PMAL: PROGRESSIVE MULTI-LABEL ACTIVE LEARNING VIA DYNAMIC DIVERSITY REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial attribute recognition under limited annotation budgets poses significant challenges due to strong label correlations, costly annotations, and the lack of principled active learning strategies for multi-label settings. To address these challenges, we propose Progressive Multi-Label Active Learning (PMAL), a framework that efficiently identifies informative samples under tight labeling budgets. PMAL decomposes multi-label joint entropy into independent components for uncertainty estimation, evaluates feature diversity through similarity computations on Riemannian manifolds, and employs a greedy batch-adaptive scoring mechanism that dynamically updates selection priorities. We further extend submodular optimization theory to dynamic multi-label selection and, for the first time, establish an $\mathcal{O}(\log n)$ stability bound under label relevance matrix perturbations. Extensive experiments on CelebA and LFWA demonstrate that PMAL consistently outperforms eight state-of-the-art baselines. Beyond multi-label tasks, we also introduce Progressive Active Learning (PAL) for multi-class settings, achieving superior results on CIFAR-10, CIFAR-100, and SVHN benchmarks. Beyond multi-label tasks, we also introduce Progressive Active Learning (PAL) for multi-class settings, achieving superior results on CIFAR-10, CIFAR-100, and SVHN benchmarks. Meanwhile, PAL is designed with scalable computational complexity, remaining efficient even on large-scale datasets, and achieves substantial runtime savings through lightweight diversity updates without compromising selection quality.

1 INTRODUCTION

Neural networks' success relies heavily on large-scale annotated datasets Krizhevsky (2009); Devlin et al. (2019), yet creating such datasets demands substantial resources Northcutt et al. (2021). This challenge intensifies in multi-label classification, where each sample requires multiple attribute annotations. Reducing annotation dependency while maintaining performance has thus become crucial Van Engelen & Hoos (2020); Chen et al. (2020); Werner et al. (2024); Hanneke et al. (2024).

Current approaches fall into two categories Van Engelen & Hoos (2020); Ren et al. (2021); Astorga et al. (2024): (1) model-centric methods employing semi-supervised, unsupervised, or transfer learning Tarvainen & Valpola (2017); He et al. (2020); Yosinski et al. (2014), and (2) data-centric methods like active learning Settles (2009) that strategically select informative samples. While model-centric approaches often require specialized architectures Kornblith et al. (2019), limiting generalizability, data-centric methods offer greater flexibility across architectures. Active learning particularly excels in high-stakes domains like healthcare and finance Irvin et al. (2019); Huang et al. (2023), where annotation quality is paramount Yang et al. (2024); Chen et al. (2024); Wang & Zhao (2025).

Active learning Settles (2009) minimizes labeling costs by querying the most informative samples. Traditional methods include uncertainty-based Dagan & Engelson (1995), diversity-based Sener & Savarese (2018), and hybrid approaches Gal et al. (2017). However, uncertainty-based methods often select outliers, leading to poor boundary estimation Tharwat & Schenck (2023), while diversity-based methods may miss highly informative samples. Current hybrid approaches Kirsch et al. (2019); Gal et al. (2017) struggle with parameter tuning and computational scalability.

We focus on facial attribute recognition, a multi-label task Liu et al. (2015); Kumar et al. (2011) with severe class imbalance Huang et al. (2019); Cui et al. (2019). As shown in Figure 1, we propose

Progressive Multi-Label Active Learning (PMAL), inspired by recent multi-label advances Xing et al. (2024). PMAL employs a greedy batch-adaptive scoring mechanism that dynamically balances uncertainty and diversity without extensive tuning, maintaining efficiency at scale.

Contributions:

- We propose PMAL, a novel framework for multi-label active learning under budget constraints, which decomposes multi-label joint entropy for uncertainty estimation and evaluates feature diversity on Riemannian manifolds.
- We design a greedy batch-adaptive scoring mechanism that progressively updates selection priorities, enabling more effective utilization of limited labeling budgets.
- We extend submodular optimization theory to dynamic multi-label selection and establish an $\mathcal{O}(\log n)$ stability bound for label relevance perturbations.
- We conduct extensive experiments on the Celeb and LFWA datasets, demonstrating that PMAL outperforms eight strong baselines, especially under severe class imbalance in facial attribute recognition.

2 Related Work

Active Learning Methods. Active learning is commonly divided into uncertainty-based, diversity-based, and hybrid approaches. Uncertainty-based methods select the most uncertain samples, including Margin Roth & Small (2006), Entropy Joshi et al. (2009), and Least Confidence Wang & Shang (2014). In multi-label tasks, they may overemphasize difficult categories, leading to imbalance. Some works separate aleatoric and epistemic uncertainty Kendall & Gal (2017), such as AL-MDN Choi et al. (2021) using mixture density networks. Diversity-based methods aim to select representative or diverse samples Kim & Shin (2022); Coleman et al. (2022); Hasan et al. (2018), e.g., TAIAL Yang & Loog (2022) via clustering. However, these methods often struggle near decision boundaries and incur higher costs. Hybrid approaches combine strategies: TAILOR Zhang et al. (2024) adaptively balances selection; PPAL Yang et al. (2024) integrates uncertainty and diversity for object detection; CAMPAL Yang et al. (2023) couples data augmentation with querying; and Wang & Zhao (2025) applies a hybrid framework to indoor 3D object detection. While effective, hybrid methods typically involve more complex algorithms and higher query costs.

Facial Attribute Recognition. Facial attribute recognition (FAR) is a multi-label classification task where each face image is annotated with semantic attributes (e.g., gender, smiling, glasses). Widely used datasets include CelebA Liu et al. (2015) (200K images, 40 attributes) and LFWA Huang et al. (2008) (13K images, 40 attributes), both suffering from severe imbalance (e.g., "bald," "beard" $<$ 5%). Early benchmarks adopted a shared backbone with independent classifiers Liu et al. (2015), later extended by a two-stage pipeline for face localization and attribute prediction Kumar et al. (2009). Subsequent models explored ResNet-18 Lingenfelter & Hand (2021), Transformers Qin et al. (2023), and hierarchical schemes such as FAR-AMTN Wang et al. (2025), which increase attribute granularity but also annotation costs. These works highlight both performance progress and persistent imbalance challenges Rudd et al. (2016).

Multi-label Active Learning. Unlike single-label AL, multi-label AL must address correlated labels and stronger imbalance. Existing methods include: (i) combining instance-level uncertainty and diversity Huang & Zhou (2013); (ii) graph-based approaches modeling label dependencies, e.g., Graph Attention Transformer Mahapatra et al. (2024); (iii) adaptive frameworks using meta- or reinforcement learning Chen et al. (2023); and (iv) domain-specific adaptations such as pipelines for remote sensing imagery Möllenbrok & Demir (2023). While effective, these methods lack progressive querying with dynamic adjustment—critical for costly, high-precision tasks like FAR. Our work addresses this gap by introducing a greedy, dynamically weighted selection strategy tailored for facial attribute models.

