A Appendix

A.1 Details of Datasets

For the experimental evaluation, the datasets were partitioned into labeled and unlabeled subsets. Specifically, for the CIFAR-100 dataset, 80% of the data was allocated to the labeled set D_S, while for the remaining datasets, this ratio was set to 50%. The labeled data was sampled from the labeled classes, and the unlabeled data D_Q was formed by merging the remaining instances. As mentioned earlier, we further split 20% of the training samples as a validation set to assist with anomaly detection. Details of the datasets are provided in Table 9 and Table 10.

Table 9: Statistical comparison of data partitions.

		CIFAR-100	ImageNet-100	CUB-200	Stanford-Cars	Herbarium19
Labeled	$\lvert\mathcal{Y}_S\rvert$	80	50	100	98	341
	$\lvert D_S\rvert$	16k	25.5k	1.2k	1.6k	1.6k
	$\lvert\mathcal{Y}_V\rvert$	80	50	100	98	341
	$\lvert D_V\rvert$	4k	6.4k	0.3k	0.4k	7.2k
Unlabeled	$\lvert\mathcal{Y}_Q\rvert$	100	100	200	196	683
	$\lvert D_Q\rvert$	30k	95.3k	4.5k	6.1k	25.4k

Table 10: Statistical comparison of data partitions.

		Fungi	Arachnida	Animalia	Mollusca	Pets	Food101
Labeled	$\lvert\mathcal{Y}_S\rvert$	61	28	39	47	341	51
	$\lvert D_S\rvert$	1.5k	1.3k	1.2k	1.6k	1.9k	15.3k
	$\lvert\mathcal{Y}_V\rvert$	61	61	39	47	341	51
	$\lvert D_V\rvert$	0.3k	0.3k	0.3k	0.4k	0.5k	3.8k
Unlabeled	$\lvert\mathcal{Y}_Q\rvert$	121	56	77	93	683	101
	$\lvert D_Q\rvert$	5.8k	4.3k	5.1k	6.1k	7.0k	56.6k

A.2 Effects of Validation Set Proportion

Regarding the proportion of validation samples, a larger validation set helps estimate a more accurate covariance matrix, while a larger training set is beneficial for learning better features. In Table 11, we compare the impact of different validation set proportions on performance. For the fine-grained CUB-200 dataset, insufficient training samples lead to a drop in accuracy due to its limited sample size. In contrast, since the Pets dataset has fewer categories, the validation split has a smaller effect on performance.

Table 11: Performance with different validation set proportion.

	CUB-200			Pets		
	All	Old	New	All	Old	New
20%	46.3	59.8	39.4	61.5	66.5	58.8
30%	45.6	57.6	39.5	59.0	66.8	54.8
40%	43.9	56.6	37.5	61.2	67.6	57.8
50%	42.6	55.9	35.9	62.3	67.6	59.6
60%	42.2	54.1	36.2	61.1	72.2	55.3

A.3 PROOF OF LEMMA AND PROPOSITION

Proof of Lemma 1. Since the feature space is whitened, we have $\boldsymbol{\Sigma} = \boldsymbol{I}$ and thus the Mahalanobis distance becomes the squared Euclidean distance. Expanding the squared distance between the means in the whitened space yields:

$$\|\boldsymbol{\mu}_C - \boldsymbol{\mu}_A\|^2 = \|\boldsymbol{\mu}_C\|^2 + \|\boldsymbol{\mu}_A\|^2 - 2\,\boldsymbol{\mu}_C^\top \boldsymbol{\mu}_A,$$
$$\|\boldsymbol{\mu}_C - \boldsymbol{\mu}_B\|^2 = \|\boldsymbol{\mu}_C\|^2 + \|\boldsymbol{\mu}_B\|^2 - 2\,\boldsymbol{\mu}_C^\top \boldsymbol{\mu}_B.$$

