Despite the widespread adoption of contrastive learning across numerous applications, developing systematic methodologies for certifying a model's ability to cluster semantically similar examples within a user-specified tolerance level remains underexplored. For instance, Fig. 1 depicts a UMAP projection [McInnes et al., 2018] of the semantic features of an input image, generated by a trained unsupervised contrastive model, in $\mathbb{R}^3$. Quantifying the model's uncertainty in distinguishing the input from its corresponding positive and negative samples in such a setting is inherently challenging. Notably, since such models map the input space to the semantic feature space, their performance is typically assessed through downstream tasks.

A natural question, then, is whether one can evaluate the uncertainty of the encoder as a standalone projector to the semantic space, independent of any downstream application. Such an intrinsic characterization of embedding quality is valuable for two key reasons. First, it serves as a diagnostic tool for model selection and hyperparameter optimization, enabling practitioners to compare projectors based on the

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:19–38.

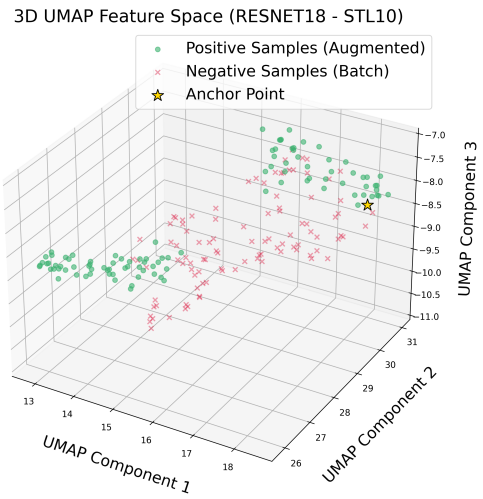

Figure 1: 3D UMAP projection of the semantic features of an STL10 anchor point and its positive and negative samples.

tightness of their covering sets at a fixed coverage level, without relying on specific downstream tasks. Second, it establishes a principled foundation for subsequent applications—such as out-of-distribution (OOD) detection or retrieval systems—which can leverage the certified embedding neighborhoods and inherit their coverage guarantees.

Quantifying the uncertainty of contrastive learners has been studied in multiple works, primarily by replacing deterministic embeddings with stochastic distributions [Oh et al., 2019] or aggregating Gaussian mappings to derive a scalar uncertainty measure from covariance norms [Wu and Goodman, 2020]. Similarly, contrastive representations have been leveraged for OOD detection using Mahalanobis distances on intermediate features [Winkens et al., 2020], novelty scores derived from shifted augmentations [Tack et al., 2020], and energy-based discriminative-generative models [Liu and Abbeel, 2020]. However, while these approaches effectively summarize uncertainty into scalar or distributional scores, they fundamentally lack the geometric structure required to provide formal, distribution-free, and user-specified inclusion guarantees for positive samples.

To address this gap, we aim to construct sets in the semantic feature space that include the embeddings of positive samples with any user-specified level, while simultaneously maximizing the exclusion of negative samples. This is inspired by the framework of conformal prediction (CP), a principled approach to uncertainty quantification that offers strong theoretical guarantees and broad applicability [Vovk et al., 2005, Angelopoulos and Bates, 2023, Angelopoulos et al., 2025, Shafer and Vovk, 2008, Messoudi et al., 2020, Quach et al., 2024, Foygel Barber et al., 2023, Lei et al., 2015]. CP has traditionally been applied to quantify uncertainty in the output space, producing prediction sets that contain the true label with a user-specified coverage guarantee in a distribution-free and model-agnostic manner. This property makes CP a versatile tool for uncertainty quantification, as it treats the underlying model as a black box [Angelopoulos and Bates, 2023, Angelopoulos et al., 2025]. Recent advancements have extended CP to leverage semantic feature space information; for example, Teng et al. [2023] utilize learned representations to improve the efficiency of output-space prediction sets, though the resulting sets remain in the output space rather than the feature space. While prior work has explored CP with multi-valued scoring functions and in output spaces [Messoudi et al., 2020, Klein et al., 2025, Diquigiovanni et al., 2022], the direct construction of CP sets within the semantic feature space remains largely uncharted.

In this work, we propose a framework that addresses this challenge. The positive inclusion guarantee is derived from the distribution-free coverage property of CP, achieved through a calibration-quantile threshold, while negative exclusion is optimized by learning the geometry of the covering set via generalized hyper-ball constraints and scaling matrices. We further establish, both theoretically and empirically, that minimizing the volume of these sets serves as a proxy for maximizing negative exclusion, enabling the framework to construct discriminative boundaries even when only positive samples are available. The proposed sets are constructed as overlays on the existing embedding space, leaving the learned representations unaltered, and can serve as a foundation for downstream tasks such as OOD detection.

Our main contributions are as follows:

1. We introduce *contrastive conformal sets*: a framework for constructing conformal prediction sets directly in the semantic feature space around anchor embeddings, designed to include positive samples while excluding negative samples.
2. We show that positive inclusion inherits the distribution-free coverage guarantee of conformal prediction at any user-specified level, while the set geometry is optimized to maximize negative exclusion through learnable norm exponents and scaling parameters.
3. We establish that, within a nested family of covering sets, minimum-volume sets maximize negative exclusion (Proposition 1), motivating a proxy objective when only positive samples are available, a common setting in contrastive learning.
4. We validate the framework empirically on simulated and real-world image benchmarks, demonstrating that it achieves the prescribed positive coverage while substantially improving negative exclusion over vanilla conformal baselines. We further show that the learned covering sets yield competitive or superior perfor-

mance on OOD detection against several dedicated methods.

2 PRELIMINARIES

We begin by reviewing the standard split CP framework, which constructs prediction sets in the output space. This serves as a foundation for several concepts that will be extended in our work.

Set-Valued Predictors and Nonconformity Scores. The goal of CP is to construct a set-valued mapping $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ that provides uncertainty quantification for model predictions. The quality of such a mapping is assessed along two key dimensions: statistical validity (coverage) and practical utility (efficiency).

CP constructs prediction sets by leveraging nonconformity scores, which quantify the degree to which a candidate label Y conforms to the model's prediction. Given a pre-trained model f_θ, the nonconformity score for an input-label pair (X, Y) is defined as $s(X, Y; \theta) = \mathsf{L}(f_\theta(X), Y)$, where L is a task-specific discrepancy measure (e.g., one minus the softmax probability of Y given X).

Given a threshold $t \in \mathbb{R}$, the prediction set is constructed by including all candidate labels whose nonconformity scores do not exceed the threshold:

$$\mathcal{C}_{\theta,t}(X) = \{y \,:\, s(X, y; \theta) \leq t\}. \tag{1}$$

The Split Conformal Guarantee. To ensure that the prediction set achieves a user-specified coverage level $1 - \alpha$, the split CP framework employs a distribution-free calibration procedure. The dataset is partitioned into three disjoint subsets: a training set $\mathcal{D}$ (used to train f_θ), a calibration set $\mathcal{D}_{\text{cal}}$ of size n_{cal}, and a test set $\mathcal{D}_{\text{test}}$.

Nonconformity scores are computed for all instances in $\mathcal{D}_{\text{cal}}$, and the critical threshold $t = \hat{q}_\alpha$ is chosen as the empirical $\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil / n_{\text{cal}}$ quantile of these scores. Assuming that the calibration and test data are exchangeable—a property typically satisfied by random splitting—this threshold guarantees marginal coverage for any test point $(X, Y) \in \mathcal{D}_{\text{test}}$ [Shafer and Vovk, 2008, Foygel Barber et al., 2023, Angelopoulos and Bates, 2023]:

$$\mathbb{P}(Y \in \mathcal{C}_{\theta,\hat{q}_\alpha}(X)) = \mathbb{P}(s(X, Y; \theta) \leq \hat{q}_\alpha) \geq 1 - \alpha. \tag{2}$$

This probability is taken jointly over the randomness of the calibration data and the test point.

Building on these foundations, the next section introduces our problem formulation: extending CP to the semantic feature space of models, with a focus on constructing contrastive sets that simultaneously optimize for the inclusion of positive samples and the exclusion of negative samples. We defer all proofs to Appendix A.

3 PROBLEM STATEMENT

Positive and Negative Samples. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label space, with (X, Y) representing a pair of random variables distributed according to the joint probability density function (pdf) $p_{X,Y}$. For a given anchor realization $(X = x_0, Y = y_0)$, the *oracle* distributions for generating positive and negative samples are defined as follows:

$$p^*_{\text{pos}}(\cdot) \propto p_{X,Y}(\cdot, y = y_0), \tag{3}$$

$$p^*_{\text{neg}}(\cdot) \propto \int_{\{y \neq y_0\}} p_{X,Y}(\cdot, y)\, \mathrm{d}y, \tag{4}$$

where p^*_{pos} generates samples conditioned on the same label as the anchor, and p^*_{neg} generates samples conditioned on labels different from that of the anchor.

In supervised settings, where label information is available [Khosla et al., 2020], these oracle distributions are well-defined and training samples are accessible. However, in unsupervised settings, where labels are unavailable [Saunshi et al., 2019], the oracle mechanisms p^*_{pos} and p^*_{neg} cannot be directly accessed. Instead, practitioners approximate these distributions using conditional mechanisms $p_{\text{pos}}(\cdot \mid X = x_0)$ and $p_{\text{neg}}(\cdot \mid X = x_0)$, respectively. The design of these approximations aims to closely align with the oracle distributions, a key desideratum in contrastive learning that has motivated extensive research [Saunshi et al., 2019, Chuang et al., 2020, Khosla et al., 2020].

While the design of these approximations is an important challenge, it is not the primary focus of this work. Instead, our objective is to develop a principled framework for constructing *contrastive conformal sets* that provide rigorous guarantees for *any given contrastive learning model*. Specifically, we aim to ensure that the embeddings of positive samples lie within a well-defined neighborhood of the anchor embedding with high probability. Our framework also provides the option of updating the contrastive learning model weights to achieve better inclusion-exclusion tradeoffs, leading to more robust contrastive embeddings.

Contrastive Conformal Learning Objective. The primary objective of this work is to certify embedding neighborhoods within a semantic feature space $\mathcal{Z}$ equipped with a scoring function g. Let $\mathcal{D}$ denote a dataset of anchor points $X \in \mathcal{X}$. For any given anchor X, let $\mathcal{X}_{\text{pos}}, \mathcal{X}_{\text{neg}} \subset \mathcal{X}$ define the subsets of its associated positive and negative samples, respectively; these associations may be established through unsupervised augmentations or supervised labels $Y \in \mathcal{Y}$. Consider an encoder $\theta : \mathcal{X} \to \mathcal{Z}$, pretrained on $\mathcal{D}$, that maps an input to its corresponding embedding $Z = \theta(X)$. A well-trained encoder is expected to map positive samples to embeddings that are proximal to the anchor's embedding, while pushing the embeddings of negative samples sufficiently far away.