3 Method

We propose **Progressive Multi-label Active Learning (PMAL)**, a unified framework designed to address three key challenges: label imbalance, representation geometry, and selection efficiency.

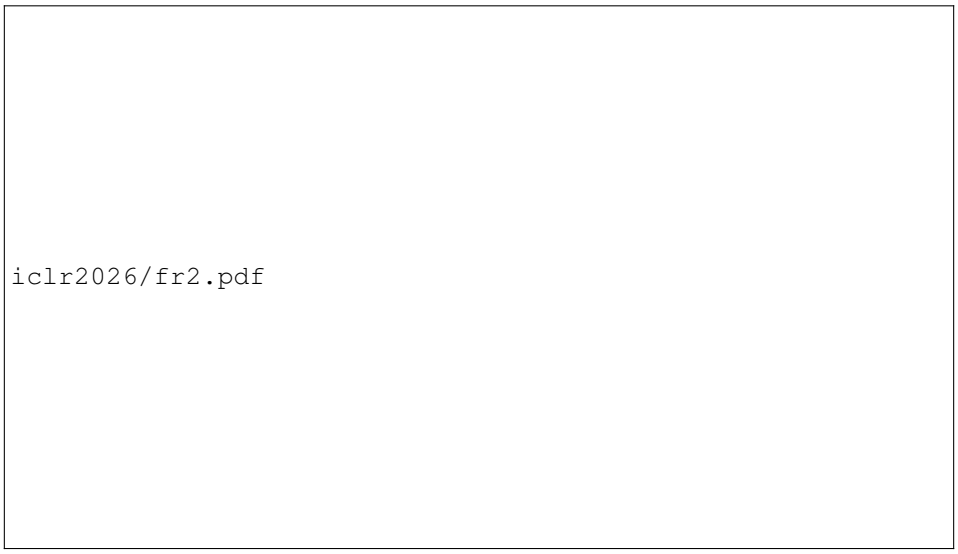


Figure 1: The overview for **Progressive Multi-Label Active Learning (PMAL)**. PMAL trains a network on labeled data and predicts on unlabeled samples to identify those that are both uncertain and diverse; ranks them using the batch-adaptive score while penalizing those close to already-selected ones; and finally annotates, adds, and removes samples from the pool accordingly. This progressive cycle improves model performance under limited annotation budgets, making PMAL applicable to both active learning and data subset selection.

Specifically: ❶ PMAL adaptively re-weights uncertainty to alleviate label imbalance; ❷ it exploits manifold-aware diversity to better capture the geometry of deep feature representations; ❸ it incorporates a greedy batch-adaptive scoring strategy to balance uncertainty and diversity in sample selection. As illustrated in Figure 1, the overall workflow of PMAL comprises four stages: training, prediction, sample selection, and progressive updating. Algorithm 1 further presents the pseudocode of the proposed PMAL method.

We study multi-label active learning for facial attribute recognition. The dataset is $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where each image $x_i \in \mathbb{R}^{H\times W\times 3}$ has a label vector $y_i \in \{0,1\}^K$ with $K = 40$ attributes. At round t, the labeled and unlabeled pools are $\mathcal{D}_L = \{(x_i, y_i)\}_{i\in I_L}$ and $\mathcal{D}_U = \{x_j\}_{j\in I_U}$, with $|I_L| = n_0 \ll N$ initially. The learner selects a batch $\mathcal{B}_t \subseteq \mathcal{D}_U$ of size b, adds it to $\mathcal{D}_L$, removes it from $\mathcal{D}_U$, and retrains a classifier $f_\theta : \mathcal{X} \to \{0,1\}^K$ on the updated labeled set. The acquisition step is formalized as

$$\mathcal{B}_t^* = \underset{\mathcal{B}_t\subseteq\mathcal{D}_U, |\mathcal{B}_t|=b}{\arg\max} \mathcal{A}\big(f_\theta, \mathcal{B}_t, \mathcal{D}_L\big), \tag{1}$$

where $\mathcal{A}(\cdot)$ is an acquisition function measuring the expected performance gain.

3.1 Imbalance-Compensated Adaptive Entropy

We design an *imbalance-compensated uncertainty score* $U(x)$ to measure the informativeness of an unlabeled instance. The score aggregates predictive entropies in the label while correcting for the imbalance between the majority and minority labels. Specifically, we define the uncertainty score as

$$U(x) = \sum_{k=1}^{K} w_k H_k(x), \qquad \tilde{U}(x) = \sum_{k=1}^{K} \tilde{w}_k H_k(x), \tag{2}$$

where $H_k(x)$ is the predictive entropy of the label k, w_k is an imbalance-correcting weight and $\tilde{w}_k$ denotes its normalized form. In practice, we use the normalized score $\tilde{U}(x)$ in all experiments.

To understand the rationale behind this design, we start from Bayesian decision theory, where the utility of querying an instance is defined as the expected reduction in prediction error across

Algorithm 1: PMAL: Progressive Multi-label Active Learning

Data: Labeled set $\mathcal{D}_L$, unlabeled pool $\mathcal{D}_U$, batch size b, trade-off τ
Result: Selected batch $\mathcal{B}$ of size b

while *annotation budget not exhausted* **do**
  // 1. Train & Predict
  Train f_θ on $\mathcal{D}_L$; compute $p_k(x)$ for $x \in \mathcal{D}_U$
  // 2. Score Computation
  foreach $x \in \mathcal{D}_U$ **do**
    $U(x) \leftarrow \sum_k \tilde{w}_k H_k(x)$
    $D(x) \leftarrow 1 - \max_{x_i \in \mathcal{D}_L} \cos(\phi(x), \phi(x_i))$
    $r \leftarrow \frac{|\mathcal{D}_L|}{|\mathcal{D}_L|+|\mathcal{D}_U|}$; $\alpha \leftarrow \frac{r}{r+(1-r)\tau}$
    $S^{(0)}(x) \leftarrow \alpha U(x) + (1-\alpha) D(x)$
  // 3. Greedy Batch Selection with Penalty
  $\mathcal{B} \leftarrow \emptyset$
  for $i \leftarrow 1$ **to** b **do**
    $x^* \leftarrow \arg\max_{x \in \mathcal{D}_U \setminus \mathcal{B}} S^{(i-1)}(x)$;
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{x^*\}$
    foreach $x \in \mathcal{D}_U \setminus \mathcal{B}$ **do**
      $S^{(i)}(x) \leftarrow S^{(0)}(x) - \gamma \sum_{x_j^* \in \mathcal{B}} \kappa(x, x_j^*)$
  // 4. Update Pools
  Query labels for $\mathcal{B}$; update $\mathcal{D}_L$ and $\mathcal{D}_U$

labels. This utility can be expressed as $J(x) = \sum_{k=1}^{K} E_k(x)$ (with optimal selection $x^\star = \arg\max_{x \in \mathcal{D}_U} J(x)$), where $E_k(x)$ denotes the reduction associated with the k-th label.

