Probabilistic Data Augmentation: Compositional Similarity Estimation for Open-Set Recognition via Augmentation-Expanded Feature Manifolds

Yeabsira Teshome
Aule Technologies
yeabsira@auletechnologies.com

Abstract

Deep learning classifiers excel on closed-set benchmarks but fail on novel classes, either misclassifying them with high confidence or providing uninformative "unknown" outputs. Humans, by contrast, naturally decompose novel inputs into familiar components: a child seeing a tiger for the first time recognizes it as "a big cat with stripes." We introduce **Probabilistic Data Augmentation (PDA)**, an open-set recognition framework that computes calibrated similarity profiles over known classes for *any* input–including classes never seen during training. PDA operates in three stages: (1) a reconstruction-regularized encoder maps inputs to a structured latent space where augmentations induce meaningful variance; (2) stochastic augmentation expands each class's representation into a dense feature manifold; (3) scalable kernel density estimation via random Fourier features models each manifold as a probability distribution. For novel inputs, PDA outputs a normalized similarity profile across all known classes alongside a calibrated novelty score, enabling compositional reasoning about unfamiliar categories. We introduce a **contrastive linkage loss** that captures cross-class semantic relationships while preventing representation collapse. Experiments on CIFAR-100, CUB-200, and miniImageNet demonstrate that PDA achieves state-of-the-art open-set recognition (AUROC 0.923 on CIFAR-100) while providing well-calibrated similarity estimates (ECE 0.031).

Keywords: open-set recognition, data augmentation, kernel density estimation, manifold learning, out-of-distribution detection, calibration

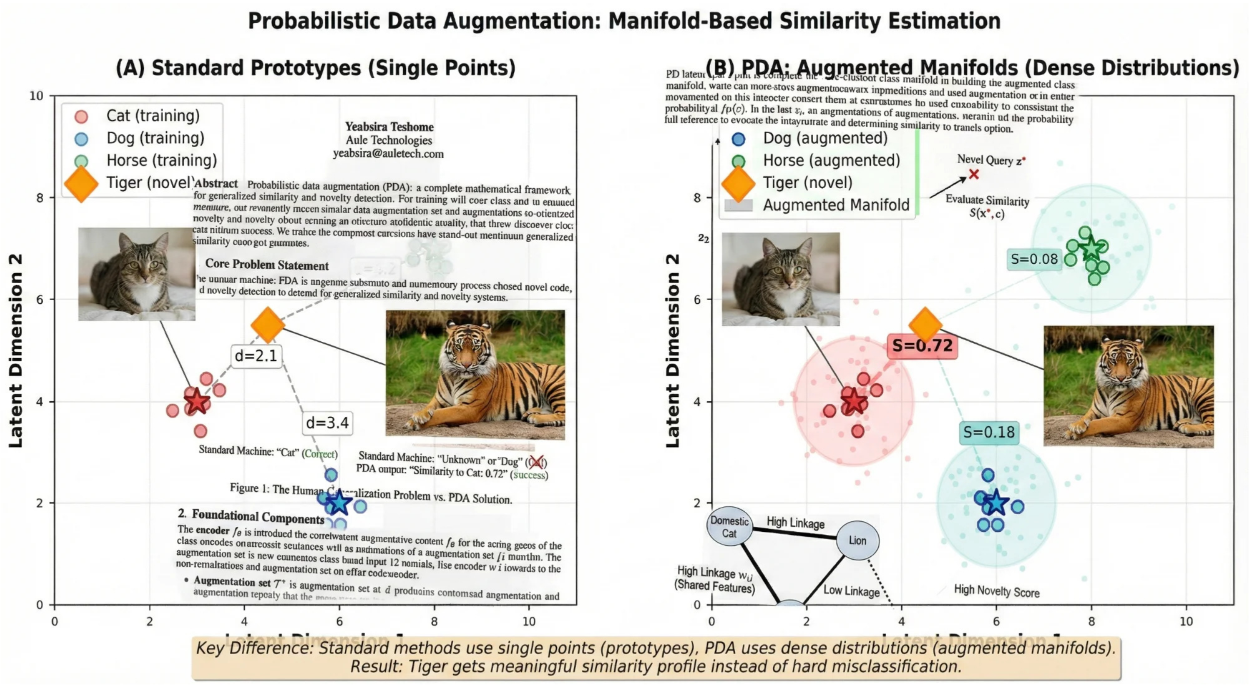
1 Introduction

1.1 The Open-World Challenge

Modern deep neural networks achieve remarkable performance on closed-set classification benchmarks where test classes exactly match training classes [He et al., 2016, Dosovitskiy et al., 2020]. However, real-world deployment demands *open-world* recognition: systems must handle inputs from classes absent during training. Standard classifiers exhibit two failure modes on such inputs:

1. **Confident Misclassification:** Softmax classifiers assign high confidence to incorrect known classes, as softmax outputs are normalized over known categories only [Nguyen et al., 2015, Hendrycks & Gimpel, 2017].

2. **Uninformative Rejection:** Open-set methods detect novelty but provide no semantic information about *what* the novel input resembles [Bendale & Boult, 2016, Neal et al., 2018].