Formally, for a given threshold $t > 0$, the neighborhood of an embedding Z is defined as:

$$\mathcal{N}(Z) \triangleq \{z \in \mathcal{Z} : g(Z, z) \leq t\}, \tag{5}$$

where $g : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ is a distance score that quantifies the proximity between embeddings. The choice of g and t determines the geometry of the neighborhood $\mathcal{N}(Z)$. An ideal encoder satisfies two key properties: (i) the embedding of any positive sample lies within the neighborhood, $Z_{\text{pos}} \in \mathcal{N}(Z)$, and (ii) the embedding of any negative sample lies outside the neighborhood, $g(Z, Z_{\text{neg}}) > t$.

In practice, these properties cannot be enforced deterministically. Instead, the goal is to construct $\mathcal{N}(Z)$ such that the embeddings of positive samples are included with high probability, controlled by a user-specified tolerance level $\alpha \in (0, 1)$, while simultaneously maximizing the probability of excluding negative embeddings. This leads to the following optimization problem:

$$\arg\max_{\mathcal{N}} \mathbb{P}(Z_{\text{neg}} \notin \mathcal{N}(Z)) \tag{6a}$$

$$\text{s. t. } \mathbb{P}(Z_{\text{pos}} \in \mathcal{N}(Z)) \geq 1 - \alpha. \tag{6b}$$

The objective is to determine the score g and the threshold t that define $\mathcal{N}(Z)$, ensuring that the positive inclusion constraint in (6b) is satisfied at the desired coverage level, while maximizing the negative exclusion probability in (6a). The optimal set $\mathcal{N}(Z)$ can then be used in downstream applications, such as OOD detection, where the inclusion of positive samples and exclusion of negative samples are critical for performance.

Corollary 1 (Oracle Inclusion-Exclusion Bounds)**.** *If the positive inclusion condition in* (6b) *is satisfied by some positive generation mechanism p_{pos} at level $1 - \alpha$, and the negative exclusion probability in* (6a) *is bounded below by $1 - \beta$ for some $\beta \in (0, 1)$, then the supervised samples satisfy positive inclusion and negative exclusion levels bounded below by*

$$1 - \frac{\alpha}{\mathbb{P}(Y_{\text{pos}} = Y)} \quad \textit{and} \quad 1 - \frac{\beta}{\mathbb{P}(Y_{\text{neg}} \neq Y)}, \tag{7}$$

respectively. Here, Y_{pos} and Y_{neg} are the labels of the positive and negative samples, respectively, and Y is the label of the anchor.

This highlights a fundamental principle of contrastive learning: the closer the alignment between the generation mechanisms $p_{\text{pos}}, p_{\text{neg}}$ and their respective oracle distributions $p^*_{\text{pos}}, p^*_{\text{neg}}$—specifically, the more accurately X_{pos} represents a sample from the same class as X and X_{neg} represents a sample from a different class—the stronger the empirical guarantees for inclusion and exclusion. It is important to emphasize that the coverage probability in (6b) is certified geometrically rather than semantically. Hence, depending on the choice of p_{pos} in practice, the deviation between the class-consistent coverage rate and the level $1 - \alpha$ can change. We empirically demonstrate in Appendix C that the coverage level carries over from augmentations to class-consistent samples.

4 METHODOLOGY

To address this problem, we begin by constructing covering sets inspired by the principles of CP. Let $\mathcal{D}_{\text{cal}}$ denote a held-out calibration set, where each sample is represented as a triplet $(X, X_{\text{pos}}, X_{\text{neg}})$. Here, X serves as the anchor, X_{pos} is a positive sample drawn from the conditional distribution $p_{\text{pos}}(\cdot \mid X)$, and X_{neg} is a negative sample drawn from $p_{\text{neg}}(\cdot \mid X)$. In the sequel, we always assume that the anchors in $\mathcal{D}_{\text{cal}}$ are exchangeable with all test anchors, given the training set. In Section 4.3, we extend this framework to scenarios where only positive samples X_{pos} are available.

For each calibration triplet, the pretrained encoder θ maps the input features into the semantic feature space $\mathcal{Z}$. Using a chosen distance function $s(\cdot, \cdot)$, we compute the distance $s(Z, Z_{\text{pos}})$ between the anchor's embedding and its conditionally generated positive counterpart.

Lemma 1 (Positive Inclusion Guarantee)**.** *Suppose $\mathcal{D}_{\text{cal}}$ consists of n_{cal} samples whose anchors X^j and their corresponding positive samples X^j_{pos} are projected to embeddings Z^j, Z^j_{pos}, respectively, for $j = 1, \ldots, n_{\text{cal}}$. Suppose the test anchor X is projected to Z, and its positive sample X_{pos} is projected to Z_{pos}. Assuming that X and the anchors X^j for $j = 1, \ldots, n_{\text{cal}}$ are exchangeable given the training set $\mathcal{D}$, the distance $s(Z, Z_{\text{pos}})$ from the test anchor to its positive sample satisfies*

$$\mathbb{P}(s(Z, Z_{\text{pos}}) \leq \hat{q}_\alpha) \geq 1 - \alpha, \tag{8}$$

where $\hat{q}_\alpha$ is the $\lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil$-th smallest value of the calibration positive distance scores, $\{s(Z^j, Z^j_{\text{pos}})\}_{j=1}^{n_{\text{cal}}}$. The probability is taken marginally over the joint distribution of the calibration data, the test point, and the positive sample.

In the Euclidean space $\mathbb{R}^d$, two common choices for any pair of vectors $u, v \in \mathbb{R}^d$ are the ℓ_2 distance and the Mahalanobis distance:

$$s_{\ell_2}(u, v) = \|u - v\|_2, \tag{9}$$

$$s_{\text{Mah}}(u, v) = \sqrt{(u - v)^\intercal \Sigma^{-1} (u - v)}. \tag{10}$$

To avoid introducing bias, the covariance matrix Σ in s_{Mah} must be estimated from a held-out split of data disjoint from $\mathcal{D}_{\text{cal}}$ [Johnstone and Cox, 2021]. Geometrically, s_{ℓ_2} yields a fixed-radius hyper-ball in the embedding space, whereas s_{Mah} produces a hyper-ellipsoid whose shape is governed by the covariance structure of the data.

Lemma 1 guarantees the positive inclusion requirement in (6b). However, fixing the scoring function $s(Z, \cdot)$ provides

no direct control over the negative exclusion objective in (6a); the exclusion probability depends entirely on how the calibration negative scores are distributed relative to the positive scores. To address this, we propose a learnable adaptive scoring function that dynamically adjusts the geometry of the conformal set to maximize the exclusion of negative samples while maintaining the desired positive inclusion rate. By parameterizing the score with learnable components, such as scaling matrices and norm exponents, we enable the conformal set to adapt to the underlying data distribution. This approach ensures that the positive inclusion constraint in (6b) is satisfied, while the negative exclusion probability in (6a) is optimized through a data-driven learning process.

4.1 LEARNABLE-METRIC CONFORMAL SETS

To simultaneously satisfy the positive inclusion constraint in (6b) and maximize the negative exclusion probability in (6a), we require covering sets whose geometry can adapt to the distribution of semantic features.

Single-norm covering sets. We define a family of candidate sets around an anchor embedding Z, parameterized by a learnable exponent $p > 0$ and a scaling matrix $M \in \mathbb{R}^{d\times d}$:

$$\mathcal{N}(Z) = \{z : \|M(Z - z)\|_p \le t\}, \tag{11}$$

for a threshold $t > 0$. When $p \ge 1$, the resulting set is convex; when $p < 1$, it becomes non-convex (star-shaped), offering greater flexibility to match non-standard feature distributions. The volume of this set can be expressed in closed form as

$$\mathrm{Vol}(\mathcal{N}) = \lambda\big(\mathbb{B}_{\|\cdot\|_p}(t)\big) \cdot \det(M)^{-1}, \tag{12}$$

where $\lambda(\cdot)$ denotes the Lebesgue measure. For the ℓ_p-ball of radius t, this takes the explicit form

$$\lambda\big(\mathbb{B}_{\|\cdot\|_p}(t)\big) = \frac{(2t)^d\,\Gamma\big(1+\frac{1}{p}\big)^d}{\Gamma\big(1+\frac{d}{p}\big)} = t^d\,\lambda\big(\mathbb{B}_{\|\cdot\|_p}(1)\big),$$

where $\Gamma(\cdot)$ is the gamma function and $\mathbb{B}_{\|\cdot\|_p}(1)$ is the unit p-norm ball [Braun et al., 2025].

Generalized hyper-ball covering sets. Although a learnable norm and scaling matrix provide considerable expressive power, semantic features may exhibit heterogeneous distributional shapes along different coordinate directions. To accommodate this, we adopt a generalized hyper-ball construction [Wang, 2005], in which each dimension is endowed with its own exponent:

$$\mathcal{N}(Z) = \left\{ z : \sum_{j=1}^{d} m_j^{p_j} \cdot \big|\{Z - z\}_j\big|^{p_j} \le t \right\}, \tag{13}$$

where $\{a\}_j$ denotes the j-th component of the vector a, each $m_j > 0$ is a learnable scale parameter, and each $p_j > 0$ is a learnable exponent. We can construct learnable vectors of them such that $p = [p_1, p_2, \ldots, p_d]^\intercal$ and $m = [m_1, m_2, \ldots, m_d]^\intercal$. Following [Wang, 2005], the volume of this generalized set admits the closed-form expression

$$\mathrm{Vol}(\mathcal{N}) = \frac{t^{\sum_{j=1}^{d} 1/p_j}}{\prod_{j=1}^{d} m_j} \cdot \frac{2^d \prod_{j=1}^{d} \Gamma\big(1+\frac{1}{p_j}\big)}{\Gamma\big(1+\sum_{j=1}^{d}\frac{1}{p_j}\big)}. \tag{14}$$

Remark 1. *The volume in* (14) *is well-behaved at both ends of the practical range. As $p_j \downarrow 1$ the unit set converges to an ℓ_1 cross-polytope (finite volume); as $p_j \uparrow \infty$ it converges to a hyper-rectangle (also finite). Pathological behavior arises only outside this range: as $p_j \downarrow 0$ the unit set degenerates onto the coordinate axes, and as $m_j \uparrow \infty$ the volume collapses to zero. In practice, one can clamp p_j from below and m_j from above or add a regularization term to the loss functions introduced in the next section.*

With these set constructions, we are now ready to tackle the objective of constructing covering sets that include positive samples with user-specified tolerance and maximum exclusion of negative samples.