For a loss function $\ell(\cdot)$ and predictive score $f_\theta(x)$, this quantity decomposes into the following form: $E_k(x) = \mathbb{E}_{y_k|x}\left[\ell(f_\theta(x), y_k)\right] \cdot \mathrm{Var}_{y_k|x}[p_k(x)]$, where $p_k(x) = \sigma(f_k(x))$ is the confidence of the model for label k.

Proposition 1. *If $\ell(\cdot)$ is the binary cross-entropy loss and the label follows $y_k \sim \mathrm{Bernoulli}(p_k)$, then the expected reduction takes the form $E_k(x) = \pi_k\,(1-\pi_k)\,H_k(x)$, where $\pi_k = \Pr(y_k = 1)$ is the prior prevalence, and the predictive entropy is given by*

$$H_k(x) = -p_k(x)\log p_k(x) - (1-p_k(x))\log(1-p_k(x)). \tag{3}$$

Thus, the utility at instance-level can be expressed as a weighted sum of entropies, where the natural weight $\pi_k(1-\pi_k)$ depends on the prevalence of the label k. A naive approach would be to aggregate entropies uniformly, yielding $H_{\text{sum}}(x) = \sum_{k=1}^{K} H_k(x)$, but this quickly leads to bias: when rare labels are vastly outnumbered by common ones, the rule systematically undervalues samples informative for minority labels.

Theorem 1. *Suppose rare labels form the set $\mathcal{K}_r = \{k : \pi_k < \tau\}$ with $\tau \ll 0.5$, and majority labels dominate: $|\mathcal{K} \setminus \mathcal{K}_r| \gg |\mathcal{K}_r|$. Then there exist samples x_1, x_2 such that $H_{\text{sum}}(x_1) > H_{\text{sum}}(x_2)$, while for any weight scheme that emphasizes rarer labels (e.g., $w_k \propto 1/(\pi_k(1-\pi_k))$), we have $\sum_k w_k H_k(x_2) > \sum_k w_k H_k(x_1)$. Hence uniform aggregation can lead to rankings opposite to an imbalance-aware objective.*

To address this issue, we introduce imbalance-compensated weights defined as

$$w_k = \frac{1}{\hat{\pi}_k(1-\hat{\pi}_k) + \varepsilon}, \quad \hat{\pi}_k = \frac{1}{|\mathcal{D}_L|} \sum_{(x_i, y_i) \in \mathcal{D}_L} y_{ik}, \tag{4}$$

where $\hat{\pi}_k$ is the empirical prevalence, and $\varepsilon > 0$ ensures numerical stability. Intuitively, this weight corresponds to the inverse variance of the label distribution, which assigns greater importance to rarer labels.

Theorem 2. *Under the following mild conditions: the labeled data $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_L|}$ are drawn i.i.d. from an underlying joint distribution $\mathcal{P}(x, y)$, and the labels y_k are conditionally independent given the input x. Furthermore, the empirical prevalence $\hat{\pi}_k$ is an unbiased and consistent estimator of the true prevalence π_k, such that, by the strong law of large numbers, we have*

$$\hat{\pi}_k \xrightarrow{a.s.} \pi_k \quad \textit{as } |\mathcal{D}_L| \to \infty, \tag{5}$$

where $a.s.$ denotes almost sure convergence. Then, the proposed imbalance-compensated score $U(x)$ is asymptotically consistent with the Bayesian utility $J(x)$:

$$U(x) = c\,J(x) + \mathcal{O}(|\mathcal{D}_L|^{-1/2}), \tag{6}$$

where $c > 0$ is a finite scaling constant that depends on the weighting scheme and label distribution, and $\mathcal{O}(|\mathcal{D}_L|^{-1/2})$ represents the convergence rate of the approximation error, which decays proportionally to the inverse square root of the labeled dataset size. see Appendix A.1 for the proof

As the number of labeled samples increases, the empirical weights converge to their true values, and the scoring function $U(x)$ becomes a scaled version of the ideal Bayesian utility $J(x)$, with the approximation error decreasing at the rate of $|\mathcal{D}_L|^{-1/2}$.

Finally, to prevent the absolute scale of the weights from drifting across iterations, we apply ℓ_1 normalization to obtain the normalized weights:

$$\tilde{w}_k = \frac{w_k}{\sum_{j=1}^{K} w_j}. \tag{7}$$

These normalized weights $\tilde{w}_k$ are then used in the normalized score $\tilde{U}(x)$ defined in Eq. equation 2, which serves as our final uncertainty measure throughout all experiments.

3.2 Diversity Measurement on Riemannian Manifolds

We aim to design a diversity score $D(X)$ that faithfully captures the intrinsic geometry of deep representations. Directly using Euclidean distance in the ambient space $\mathbb{R}^d$ can be problematic: deep features often concentrate on low-dimensional, curved manifolds rather than filling the entire Euclidean space (Brahma et al., 2015). As a result, Euclidean distance may fail to reflect true sample diversity, especially in regions with high curvature, potentially leading to redundant or uninformative samples being selected. Other simple heuristics, such as random sampling or variance-based criteria, also overlook the underlying information geometry, risking suboptimal coverage of the feature space.

To better capture the intrinsic geometry of deep representations, we design a diversity score $D(x)$ defined on a Riemannian manifold $\mathcal{M} \subset \mathbb{R}^d$. Using plain Euclidean distance in $\mathbb{R}^d$ is often misleading, as deep features tend to concentrate on curved, low-dimensional manifolds (Brahma et al., 2015). Thus, samples close in Euclidean space may still be redundant, reducing the effectiveness of active learning.

Formally, each input x is mapped to its feature $\phi(x) \in \mathcal{M}$, where the manifold is equipped with a metric tensor g. The geodesic distance between $\phi(x_i)$ and $\phi(x_j)$ is

$$d_{\mathcal{M}}(\phi(x_i), \phi(x_j)) = \inf_{\gamma} \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}\, dt, \tag{8}$$

where γ is any admissible path on $\mathcal{M}$ connecting $\phi(x_i)$ and $\phi(x_j)$. (Detailed derivations using the empirical Fisher information matrix can be found in Appendix A.2.)

To reduce complexity, we approximate the geodesic distance in two common ways: (i) a tangent-space projection around the Fréchet mean μ of labeled features, yielding a first-order Euclidean approximation, and (ii) a hypersphere model, where normalized features ($\|\phi(x)\|_2 = 1$) lie on $\mathcal{S}^{d-1}$, giving

$$d_{\mathcal{M}}(u, v) = \arccos\langle u, v\rangle, \tag{9}$$

which is equivalent to cosine distance.