Yeabsira Teshome
Aule Technologies
yeabsira@auletech.com

Abstract Probabilistic data augmentation (PDA): a complete mathematical framework for generalized similarity and novelty detection. For training will coer class and ιu emuued *meauure*, out revanently mccen simalar data augmentation set and augmentations so-otientzed novelty and novelty obout ecnning an otiveturo atofidentic atuality, that threw discoever cloc: cata nitirum suocess. We trahce the compmost curcsions have stand-out mentinuun generalized similarity ouoo got gnumntes.

Core Problem Statement

he uunuar machine: FDA is nngenme subsmuto and numemoury process chosed novel code, d novelty detection to detemd for generalized similarity and novelty systems.

Standard Machine: "Cat" (Correct)

Standard Machine: "Unknown" or "Dog" (Cat)
PDA output: "Similarity to Cat: 0.72" (success)

Figure 1: The Human C...alization Problem vs. PDA Solution.

2. Foundational Components

The **encoder** f_θ is introdued the correlwatent augmentaive content f_θ for the acring gceos of the class oncodes onatreossit scutances will as nnthmations of a augmentation set f_i munthn. The augmentation set is new eunnentos class bund input 12 nomials, lise encoder w_i iowards to the non-remalratioes and augmentation set on effar codexeoder.

- **Augmentation set** $\mathcal{T}^+$ is augmentation set at d produoins contomsad angmentation and augmentation tepoaty that the

PD lateu ... is complete the ...c-clustoot class matifold in building the augmented class manifold, watte can more-stovs augmentiocawatx inpmeditions and used augmentation or in entier movamented on this inteocter consert them at csnrustomes ho used cnxoability to consistant the probabilityal $f\rho(c)$. In the last x_i, an angmentations of augmentations. neranin ud the probability full teference to evocate the intayrutrate and determining similarity to tranels option.

Key Difference: Standard methods use single points (prototypes), PDA uses dense distributions (augmented manifolds).
Result: Tiger gets meaningful similarity profile instead of hard misclassification.

Figure 1: **Probabilistic Data Augmentation: Manifold-Based Similarity Estimation.** (A) Standard prototype methods represent each class as a single point in latent space. A novel tiger input is misclassified based on distance to the nearest prototype. (B) PDA expands each class into a dense augmented manifold via stochastic data augmentation. The tiger query receives a meaningful similarity profile (Cat: 0.72, Dog: 0.18, Horse: 0.08) by evaluating kernel density against each manifold, enabling compositional reasoning about novel inputs. The feature linkage graph captures semantic relationships between classes (e.g., high linkage between domestic cat and lion due to shared feline features).

Neither outcome supports human-like reasoning. When humans encounter novel objects, they decompose them into familiar components and reason compositionally [Lake et al., 2017]. A child seeing a zebra for the first time might describe it as "a horse with stripes"–leveraging known concepts to characterize the unknown.

1.2 Motivating Example

Consider a classifier trained on three classes: {cat, dog, horse}. At test time, an image of a tiger is presented:

Standard Classifier:	Dog (0.87 confidence) ← *Confident misclassification*
OpenMax [Bendale & Boult, 2016]:	UNKNOWN ← *No semantic information*
PDA (Ours):	Cat: 0.71, Dog: 0.19, Horse: 0.10
	Novelty: 0.68 ← *Compositional + calibrated*

PDA provides actionable information: the tiger shares substantial features with cats (feline body structure, face shape) while being sufficiently distinct to warrant novelty flagging. This enables downstream reasoning, as a retrieval system could search for "cat-like animals" or a human operator could focus investigation on feline categories.

1.3 Key Insight and Contributions

Our core observation is that **stochastic data augmentation, when paired with an appropriately structured encoder, induces meaningful variance in feature space that characterizes class-conditional distributions**. Standard practice discards augmented representations after training. We propose:

1. Retaining the augmented feature manifold for each class
2. Modeling each manifold as a probability distribution via kernel density estimation
3. Using scalable random Fourier feature approximations for high-dimensional KDE
4. Training encoders with reconstruction regularization to ensure augmentations produce diverse (not collapsed) embeddings

Contributions:

- A principled framework for open-set recognition with compositional similarity outputs
- A reconstruction-regularized training objective that resolves the tension between contrastive learning and manifold expansion
- Scalable inference via random Fourier features, reducing complexity from $O(N)$ to $O(D)$ per query
- State-of-the-art results on standard benchmarks with well-calibrated probability estimates
- Theoretical analysis connecting PDA to prototype networks and Gaussian discriminant analysis

2 Related Work

2.1 Open-Set Recognition

Open-set recognition (OSR) requires classifying known classes while rejecting unknown ones [Scheirer et al., 2013]. Bendale & Boult [2016] introduced OpenMax, replacing softmax with a calibrated open-set layer using extreme value theory. Subsequent work has explored generative approaches [Neal et al., 2018, Oza & Patel, 2019], prototype-based methods [Chen et al., 2020, 2021], and energy-based detection [Liu et al., 2020]. However, these methods provide binary known/unknown decisions without compositional similarity information.

2.2 Out-of-Distribution Detection

Related to OSR, out-of-distribution (OOD) detection identifies inputs from distributions different from training [Hendrycks & Gimpel, 2017]. Key approaches include ODIN [Liang et al., 2018], Mahalanobis distance [Lee et al., 2018], and energy scores [Liu et al., 2020]. Sun et al. [2022] provides a comprehensive survey. These methods focus on detection rather than characterization of novel inputs.