4.2 OPTIMIZATION OBJECTIVE

Let $\mathcal{D}_{\text{train}}$ denote a training set of size n_{train} (not to be confused with $\mathcal{D}$, the data used to pretrain the encoder $\theta(\cdot)$). To approximate the probabilities in (6a) and (6b), we reformulate the optimization problem as follows:

$$\arg\max_{M,p} \sum_{i=1}^{n_{\text{train}}} \sum_{j=1}^{K} \mathbf{1}_{\{g_{M,p}(Z^{i,j}_{\text{neg}}, Z^i) > t\}}, \tag{15a}$$

$$\text{s.t.} \quad \frac{1}{n_{\text{train}} K} \sum_{i=1}^{n_{\text{train}}} \sum_{j=1}^{K} \mathbf{1}_{\{g_{M,p}(Z^{i,j}_{\text{pos}}, Z^i) \le t\}} \ge 1 - \alpha, \tag{15b}$$

where $g_{M,p}(\cdot, Z)$ represents either the single-norm metric or the generalized hyper-ball score, the superscripts i and j index individual anchor instances and their corresponding positive or negative samples, respectively, and K is the number of positive (or negative) samples per anchor point. To enable optimization via first-order methods, we introduce the following modifications:

1. The threshold t is set to the empirical $\lceil (1-\alpha)(n_{\text{train}}+1)\rceil / n_{\text{train}}$ quantile of the positive distances $g_{M,p}(Z_{\text{pos}}, Z)$ at each iteration, denoted as $\hat{q}_\alpha$. This ensures that the constraint in (15b) is satisfied implicitly by Lemma 1. Since the quantile computation involves a sorting operation, which is non-differentiable, $\hat{q}_\alpha$ is detached from the computational graph during the backward pass.

2. The non-differentiable indicator function $\mathbf{1}_{\{c>t\}}$ is replaced with a smooth sigmoid approximation, defined as $\sigma_T(c-t) = \frac{1}{1+e^{-T(c-t)}}$, where T is a temperature parameter. Increasing T improves the sharpness of the approximation but may lead to vanishing gradients.
3. To enforce positivity constraints on the parameters without resorting to constrained optimization solvers, we optimize over element-wise positive reparameterizations for the generalized hyper-ball score (squares for m and absolute values for p). For the single-norm metric, we parameterize M as $M = AA^\intercal$, where A is a learnable matrix. This ensures that M remains positive semi-definite throughout the optimization process.

The relaxed optimization problem reduces to:

$$\arg\max_{M,p} \; J_{\text{neg}} \triangleq \sum_{i=1}^{n_{\text{train}}} \sum_{j=1}^{K} \sigma_T(g_{M,p}(Z_{\text{neg}}^{i,j}, Z^i) - \hat{q}_\alpha). \tag{16}$$

This objective directly focuses on maximizing the exclusion of negative samples. However, relying exclusively on this formulation assumes the ready availability of negative pairs, which is not always guaranteed. Furthermore, to be practically informative, the resulting conformal sets must be structurally efficient, i.e., possessing a small Lebesgue measure.

4.3 POSITIVE-SAMPLE-ONLY MINIMUM VOLUME SETS

In many modern representation learning frameworks, accessing explicit negative samples can be challenging or architecturally unsupported. Under such positive-sample-only regimes, we must still construct conformal sets that achieve high negative exclusion without directly optimizing against negative instances.

Fortunately, the geometric size of the covering set can motivate a heuristic proxy. The following proposition demonstrates that among all nested sets satisfying the positive inclusion constraint in (6b), the one with the smallest volume maximizes the empirical exclusion rate of the negative distribution.

Proposition 1. *Let* Vol *denote the reference volume measure on $\mathcal{Z}$. Fix $\alpha \in (0,1)$ and consider a family of measurable sets*

$$\{\mathcal{B}_t \,:\, t \in T\}, \quad \mathcal{B}_t \subseteq \mathcal{Z}, \tag{17}$$

that is nested in t, i.e.,

$$t_1 \le t_2 \;\Rightarrow\; \mathcal{B}_{t_1} \subseteq \mathcal{B}_{t_2} \text{ and } \text{Vol}\left(\mathcal{B}_{t_1}\right) \le \text{Vol}\left(\mathcal{B}_{t_2}\right).$$

Denote by $Z_{\text{pos}} = \theta(X_{\text{pos}})$ the embedding of positive samples. Let

$$t^* \in \arg\min_{t \in T} \left\{ \text{Vol}\left(\mathcal{B}_t\right) :\; \mathbb{P}(Z_{\text{pos}} \in \mathcal{B}_t) \ge 1-\alpha \right\}, \tag{18}$$

and set $\mathcal{B}_\alpha^* \triangleq \mathcal{B}_{t^*}$. Then $\mathcal{B}_\alpha^*$ satisfies the positive inclusion requirement (6b) *and maximizes the negative exclusion probability* (6a) *within the family $\{\mathcal{B}_t : t \in T\}$. Equivalently, $\mathcal{B}_\alpha^*$ is a minimum-volume set in this family among all sets that satisfy* (6b)*.*

We call such a set $\mathcal{B}_\alpha^*$ a *Minimum-Volume α-Inclusion Set* (MVαIS).

Guided by Proposition 1, if no negative samples are available, then we might rely on only positive samples by isolating the volume as our primary minimization objective:

$$\arg\min_{M,p} \; J_{\text{vol}} \triangleq \sum_{i=1}^{n_{\text{train}}} \log \text{Vol}(\mathcal{N}_{\hat{q}_\alpha}(Z^i)), \tag{19}$$

where the threshold $\hat{q}_\alpha$ is implicitly tied to the empirical $\lceil (1-\alpha)(n_{\text{train}}+1) \rceil / n_{\text{train}}$-quantile of the positive distances at each iteration, satisfying the inclusion constraint by design. Finding a clean result that generalizes Proposition 1 to non-nested families is left to future work.

On the other hand, when negative samples are fully available, the log-volume acts as a potent geometric regularizer alongside the direct exclusion objective. Utilizing the log-volume rather than the raw Lebesgue measure ensures numerical stability against the $[0,1]$ bounded sigmoid term. The combined objective becomes:

$$\arg\max_{M,p} J_{\text{neg+vol}} \triangleq \sum_{i=1}^{n_{\text{train}}} \sum_{j=1}^{K} \Big[(1-\lambda)\sigma_T(g_{M,p}(Z_{\text{neg}}^{i,j}, Z^i) - \hat{q}_\alpha) - \lambda \log \text{Vol}(\mathcal{N}_{\hat{q}_\alpha}(Z^i)) \Big], \tag{20}$$

where $\lambda \in [0,1]$ is a regularization constant. By controlling λ, we can learn covering sets that strictly minimize the volume, strictly maximize the negative exclusion rate, or smoothly balance both. Notice that setting $\lambda = 1$ gracefully recovers the positive-sample-only objective J_{vol} (formulated equivalently as a maximization of the negative log-volume).

To further enhance the adaptability of the contrastive conformal sets, we propose incorporating the model weights into the optimization process. This approach allows the feature space to dynamically adjust in conjunction with the learned geometry. In practice, contrastive models are often trained using the InfoNCE loss [van den Oord et al., 2018], which is defined as:

$$L_{\text{InfoNCE}} = -\log \frac{e^{\text{sim}(Z, Z_{\text{pos}})/\tau}}{e^{\text{sim}(Z, Z_{\text{pos}})/\tau} + \sum_{Z_{\text{neg}}} e^{\text{sim}(Z, Z_{\text{neg}})/\tau}}, \tag{21}$$

where $\mathrm{sim}(u, v)$ denotes a similarity measure between u and v (commonly the cosine similarity, $\mathrm{sim}(u, v) = u^\intercal v$ for ℓ_2-normalized embeddings), and τ is a temperature parameter.

To integrate this into our framework, we augment the optimization objective by including the InfoNCE loss as a regularization term. The resulting objective is given by:

$$\underset{M,p,\theta}{\arg\max}\ J_{\text{neg+vol}} - \lambda_{\text{InfoNCE}} L_{\text{InfoNCE},\theta}, \qquad (22)$$

where θ represents the model's weights or a selected subset thereof, and λ_{InfoNCE} is a hyperparameter. In our experiments, we add one fully connected layer and train its weights under the InfoNCE loss while completely freezing the backbone. This formulation enables the simultaneous optimization of the contrastive conformal set geometry and the underlying feature representations, thereby improving the alignment of the embedding space with the desired inclusion-exclusion properties.

To guarantee the positive inclusion rate as in (6b), we use a held-out calibration set $\mathcal{D}_{\text{cal}}$ to find the final threshold $\hat{q}_\alpha$ using the learned score. We use that threshold in constructing the final test contrastive conformal set.

The preceding discussion leaves us with multiple variants of the method depending on what data the user has access to and what updates they are willing to perform on their model. If the user has access to negative samples, then optimizing the negative exclusion rate directly is the natural choice. Otherwise, volume minimization serves as a proxy, aiming for higher negative exclusion under the specified positive coverage. If the user is willing to update the model's parameters, contrastive learning updates can additionally help produce a geometric representation with smaller covering sets and higher negative exclusion rates.

Remark 2. *Our framework can be applied to any Euclidean semantic feature space, irrespective of the underlying pre-training paradigm. The designation "contrastive" does not require the backbone model itself to be trained via contrastive learning; rather, it describes our core objective of explicitly contrasting positive samples (for inclusion) against negative samples (for exclusion) to construct the covering sets.*

While the primary focus of this work is constructing contrastive conformal sets that guarantee positive sample inclusion while maximizing negative sample exclusion, we also explore their utility beyond strict uncertainty quantification. Specifically, we ask: *can contrastive conformal sets be employed off-the-shelf for downstream tasks without retraining?* To answer this, we evaluate our learned geometric sets on a standard application in robust representation learning: out-of-distribution (OOD) detection.

OOD Detection. Metric sets trained on in-distribution (ID) data are specifically optimized to exclude ID negative samples. However, when these learned parameters are applied to OOD data, the separation between positive and negative samples deteriorates. This breakdown is reflected in a significantly reduced negative exclusion rate when evaluating OOD samples against the fixed, ID-calibrated threshold $\hat{q}_\alpha$. To exploit this phenomenon for OOD detection, we define an anomaly score as:

$$\text{OOD Score} = -\frac{s_n}{\hat{q}_\alpha}, \qquad (23)$$

where s_n is the mean of the negative samples' distances to the anchor under the trained score. For ID data, the high negative exclusion rate enforced during training results in a low anomaly score. In contrast, the diminished exclusion rate for OOD data means smaller distances between the anchor and its negative samples and leads to a higher anomaly score, providing a robust and interpretable criterion for distinguishing OOD samples from ID samples.