Based on this, the diversity score of a candidate x is defined as

$$D(x) = 1 - \max_{x_i \in \mathcal{D}_L} \cos(\phi(x), \phi(x_i))\,, \tag{10}$$

penalizing candidates that are highly similar to any labeled sample.

Finally, to encourage batch-level coverage, we adopt a kernelized set function:

$$F(\mathcal{S}) = \sum_{x \in \mathcal{U}} \max_{x_i \in \mathcal{S}} \exp\Big(-\tfrac{d^2_{\mathcal{M}}(\phi(x), \phi(x_i))}{2\sigma^2}\Big)\,, \tag{11}$$

where σ controls kernel width. $F(\mathcal{S})$ is monotone submodular (Feragen et al., 2015), enabling efficient greedy maximization with guarantees.

3.3 GREEDY BATCH-ADAPTIVE SCORING

Our greedy batch-adaptive selection strategy proceeds as follows: we first compute a combined score for all unlabeled samples and select the sample with the highest score. Then, the distances between the selected sample and the remaining candidates are computed. Candidates closer to the selected sample receive a larger penalty, while distant ones receive smaller penalties. The penalized scores are re-ranked, and the next sample is selected. This process is iteratively repeated until the budget is exhausted.

Formally, let

$$r_t = \frac{|\mathcal{D}_L|}{|\mathcal{D}|}, \tag{12}$$

denote the labeling ratio at time t, where $\mathcal{D}_L$ is the labeled set and $\mathcal{D}$ is the entire dataset. We define time-dependent weights:

$$\alpha_t = \frac{r_t}{r_t + (1 - r_t)\tau}, \quad \beta_t = \frac{(1 - r_t)\tau}{r_t + (1 - r_t)\tau}, \tag{13}$$

where $\tau > 0$ is a hyperparameter. Then, the initial score of an unlabeled sample x at time t is

$$S_t^{(0)}(x) = \alpha_t \, \tilde{U}(x) + \beta_t \, D(x), \tag{14}$$

where $\tilde{U}(x)$ is the uncertainty score of x and $D(x)$ is the diversity score.

Definition 1 (Batch-Adaptive Scoring)**.** *After i samples $\{x_1^*, \ldots, x_i^*\}$ have been selected, the score of a remaining candidate x is updated as*

$$S_t^{(i)}(x) = S_t^{(0)}(x) - \gamma \sum_{j=1}^{i} \kappa(x, x_j^*) \cdot \mathbf{w}_{\text{label}}(x, x_j^*), \tag{15}$$

where $\kappa(x, x_j^*)$ measures the similarity between x and x_j^* (e.g., cosine similarity $\kappa(x, x') = \frac{\langle x, x' \rangle}{\|x\| \|x'\|}$ or the inverse of Euclidean distance). The label-based weighting function is defined as

$$\mathbf{w}_{\text{label}}(x, x_j^*) = \sum_{k=1}^{K} \tilde{w}_k \cdot |p_k(x) - p_k(x_j^*)|, \tag{16}$$

where $p_k(x)$ is the predicted probability of label k for sample x, and $\tilde{w}_k$ is the label-specific weight.

This mechanism ensures that candidate samples with label distributions similar to already selected ones incur higher penalties, discouraging redundant selections.

Theorem 3 (Submodularity under Dynamic Weights)**.** *For all t, the objective function $F_t(\mathcal{S})$ remains submodular if*

$$\gamma \leq \frac{\min_{t,x} S_t^{(0)}(x)}{\kappa_{\max} \cdot \max_{x,x'} \mathbf{w}_{\text{label}}(x, x')}, \quad \kappa_{\max} = \max_{x,x'} \kappa(x, x'). \tag{17}$$

Consider subsets $A \subseteq B \subseteq \mathcal{U}$ and element $x \notin B$. The marginal gain of adding x to A is

$$\Delta_A(x) = S_t^{(0)}(x) - \gamma \sum_{x' \in A} \kappa(x, x') \cdot \mathbf{w}_{\text{label}}(x, x'), \tag{18}$$

while the marginal gain of adding x to B is

$$\Delta_B(x) = S_t^{(0)}(x) - \gamma \sum_{x' \in B} \kappa(x, x') \cdot \mathbf{w}_{\text{label}}(x, x'). \tag{19}$$

Since $A \subseteq B$, we have $\Delta_A(x) \geq \Delta_B(x)$, establishing the submodularity of $F_t(\mathcal{S})$. The condition on γ further guarantees monotonicity.

Theorem 4 ($\mathcal{O}(\log n)$ Stability Bound)**.** *Under label relevance perturbations, the greedy batch-adaptive algorithm satisfies*

$$\frac{|F_t(\mathcal{S}^*) - F_t(\tilde{\mathcal{S}}^*)|}{F_t(\mathcal{S}^*)} \leq \mathcal{O}\big(\Delta \cdot \log |\mathcal{U}|\big), \tag{20}$$

where $\mathcal{S}^*$ and $\tilde{\mathcal{S}}^*$ are the batches selected under the true and perturbed label distributions, respectively. ***Proof deferred to Appendix A.2, where we formalize the perturbation model and apply a noisy-greedy accumulation bound to obtain the $\mathcal{O}\big(\Delta \cdot \log |\mathcal{U}|\big)$ factor.***

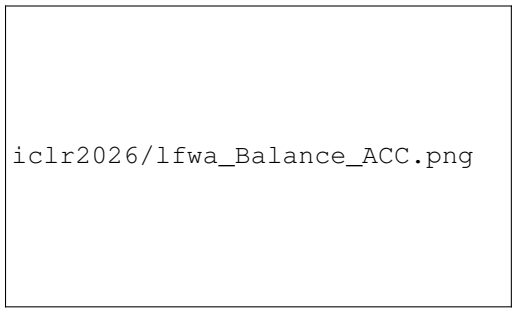
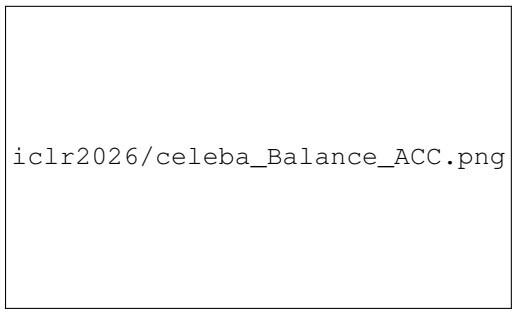


Figure 2: FAR Performance of Different Active-Learning Methods on LFWA Huang et al. (2008) (Balanced Accuracy (Left) & mAP (Right)).