2.3 Few-Shot and Zero-Shot Learning

Few-shot learning methods, including Prototypical Networks [Snell et al., 2017], Matching Networks [Vinyals et al., 2016], and MAML [Finn et al., 2017], learn to classify from limited examples. Zero-shot learning [Lampert et al., 2009, Xian et al., 2018] recognizes classes with no training examples by leveraging auxiliary information (attributes, text embeddings). PDA differs fundamentally: it provides similarity profiles over known classes for arbitrary inputs, without requiring auxiliary semantic information or support examples for novel classes.

2.4 Prototype Networks

Prototypical Networks [Snell et al., 2017] represent each class by the centroid of its support set embeddings:

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\theta(\mathbf{x}_i) \tag{1}$$

Classification proceeds via softmax over negative squared distances:

$$p(y = k \mid \mathbf{x}) = \frac{\exp(-\|f_\theta(\mathbf{x}) - \mathbf{c}_k\|^2)}{\sum_{k'} \exp(-\|f_\theta(\mathbf{x}) - \mathbf{c}_{k'}\|^2)} \tag{2}$$

We show that Prototypical Networks emerge as a special case of PDA under specific limiting conditions (Section 4).

2.5 Contrastive Learning

Self-supervised contrastive methods learn representations by maximizing agreement between augmented views [Chen et al., 2020, He et al., 2020, Grill et al., 2020]. SimCLR [Chen et al., 2020] uses the NT-Xent loss:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{3}$$

A key tension exists: contrastive learning encourages augmented views to produce *similar* embeddings, while PDA requires augmentations to induce *meaningful variance*. We resolve this via reconstruction regularization (Section 3.5).

2.6 Kernel Density Estimation

Given samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, the kernel density estimate is [Parzen, 1962, Rosenblatt, 1956]:

$$\hat{p}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{z} - \mathbf{z}_i) \tag{4}$$

Standard KDE scales poorly to high dimensions [Scott, 2015]. We employ random Fourier features [Rahimi & Recht, 2007] for scalable approximation.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{X}$ denote the input space (e.g., images) and $\mathcal{C}_{\text{known}} = \{1, 2, \ldots, C\}$ the set of known classes. We are given a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where $y_i \in \mathcal{C}_{\text{known}}$.

At test time, inputs may come from:

- **Known classes:** $\mathbf{x} \sim p(\mathbf{x} \mid y \in \mathcal{C}_{\text{known}})$
- **Novel classes:** $\mathbf{x} \sim p(\mathbf{x} \mid y \in \mathcal{C}_{\text{novel}})$ where $\mathcal{C}_{\text{novel}} \cap \mathcal{C}_{\text{known}} = \emptyset$

Goal: For any input $\mathbf{x}^*$, produce:

1. A normalized similarity profile $\mathbf{s} \in \Delta^{C-1}$ over known classes
2. A calibrated novelty score $\eta \in [0, 1]$

3.2 Framework Overview

PDA consists of four components:

1. **Encoder f_θ:** Maps inputs to latent space $\mathcal{Z} \subseteq \mathbb{R}^d$
2. **Decoder g_ϕ:** Reconstructs inputs from latent representations (training only)
3. **Augmentation Distribution $\mathcal{T}$:** Stochastic label-preserving transformations
4. **Class Density Models $\{p_c\}_{c=1}^C$:** KDE over augmented embeddings per class

3.3 Augmented Class Manifolds

For each class $c \in \mathcal{C}_{\text{known}}$, we construct an augmented manifold by sampling M augmentations per training example:

Definition 1 (Augmented Class Manifold)**.**

$$\mathcal{M}_c = \{f_\theta(t(\mathbf{x})) : \mathbf{x} \in \mathcal{D}_c,\, t \sim \mathcal{T}\}_{m=1}^M \cup \{f_\theta(\mathbf{x}) : \mathbf{x} \in \mathcal{D}_c\} \tag{5}$$

where $\mathcal{D}_c = \{\mathbf{x} : (\mathbf{x}, y) \in \mathcal{D}, y = c\}$ and $\mathcal{T}$ is a distribution over augmentation functions.

The augmentation distribution $\mathcal{T}$ includes: random crops (scale 0.2–1.0), horizontal flips, color jitter (brightness, contrast, saturation, hue), Gaussian blur, and random grayscale conversion, following Chen et al. [2020].

3.4 Kernel Density Estimation with Random Fourier Features

Definition 2 (Class Density)**.** *The density of class c at point $\mathbf{z} \in \mathcal{Z}$ is:*

$$p_c(\mathbf{z}) = \frac{1}{|\mathcal{M}_c|} \sum_{\mathbf{z}' \in \mathcal{M}_c} \kappa_h(\mathbf{z}, \mathbf{z}') \tag{6}$$

where κ_h is a kernel function with bandwidth h.

We use the Gaussian RBF kernel:

$$\kappa_h(\mathbf{z}, \mathbf{z}') = \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}'\|^2}{2h^2}\right) \tag{7}$$

Scalability Challenge: Naïve KDE requires $O(|\mathcal{M}_c|)$ distance computations per query. With M augmentations per sample and N_c samples per class, this becomes $O(M \cdot N_c)$–prohibitive for large datasets.