Taking expectations over class means in groups A and B respectively:

$$\mathbb{E}[\boldsymbol{d}_A] = \|\boldsymbol{\mu}_C\|^2 + \mathbb{E}\|\boldsymbol{\mu}_A\|^2 - 2\,\boldsymbol{\mu}_C^\top \mathbb{E}[\boldsymbol{\mu}_A],$$
$$\mathbb{E}[\boldsymbol{d}_B] = \|\boldsymbol{\mu}_C\|^2 + \mathbb{E}\|\boldsymbol{\mu}_B\|^2 - 2\,\boldsymbol{\mu}_C^\top \mathbb{E}[\boldsymbol{\mu}_B].$$

Subtracting the two gives:

$$\mathbb{E}[\boldsymbol{d}_B] - \mathbb{E}[\boldsymbol{d}_A] = \left(\mathbb{E}\|\boldsymbol{\mu}_B\|^2 - \mathbb{E}\|\boldsymbol{\mu}_A\|^2\right) - 2\,\boldsymbol{\mu}_C^\top \left(\mathbb{E}[\boldsymbol{\mu}_B] - \mathbb{E}[\boldsymbol{\mu}_A]\right).$$

Under the stated assumptions:

$$\boldsymbol{\mu}_C^\top \left(\mathbb{E}[\boldsymbol{\mu}_A] - \mathbb{E}[\boldsymbol{\mu}_B]\right) > 0, \quad \mathbb{E}\|\boldsymbol{\mu}_A\|^2 - \mathbb{E}\|\boldsymbol{\mu}_B\|^2 \geq 0,$$

which implies that

$$\mathbb{E}[\boldsymbol{d}_B] - \mathbb{E}[\boldsymbol{d}_A] > 0,$$

and thus

$$\mathbb{E}[\boldsymbol{d}_A] < \mathbb{E}[\boldsymbol{d}_B].$$

□

Proof of Proposition 1. Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_C, \boldsymbol{\Sigma})$ be a sample from class C. The log-likelihoods under Gaussian distributions with means $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ (and shared covariance $\boldsymbol{\Sigma}$) are:

$$\log p_A(\boldsymbol{X}) = -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_A)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_A) + \text{const},$$
$$\log p_B(\boldsymbol{X}) = -\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_B)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_B) + \text{const}.$$

Hence, the classification rule reduces to comparing Mahalanobis distances:

$$\boldsymbol{d}_A = (\boldsymbol{X} - \boldsymbol{\mu}_A)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_A),$$
$$\boldsymbol{d}_B = (\boldsymbol{X} - \boldsymbol{\mu}_B)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_B).$$

Let $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}_C)$, so that $\boldsymbol{Y} \sim \mathcal{N}(0, I)$. Define:

$$a = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_C - \boldsymbol{\mu}_A), \quad b = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_C - \boldsymbol{\mu}_B).$$

Then we can rewrite:

$$\boldsymbol{d}_A = \|\boldsymbol{Y} + a\|^2, \quad \boldsymbol{d}_B = \|\boldsymbol{Y} + b\|^2.$$

Thus, the condition $\boldsymbol{d}_A < \boldsymbol{d}_B$ becomes:

$$\|\boldsymbol{Y} + a\|^2 < \|\boldsymbol{Y} + b\|^2 \quad \Longleftrightarrow \quad 2(b - a)^\top \boldsymbol{Y} > \|a\|^2 - \|b\|^2.$$

Since $\boldsymbol{Y} \sim \mathcal{N}(0, I)$, the projection $(b-a)^\top \boldsymbol{Y}$ is a scalar Gaussian random variable with zero mean and symmetric distribution. If $\|a\|^2 < \|b\|^2$, then the threshold on the right-hand side is negative, and we have:

$$P(\boldsymbol{d}_A < \boldsymbol{d}_B) = P\left((b - a)^\top \boldsymbol{Y} > \frac{1}{2}(\|a\|^2 - \|b\|^2)\right) > 0.5.$$

This implies that a sample from class C is more likely to be classified as A (non-B) than as B under the Mahalanobis distance-based decision rule.