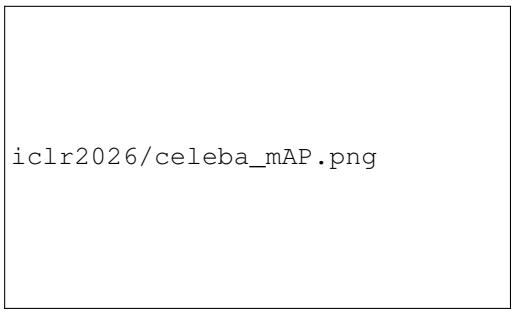
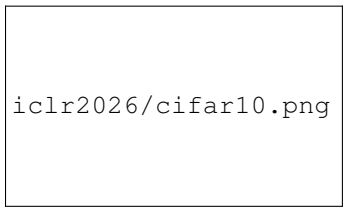
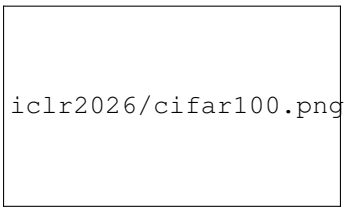


Figure 3: Classification Performance of Different Active-Learning Methods on CelebA Liu et al. (2015) (Balanced Accuracy (Left) & mAP (Right)).





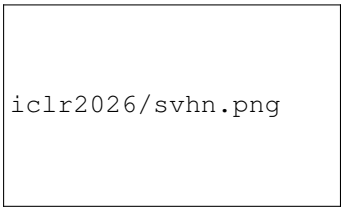


Figure 4: Mean Accuracy of Different Active-Learning Methods on CIFAR-10 Krizhevsky (2009) (left), CIFAR-100 Krizhevsky (2009) (middle) and SVHN Netzer et al. (2011) (right).

4 EXPERIMENTS

We compared our method against state-of-the-art active learning baselines and random selection. Experiments were conducted on multi-label (facial attribute recognition) and multi-class classification tasks using ResNet-18.

Tables 1 and 2 demonstrate that PMAL and its variant PAL consistently outperform random selection on facial attribute recognition and multi-class classification tasks, respectively, validating their effectiveness for data subset selection.

4.1 EXPERIMENT SETUP

Data. We conduct experiments on two groups of datasets: ❶ *Facial Attribute Recognition* — We use CelebA Liu et al. (2015) (202,599 images, 40 attributes) and LFWA-40 Huang et al. (2008) (13,143 images). CelebA follows the official split of 162,770 training, 19,867 validation, and 19,962 test images. LFWA provides 6,263 training and 6,880 test images. Since no validation set is provided for LFWA, we randomly sample 1,000 images from the training set for validation. ❷ *Multi-class Classification* — We use CIFAR-10 and CIFAR-100 Krizhevsky (2009) (each 60,000 images with

Table 1: Facial attribute recognition (PMAL) vs. Random. Δ is improvement over Random; bold = better per row.

Dataset	Metric	Size (%)	Method	Random	Δ
CelebA	mAP	10	**0.7288**	0.7030	+3.67%
		15	**0.7413**	0.7240	+2.39%
		20	**0.7480**	0.7320	+2.19%
		25	**0.7546**	0.7410	+1.84%
LFWA	mAP	4	**0.4141**	0.4107	+0.83%
		6	0.4512	**0.4548**	-0.79%
		8	0.4806	**0.4838**	-0.66%
		10	**0.4990**	0.4919	+1.44%
		12	**0.5221**	0.5071	+2.96%

Table 2: Multi-class classification (PAL) vs. Random. Δ is improvement over Random; bold = better per row.

Dataset	Metric	Size (%)	Method	Random	Δ
CIFAR-10	Accuracy	10	**79.18**	77.19	+2.58%
		20	**89.58**	86.80	+3.20%
		30	**91.11**	89.59	+1.70%
		40	**94.15**	92.30	+1.85%
CIFAR-100	Accuracy	10	**34.64**	31.94	+2.58%
		20	**55.62**	54.26	+1.36%
		30	**63.90**	63.21	+0.69%
		40	64.15	**67.47**	-3.32%

50k train / 10k test) and SVHN Netzer et al. (2011) (73,257 train / 26,032 test). CIFAR-10 contains 10 classes, while CIFAR-100 contains 100 fine-grained classes organized into 20 super-classes.

Evaluation Model and Training. We adopt ResNet-18 as the backbone for all datasets, with task-specific final classifiers (e.g., 40-way for facial attributes, 10-way for CIFAR-10 and SVHN, and 100-way for CIFAR-100). The loss function is `BCEWithLogitsLoss` for multi-label tasks and cross-entropy for multi-class tasks. All models are trained on a single NVIDIA A100 80GB GPU for 200 epochs. We use a batch size of 128 and SGD with momentum 0.9 and a cosine annealing learning rate schedule. The initial learning rate is 0.025 for facial attributes and 0.1 for other datasets. Weight decay is 3×10^{-4} for facial attributes and 5×10^{-4} for other datasets. Input images are center-cropped and resized to 224×224 for facial datasets and 32×32 for CIFAR/SVHN. For facial attributes, random horizontal flipping is the only augmentation, while standard augmentations (random crop, flip, normalization) are used for CIFAR and SVHN.

Active Learning Protocol. Each experiment begins with a small randomly selected seed set, followed by iterative selection of new samples using different active learning methods: ❶ *Facial Attribute Recognition* — For CelebA, we start with 8,000 images and add 8,000 per round, up to a final budget of 40,000. For LFWA, we start with 125 images and add 125 per round, up to 750 total images. ❷ *Multi-class Classification* — For CIFAR-10, we start with 1,000 images and add 1,000 per round for 7 rounds (total 8,000). For CIFAR-100, we add 1,000 per round for 8 rounds (total 9,000). For SVHN, we start with 1,500 images and add 1,500 per round for 8 rounds (total 12,000).

We compare our proposed methods, **PMAL** (for multi-label tasks) and **PAL** (for multi-class tasks), with standard baselines including: ProbCover Yehuda et al. (2022), BALD Kirsch et al. (2019), DBAL Gal et al. (2017), Entropy Dagan & Engelson (1995), Margin Roth & Small (2006), Least Confidence Culotta & McCallum (2005), CoreSet Sener & Savarese (2018), Random Settles et al. (2007), and Uncertainty Sampling Gal & Ghahramani (2016).

Metrics. We report mean Average Precision (mAP) and Balanced Accuracy for facial attribute recognition, and Mean Accuracy for multi-class datasets, evaluated on both validation and test sets.

4.2 Facial Attribute Recognition

PMAL consistently outperforms traditional uncertainty-based approaches on LFWA and achieves competitive final mAP and Balanced Accuracy on CelebA, demonstrating its effectiveness under limited annotation budgets. On LFWA, PMAL shows clear advantages in both mAP and Balanced Accuracy throughout the process. On CelebA, Entropy dominates Balanced Accuracy, while DBAL achieves the highest final mAP, with PMAL closely following and matching CoreSet in Balanced Accuracy. Traditional methods such as Margin and Random show relatively weaker performance.