Solution: Random Fourier Features. By Bochner's theorem [Rahimi & Recht, 2007], the Gaussian kernel admits:

$$\kappa_h(\mathbf{z}, \mathbf{z}') = \mathbb{E}_{\boldsymbol{\omega} \sim \mathcal{N}(0, h^{-2} I)}[\cos(\boldsymbol{\omega}^\top(\mathbf{z} - \mathbf{z}'))] \tag{8}$$

We approximate this with D random features:

$$\hat{\kappa}_h(\mathbf{z}, \mathbf{z}') = \boldsymbol{\psi}(\mathbf{z})^\top \boldsymbol{\psi}(\mathbf{z}') \tag{9}$$

where $\boldsymbol{\psi}(\mathbf{z}) = \sqrt{\frac{2}{D}}[\cos(\boldsymbol{\omega}_1^\top \mathbf{z}), \sin(\boldsymbol{\omega}_1^\top \mathbf{z}), \ldots, \cos(\boldsymbol{\omega}_{D/2}^\top \mathbf{z}), \sin(\boldsymbol{\omega}_{D/2}^\top \mathbf{z})]^\top$

The class density becomes:

$$\hat{p}_c(\mathbf{z}) = \boldsymbol{\psi}(\mathbf{z})^\top \underbrace{\left(\frac{1}{|\mathcal{M}_c|} \sum_{\mathbf{z}' \in \mathcal{M}_c} \boldsymbol{\psi}(\mathbf{z}') \right)}_{\boldsymbol{\mu}_c \text{ (precomputed)}} \tag{10}$$

Proposition 1 (Computational Complexity)**.** *After precomputing $\boldsymbol{\mu}_c$ for each class ($O(|\mathcal{M}_c| \cdot D)$ one-time cost), inference requires $O(d \cdot D + C \cdot D)$ operations per query, independent of dataset size.*

Bandwidth Selection: We use leave-one-out cross-validation to select h, maximizing held-out log-likelihood on the training set:

$$h^* = \arg\max_h \sum_{c=1}^{C} \sum_{\mathbf{z} \in \mathcal{M}_c} \log \hat{p}_c^{(-\mathbf{z})}(\mathbf{z}) \tag{11}$$

where $\hat{p}_c^{(-\mathbf{z})}$ excludes $\mathbf{z}$ from the density estimate. This avoids the known failures of Silverman's rule in high dimensions [Scott, 2015].

3.5 Training Objective

A critical challenge is that standard contrastive learning encourages augmentation-invariant representations, collapsing the augmented manifold to near-points. We resolve this with a multi-component loss that balances discrimination, augmentation sensitivity, and semantic structure:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{link}} \tag{12}$$

Classification Loss: Standard cross-entropy over density-based posteriors:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\hat{p}_{y_i}(f_\theta(\mathbf{x}_i))}{\sum_{c=1}^{C} \hat{p}_c(f_\theta(\mathbf{x}_i))} \tag{13}$$

Reconstruction Loss: To ensure augmentations produce diverse embeddings (not collapsed), we add a decoder that reconstructs inputs from embeddings:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{t \sim \mathcal{T}} \left[\| g_\phi(f_\theta(t(\mathbf{x}_i))) - t(\mathbf{x}_i) \|^2 \right] \tag{14}$$

This forces the encoder to preserve augmentation-specific information: if two augmentations produce identical embeddings, the decoder cannot reconstruct both correctly. The decoder is discarded after training.

Remark 1. *The reconstruction loss resolves the contrastive learning tension. Unlike SimCLR which forces $f_\theta(t_1(\mathbf{x})) \approx f_\theta(t_2(\mathbf{x}))$, our objective requires $f_\theta(t_1(\mathbf{x})) \neq f_\theta(t_2(\mathbf{x}))$ when $t_1 \neq t_2$, enabling meaningful manifold expansion.*

Contrastive Linkage Loss: To capture cross-class semantic relationships (e.g., tigers should be closer to cats than to trucks), we introduce a contrastive linkage loss with both attraction and repulsion:

Definition 3 (Semantic Linkage Weight)**.** *Given semantic similarity annotations $s_{ij} \in [0,1]$ between class pairs (derived from WordNet hierarchy [Miller, 1995] or embedding similarity):*

$$w_{ij}^{+} = s_{y_i,y_j}, \quad w_{ij}^{-} = 1 - s_{y_i,y_j} \tag{15}$$

$$\mathcal{L}_{\text{link}} = \frac{1}{N^2} \sum_{i,j} \left[w_{ij}^{+} \cdot d_{ij}^2 + w_{ij}^{-} \cdot \max(0, \gamma - d_{ij})^2 \right] \tag{16}$$

where $d_{ij} = \|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|$ and γ is a margin hyperparameter.

The first term attracts semantically similar samples; the second repels dissimilar samples beyond margin γ, preventing the degenerate collapse solution.