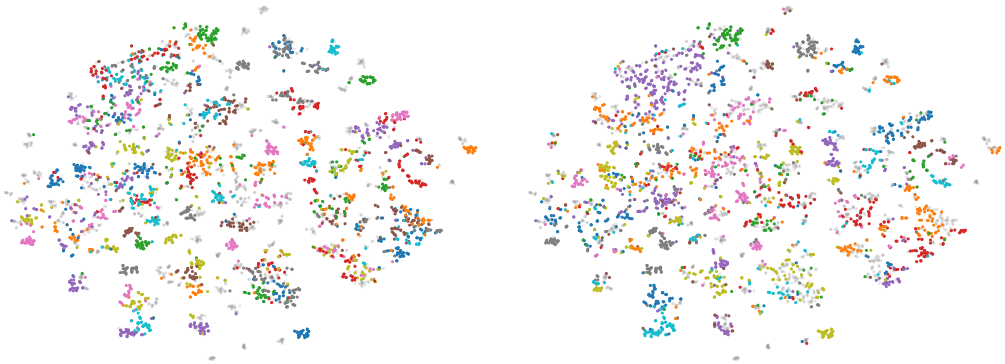
Therefore, when $\boldsymbol{\mu}_C$ is closer to $\boldsymbol{\mu}_A$ than to $\boldsymbol{\mu}_B$ in the Mahalanobis sense, using class A as a non-B background model leads to a correct classification (as non-B) with probability greater than 0.5, making it a justified and effective open-set prior. □

A.4 VISUALIZATION

To better illustrate the superiority of AGE and its practical benefits, we provide t-SNE visualizations of both known and novel classes from the CUB-200 dataset. Fig. 7 depicts the feature distribution of known classes, where translucent gray points represent samples from both ground-truth novel classes and predicted novel classes. Samples with the same color correspond to the same predicted class. It is clear that our method not only accurately recognizes known classes but also yields reliable predictions. Fig. 6 presents the feature distribution of novel classes, with translucent gray points again indicating novel class samples. AGE demonstrates a pronounced advantage in clustering novel samples: instances from the same predicted class form compact clusters in the feature space while maintaining clear separations from other classes. These results highlight ability to effectively distinguish and structure novel class features without any prior category information, providing strong evidence of its effectiveness in novel class identification and modeling under the open-set recognition setting.

(a) Ground Truth (b) Prediction

Figure 5: The t-SNE visualization of known classes on the CUB-200 dataset.



(a) Ground Truth (b) Prediction

Figure 6: The t-SNE visualization of novel classes on the CUB-200 dataset.

For an intuitive visualization of how categories emerge throughout the dynamic discovery process, we plot a series of t-SNE maps on the Pets dataset in a continual-learning style. (a) shows only the old classes; (b–d) illustrate the progressive emergence of new classes; (e) presents the final predictions for all newly discovered classes; (f) provides the corresponding ground truth.