5 EXPERIMENTS

We evaluate the proposed learned conformal sets on both simulated data and real-world image datasets. We assess performance based on three primary metrics: the coverage rate of positive samples, the exclusion rate of negative samples, and the log volume of the covering sets. As **no prior work has addressed this formulation**, we establish baselines using vanilla CP procedures utilizing the ℓ_2 norm and the Mahalanobis distance in (9) and (10). To ensure a fair comparison and avoid bias in the Mahalanobis baseline, we split the calibration data equally into disjoint sets: one half is used to estimate the covariance matrix, while the remaining half is used to compute the conformity scores and empirical quantiles. We use the same calibration set to recalibrate the threshold for all methods post training.

Simulated Data. We first perform experiments on 3-dimensional simulated data drawn from various distributions. We compare our approach under three optimization objectives: minimizing only the log volume, maximizing only the approximate negative exclusion term, and optimizing the regularized objective introduced in (20). All simulated experiments are averaged over 20 random seeds.

Image Classification. For real-world evaluation, we utilize the CIFAR100 dataset [Krizhevsky, 2009]. We employ a RepVGG-A2 architecture [Ding et al., 2021], pre-trained in a supervised manner with standard normalization techniques. We use the pre-trained weights obtained by Chen and made available here. We restrict our training to a random fraction of the CIFAR100 training split (see Appendix B). All experiments are averaged over 6 random seeds.

OOD Detection. Furthermore, we evaluate the models' robustness in an OOD detection setting. We treat CIFAR100

as the in-distribution (ID) dataset, and CIFAR-10 and SVHN as the OOD datasets. We compare against commonly used OOD detection baselines: Maximum Softmax Probability (MSP) [Hendrycks and Gimpel, 2017], Energy (with temperature $T = 1$) [Liu et al., 2020], K-Nearest Neighbors (KNN) [Sun et al., 2022], Mahalanobis distance [Lee et al., 2018], COMBOOD [Rajasekaran et al., 2024], D-KNN [Peng et al., 2026], CIDER [Ming et al., 2023], and ViM [Wang et al., 2022]. Note that we perform OOD detection using only augmented data samples without access to labels. More on the implementation details can be found in Appendix B.

Evaluation metrics. *Contrastive Conformal Sets:* We report (1) the coverage rate of positive samples, (2) the exclusion rate of negative samples, and (3) the LogVol of the produced set divided by the feature dimensionality d. *OOD:* We report (1) the false positive rate (FPR95) of OOD samples when the true positive rate of ID samples is at 95% and (2) the AUROC.

Implementation Details. We optimize the learnable parameters using Stochastic Gradient Descent (SGD). During training, we sample k positive and k negative instances without replacement for each anchor to ensure a robust gradient signal ($k = 50$ for CIFAR100, $k = 20$ for STL10 and ImageNet100, and $k = 500$ in simulation). When the InfoNCE loss is included, we jointly optimize the representations by alternating updates between the CCS parameters and the added projection-layer weights at each epoch. For all methods, we perform model selection by evaluating the checkpoint that achieves the highest negative exclusion rate on the validation split. The conformal miscoverage tolerance is strictly fixed at $\alpha = 0.05$. All dataset results are reported as the mean $\pm$ standard deviation across six random seeds, whereas the simulation results are averaged over 20 seeds. Exhaustive details regarding hyperparameters, data splits, baseline set up, and hardware configuration are deferred to Appendix B.

6 RESULTS AND DISCUSSION

We evaluate our proposed methods on simulated 3D data, reporting positive coverage, negative exclusion, and log-volume divided by the feature dimension d (LogVol/d). We also assess downstream OOD detection using CIFAR100 (ID) against CIFAR-10 (near-OOD) and SVHN (far-OOD). Extended metrics for CIFAR100, STL10, and ImageNet100, additional OOD evaluations, simulation implementation details, and a sensitivity analysis are deferred to the appendices (Appendices B, C and C.1). In all tables, the top three results are highlighted as **red**, **blue**, and **cyan**.

Table 1 presents the simulation results comparing the vanilla CP baselines with our proposed learned-geometry methods. Note that the learned-geometry conformal sets achieve substantially higher negative exclusion rates and lower log-volumes (LogVol). For example, the highest negative exclusion rate is achieved by the single-norm objective optimized for volume (Single (Vol)) at $\approx 73.4\%$. This is higher than the Mahalanobis Ellipsoid method's exclusion rate by $\approx 12\%$, which is the strongest of the two vanilla CP baselines in this regard. The other variants including the generalized hyper-ball also outperform the vanilla CP baselines by a large margin in terms of the negative exclusion rates.

The average minimum LogVol/d, on the other hand, is achieved by the generalized hyper-ball method optimized for volume (Generalized (Vol)) at ≈ -4.13. This is a significantly smaller volume compared to the Mahalanobis ellipsoid (LogVol/d ≈ 2.28), which yields the smallest volume among the vanilla baselines. In general, single-norm variants produce higher negative-exclusion rates due to more stable gradient convergence, while generalized hyper-ball variants achieve lower LogVol, likely because the per-dimension exponents p_j can independently compress the bounding volume.

Table 1: Simulated 3D clusters (mixed): $n_{\text{clusters}} = 5$, $n_{\text{per\_cluster}} = 5000$, separation$= 4$, $k = 500$, $\alpha = 0.05$, $n_{\text{test}} = 5000$.

Method	Coverage	Exclusion↑	LogVol/d↓
Conformal Ball (ℓ_2)	$94.7 \pm 0.0\%$	$15.3 \pm 0.1\%$	2.74 ± 0.00
Mahalanobis Ellipsoid	$94.6 \pm 0.0\%$	$61.8 \pm 0.1\%$	2.28 ± 0.00
Single (Neg)	$94.7 \pm 0.1\%$	$66.9 \pm 4.8\%$	5.87 ± 1.52
Generalized (Neg)	$94.9 \pm 0.0\%$	$66.3 \pm 3.4\%$	4.93 ± 4.00
Single (Vol)	$94.7 \pm 0.0\%$	$73.4 \pm 2.2\%$	2.24 ± 0.04
Generalized (Vol)	$94.7 \pm 0.0\%$	$52.6 \pm 2.7\%$	-4.13 ± 0.46
Single (Neg,Vol)	$94.7 \pm 0.1\%$	$70.3 \pm 3.7\%$	3.41 ± 0.31
Generalized (Neg,Vol)	$94.9 \pm 0.0\%$	$66.7 \pm 2.1\%$	3.93 ± 2.76

Table 2 demonstrates similar results for CIFAR100. Most learned-geometry methods produce CCSs with higher negative exclusion rates and smaller volumes compared to naive ℓ_2 norm or Mahalanobis scores. Here the largest margin of $\approx 39.4\%$ is achieved by the generalized hyper-ball optimized for negative exclusion. This emphasizes that the geometry of the data might not always be assumed isotropic. Moreover, learning the geometry of samples' features helps in creating covering sets that are highly contrastive in the objective of (6a) and (6b).

In Table 3, we compare the OOD detection results of our learned-geometry contrastive conformal sets against eight baselines. To utilize our learned-geometry sets for this task, we sample positive and negative instances for the given dataset in an unsupervised manner, construct the covering sets using the learned parameters, and compute an anomaly score defined as in (23).

For CIFAR100 (ID), this score is appropriately low since the exclusion of negatives is high. For CIFAR-10 and SVHN

Table 2: CIFAR100, model=RepVGG-A2, $k = 50$, $\alpha = 0.05$, $n_{\text{test}} = 5000$.

Method	Coverage	Exclusion↑	LogVol/d↓
Conformal Ball (ℓ_2)	$94.9 \pm 0.3\%$	$43.1 \pm 0.8\%$	0.45 ± 0.0
Mahalanobis Ellipsoid	$95.1 \pm 0.4\%$	$16.8 \pm 1.1\%$	-1.88 ± 0.01
Single (Neg)	$94.8 \pm 0.3\%$	$70.3 \pm 2.2\%$	2.96 ± 0.25
Generalized (Neg)	$95.1 \pm 0.4\%$	$82.5 \pm 0.6\%$	-1.30 ± 0.41
Single (Vol)	$95.1 \pm 0.3\%$	$58.0 \pm 1.2\%$	-0.08 ± 0.01
Generalized (Vol)	$95.0 \pm 0.3\%$	$51.0 \pm 0.7\%$	-0.36 ± 0.09
Single (Neg,Vol)	$95.1 \pm 0.3\%$	$69.0 \pm 1.6\%$	0.06 ± 0.02
Generalized (Neg,Vol)	$95.0 \pm 0.3\%$	$63.0 \pm 0.8\%$	-0.16 ± 0.12
Single (Vol,InfoNCE)	$94.9 \pm 0.4\%$	$34.9 \pm 3.4\%$	-0.08 ± 0.32
Generalized (Vol,InfoNCE)	$95.1 \pm 0.2\%$	$39.2 \pm 1.8\%$	-1.83 ± 0.41
Single (Neg,InfoNCE)	$94.7 \pm 0.2\%$	$77.8 \pm 1.1\%$	2.21 ± 0.09
Generalized (Neg,InfoNCE)	$95.0 \pm 0.3\%$	$71.8 \pm 4.8\%$	-0.18 ± 0.34
Single (Neg,Vol,InfoNCE)	$95.0 \pm 0.4\%$	$27.6 \pm 1.0\%$	1.57 ± 0.10
Generalized (Neg,Vol,InfoNCE)	$94.7 \pm 0.4\%$	$62.8 \pm 1.3\%$	-0.38 ± 0.18

(OOD), the exclusion rate of negatives is conversely low (the negative distances fall largely below the ID-calibrated threshold), and hence the anomaly scores are very high. This separation in the derived score makes both learned-geometry variants substantially outperform the baselines across both OOD datasets in AUROC and FPR95.

Table 3: OOD detection on CIFAR100 (ID). Backbone: RepVGG-A2.