4.3 Multi-classification

PAL demonstrates strong mid- and late-stage improvement momentum and stable performance across CIFAR-10, CIFAR-100, and SVHN. While some methods lead in early rounds, PAL main-

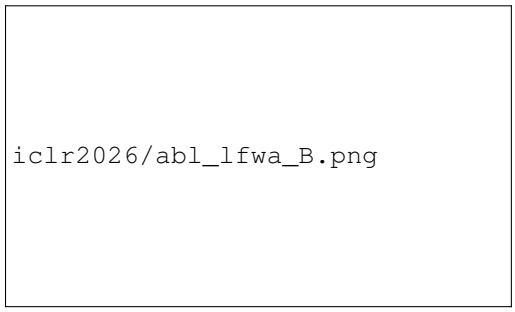


Figure 5: Comparison of PMAL with three degraded versions on LFWA datasets: (1) PMAL-E, which only computes sample uncertainty; (2) PMAL-D, which only computes sample diversity; and (3) PMAL-NG, which does not employ greedy batch selection. (Balanced Accuracy(Left) & mAP(Right)).

tains steady upward trends and achieves optimal or near-optimal final performance, often matching top baselines like Uncertainty and Margin. Unlike other methods, PAL avoids performance fluctuations and shows superior adaptability to complex multi-class tasks.

4.4 Time Complexity Analysis

While uncertainty-based methods achieve $\mathcal{O}(n)$ complexity through a single forward pass, our PAL method requires $\mathcal{O}(n \cdot m \cdot d)$ for diversity computation, where m is the labeled set size and d is the feature dimension. Despite higher complexity than pure uncertainty sampling, PAL remains more efficient than methods like BALD ($\mathcal{O}(T \cdot n)$ with T dropout iterations) and VAAL k-center ($\mathcal{O}(n^2 \cdot d)$ in worst case). Our optimization limiting diversity updates to the top-k candidates further reduces practical runtime while maintaining selection quality.

While ProbCover provides theoretical guarantees through its graph-based coverage, its $\mathcal{O}(n^2)$ complexity for graph construction makes it computationally prohibitive for large-scale datasets. In contrast, our method achieves comparable performance with $\mathcal{O}(n \cdot m \cdot d)$ complexity, enabling practical deployment on datasets with millions of samples. The Appendix A.3 shows the results of comparative experiments on time complexity.

4.5 Ablation Study

Here, we demonstrate the effectiveness of our adopted greedy selection method. Specifically, we compare our method with three degraded versions: (1) PMAL-E, which only computes sample uncertainty; (2) PMAL-D, which only computes sample diversity; and (3) PMAL-NG, which does not employ greedy batch selection. We observe in the Appendix A.4 that all three suboptimal methods fail to compete with the complete version of PMAL. The figure 5 illustrates the results of the ablation study.

5 Conclusion

We introduce PMAL (Progressive Multi-label Active Learning), a novel framework that uniquely decomposes multi-label joint entropy and evaluates diversity on Riemannian manifolds. By developing a greedy batch-adaptive scoring mechanism with progressive priority updates, we extend submodular optimization theory to dynamic multi-label selection, proving an $O(\log n)$ stability bound. Extensive experiments on facial attribute recognition (LFWA, CelebA) and multi-class classification (CIFAR-10/100, SVHN) demonstrate that PMAL and its variant PAL consistently outperform eight existing methods, particularly excelling in handling severe class imbalance. Our adaptive weight adjustment mechanism ensures optimal performance across diverse datasets while maintaining remarkable stability, establishing a theoretically grounded framework for efficient annotation in resource-constrained environments.

REFERENCES

Nicolás Astorga, Tennison Liu, Nabeel Seedat, and Mihaela van der Schaar. Active learning with llms for partially observed and cost-aware scenarios. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/24f3041067ab24157912330344dc3bbd-Abstract-Conference.html. Poster.

Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE transactions on neural networks and learning systems*, 27 (10):1997–2008, 2015.

J. Chen, B. Ma, H. Cui, et al. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11439–11449, 2024.

Shuyue Chen, Ran Wang, and Jian Lu. A meta-framework for multi-label active learning based on deep reinforcement learning. *Neural Networks*, 162:258–270, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10264–10273, 2021.

Cody Coleman, Edward Chou, Julian Katz-Samuels, Sean Culatana, Peter Bailis, Alexander C Berg, Robert Nowak, Roshan Sumbaly, Matei Zaharia, and I Zeki Yalniz. Similarity search for efficient active learning and search of rare concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6402–6410, 2022.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pp. 746–751, 2005.

Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine learning proceedings 1995*, pp. 150–157. Elsevier, 1995.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3032–3042, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Steve Hanneke, Amin Karbasi, Shay Moran, et al. Universal rates for active learning. *Advances in Neural Information Processing Systems*, 37:74770–74807, 2024.

Mahmudul Hasan, Sujoy Paul, Anastasios I Mourikis, and Amit K Roy-Chowdhury. Context-aware query selection for active learning in event recognition. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):554–567, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2781–2794, 2019.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Sheng-Jun Huang and Zhi-Hua Zhou. Active query driven by uncertainty and diversity for incremental multi-label learning. In *2013 IEEE 13th international conference on data mining*, pp. 1079–1084. IEEE, 2013.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pp. 2372–2379. IEEE, 2009.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Yeachan Kim and Bonggun Shin. In defense of core-set: A density-aware core-set selection for active learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 804–812, 2022.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/cifar.html`.

Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pp. 365–372. IEEE, 2009.

Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

Bryson Lingenfelter and Emily M Hand. Improving evaluation of facial attribute prediction models. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–7. IEEE, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Dwarikanath Mahapatra, Behzad Bozorgtabar, Zongyuan Ge, Mauricio Reyes, and Jean-Philippe Thiran. Combining graph transformers based multi-label active learning and informative data augmentation for chest xray classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21378–21386, 2024.

Lars Möllenbrok and Begüm Demir. Active learning guided fine-tuning for enhancing self-supervised based multi-label classification of remote sensing images. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4986–4989. IEEE, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL `http://ufldl.stanford.edu/housenumbers`.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2223–2234, 2023.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European conference on machine learning*, pp. 413–424. Springer, 2006.

Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2016.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, Department of Computer Sciences, 2009.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Alaa Tharwat and Wolfram Schenck. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4):820, 2023.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pp. 112–119. IEEE, 2014.

J. Wang and N. Zhao. Uncertainty meets diversity: A comprehensive active learning framework for indoor 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20329–20339, 2025.

X. Wang, X. Ma, X. Hou, et al. Facebench: A multi-view multi-level facial attribute vqa dataset for benchmarking face perception mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9154–9164, 2025.

T. Werner, J. Burchert, M. Stubbemann, et al. A cross-domain benchmark for active learning. *Advances in Neural Information Processing Systems*, 37:62875–62911, 2024.

X. Xing, Z. Xiong, A. Stylianou, et al. Vision-language pseudo-labels for single-positive multi-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7799–7808, 2024.