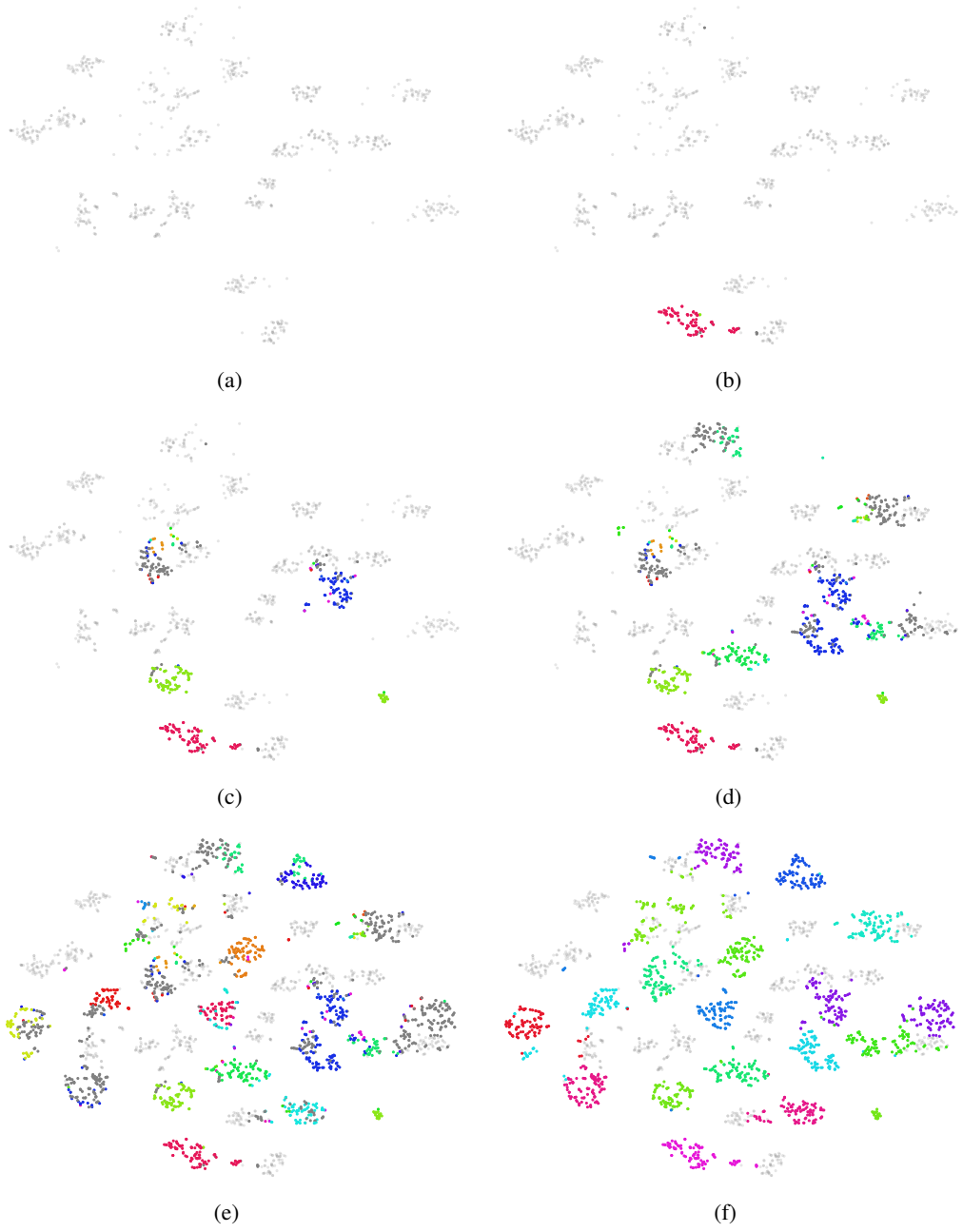
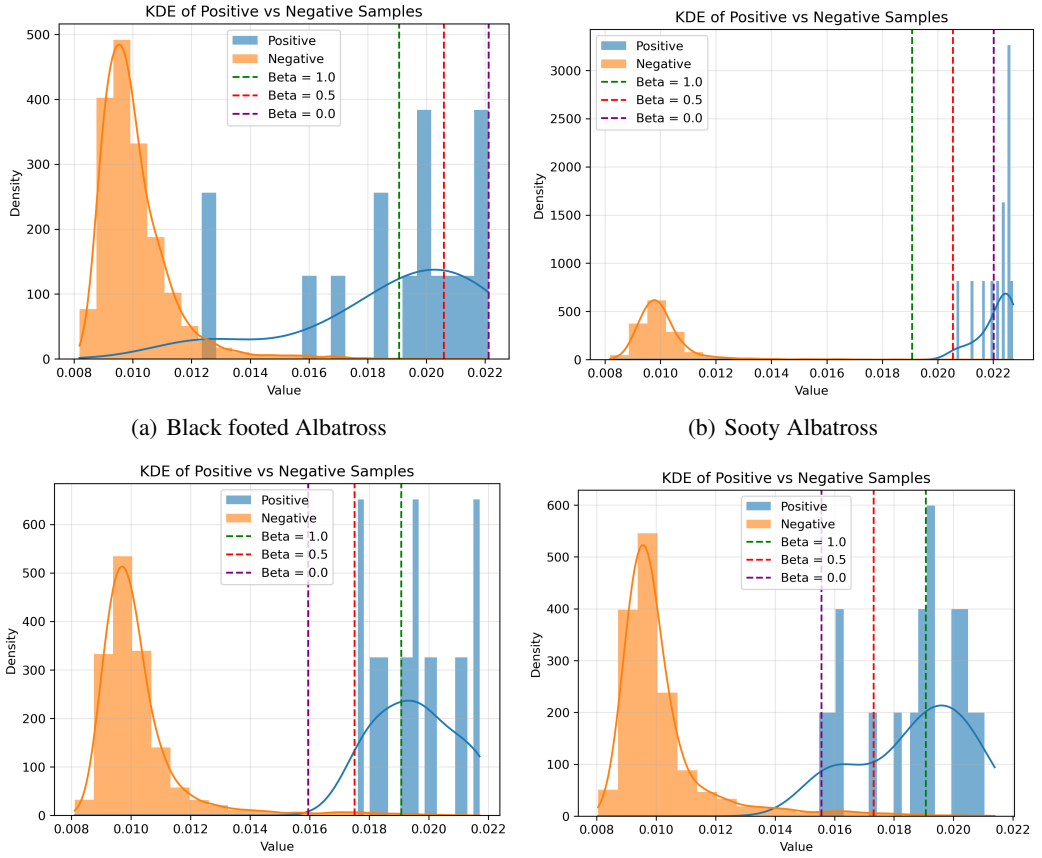