Method	SVHN		CIFAR10	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP	78.3 ± 0.1	54.4 ± 0.3	79.7 ± 0.0	56.5 ± 0.1
Energy (T=1)	77.6 ± 0.2	54.1 ± 0.4	79.8 ± 0.1	56.7 ± 0.1
Mahalanobis	81.8 ± 0.0	54.4 ± 0.0	71.4 ± 0.0	72.3 ± 0.0
KNN (ℓ_2)	32.0 ± 1.1	91.8 ± 0.5	20.0 ± 0.7	98.6 ± 0.2
COMBOOD	90.4 ± 0.0	41.5 ± 0.0	78.7 ± 0.0	60.7 ± 0.0
D-KNN	80.5 ± 0.0	55.6 ± 0.0	78.1 ± 0.0	62.6 ± 0.0
CIDER	77.2 ± 0.6	55.5 ± 1.3	77.7 ± 0.1	59.4 ± 0.5
ViM	83.4 ± 0.0	50.5 ± 0.0	72.9 ± 0.0	65.2 ± 0.0
Single (Neg)	99.8 ± 0.1	0.1 ± 0.1	94.9 ± 0.9	30.7 ± 4.8
Generalized (Neg)	99.8 ± 0.0	0.5 ± 0.2	85.9 ± 0.7	52.7 ± 2.8

One could propose using a simple baseline for covering sets by constructing a convex hull using the positive samples and shrinking it until 95% (or any required coverage level) of those samples are included. However, this has a fundamental issue: its coverage guarantee applies only to the positives used to construct the hull, not to held-out positives drawn from $p_{\text{pos}}(\cdot \mid X)$ at test time. In the CCS setting, neither the test point's positive samples nor the positive-sampling mechanism is available. The calibrated threshold must be transferred across anchors, which the hull fails to achieve. Table 4 illustrates this under the same setting of Table 2: the convex hull is constructed in the 8D-PCA space from 25 positives, and coverage is measured on a separate set of 50 positives. For fair comparison, the two CCS variants shown are calibrated on 25 samples and tested with 50 positives per anchor. The hull achieves only $\approx 49\%$ coverage. A similar failure appears when applying uncertainty quantification methods that do not certify coverage. For example, HIB [Oh et al., 2019], shown in the same table, severely underincludes positives. This confirms the gap that CCSs fill by providing guaranteed coverage as formalized in Lemma 1, in contrast to other uncertified uncertainty measures.

Table 4: Positive sample coverage for CIFAR100 with RepVGG-A2. α=0.05.

Method	Coverage
Convex Hull	$49.4 \pm 0.5\%$
HIB	$69.1 \pm 0.2\%$
HIB-MoG	$78.9 \pm 0.1\%$
Single (Neg)	$95.1 \pm 0.4\%$
Generalized (Neg)	$95.0 \pm 0.2\%$

7 CONCLUSION

This work extends the principles of CP to semantic feature spaces by introducing contrastive conformal sets. These sets are designed as minimum-volume, generalized hyper-ball regions that guarantee the inclusion of positive samples with a user-specified probability, while simultaneously maximizing the exclusion of negative samples. The proposed framework is broadly applicable to any Euclidean semantic feature space, including those generated by pretrained contrastive models. These certified covering sets provide a principled mechanism for quantifying the uncertainty of semantic representations (e.g., self-supervised embeddings) and enable robust OOD detection.

Future work will explore leveraging the geometric properties of these certified sets for downstream tasks directly within their boundaries. Additionally, the current formulation is limited to Euclidean spaces and does not address Riemannian manifold geometries or other non-Euclidean spaces. Extending the framework to accommodate such complex geometries represents a promising direction for further research.

ACKNOWLEDGMENT

We thank Benedict Lin for his useful comments on an earlier version of the manuscript.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction*. 2025. URL https://arxiv.org/abs/2411.11824. Forthcoming, Cambridge University Press. arXiv:2411.11824v3.

Shervin Ardeshir and Navid Azizan. Embedding reliability: On the predictability of downstream performance. In *ML Safety Workshop, Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://openreview.net/forum?id=TedqYedIERd. Workshop paper.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Workshop on Computational Learning Theory (COLT)*, COLT '92, pages 144–152, Pittsburgh, PA, USA, July 1992. ACM Press. doi: 10.1145/130385.130401. URL https://doi.org/10.1145/130385.130401.

Sacha Braun, Liviu Aolaritei, Michael I. Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*, 2025. doi: 10.48550/arXiv.2503.19068.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Yaofo Chen. Pytorch cifar models. https://github.com/chenyaofo/pytorch-cifar-models. Accessed: 2026-01-15.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13733–13742, June 2021.

Tai Dinh, Hauchi Wong, Daniil Lisik, Michal Koren, Dat Tran, Philip S. Yu, and Joaquín Torres-Sospedra. Data clustering: a fundamental method in data science and management. *Data Science and Management*, 9(1):100161, 2026. ISSN 2666-7649. doi: 10.1016/j.dsm.2025.08.001. URL https://doi.org/10.1016/j.dsm.2025.08.001.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022. doi: 10.1016/j.jmva.2021.104879.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://projecteuclid.org/journals/annals-of-statistics/volume-51/issue-2/Conformal-prediction-beyond-exchangeability/10.1214/23-AOS2276.full.

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.

Edwin Hewitt and Leonard J. Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955. ISSN 00029947, 10886850. URL http://www.jstor.org/stable/1992999.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen, editors, *Symposium on Conformal and Probabilistic Prediction and Applications (COPA)*, volume 152 of *Proceedings of Machine Learning Research*, pages 72–90. PMLR, 08–10 Sep 2021.

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, 2020.

Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, February 2025. URL https://arxiv.org/abs/2502.03609.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

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.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1):29–43, 2015. doi: 10.1007/s10472-013-9366-6.

Hao Liu and Pieter Abbeel. Hybrid discriminative-generative training via contrastive learning. *arXiv preprint arXiv:2007.09070*, 2020.

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, pages 21464–21475, 2020.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29): 861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Conformal multi-target regression using neural networks. In Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin, editors, *Symposium on Conformal and Probabilistic Prediction and Applications (COPA)*, volume 128 of *Proceedings of Machine Learning Research*, pages 65–83. PMLR, 09–11 Sep 2020. URL https://proceedings.mlr.press/v128/messoudi20a.html.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=aEFaE0W5pAd.

Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *International Conference on Learning Representations (ICLR)*, 2019.

Ningkang Peng, Xiaoqian Peng, Yuhao Zhang, Qianfeng Yu, Feng Xing, Peirong Ma, Xichen Yang, Yi Chen, Tingyu Lu, and Yanhui Gu. Breaking semantic hegemony: Decoupling principal and residual subspaces for generalized ood detection, 2026. URL https://arxiv.org/abs/2602.05360.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Magesh Rajasekaran, Md Saiful Islam Sajol, Frej Berglind, Supratik Mukhopadhyay, and Kamalika Das. *COMBOOD: A Semiparametric Approach for Detecting Out-of-distribution Data for Image Classification*, pages 643–651. 2024. doi: 10.1137/1.9781611978032.74. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611978032.74.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR, pages 815–823. IEEE, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Schroff_FaceNet_A_Unified_2015_CVPR_paper.pdf.

Glenn Shafer and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. Elsevier, 2008. ISBN 9780123985378.

Ambesh Shekhar. Imagenet100: A sample of imagenet classes. Kaggle Dataset, August 2021. URL https://www.kaggle.com/datasets/ambityga/imagenet100.

Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, pages 20827–20840. PMLR, 2022.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive inference with feature conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=0uRm1YmFTu.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-Of-Distribution with Virtual-logit Matching . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4911–4920,

Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00487. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00487.

Xianfu Wang. Volumes of generalized unit balls. *Mathematics Magazine*, 78(5):390–395, 2005. ISSN 0025570X, 19300980. URL http://www.jstor.org/stable/30044198.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, pages 499–515, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Mike Wu and Noah Goodman. A Simple Framework for Uncertainty in Contrastive Learning, October 2020. URL http://arxiv.org/abs/2010.02038.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, June 2020. URL https://arxiv.org/abs/2006.04131.

Contrastive Conformal Sets (Supplementary Material)

Yahya Alkhatib[1] **Wee Peng Tay**[1]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

A PROOF OF THEORETICAL RESULTS AND COMPLEXITY ANALYSIS

This section presents proofs of the theoretical results established in the main paper, alongside a comprehensive computational complexity analysis of the proposed methodology.

A.1 PROOF OF Corollary 1

Let $\mathcal{E}_{\text{pos}} \triangleq \{Z_{\text{pos}} \in \mathcal{N}(Z)\}$, $\mathcal{E}_{\text{neg}} \triangleq \{Z_{\text{neg}} \notin \mathcal{N}(Z)\}$, and define the event that the positive sample shares the anchor's class as $E \triangleq \{Y_{\text{pos}} = Y\}$.

From (6b), we have $\mathbb{P}\big(\mathcal{E}_{\text{pos}}^c\big) \leq \alpha$. Applying the Law of Total Probability, we can expand this as:

$$\begin{aligned}\alpha \geq \mathbb{P}\big(\mathcal{E}_{\text{pos}}^c\big) &= \mathbb{P}(E)\,\mathbb{P}\big(\mathcal{E}_{\text{pos}}^c \,\big|\, E\big) + \mathbb{P}(E^c)\,\mathbb{P}\big(\mathcal{E}_{\text{pos}}^c \,\big|\, E^c\big) \\ &\geq \mathbb{P}(E)\,\mathbb{P}\big(\mathcal{E}_{\text{pos}}^c \,\big|\, E\big). \end{aligned} \tag{24}$$

Rearranging the terms to isolate the conditional probability yields the oracle positive inclusion bound:

$$\mathbb{P}(\mathcal{E}_{\text{pos}} \,|\, E) \geq 1 - \frac{\alpha}{\mathbb{P}(E)} = 1 - \frac{\alpha}{\mathbb{P}(Y_{\text{pos}} = Y)}. \tag{25}$$

Similarly, for the negative exclusion rate, let $E' \triangleq \{Y_{\text{neg}} \neq Y\}$. Given that the exclusion probability satisfies $\mathbb{P}(\mathcal{E}_{\text{neg}}) \leq \beta$, an analogous expansion gives:

$$\beta \geq \mathbb{P}(\mathcal{E}_{\text{neg}}) \geq \mathbb{P}(E')\,\mathbb{P}(\mathcal{E}_{\text{neg}} \,|\, E'), \tag{26}$$

which directly implies the oracle negative exclusion bound:

$$\mathbb{P}\big(\mathcal{E}_{\text{neg}}^c \,\big|\, E'\big) \geq 1 - \frac{\beta}{\mathbb{P}(E')} = 1 - \frac{\beta}{\mathbb{P}(Y_{\text{neg}} \neq Y)}. \tag{27}$$

This completes the proof.

A.2 PROOF OF Lemma 1

Let $V^j = (Z^j, Z^j_{\text{pos}})$ for $j = 1, \ldots, n_{\text{cal}}$ denote the pairs of embeddings corresponding to the calibration anchors and their associated positive samples. Similarly, let $V^{n_{\text{cal}}+1} = (Z, Z_{\text{pos}})$ denote the pair of embeddings for the test anchor and its associated positive sample. It suffices to show that the set of pairs $\{V^1, \ldots, V^{n_{\text{cal}}+1}\}$ is exchangeable. The result then follows from the standard conformal prediction coverage guarantee [Angelopoulos et al., 2025].