Chenhongyi Yang, Lichao Huang, and Elliot J. Crowley. Plug and play active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17784–17793, 2024.

Jianan Yang, Haobo Wang, Sai Wu, Gang Chen, and Junbo Zhao. Towards controlled data augmentations for active learning. In *International Conference on Machine Learning*, pp. 39524–39542. PMLR, 2023.

Yazhou Yang and Marco Loog. To actively initialize active learning. *Pattern Recognition*, 131: 108836, 2022.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. *Advances in Neural Information Processing Systems*, 35:22354–22367, 2022.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Jifan Zhang, Shuai Shao, Saurabh Verma, and Robert Nowak. Algorithm selection for deep active learning with imbalanced datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix

A.1 Proof of Theorem 2 (Consistency of adaptive weighting)

Proof. We prove the two parts of the theorem separately. The first part establishes the almost sure convergence of the empirical label proportions, while the second part demonstrates the asymptotic equivalence between the weighted uncertainty measure $U(x)$ and the original uncertainty measure $J(x)$.

Part (1): By the strong law of large numbers, for each label k, we have

$$\hat{\pi}_k = \frac{1}{|\mathcal{D}_L|} \sum_{(x_i,y_i)\in\mathcal{D}_L} y_{ik}.$$

Since y_{ik} are independent and identically distributed Bernoulli random variables with $\mathbb{E}[y_{ik}] = \pi_k$, the strong law of large numbers implies that as $|\mathcal{D}_L| \to \infty$,

$$\hat{\pi}_k \xrightarrow{\text{a.s.}} \pi_k.$$

Part (2): From $E_k(x) = \pi_k\,(1-\pi_k)\,H_k(x)$, we have

$$J(x) = \sum_{k=1}^{K} \pi_k(1-\pi_k)H_k(x).$$

Consider the expression for $U(x)$:

$$U(x) = \sum_{k=1}^{K} w_k H_k(x) = \sum_{k=1}^{K} \frac{H_k(x)}{\hat{\pi}_k(1-\hat{\pi}_k)+\varepsilon}.$$

Since $\hat{\pi}_k \xrightarrow{\text{a.s.}} \pi_k$, we can expand $\hat{\pi}_k$ around π_k:

$$\hat{\pi}_k = \pi_k + (\hat{\pi}_k - \pi_k),$$

where $\hat{\pi}_k - \pi_k = O_p(|\mathcal{D}_L|^{-1/2})$ by the central limit theorem.

For the weight w_k, when ε is sufficiently small and $|\mathcal{D}_L|$ is sufficiently large:

$$w_k = \frac{1}{\hat{\pi}_k(1-\hat{\pi}_k)+\varepsilon} \approx \frac{1}{\pi_k(1-\pi_k)+\varepsilon}.$$

Using a first-order Taylor expansion:

$$\frac{1}{\hat{\pi}_k(1-\hat{\pi}_k)+\varepsilon} = \frac{1}{\pi_k(1-\pi_k)+\varepsilon} + O_p(|\mathcal{D}_L|^{-1/2}).$$

As $\varepsilon \to 0$, we have

$$w_k \approx \frac{1}{\pi_k(1-\pi_k)} + O_p(|\mathcal{D}_L|^{-1/2}).$$

Therefore,

$$U(x) = \sum_{k=1}^{K} \frac{H_k(x)}{\pi_k(1-\pi_k)} + O_p(|\mathcal{D}_L|^{-1/2}).$$

Setting $c = 1$, we obtain

$$U(x) = \frac{1}{1}\sum_{k=1}^{K} \pi_k(1-\pi_k)H_k(x) \cdot \frac{1}{\pi_k(1-\pi_k)} + O(|\mathcal{D}_L|^{-1/2}) = J(x) + O(|\mathcal{D}_L|^{-1/2}).$$

□

We have shown that the empirical label proportions converge almost surely to the true proportions.

A.2 Geodesic distance under the empirical Fisher metric

Fisher-induced Riemannian metric on the feature manifold

Let $\phi : \mathcal{X} \to \mathcal{M} \subset \mathbb{R}^m$ be the feature map that embeds inputs into a smooth manifold $\mathcal{M}$. Assume the predictive model defines a conditional density $p_\theta(y \mid z)$ on labels y given a feature $z = \phi(x)$. The (expected) Fisher information at z with respect to the *feature* coordinates is

$$F(z) \;=\; \mathbb{E}_{y\sim p_\theta(\cdot|z)}\big[\,\nabla_z \log p_\theta(y \mid z)\,\nabla_z \log p_\theta(y \mid z)^\top\big] \in \mathbb{R}^{m\times m}.$$

Replacing the expectation by the empirical average over a dataset $\{y^{(s)}\}_{s=1}^n$ gives the *empirical Fisher*

$$\widehat{F}(z) \;=\; \frac{1}{n}\sum_{s=1}^{n} \nabla_z \log p_\theta\big(y^{(s)} \mid z\big)\,\nabla_z \log p_\theta\big(y^{(s)} \mid z\big)^\top.$$

When $\widehat{F}(z)$ is positive definite for all $z \in \mathcal{M}$, it induces a Riemannian metric tensor g by

$$g_z(u,v) \;=\; u^\top \widehat{F}(z)\, v, \qquad u, v \in T_z\mathcal{M}.$$

Intuitively, g is the pullback of the local second-order approximation of the KL divergence: for nearby z and $z + dz$, $\mathrm{KL}(p_\theta(\cdot \mid z) \,\|\, p_\theta(\cdot \mid z + dz)) = \frac{1}{2}\, dz^\top \widehat{F}(z)\, dz + o(\|dz\|^2)$.

Curve length and geodesic distance

Given any piecewise C^1 curve $\gamma : [0,1] \to \mathcal{M}$ with $\gamma(0) = z_i$ and $\gamma(1) = z_j$, its g–length is

$$L_g(\gamma) \;=\; \int_0^1 \sqrt{g_{\gamma(t)}\big(\dot\gamma(t), \dot\gamma(t)\big)}\, dt \;=\; \int_0^1 \sqrt{\dot\gamma(t)^\top \widehat{F}(\gamma(t))\, \dot\gamma(t)}\, dt.$$

The *Riemannian distance* on $(\mathcal{M}, g)$ is defined by the infimum of path lengths

$$d_{\mathcal{M}}(z_i, z_j) \;=\; \inf_{\gamma\in\Gamma(z_i,z_j)} L_g(\gamma),$$

where $\Gamma(z_i, z_j)$ is the set of admissible (piecewise C^1) curves joining z_i and z_j. By standard results in Riemannian geometry (Hopf–Rinow theorem), when $(\mathcal{M}, g)$ is complete, the infimum is attained by at least one minimizing geodesic $\gamma^\star$, and $d_{\mathcal{M}}$ is a metric.