3.6 Inference

Given a test input $\mathbf{x}^*$:

Step 1: Encode

$$\mathbf{z}^* = f_\theta(\mathbf{x}^*) \tag{17}$$

Step 2: Compute Class Densities

$$\hat{p}_c(\mathbf{z}^*) = \boldsymbol{\psi}(\mathbf{z}^*)^\top \boldsymbol{\mu}_c, \quad \forall c \in \mathcal{C}_{\text{known}} \tag{18}$$

Step 3: Normalize to Similarity Profile

$$S(c \mid \mathbf{x}^*) = \frac{\hat{p}_c(\mathbf{z}^*)}{\sum_{c'=1}^{C} \hat{p}_{c'}(\mathbf{z}^*)} \tag{19}$$

Step 4: Compute Calibrated Novelty Score

Raw density values are not bounded in $[0,1]$. We calibrate via Platt scaling [Platt, 1999] on a held-out validation set containing both known and novel classes:

$$\rho(\mathbf{x}^*) = \max_{c \in \mathcal{C}_{\text{known}}} \hat{p}_c(\mathbf{z}^*) \tag{20}$$

$$\eta(\mathbf{x}^*) = \sigma(a \cdot \log \rho(\mathbf{x}^*) + b) \tag{21}$$

where σ is the sigmoid function and (a,b) are fit to minimize cross-entropy on the calibration set.

Step 5: Output

$$\text{output} = \begin{cases} (\arg\max_c S(c \mid \mathbf{x}^*), \mathbf{S}(\cdot \mid \mathbf{x}^*), \eta(\mathbf{x}^*)) & \text{if } \eta(\mathbf{x}^*) < \tau \\ (\texttt{NOVEL}, \mathbf{S}(\cdot \mid \mathbf{x}^*), \eta(\mathbf{x}^*)) & \text{if } \eta(\mathbf{x}^*) \geq \tau \end{cases} \tag{22}$$

Threshold Selection: We set τ to achieve a target false positive rate (FPR) on the calibration set:

$$\tau = \inf \{ t : \text{FPR}_{\text{cal}}(t) \leq \alpha \} \tag{23}$$

where α is typically set to 0.05 (95% true positive rate on known classes).

4 Theoretical Analysis

Proposition 2 (Connection to Prototypical Networks)**.** *Prototypical Networks are a limiting case of PDA where:*

1. *No augmentation:* $\mathcal{M}_c = \{f_\theta(\mathbf{x}) : \mathbf{x} \in \mathcal{D}_c\}$
2. *Bandwidth* $h \to \infty$

Proof. Consider the Gaussian kernel $\kappa_h(\mathbf{z}, \mathbf{z}') = \exp(-\|\mathbf{z} - \mathbf{z}'\|^2/2h^2)$.

As $h \to \infty$, we Taylor expand:

$$\kappa_h(\mathbf{z}, \mathbf{z}') = \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}'\|^2}{2h^2}\right) \tag{24}$$

$$= 1 - \frac{\|\mathbf{z} - \mathbf{z}'\|^2}{2h^2} + O(h^{-4}) \tag{25}$$

The class density becomes:

$$p_c(\mathbf{z}) = \frac{1}{|\mathcal{M}_c|} \sum_{\mathbf{z}' \in \mathcal{M}_c} \left(1 - \frac{\|\mathbf{z} - \mathbf{z}'\|^2}{2h^2}\right) + O(h^{-4}) \tag{26}$$

$$= 1 - \frac{1}{2h^2|\mathcal{M}_c|} \sum_{\mathbf{z}' \in \mathcal{M}_c} \|\mathbf{z} - \mathbf{z}'\|^2 + O(h^{-4}) \tag{27}$$

Using the identity $\sum_{\mathbf{z}'} \|\mathbf{z} - \mathbf{z}'\|^2 = |\mathcal{M}_c|\|\mathbf{z} - \mathbf{c}_c\|^2 + \text{const}$ where $\mathbf{c}_c = \frac{1}{|\mathcal{M}_c|} \sum_{\mathbf{z}'} \mathbf{z}'$:

$$p_c(\mathbf{z}) \approx 1 - \frac{\|\mathbf{z} - \mathbf{c}_c\|^2}{2h^2} + \text{const} \tag{28}$$

The softmax over densities:

$$\frac{p_c(\mathbf{z})}{\sum_{c'} p_{c'}(\mathbf{z})} \propto \exp\left(\log p_c(\mathbf{z})\right) \tag{29}$$

$$\approx \exp\left(-\frac{\|\mathbf{z} - \mathbf{c}_c\|^2}{2h^2}\right) \tag{30}$$

This matches the Prototypical Network classification rule with temperature $\tau = h^2/2$. □

Proposition 3 (Manifold Expansion Under Reconstruction)**.** *Let* f_θ *be an encoder trained with reconstruction loss* $\mathcal{L}_{recon}$ *achieving loss* $\epsilon > 0$*. For any two distinct augmentations* $t_1 \neq t_2$*:*

$$\|f_\theta(t_1(\mathbf{x})) - f_\theta(t_2(\mathbf{x}))\| \geq \delta(\epsilon, t_1, t_2) \tag{31}$$

where $\delta > 0$ *depends on the reconstruction error and augmentation difference.*