Denote $X^{n_{\text{cal}}+1} = X$. From de Finetti's theorem [Hewitt and Savage, 1955], since $(X^1, \ldots, X^{n_{\text{cal}}+1})$ are exchangeable, there exists a latent variable W such that the sequence is conditionally independent and identically distributed (i.i.d.) given W. We therefore have:

$$
\begin{aligned}
&p_{V^1,\ldots,V^{n_{\text{cal}}+1}|W}(v^1, \ldots, v^{n_{\text{cal}}+1} \mid w) \\
&= \int p(v^1, \ldots, v^{n_{\text{cal}}+1} \mid w, x^1, \ldots, x^{n_{\text{cal}}+1}) p(x^1, \ldots, x^{n_{\text{cal}}+1} \mid w)\, \mathrm{d}x^1 \ldots \mathrm{d}x^{n_{\text{cal}}+1} \\
&= \int \prod_{j=1}^{n_{\text{cal}}+1} p(v^j \mid w, x^j) \prod_{j=1}^{n_{\text{cal}}+1} p(x^j \mid w)\, \mathrm{d}x^1 \ldots \mathrm{d}x^{n_{\text{cal}}+1} \\
&= \int \prod_{j=1}^{n_{\text{cal}}+1} p(v^j, x^j \mid w)\, \mathrm{d}x^1 \ldots \mathrm{d}x^{n_{\text{cal}}+1} \\
&= \prod_{j=1}^{n_{\text{cal}}+1} \int p(v^j, x^j \mid w)\, \mathrm{d}x^j \\
&= \prod_{j=1}^{n_{\text{cal}}+1} p(v^j \mid w),
\end{aligned}
$$

where the second equality follows from the fact that given W, the anchors $(X^1, \ldots, X^{n_{\text{cal}}+1})$ are independent and the generation mechanism of X^j_{pos} depends only on X^j. Therefore, the sequence of pairs $(V^1, \ldots, V^{n_{\text{cal}}+1})$ is conditionally i.i.d. given W. This implies that the sequence is exchangeable, which completes the proof.

A.3 PROOF OF Proposition 1

Recall that $Z_{\text{pos}} = \theta(X_{\text{pos}})$ and $Z_{\text{neg}} = \theta(X_{\text{neg}})$, where X_{pos} and X_{neg} are the positive and negative samples associated with X under the mechanisms $p_{\text{pos}}(\cdot \mid X)$ and $p_{\text{neg}}(\cdot \mid X)$, respectively. For $t \in T$, define

$$
p_+(t) \triangleq \mathbb{P}(Z_{\text{pos}} \in \mathcal{B}_t), \qquad p_-(t) \triangleq \mathbb{P}(Z_{\text{neg}} \in \mathcal{B}_t), \qquad q_-(t) \triangleq \mathbb{P}(Z_{\text{neg}} \notin \mathcal{B}_t) = 1 - p_-(t).
$$

By the nesting assumption, if $t_1 \leq t_2$ then

$$
\mathcal{B}_{t_1} \subseteq \mathcal{B}_{t_2},
$$

hence

$$
\{Z_{\text{pos}} \in \mathcal{B}_{t_1}\} \subseteq \{Z_{\text{pos}} \in \mathcal{B}_{t_2}\}, \qquad \{Z_{\text{neg}} \in \mathcal{B}_{t_1}\} \subseteq \{Z_{\text{neg}} \in \mathcal{B}_{t_2}\}.
$$

By monotonicity of probability,

$$
p_+(t_1) \leq p_+(t_2) \quad \text{and} \quad p_-(t_1) \leq p_-(t_2),
$$

so p_+ and p_- are non-decreasing in t. Consequently,

$$
q_-(t) = 1 - p_-(t)
$$

is non-increasing in t. Next, consider the set of indices for which the positive inclusion requirement is satisfied:

$$
T_\alpha \triangleq \big\{t \in T : p_+(t) \geq 1 - \alpha\big\}.
$$

Any $t \in T_\alpha$ corresponds to a set $\mathcal{B}_t$ satisfying (6b). Since p_+ is non-decreasing, T_α is either empty or an upper set in T (if $t_0 \in T_\alpha$ and $t \geq t_0$, then $t \in T_\alpha$). Assume $T_\alpha \neq \varnothing$ and let t^* be any solution of (18). In particular, $t^* \in T_\alpha$, so $\mathcal{B}^*_\alpha = \mathcal{B}_{t^*}$ satisfies the positive inclusion requirement (6b). Now take any $t \in T_\alpha$. By the definition of t^* as a minimizer of $\mathrm{Vol}(\mathcal{B}_t)$ over T_α, and by the monotonicity of $\mathrm{Vol}(\mathcal{B}_t)$ in t, we must have $t \geq t^*$. Since $q_-(t)$ is non-increasing in t, it follows that

$$
q_-(t) \leq q_-(t^*) \qquad \text{for all } t \in T_\alpha.
$$

Equivalently,

$$
\mathbb{P}(Z_{\text{neg}} \notin \mathcal{B}_t) \leq \mathbb{P}(Z_{\text{neg}} \notin \mathcal{B}_{t^*}) \qquad \forall\, t \in T_\alpha.
$$

Thus, among all sets $\mathcal{B}_t$ that satisfy the positive inclusion constraint, $\mathcal{B}^*_\alpha$ attains the largest negative exclusion probability (6a). Finally, by construction in (18), $\mathcal{B}^*_\alpha$ also minimizes $\mathrm{Vol}(\mathcal{B}_t)$ over T_α, i.e. it has minimum volume among all sets in the family $\{\mathcal{B}_t : t \in T\}$ that satisfy (6b). This proves the claim.

A.4 COMPLEXITY ANALYSIS

In this analysis, we assume semantic feature vectors of dimensionality d.

Vanilla CP Complexity. Assume a calibration set $\mathcal{D}_{\text{cal}}$ of size $n_{\text{cal}} \gg d$. For the vanilla CP baselines, we compute the positive distances for all calibration points and extract the empirical $\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil / n_{\text{cal}}$ quantile. In the ℓ_p norm case, computing the d-dimensional differences takes $\mathcal{O}(n_{\text{cal}} d)$, and sorting them to find the quantile takes $\mathcal{O}(n_{\text{cal}} \log n_{\text{cal}})$. Thus, the overall complexity is $\mathcal{O}(n_{\text{cal}}(d + \log n_{\text{cal}}))$. For the Mahalanobis distance baseline, we must first estimate the $d \times d$ covariance matrix, which requires $\mathcal{O}(n_{\text{cal}} d^2)$ operations, and then invert it, taking $\mathcal{O}(d^3)$ operations. Computing the distances for the calibration points then takes an additional $\mathcal{O}(n_{\text{cal}} d^2)$. Since $n_{\text{cal}} \gg d$, the covariance estimation and distance computations dominate the matrix inversion, resulting in a total complexity of $\mathcal{O}(n_{\text{cal}} d^2 + n_{\text{cal}} \log n_{\text{cal}})$.

Learnable-Score Conformal Sets Complexity. Assume a training set $\mathcal{D}_{\text{train}}$ of size $n_{\text{train}} \gg d$ and a training duration of k epochs. For the single-norm sets, computing the distances requires a matrix-vector multiplication for each sample, incurring a cost of $\mathcal{O}(n_{\text{train}} d^2)$ per epoch. Additionally, calculating the empirical $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the positive distances requires sorting at each epoch, which adds $\mathcal{O}(n_{\text{train}} \log n_{\text{train}})$. Thus, over k epochs, the total training complexity is $\mathcal{O}(k n_{\text{train}}(d^2 + \log n_{\text{train}}))$. Because the squared dimensionality d^2 is typically large in deep learning semantic spaces, the matrix multiplication term $\mathcal{O}(k n_{\text{train}} d^2)$ dominates the sorting term. Conversely, the generalized hyper-ball sets only require element-wise scaling and power operations, reducing the distance computation per epoch to $\mathcal{O}(n_{\text{train}} d)$. The overall training complexity for the generalized hyper-ball is therefore bounded by $\mathcal{O}(k n_{\text{train}}(d + \log n_{\text{train}}))$, making it significantly more computationally efficient than both the single-norm approach and the vanilla Mahalanobis baseline.

B REPRODUCIBILITY DETAILS

3D Simulation Setup (Mixed Distribution). To evaluate the learned covering sets on complex, non-spherical geometries, we simulate a 3D dataset ($d = 3$) comprising 5 distinct classes, each containing 5,000 points. The cluster centers are uniformly distributed along a circle of radius 4 with a slight sinusoidal elevation in the z-axis. To create the "mixed" geometry, the first two classes are generated as anisotropic Gaussians by stretching a random covariance matrix along a randomly selected axis by a factor of $U(3, 5)$. The remaining three classes are generated as "banana" (crescent) distributions by sampling points along 3D arcs (base radius 1.5), applying random orthogonal rotations, and injecting isotropic Gaussian noise ($\sigma = 1.0$). See Fig. 2.

For the experimental pipeline, we bypass the neural encoder and optimize the set parameters directly on the 3D coordinates. The dataset is partitioned into 60% for training and 40% for conformal calibration and testing (split equally). During training, we sample $k = 500$ positive and negative instances per anchor. The set parameters are optimized for 150 epochs using SGD with a learning rate of 0.3, a batch size of 256, and gradient clipping at 1.0. For methods incorporating volume regularization, the balancing weight is set to $\lambda = 0.001$. The simulated experiments are averaged over 20 seeds (0-19), and the data is generated in the first seed and then held fixed in the following seeds.

Datasets and Pretrained Models. We conduct experiments on CIFAR100 [Krizhevsky, 2009], accessed directly via `torchvision`, and ImageNet100 [Shekhar, 2021], a 100-class subset of ImageNet-1K available via Kaggle, as well as STL10, a partially labeled image dataset with most training data unlabeled. All experiments are averaged over 6 random seeds (42–47). For our backbones, we utilize a RepVGG-A2 model with pretrained weights downloaded from this repository[1], a ResNet-18 model pretrained on ImageNet-1K (loaded via `torchvision`), and a ResNet18-SimCLR model trained using SimCLR on partially labeled STL10.

Data Splitting and Calibration. To learn the conformal sets, we utilize a random 20% subset of the CIFAR100 and ImageNet100 training split, while using the full labeled training split for STL10, although we do not use the label information (this split consists of only 5000 images, therefore it is already a small training split). For the conformal calibration phase (i.e., determining the quantile threshold $\hat{q}_\alpha$), we use 50% of the validation/test splits, reserving the remaining 50% (disjoint) strictly for final evaluation. For all variants of our method, we sample k positive augmented samples and k negative samples from batch instances ($k = 50$ for CIFAR100 and $k = 20$ for STL10 and ImageNet100), adhering to the unsupervised setting. The conformal miscoverage tolerance is fixed at $\alpha = 0.05$ across all experiments, following standard practice.