Figure 7: The t-SNE visualization on the Pets.

A.5 EFFETCS OF β

In our experiments, we observed that using either a global threshold or a per-class threshold can lead to significant bias. A global threshold may overlook certain low-confidence classes, while per-class thresholds can be excessively noisy when the validation set is small. In Fig.8, we visualized the confidence distributions for several classes on CUB-200 along with thresholds under different β values. The results show that the soft-thresholding strategy effectively mitigates such bias.



(a) Black footed Albatross (b) Sooty Albatross

(c) Rusty Blackbird (d) Pelagic Cormorant

Figure 8: Confidence distributions of selected classes on CUB-200

In Fig. 9, we summarize the impact of different values of the fusion coefficient β on the performance of anomaly detection. The evaluation metric is the accuracy of OOD detection, defined as the proportion of samples correctly identified as either in-distribution or out-of-distribution.

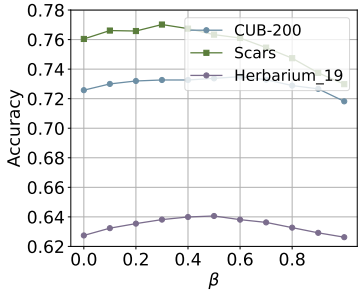


Figure 9: Performance with different β.

A.6 PSEUDO-CODE

To improve clarity and ensure reproducibility, we provide the pseudocode of the proposed Adaptive Gaussian Expansion (AGE) inference procedure in Algorithm 1. This offers a concise summary of the key steps involved, particularly in the streaming setting where real-time category discovery is required.

Algorithm 1 Adaptive Gaussian Expansion

Input: Validation set D_V, query stream D_Q, encoder E, projection head H, classifier C
Parameter: Threshold smoothing β, soft prior s, stability ε
Output: Predicted labels $\hat{\boldsymbol{Y}}_Q$