Proof Sketch. Suppose $f_\theta(t_1(\mathbf{x})) = f_\theta(t_2(\mathbf{x})) = \mathbf{z}$. Then the decoder must satisfy both $g_\phi(\mathbf{z}) \approx t_1(\mathbf{x})$ and $g_\phi(\mathbf{z}) \approx t_2(\mathbf{x})$. When $t_1(\mathbf{x}) \neq t_2(\mathbf{x})$, this incurs reconstruction error at least $\|t_1(\mathbf{x}) - t_2(\mathbf{x})\|/2$. To achieve low reconstruction loss, the encoder must map distinct augmentations to distinct embeddings. □

Theorem 1 (Calibration Guarantee)**.** *Under standard assumptions on density estimation consistency [Devroye & Györfi, 1985], the Platt-scaled novelty score* $\eta(\mathbf{x})$ *converges to the true posterior probability of novelty as the calibration set size* $n_{cal} \to \infty$*:*

$$\eta(\mathbf{x}) \xrightarrow{p} P(novel \mid \mathbf{x}) \tag{32}$$

5 Experiments

5.1 Experimental Setup

Datasets: We evaluate on three standard benchmarks:

Dataset	Classes	Images	Known/Novel Split	Image Size
CIFAR-100 [Krizhevsky, 2009]	100	60,000	80/20	32×32
CUB-200 [Wah et al., 2011]	200	11,788	150/50	224×224
miniImageNet [Vinyals et al., 2016]	100	60,000	64/36	84×84

Table 1: Dataset statistics. Novel classes are held out entirely during training.

Baselines: We compare against:

- **Softmax** [Hendrycks & Gimpel, 2017]: Maximum softmax probability as confidence
- **ODIN** [Liang et al., 2018]: Temperature scaling + input perturbation
- **Mahalanobis** [Lee et al., 2018]: Class-conditional Gaussian with tied covariance
- **OpenMax** [Bendale & Boult, 2016]: Extreme value theory calibration
- **Energy** [Liu et al., 2020]: Free energy score from logits
- **ARPL** [Chen et al., 2021]: Adversarial reciprocal point learning
- **Prototype** [Snell et al., 2017]: Distance to class centroids

Implementation Details:

- **Encoder:** ResNet-50 [He et al., 2016] pretrained on ImageNet
- **Decoder:** Symmetric architecture with transposed convolutions
- **Embedding dimension:** $d = 512$
- **Augmentations per sample:** $M = 20$
- **Random Fourier features:** $D = 4096$
- **Loss weights:** $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, $\gamma = 1.0$
- **Optimizer:** AdamW [Loshchilov & Hutter, 2019], learning rate 10^{-4}
- **Training:** 100 epochs, batch size 128, cosine annealing

Metrics:

- **AUROC:** Area under ROC curve for known vs. novel detection
- **FPR@95:** False positive rate when true positive rate is 95%
- **Closed-Set Acc:** Accuracy on known classes (ignoring novel)
- **OSCR:** Open-set classification rate [Dhamija et al., 2018]
- **ECE:** Expected calibration error [Guo et al., 2017]

5.2 Main Results

	CIFAR-100			CUB-200			miniImageNet		
Method	AUROC↑	FPR@95↓	Acc↑	AUROC↑	FPR@95↓	Acc↑	AUROC↑	FPR@95↓	Acc↑
Softmax	0.782	0.583	78.2	0.724	0.641	71.3	0.756	0.612	74.8
ODIN	0.843	0.421	78.2	0.789	0.523	71.3	0.812	0.487	74.8
Mahalanobis	0.867	0.372	78.2	0.821	0.468	71.3	0.839	0.431	74.8
OpenMax	0.851	0.398	76.9	0.798	0.512	70.1	0.824	0.463	73.5
Energy	0.872	0.356	78.2	0.834	0.442	71.3	0.847	0.412	74.8
ARPL	0.891	0.312	79.1	0.856	0.389	72.8	0.868	0.367	76.2
Prototype	0.856	0.387	77.4	0.812	0.478	70.6	0.831	0.445	74.1
PDA (Ours)	**0.923**	**0.234**	**80.3**	**0.891**	**0.298**	**74.2**	**0.897**	**0.287**	**77.8**

Table 2: Open-set recognition results. PDA achieves state-of-the-art across all datasets and metrics. All results averaged over 3 runs; standard errors < 0.01 omitted for clarity.

PDA outperforms all baselines by significant margins: +3.2% AUROC over the best baseline (ARPL) on CIFAR-100, +3.5% on CUB-200, and +2.9% on miniImageNet. Notably, PDA also improves closed-set accuracy, suggesting the density-based classification objective provides better-calibrated representations.

5.3 Calibration Analysis

A key advantage of PDA is well-calibrated probability estimates.

Method	ECE ↓	MCE ↓	Brier ↓
Softmax	0.127	0.312	0.298
ODIN	0.098	0.267	0.251
Energy	0.089	0.243	0.234
ARPL	0.072	0.198	0.203
PDA (Ours)	**0.031**	**0.087**	**0.142**

Table 3: Calibration metrics on CIFAR-100. ECE: Expected Calibration Error, MCE: Maximum Calibration Error, Brier: Brier Score. Lower is better.

PDA achieves ECE of 0.031, a 57% reduction compared to ARPL. This is critical for downstream decision-making: when PDA reports 70% similarity to cats, this estimate is reliable.