[1] https://github.com/chenyaofo/pytorch-cifar-models

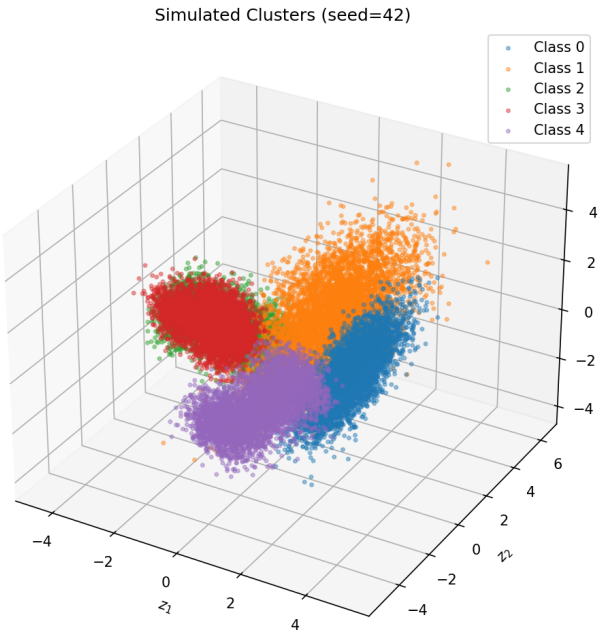

Figure 2: 3D Simulation Mixed Data Distribution

Optimization and Hyperparameters. We train the conformal sets using Stochastic Gradient Descent (SGD) with a learning rate of 0.1 for all three datasets, a momentum of 0.9, and a weight decay of 5×10^{-4}, annealed using a `CosineAnnealingLR` scheduler. The batch size is set to 256. To stabilize training, gradients are clipped at a maximum norm of 1.0, and the sigmoid temperature for the negative exclusion loss is fixed at 7.0. We train all variants for 150 epochs. When InfoNCE is used, we restrict the model updates to an additional fully connected layer with the same feature dimension. For methods incorporating volume regularization (Neg,Vol), the balancing weight is set to $\lambda = 0.001$, with the negative objective weighted by $1 - \lambda$ as in (20). The InfoNCE weight is $\lambda_{\text{InfoNCE}} = 1.0$, with the InfoNCE temperature fixed at $\tau = 0.1$. The exponents p are clamped to $[0.1, 10]$ and the scales m from below at 10^{-3}, as mentioned under Remark 1. Regardless of the specific optimization objective, we always checkpoint and evaluate the model that achieves the highest negative exclusion rate on the validation split, with early stopping if no improvement on the negative exclusion rate is observed for 8 consecutive epochs. The positive and negative sampling is performed only once in each seed and fixed during training.

Baseline Configurations. For the out-of-distribution (OOD) detection baselines, the Energy score temperature is fixed at $T = 1.0$. For the remaining baselines, we follow the hyperparameter configurations recommended in their respective original papers:

- **COMBOOD:** $k = 50$, $C = 1.0$. We extract the maximum activations exclusively from the ReLU layers, as these constitute the vast majority of activations in both ResNet-18 and RepVGG-A2 architectures.
- **D-KNN:** $k = 50$, $d = 0.95$, $\alpha = 0.5$.
- **CIDER:** Projection dimension $= 128$, epochs $= 100$, learning rate $= 0.5$, $\tau = 0.1$, $\lambda_c = 2.0$, $\alpha = 0.999$. To ensure a fair comparison with our method, we freeze the backbone and only train the projection head, as the other methods do not train the backbone.

Hardware Configuration. Experiments were distributed across four NVIDIA GeForce RTX 3090 GPUs using `torch.nn.DataParallel`.

Code Availability. To ensure full reproducibility, the complete source code, including data preprocessing scripts, training routines, and evaluation protocols, can be found on this GitHub repo[2].

[2]https://github.com/Y-kht/ccs_official

C MORE RESULTS

Tables 5 and 6 report coverage, exclusion, and LogVol/d results on the image benchmarking datasets ImageNet100 and STL10 using the pretrained models ResNet18 (ImageNet-1K) and ResNet18-SimCLR, respectively. We can see a somewhat similar trend to the simulation data. Several learned-geometry methods surpass the two vanilla CP baselines in negative exclusion, LogVol, or both. This suggests that by optimizing directly for the negative exclusion rate or the volume of the conformal covering sets, one can achieve a superior balance of the two objectives. Note that optimizing the generalized hyper-ball score yields relatively smaller LogVol/d for the resulting sets. As introduced in our methodology, this volume reduction is likely driven by the optimizer increasing the dimension-wise exponents p_j. As $p_j \gg 2$, the generalized hyper-ball sets transition from smooth, spherical boundaries to sharper, hyper-rectangular geometries. Combined with dimension-specific scaling, this allows the sets to tightly pack around the data distribution, discarding the wasted empty volume typical of smooth ellipsoids.

Table 5: STL10, model=ResNet18-SimCLR, $k = 20$, $\alpha = 0.05$, $n_{\text{test}} = 4000$.

Method	**Coverage**	**Exclusion ↑**	**LogVol/d ↓**
Conformal Ball (ℓ_2)	$94.9 \pm 0.4\%$	$27.0 \pm 1.3\%$	1.52 ± 0.01
Mahalanobis Ellipsoid	$94.9 \pm 0.3\%$	$87.1 \pm 0.6\%$	0.02 ± 0.0
Single (Neg)	$95.0 \pm 0.4\%$	$44.7 \pm 2.9\%$	4.09 ± 0.03
Generalized (Neg)	$95.1 \pm 0.2\%$	$59.0 \pm 0.7\%$	3.05 ± 0.22
Single (Vol)	$95.0 \pm 0.5\%$	$22.5 \pm 1.4\%$	1.32 ± 0.02
Generalized (Vol)	$95.0 \pm 0.5\%$	$29.2 \pm 2.3\%$	0.81 ± 0.23
Single (Neg,Vol)	$95.0 \pm 0.6\%$	$35.9 \pm 3.6\%$	2.02 ± 0.36
Generalized (Neg,Vol)	$95.0 \pm 0.5\%$	$50.9 \pm 2.1\%$	1.48 ± 0.04
Single (Vol,InfoNCE)	$94.8 \pm 0.4\%$	$40.5 \pm 3.2\%$	0.54 ± 0.22
Generalized (Vol,InfoNCE)	$94.9 \pm 0.3\%$	$44.8 \pm 3.3\%$	-0.85 ± 0.97
Single (Neg,InfoNCE)	$95.0 \pm 0.3\%$	$66.1 \pm 6.3\%$	2.19 ± 0.10
Generalized (Neg,InfoNCE)	$95.2 \pm 0.5\%$	$89.8 \pm 1.3\%$	-0.07 ± 0.04
Single (Neg,Vol,InfoNCE)	$95.1 \pm 0.4\%$	$64.1 \pm 9.0\%$	1.23 ± 0.35
Generalized (Neg,Vol,InfoNCE)	$95.1 \pm 0.4\%$	$85.5 \pm 4.1\%$	-0.23 ± 0.14

Table 6: IMAGENET100, model=ResNet18, $k = 20$, $\alpha = 0.05$, $n_{\text{test}} = 2500$.

Method	**Coverage**	**Exclusion ↑**	**LogVol/d ↓**
Conformal Ball (ℓ_2)	$95.1 \pm 0.7\%$	$65.6 \pm 1.7\%$	1.60 ± 0.0
Mahalanobis Ellipsoid	$94.9 \pm 0.6\%$	$76.3 \pm 1.7\%$	1.08 ± 0.01
Single (Neg)	$94.9 \pm 0.5\%$	$75.3 \pm 3.4\%$	11.00 ± 3.85
Generalized (Neg)	$95.2 \pm 0.5\%$	$90.2 \pm 0.5\%$	-1.91 ± 4.13
Single (Vol)	$95.3 \pm 0.4\%$	$81.4 \pm 0.8\%$	1.46 ± 0.01
Generalized (Vol)	$95.2 \pm 0.3\%$	$62.4 \pm 1.1\%$	-0.71 ± 0.02
Single (Neg,Vol)	$95.1 \pm 0.2\%$	$85.6 \pm 1.3\%$	1.65 ± 0.03
Generalized (Neg,Vol)	$95.2 \pm 0.4\%$	$85.7 \pm 0.5\%$	0.68 ± 0.17
Single (Vol,InfoNCE)	$94.9 \pm 0.5\%$	$69.8 \pm 1.8\%$	0.67 ± 0.06
Generalized (Vol,InfoNCE)	$95.0 \pm 0.4\%$	$76.9 \pm 3.1\%$	-5.16 ± 1.03
Single (Neg,InfoNCE)	$95.2 \pm 0.5\%$	$96.3 \pm 0.2\%$	2.77 ± 0.23
Generalized (Neg,InfoNCE)	$95.0 \pm 0.4\%$	$96.7 \pm 0.2\%$	-1.83 ± 0.76
Single (Neg,Vol,InfoNCE)	$95.2 \pm 0.4\%$	$95.8 \pm 0.2\%$	0.60 ± 0.21
Generalized (Neg,Vol,InfoNCE)	$94.9 \pm 0.2\%$	$96.3 \pm 0.1\%$	0.40 ± 0.18

Table 7 reports the OOD detection results on STL10 (ID) against SVHN (far-OOD) and CIFAR10 and CIFAR100 (near-OOD). These results demonstrate that the learned contrastive conformal sets can be effectively utilized for anomaly detection even on highly complex image distributions. Furthermore, the utilized variants are directly optimized for negative exclusion and do not require InfoNCE training, confirming that CCSs can be readily used as an overlay construction for the OOD detection objective without label information.

Table 7: OOD detection on STL10 (ID). Backbone: ResNet18-SimCLR.