Extract features $\boldsymbol{f}_i = E(x_i)$ for all $x_i \in D_V$
Apply PCA to reduce feature dimensionality to d
Compute class-wise confidence means $\boldsymbol{\mu}_k$ and stds $\boldsymbol{\Sigma}_k$ on D_V
Compute thresholds: $\tau_k = \boldsymbol{\mu}_k - \boldsymbol{\Sigma}_k$, global $\tau_g = \frac{1}{|\boldsymbol{Y}_S|} \sum_k \tau_k$
Final threshold: $\tau'_k = \beta \cdot \tau_k + (1 - \beta) \cdot \tau_g$
Estimate $\boldsymbol{\mu}_k$, covariance $\boldsymbol{\Sigma}_k$ for each class using Ledoit–Wolf shrinkage
for each query sample $x_i \in D_Q$ **do**
    Extract and project feature $\boldsymbol{f}_i = \text{PCA}(E(x_i))$
    Compute prediction $p_i = \text{Softmax}(C(H(\boldsymbol{f}_i)))$
    if $\max(p_i) > \tau'_{\arg\max(p_i)}$ **then**
        Assign label $\hat{y}_i = \arg\max(p_i)$
    else
        Compute log-PDFs $\log \boldsymbol{p}_k$ using Eq. 13 for all novel clusters
        if predicted as known **then**
            Create new cluster with mean $\boldsymbol{f}_i$, covariance $\boldsymbol{\Sigma}_{\text{prior}}$
        else
            Assign to cluster $k^* = \arg\max \log \boldsymbol{p}_k$
            Update $\boldsymbol{\mu}_{k^*}$ and $\boldsymbol{\Sigma}_{k^*}$ with Ledoit–Wolf
        end if
    end if
end for
return predicted labels $\hat{\boldsymbol{Y}}_Q$

A.7 Stability Analysis of AGE

To evaluate the stability of AGE, we conducted five independent runs and reported the mean and standard deviation of the results, as shown in Table 12. The results demonstrate that AGE consistently exhibits low variance across multiple datasets, indicating strong stability and robustness under different random initializations. AGE not only achieves higher average performance in most scenarios but also significantly reduces performance fluctuations, further validating its reliability and robustness in practical applications.

Table 12: Stability evaluation of AGE across five runs. Results are reported as mean ± standard deviation.

Dataset	All	Old	New
CIFAR-100	60.24±0.17	75.63±0.46	29.46±0.85
ImageNet-100	48.16±0.45	87.08±0.03	28.59±0.63
CUB-200	46.20±0.52	61.57±0.40	38.51±0.67
Scars	35.01±0.10	63.05±0.51	21.47±0.48
Pets	61.58±1.77	66.60±0.01	58.94±2.71
Herbarium19	30.93±0.25	54.57±0.08	18.20±0.40
Fungi	39.39±0.73	64.90±4.44	28.08±2.60
Arachnida	43.11±0.60	73.78±0.15	23.82±0.97
Animalia	44.33±0.44	66.78±0.18	35.03±0.58
Mollusca	43.23±0.40	69.95±0.04	29.01±0.62
Food101	30.16±0.12	69.95±0.05	9.87±0.33

A.8 Comparison with K-means Clustering

We further conducted comparative experiments with the classical clustering method K-means. Specifically, we directly performed clustering on all samples in the query set D_Q during the inference stage. Since K-means relies on the global distribution of all samples during inference, it may have certain advantages in specific scenarios. We report the classification performance across multiple datasets and provide a comprehensive comparison with our proposed AGE method, as shown in Table 13.

As shown in the table, K-means achieves slightly better novel class recognition performance than AGE on some datasets, such as CIFAR-100 and ImageNet-100. However, this advantage mainly stems from its dependence on global distribution information, making it less suitable for online or incremental category discovery tasks. In contrast, AGE demonstrates stronger known-class retention and more stable novel-class discovery performance on certain datasets, particularly exhibiting significant advantages on fine-grained datasets such as Herbarium19 and Scars. This validates that AGE can achieve robust and generalizable category discovery without access to the global sample distribution.

Table 13: Comparison with the inductive K-means.