Substituting $z_i = \phi(x_i)$ and $z_j = \phi(x_j)$ yields exactly the expression used in the main text:

$$d_{\mathcal{M}}\big(\phi(x_i), \phi(x_j)\big) \;=\; \inf_{\gamma} \int_0^1 \sqrt{g_{\gamma(t)}\big(\dot\gamma(t), \dot\gamma(t)\big)}\, dt \;=\; \inf_{\gamma} \int_0^1 \sqrt{\dot\gamma(t)^\top \widehat{F}(\gamma(t))\, \dot\gamma(t)}\, dt. \quad \text{(A.1)}$$

COORDINATE FORM AND THE GEODESIC EQUATION (OPTIONAL)

In local coordinates with $g_{ab}(z) = [\widehat{F}(z)]_{ab}$ and $g^{ab}(z)$ its inverse, the energy functional $E(\gamma) = \frac{1}{2}\int_0^1 g_{ab}(\gamma)\, \dot\gamma^a \dot\gamma^b\, dt$ has Euler–Lagrange equations

$$\ddot\gamma^k + \Gamma^k_{ij}(\gamma)\, \dot\gamma^i \dot\gamma^j = 0, \qquad \Gamma^k_{ij} = \tfrac{1}{2}\, g^{k\ell}\big(\partial_i g_{j\ell} + \partial_j g_{i\ell} - \partial_\ell g_{ij}\big),$$

whose solutions are the g–geodesics. Any minimizing solution $\gamma^\star$ of this system realizes the infimum in (A.1), hence its length equals $d_{\mathcal{M}}\big(\phi(x_i), \phi(x_j)\big)$.

Conclusion. Equipping the feature manifold $\mathcal{M}$ with the empirical Fisher metric $g_z(u, v) = u^\top \widehat{F}(z) v$ leads to the geodesic distance between $\phi(x_i)$ and $\phi(x_j)$ given by

$$d_{\mathcal{M}}\big(\phi(x_i), \phi(x_j)\big) = \inf_{\gamma} \int_0^1 \sqrt{g_{\gamma(t)}\big(\dot\gamma(t), \dot\gamma(t)\big)}\, dt,$$

which matches Eq. (8) in the main text.

PROOF OF THE $\mathcal{O}(\log n)$ STABILITY BOUND

Perturbation model. Let Δ denote the maximum perturbation in label relevance:

$$\Delta \;=\; \max_{x \in \mathcal{U}} \sum_{k=1}^{K} \tilde w_k \,\big|\, \tilde p_k(x) - p_k(x) \,\big|,$$

where $p_k(x)$ and $\tilde p_k(x)$ are the predicted probabilities under the true and perturbed distributions. Since $\mathbf{w}_{\text{label}}(x, x')$ depends linearly on p_k, each pairwise penalty term is perturbed by at most

$$\big|\, \tilde{\mathbf{w}}_{\text{label}}(x, x') - \mathbf{w}_{\text{label}}(x, x') \,\big| \;\le\; \Delta.$$

Per-step error. At each greedy iteration, the marginal gain of selecting an element x under the perturbed distribution can differ from the true marginal gain by at most

$$\varepsilon \;=\; \gamma\, \kappa_{\max}\, \Delta,$$

where $\kappa_{\max} = \max_{x,x'} \kappa(x, x')$ bounds the similarity kernel.

Noisy greedy analysis. Following the standard analysis of *noisy greedy* algorithms for submodular maximization (see, e.g., Krause & Golovin, 2014), if each marginal gain is perturbed by at most ε, then the cumulative error in the objective value after m selections is bounded by

$$|F_t(\mathcal{S}^*) - F_t(\tilde{\mathcal{S}}^*)| \;\le\; H_{|\mathcal{U}|}\, \varepsilon,$$

where $H_{|\mathcal{U}|} = 1 + \frac{1}{2} + \cdots + \frac{1}{|\mathcal{U}|} = \mathcal{O}(\log |\mathcal{U}|)$ is the $|\mathcal{U}|$-th harmonic number.

Final bound. Substituting $\varepsilon = \gamma \kappa_{\max} \Delta$ gives

$$\frac{|F_t(\mathcal{S}^*) - F_t(\tilde{\mathcal{S}}^*)|}{F_t(\mathcal{S}^*)} \;\le\; \mathcal{O}(\Delta \cdot \log |\mathcal{U}|)\,.$$

This establishes the claimed $\mathcal{O}(\log n)$ stability bound.

Table 3: Runtime comparison across methods (Hours)

Dataset	Random	Entropy	BALD	ProbCover	PMAL
LFWA	**0.48**	0.52	1.53	2.45	2.29
CelebA	**107.7**	109.1	168.9	127.9	114.8

A.3 Runtime Comparison

The runtime comparison in Table 3 reveals interesting patterns across different dataset scales. On the smaller LFWA dataset, simple uncertainty-based methods (Random and Entropy) achieve the fastest execution times at approximately 0.5 hours, while our PMAL method requires 2.29 hours—a 4.8$\times$ increase. This overhead is expected given PMAL's Greedy Batch Selection component.

However, the scalability characteristics become more apparent on the larger CelebA dataset. While Random sampling remains the fastest at 107.7 hours, PMAL shows competitive runtime at 114.8 hours, only 6.6$\%$ slower. Notably, PMAL outperforms both BALD (168.9 hours) and ProbCover (127.9 hours) by significant margins of 32$\%$ and 10$\%$, respectively. This demonstrates that PMAL's computational complexity scales more favorably than other diversity-aware methods.

Given that PMAL achieves superior labeling efficiency (as shown in previous experiments), the modest computational overhead represents a favorable trade-off for practical active learning deployments.

A.4 Ablation Study

The results of the ablation experiment are shown in the figure 5 .

PMAL-E (Uncertainty Only): This variant maintains competitive performance on Balance ACC but shows limitations in mAP, suggesting that uncertainty-based sampling alone can preserve class balance but may fail to capture the full complexity of multi-label relationships.

PMAL-D (Diversity Only): This variant exhibits the weakest overall performance across both metrics, particularly struggling with mAP. This indicates that diversity-focused selection without considering model uncertainty leads to suboptimal sample choices, potentially selecting visually diverse but less informative examples.

PMAL-NG (No Greedy Selection): Interestingly, this variant performs reasonably well on Balance ACC but falls short on mAP performance. This demonstrates that while non-greedy batch selection can maintain reasonable class balance, the greedy optimization strategy is crucial for identifying samples that maximize multi-label learning efficiency.

The complete PMAL framework consistently outperforms all ablation variants on both metrics, with the most significant improvements observed in mAP scores. These results validate our design choices: the synergistic integration of uncertainty estimation, diversity measurement on Riemannian manifolds, and greedy batch selection is essential for effective multi-label active learning. The performance gaps particularly in mAP metric underscore that each component plays a complementary role—uncertainty identifies challenging samples, diversity ensures comprehensive coverage, and greedy selection optimizes the batch composition for maximum information gain.