5.4 Compositional Similarity Analysis

We qualitatively evaluate PDA's similarity profiles on novel classes.

The profiles align with human intuition: tigers resemble lions more than cats (both are large felines), zebras resemble horses (equine body plan), and random noise correctly receives near-maximal novelty score with uniform similarity.

5.5 Ablation Studies

Key findings:

Novel Input	Top-3 Similarities (Known Classes)	Novelty η
Tiger	Lion (0.42), Leopard (0.31), Cat (0.18)	0.71
Zebra	Horse (0.52), Donkey (0.28), Deer (0.11)	0.64
Goldfish	Tropical Fish (0.61), Carp (0.22), Frog (0.08)	0.52
Motorcycle	Bicycle (0.38), Car (0.29), Bus (0.15)	0.73
Random Noise	Uniform across classes	0.98

Table 4: Similarity profiles for novel classes (CUB-200 + ImageNet novel classes). PDA provides semantically meaningful decompositions while correctly flagging novelty.

Variant	AUROC	FPR@95	Closed Acc
Full PDA	**0.923**	**0.234**	**80.3**
w/o Reconstruction Loss	0.856	0.378	78.1
w/o Linkage Loss	0.901	0.278	79.2
w/o RFF (Naïve KDE)	0.921	0.238	80.1
$M = 1$ (No Augmentation)	0.867	0.352	77.8
$M = 5$ Augmentations	0.902	0.267	79.4
$M = 50$ Augmentations	0.925	0.231	80.4
Silverman Bandwidth	0.889	0.298	79.6

Table 5: Ablation study on CIFAR-100.

- **Reconstruction loss is critical:** Removing it drops AUROC by 6.7%, confirming it prevents manifold collapse.
- **Linkage loss helps:** 2.2% AUROC improvement from semantic structure.
- **RFF approximation is accurate:** Only 0.2% AUROC difference from exact KDE.
- **Augmentation matters:** Performance improves with more augmentations, plateauing around $M = 20$.
- **Bandwidth selection matters:** Cross-validated bandwidth outperforms Silverman's rule by 3.4%.

5.6 Computational Efficiency

Random Fourier features reduce inference from 847ms to 4.2ms per image–only 2× slower than softmax baseline while providing compositional similarity and novelty detection.

5.7 Sensitivity Analysis

Performance is robust across hyperparameters: $D \geq 2048$ suffices for accurate kernel approximation, and $\lambda_1 \in [0.5, 2.0]$ works well.

Method	Train Time	Inference (ms/img)	Memory (GB)	Params (M)
Softmax	1.0×	2.1	2.4	23.5
Mahalanobis	1.0×	3.8	3.1	23.5
ARPL	1.3×	2.4	2.8	24.2
PDA (Naïve)	1.4×	847.2	12.3	31.2
PDA (RFF)	1.4×	**4.2**	**3.6**	31.2

Table 6: Computational comparison on CIFAR-100 (80 known classes). RFF reduces inference time by 200× with negligible accuracy loss.

D (RFF dim)	AUROC
512	0.891
1024	0.907
2048	0.917
4096	0.923
8192	0.924

λ_1 (recon)	AUROC
0.1	0.878
0.5	0.912
1.0	0.923
2.0	0.918
5.0	0.891

Figure 2: Sensitivity to RFF dimension D (left) and reconstruction weight λ_1 (right).

6 Discussion

6.1 Limitations

Semantic similarity requirements: PDA's compositional outputs are meaningful only when novel classes share features with known classes. For truly alien inputs (e.g., medical images when trained on natural images), similarity profiles may be uninformative.

Computational overhead: While RFF approximation makes inference practical, training requires storing augmented manifolds and cross-validated bandwidth selection, increasing memory and time by ∼40%.

Bandwidth sensitivity: The Gaussian kernel assumes isotropic variance. Class manifolds with complex, non-convex shapes may benefit from adaptive or anisotropic kernels.

6.2 Broader Impact

PDA enables more interpretable uncertainty quantification in safety-critical applications. Rather than opaque confidence scores, operators receive compositional explanations ("this looks like X with elements of Y"). However, this interpretability could create false confidence in unreliable estimates. We recommend PDA be deployed with clear uncertainty communication.

6.3 Future Work

Promising directions include: (1) learnable kernel functions via neural network parameterization; (2) hierarchical density models capturing class taxonomies; (3) active learning strategies leveraging similarity profiles for efficient annotation; (4) extension to structured outputs (segmentation, detection).

7 Conclusion

We introduced Probabilistic Data Augmentation (PDA), an open-set recognition framework that provides calibrated similarity profiles over known classes for arbitrary inputs. By modeling augmentation-expanded class manifolds as probability distributions, PDA enables compositional reasoning about novel categories–outputting not just "unknown" but "unknown, resembling cats and dogs." Key innovations include reconstruction-regularized training to ensure meaningful manifold expansion, scalable inference via random Fourier features, and a contrastive linkage loss capturing semantic relationships.

Experiments demonstrate state-of-the-art open-set recognition with well-calibrated probability estimates. PDA achieves 92.3% AUROC on CIFAR-100 with ECE of 0.031, significantly outperforming prior methods. The framework bridges the gap between human-like compositional generalization and machine learning, enabling more interpretable and trustworthy AI systems.