	CIFAR-100			ImageNet-100			CUB-200			Scars			Herbarium19		
Method	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
K-means	65.9	77.8	42.3	66.6	71.6	50	53.0	52.8	53.1	33.7	36.6	31.6	26.5	33.8	22.6
AGE	60.8	75.8	30.7	48.2	87.1	28.6	46.3	59.8	39.4	34.8	62.7	21.3	30.9	54.7	18.1
	Fungi			Arachnida			Animalia			Mollusca			Pets		
Method	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
K-means	40.0	55.9	33.0	45.3	65.5	32.6	44.9	53.8	41.2	44.7	68.3	32.2	69.0	52.8	77.6
AGE (Ours)	40.0	66.1	28.5	42.7	73.7	23.3	44.1	68.3	34.2	43.0	70.0	28.7	61.5	66.5	58.8

A.9 TIME EFFICIENCY

To compare the time complexity of AGE with that of existing methods, we measured the inference time across multiple datasets, where the reported cost for AGE includes the threshold estimation on the validation set. As shown in Table 14, AGE incurs a slightly higher time cost than SMILE and PHE, yet achieves a substantial improvement in accuracy. It is worth noting that AGE does not require training on the validation set during the training phase, leading to a slightly reduced overall training time compared with prior methods.

Table 14: Time efficiency (seconds) of different methods.

	CUB-200	Scars	Pets	ImageNet-100	Herbarium9	Fungi
SMILE	34	59	29	745	330	78
PHE	34	60	29	747	331	78
AGE (Ours)	37	68	32	794	368	85

A.10 EFFECTS OF TEST SAMPLE ORDERING

Since the model needs to retain information from previously observed test samples during inference, which inherently depends on their content, we compare the performance of AGE under different scenarios. In previous studies, test datasets are typically presented in an unshuffled order to better simulate real-world conditions, where novel classes often appear in succession—an observation aligned with the locality principle in computer memory access. Table 15 presents the impact of shuffling test samples on performance, showing that AGE exhibits low sensitivity to the temporal order of data, thereby highlighting its robustness.

Table 15: Performance with different ordering.

	CUB-200			Pets		
	All	Old	New	All	Old	New
Order	46.3	59.8	39.4	61.5	66.5	58.8
Shuffle	46.4	62.8	38.1	62.1	66.6	59.8

A.11 EFFECTS OF SLIDING-WINDOW SIZE

To examine the influence of sliding-window size on model performance, we evaluate two datasets where the effect is particularly evident. As shown in Table 16, accuracy improves steadily with increasing window size, suggesting that a larger window yields a more stable and reliable covariance estimation.

Table 16: Performance with different window size.

	CUB-200			Fungi		
Size	All	Old	New	All	Old	New
2	32.7	63.2	18.0	37.3	67.4	23.9
4	34.6	63.7	20.6	38.0	66.6	25.4
8	35.2	63.1	21.7	38.6	67.3	25.6
16	35.2	62.7	21.9	39.5	67.3	27.3
32	35.3	63.1	21.9	39.9	67.3	27.9

A.12 FURTHER DISCUSSION

Compared to existing methods, the advantages of AGE are not limited to its superior performance; more importantly, it demonstrates greater practicality in real-world applications. In approaches such as SMILE and PHE, identifying previously seen classes requires retrieving representative hash codes from the old class samples. However, due to the inherently sensitive nature of hash codes, these methods often suffer from low classification accuracy.

The proposed method effectively addresses the long-standing issue of suboptimal performance in feature-level clustering approaches. By using features as unique identifiers for samples during training, AGE eliminates the need for additional parameters or loss terms, thereby introducing no extra training computational overhead.

A.13 FUTURE WORK

The current outlier detection module in AGE adopts a relatively simple thresholding strategy and does not leverage recent advances in the Open Set Recognition domain. Future research could explore deeper integration with state-of-the-art Open Set Recognition techniques to enhance outlier detection capabilities.

Moreover, under the constraint of streaming data, AGE has demonstrated even higher performance than K-means on certain datasets. This suggests the potential of integrating AGE into tasks such as Novel Category Discovery and Generalized Category Discovery.

Finally, we believe that incorporating language models to guide OCD can effectively alleviate the bias caused by the inherent invisibility of novel classes. Future work may consider exploring multi-modal approaches to further improve performance.