Method	**SVHN**		**CIFAR10**		**CIFAR100**	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP	71.7 ± 0.5	59.4 ± 0.9	71.7 ± 0.1	66.3 ± 0.3	69.7 ± 0.2	68.5 ± 0.4
Energy (T=1)	70.5 ± 1.9	58.1 ± 1.7	70.4 ± 1.2	66.0 ± 1.0	69.0 ± 1.2	69.1 ± 1.4
Mahalanobis	61.6 ± 0.0	76.2 ± 0.0	48.7 ± 0.0	89.5 ± 0.0	49.8 ± 0.0	88.1 ± 0.0
KNN (ℓ_2)	57.2 ± 1.3	76.8 ± 1.8	57.2 ± 0.6	85.2 ± 0.5	53.7 ± 0.7	88.1 ± 0.9
COMBOOD	96.6 ± 0.0	18.3 ± 0.0	83.8 ± 0.0	49.9 ± 0.0	84.9 ± 0.0	49.1 ± 0.0
D-KNN	90.1 ± 0.0	29.9 ± 0.0	76.9 ± 0.0	56.3 ± 0.0	78.6 ± 0.0	53.8 ± 0.0
CIDER	67.1 ± 2.7	67.6 ± 2.9	71.1 ± 3.1	70.4 ± 3.4	68.4 ± 0.7	73.0 ± 1.1
Single (Neg)	96.2 ± 0.7	23.0 ± 6.9	90.1 ± 2.0	54.5 ± 5.4	92.3 ± 1.6	47.5 ± 6.3
Generalized (Neg)	99.2 ± 0.3	2.7 ± 1.7	89.9 ± 1.7	50.1 ± 5.5	92.0 ± 1.6	41.8 ± 6.7

C.1 SENSITIVITY ANALYSIS

We present a sensitivity analysis for a selected variant of our method in Fig. 3. Specifically, we evaluate the generalized hyper-ball trained using the joint objective defined in (22). We systematically vary the miscoverage tolerance α, the steepness of the sigmoid surrogate T, the contrastive loss weight λ_{InfoNCE}, the InfoNCE temperature τ, the volume regularization constant λ, and the number of sampled instances k. Unless otherwise specified, the default parameters are $\alpha = 0.05$, $T = 7.0$, $\lambda_{\text{InfoNCE}} = 1.0$, $\tau = 0.1$, $\lambda = 0.05$, and $k = 20$.

As expected, increasing α strictly decreases the target positive inclusion probability $(1 - \alpha)$, resulting in a linear decline in empirical coverage. Correspondingly, a larger α yields a smaller quantile threshold $\hat{q}_\alpha$, shrinking the volume of the covering sets. Following Proposition 1, this reduced volume naturally translates to an increased negative exclusion rate. Across all other parameter sweeps, the empirical coverage remains stably bounded near 0.95, validating the distribution-free guarantee established in Lemma 1.

Modifying the steepness parameter T of the sigmoid approximation does not significantly impact the exclusion rate, while the set volume varies non-monotonically with T. Although sharper sigmoids approximate the hard indicator function more tightly, extreme values can induce vanishing gradients around the threshold boundary, potentially destabilizing optimization.

The coefficient λ_{InfoNCE} controls the influence of the contrastive loss. A stronger contrastive penalty enforces better semantic separation, leading to higher negative exclusion. However, the set volume fluctuates under this parameter: excessively high contrastive weights overshadow the volume regularization term in (22), resulting in larger volumes. Conversely, if λ_{InfoNCE} is too small, the encoder struggles to cluster positive samples tightly, forcing the conformal sets to expand to satisfy the required coverage level.

A higher InfoNCE temperature τ sharpens the separation between positive and negative samples, directly improving the exclusion rate. On the other hand, increasing the volume regularization constant λ inherently reduces the relative weight of the negative exclusion term in $J_{\text{neg+vol}}$, yielding a slightly lower exclusion rate without a consistent reduction in volume.

Finally, the exclusion rate demonstrates remarkably low sensitivity to the number of generated positive and negative samples, k. The positive-coverage guarantee in Lemma 1 is dimension-free. The calibration step ensures $1 - \alpha$ marginal coverage regardless of d, which is verified empirically. The negative exclusion rate, by contrast, is computed empirically over a finite negative sample. Thus, its approximation is expected to be affected by the curse of dimensionality: as d grows, the negatives concentrate on a narrow band of directions in score space, while the volume formula still integrates over all directions. The two can therefore decouple, which is exactly what Fig. 3 and Table 2 show: coverage and exclusion remain stable while LogVol swings widely along directions where no negative mass lies. This does not contradict the coverage guarantee (which

is marginal on the positive side), but it is a real limitation: faithful exclusion estimation likely requires a number of negatives that grows with d, and we leave analysing that to future work.

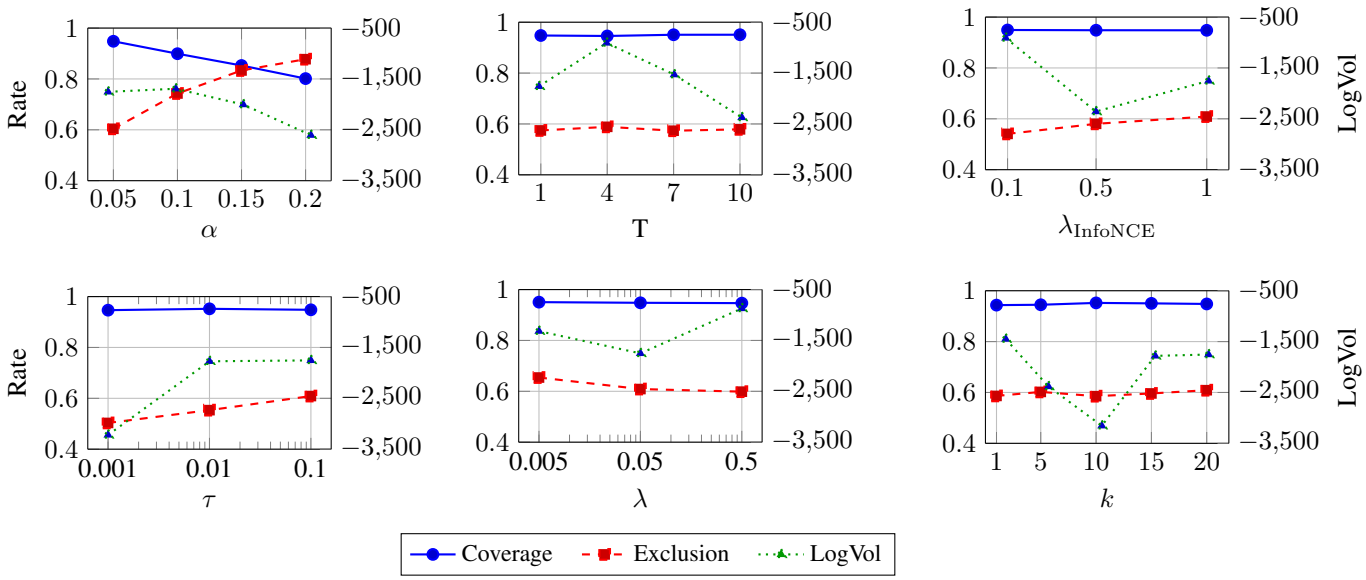

Figure 3: Sensitivity analysis results. (Top): Sweeping α, steepness of sigmoid T, and λ_{InfoNCE}. (Bottom): Sweeping InfoNCE temperature τ, regularization constant λ, and number of generated samples k. Coverage and Exclusion rates are plotted against the left axis, while Log Volume is plotted against the right axis.

Table 8 shows the coverage level transfer from augmentation-based CCSs to class-consistent positive samples under the setting of Table 2. We notice that although we construct the CCS sets using augmentation positive samples, the coverage level does not generally deviate significantly on class-consistent samples. However, CCSs are geometric rather than semantic by construction. If the sample generation mechanisms are close to the supervised ones, then we expect the CCSs to generalize well to class-consistent samples under Corollary 1.

Table 8: Coverage transfer from augmentation samples to class-consistent samples in CIFAR100 with RepVGG-A2, $k = 50$, $\alpha = 0.05$, $n_{\text{test}} = 5000$.

Method	**Augmentation Coverage**	**Class-Consistent Coverage**
Conformal Ball (ℓ_2)	$94.9 \pm 0.3\%$	$98.5 \pm 0.1\%$
Mahalanobis Ellipsoid	$95.1 \pm 0.4\%$	$91.5 \pm 0.5\%$
Single (Neg)	$94.8 \pm 0.3\%$	$94.9 \pm 0.9\%$
Generalized (Neg)	$95.1 \pm 0.4\%$	$88.4 \pm 0.3\%$
Single (Vol)	$95.1 \pm 0.3\%$	$93.5 \pm 0.4\%$
Generalized (Vol)	$95.0 \pm 0.3\%$	$97.5 \pm 0.1\%$
Single (Neg,Vol)	$95.1 \pm 0.3\%$	$92.9 \pm 0.6\%$
Generalized (Neg,Vol)	$95.0 \pm 0.3\%$	$95.8 \pm 0.1\%$
Single (Vol,InfoNCE)	$94.9 \pm 0.4\%$	$97.1 \pm 0.4\%$
Generalized (Vol,InfoNCE)	$95.1 \pm 0.2\%$	$97.2 \pm 0.2\%$
Single (Neg,InfoNCE)	$94.7 \pm 0.2\%$	$93.0 \pm 0.3\%$
Generalized (Neg,InfoNCE)	$95.0 \pm 0.3\%$	$94.2 \pm 1.8\%$
Single (Neg,Vol,InfoNCE)	$95.0 \pm 0.4\%$	$96.0 \pm 0.8\%$
Generalized (Neg,Vol,InfoNCE)	$94.7 \pm 0.4\%$	$95.8 \pm 0.4\%$

We next compare an anisotropic set with an isotropic set to see how the exclusion rate is influenced by the shape in Table 9. An isotropic-vs-anisotropic comparison must be controlled carefully. We can only fix one of {volume, coverage}: fixing coverage at $1 - \alpha$ lets the volume vary, while fixing the volume forces a different threshold and breaks the calibrated coverage. We performed both on the simulated data. At fixed coverage, the anisotropic shape attains a smaller volume

and higher negative exclusion than the isotropic one. At fixed volume, the anisotropic sets become conservative (higher coverage) and exclusion is non-monotone in p, with the most anisotropic shapes ($p = 0.5$) excluding the most negatives (see Table 9). On CIFAR100 (not shown) the pattern reversed: at fixed coverage exclusion decreased with anisotropy, while at fixed volume the trend was again non-uniform. This suggests the optimal anisotropy depends on the local geometry of the embedding distribution.

Table 9: Effect of the p-norm on coverage, exclusion, threshold, and log-volume. The left panel fixes the log-volume at 7.30 and reports the resulting coverage, exclusion, and threshold; the right panel instead fixes the target coverage at 95.0% and reports the resulting exclusion and log-volume.

	Fixed Log-Volume (7.30)			Fixed Coverage (95.0%)	
p	Coverage	Exclusion	Threshold	Exclusion	LogVol
2	95.0%	14.5%	7.0673	14.5%	7.30
1	96.2%	11.9%	10.3508	17.6%	7.04
0.9	96.2%	11.9%	11.3687	18.1%	7.02
0.8	96.3%	12.1%	12.8215	18.6%	7.02
0.7	96.3%	12.6%	15.0267	19.2%	7.02
0.6	96.2%	13.5%	18.6739	19.7%	7.04
0.5	96.0%	15.1%	25.5272	20.3%	7.09