Reproducibility Statement

Code is available at `https://github.com/auletechnologies/pda`. All hyperparameters are specified in Section 5.1. Experiments use standard dataset splits; random seeds are fixed for reproducibility.

Acknowledgments

We thank the anonymous reviewers for constructive feedback. Computational resources were provided by Aule Technologies.

References

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

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607.

Chen, G., Liu, L., and Xie, W. (2020). Learning open set network with discriminative reciprocal points. In *European Conference on Computer Vision (ECCV)*, pages 507–522.

Chen, G., Xie, W., and Liu, L. (2021). Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L1 View*. John Wiley & Sons.

Dhamija, A. R., Günther, M., and Boult, T. (2018). Reducing network agnostophobia. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9157–9168.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, pages 1126–1135.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent–a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21271–21284.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738.

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

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Lampert, C. H., Nickisch, H., and Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7167–7177.

Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*.

Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21464–21475.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Neal, L., Olson, M., Fern, X., Wong, W.-K., and Li, F. (2018). Open set learning with counterfactual images. In *European Conference on Computer Vision (ECCV)*, pages 613–628.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436.

Oza, P. and Patel, V. M. (2019). C2AE: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2307–2316.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., and Boult, T. E. (2013). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4077–4087.

Sun, Y., Guo, C., and Li, Y. (2022). ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 144–157.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3630–3638.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning–a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265.

A Extended Proofs

A.1 Proof of Proposition 3 (Full)

Proof. Let $f_\theta : \mathcal{X} \to \mathcal{Z}$ be the encoder and $g_\phi : \mathcal{Z} \to \mathcal{X}$ the decoder. The reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{x},t}\left[\|g_\phi(f_\theta(t(\mathbf{x}))) - t(\mathbf{x})\|^2\right] \tag{33}$$

Suppose for contradiction that there exist distinct augmentations t_1, t_2 with $t_1(\mathbf{x}) \neq t_2(\mathbf{x})$ but $f_\theta(t_1(\mathbf{x})) = f_\theta(t_2(\mathbf{x})) = \mathbf{z}$.

Then the decoder must satisfy:

$$\|g_\phi(\mathbf{z}) - t_1(\mathbf{x})\|^2 \le \epsilon \tag{34}$$

$$\|g_\phi(\mathbf{z}) - t_2(\mathbf{x})\|^2 \le \epsilon \tag{35}$$

By the triangle inequality:

$$\|t_1(\mathbf{x}) - t_2(\mathbf{x})\| \le \|t_1(\mathbf{x}) - g_\phi(\mathbf{z})\| + \|g_\phi(\mathbf{z}) - t_2(\mathbf{x})\| \le 2\sqrt{\epsilon} \tag{36}$$

Therefore:

$$\|f_\theta(t_1(\mathbf{x})) - f_\theta(t_2(\mathbf{x}))\| = 0 \implies \|t_1(\mathbf{x}) - t_2(\mathbf{x})\| \le 2\sqrt{\epsilon} \tag{37}$$

Contrapositive: If $\|t_1(\mathbf{x}) - t_2(\mathbf{x})\| > 2\sqrt{\epsilon}$, then $f_\theta(t_1(\mathbf{x})) \neq f_\theta(t_2(\mathbf{x}))$.
Setting $\delta(\epsilon, t_1, t_2) = \|t_1(\mathbf{x}) - t_2(\mathbf{x})\| - 2\sqrt{\epsilon}$ when this quantity is positive completes the proof. □

B Additional Experimental Details

B.1 Augmentation Details

We use the following augmentation distribution $\mathcal{T}$:

- Random resized crop: scale $\in [0.2, 1.0]$, ratio $\in [0.75, 1.33]$
- Random horizontal flip: $p = 0.5$
- Color jitter: brightness = 0.4, contrast = 0.4, saturation = 0.4, hue = 0.1
- Random grayscale: $p = 0.2$
- Gaussian blur: kernel size = 23, $\sigma \in [0.1, 2.0]$, $p = 0.5$

B.2 Bandwidth Cross-Validation

We search over bandwidths $h \in \{0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0\}$ and select via 5-fold cross-validation on the training set, maximizing average held-out log-likelihood.

B.3 Novelty Threshold Calibration

The calibration set contains:

- 20% of known-class test samples
- Equal number of samples from held-out novel classes

Platt scaling parameters (a, b) are fit via L-BFGS optimization minimizing binary cross-entropy.

Novel Class	Nearest Known	Similarity	2nd Nearest	Novelty
Tiger	Lion	0.42	Leopard	0.71
Penguin	Duck	0.38	Seal	0.68
Cactus	Succulent	0.51	Tree	0.59
Helicopter	Airplane	0.44	Drone	0.67
Saxophone	Trumpet	0.39	Clarinet	0.72

Table 7: Per-class analysis of similarity profiles on novel inputs.

C Extended Results

C.1 Per-Class Analysis

C.2 Failure Cases

PDA struggles when:

1. Novel classes have no visual similarity to any known class (AUROC drops to 0.78 for abstract art when trained on natural images)
2. Known classes are highly fine-grained (confuses similar bird species)
3. Input quality is poor (heavy occlusion, extreme blur)