Combining Discrepancy-Confusion Uncertainty and Calibration Diversity for Active Fine-Grained Image Classification

Yinghao Jin and Xi Yang*

Jilin University, China

Abstract. Active learning (AL) aims to build high-quality labeled datasets by iteratively selecting the most informative samples from an unlabeled pool under limited annotation budgets. However, in fine-grained image classification, assessing this informativeness reliably is especially challenging due to subtle differences between classes. In this paper, we introduce a novel active learning method, combining **D**iscr**E**pancy-**C**onfusion unc**E**rtainty and calib**R**atio**N** diversity for active fine-grained image classification (**DECERN**), to effectively perceive the distinctiveness between fine-grained images and evaluate the sample value. DECERN introduces a multifaceted informativeness measure that combines discrepancy-confusion uncertainty and calibration diversity. The discrepancy-confusion uncertainty quantifies the structural stability and category directionality of fine-grained unlabeled data during local feature fusion. Subsequently, uncertainty-weighted clustering is performed to diversify the uncertainty samples. Then we calibrate the diversity to maximize the global diversity of the selected sample while maintaining its local representativeness. Extensive experiments conducted on 7 fine-grained image datasets across 39 distinct experimental settings demonstrate that our method achieves superior performance compared to state-of-the-art methods.

Keywords: Active learning · Fine-grained image classification

1 Introduction

The success of deep neural networks is highly dependent on large-scale annotated data. However, a large amount of data is unlabeled, and obtaining high-quality data annotation is a time-consuming task that requires expert knowledge [25, 34, 67]. For fine-grained images, as well as images in specialized fields, such as archaeology, medicine, and natural species, the budget for expert annotation is even more expensive. Active learning (AL) [2, 4, 27, 29, 42, 47, 52, 61, 62] methods have been proposed to achieve the construction of high-quality labeled datasets with limited budgets and maximize model training performance by iteratively selecting informative samples, rather than the entire data, to be annotated during the training process.

* Corresponding Author

Various AL methods have been proposed and can be broadly categorized into uncertainty-based, diversity-based, and hybrid methods. Uncertainty-based methods [19,28,39] select uncertain samples near decision boundaries for annotation as those that have the most information. Diversity-based methods [9, 46, 58] select samples with maximum diversity to cover the distribution of unlabeled data without redundancy. Hybrid methods [1, 3, 39, 41] exploit both uncertainty and diversity. These methods typically rely on reliable feature representations to estimate sample informativeness by analyzing the underlying data distribution [64]. However, in fine-grained image classification, where images represent subcategories under the same base category, visual similarity at shallow levels results in highly shared semantic features across deep network representations [32, 49, 55]. The overlap in feature distributions leads to unreliable uncertainty values estimates and reduces the effectiveness of diversity sampling, since samples from different subcategories may appear close in the feature space while being semantically distinct.

Recently, data augmentation methods [18,39,64] have been able to efficiently explore the neighborhood of unlabeled samples by constructing convex combinations of features through feature fusion, amplifying subtle differences between representations. This has led to approaches such as ALFA-Mix [39], which identify informative samples by evaluating the label variability of perturbed versions of unlabeled samples. However, the difficulty of changing labels varies across samples, and ALFA-Mix's approach of relying only on label variability to identify informative samples may overlook fine-grained samples that have the same pattern across multiple categories.

In this paper, we propose DECERN, a novel AL method for active fine-grained image classification. As shown in Fig. 1, we first strategically deploy local feature fusion mechanisms to disentangle the structural stability and the category directionality of unlabeled samples. Notably, low semantic-inconsistent perturbations trigger significant distributional shifts, exhibiting poor structural stability of features. Meanwhile, high semantic-consistent perturbations induce knowledge confusion in samples, eroding the category directionality of features. Then the discrepancy-confusion uncertainty takes into account both the poor structural stability and the low category directionality. Subsequently, we select candidates with high uncertainty scores from the unlabeled data. After performing uncertainty-weighted clustering on the candidates, we utilize calibration diversity to enrich the diversity of the selected uncertainty samples. On the one hand, we assume that samples proximal to uncertainty-weighted cluster centroids have larger local representativeness. These instances containing prototypical features or critical patterns of unlabeled data serve to consolidate the model's comprehension of pivotal regions within the real distribution. On the other hand, samples far from the labeled data centroids, called anchors, demonstrate more global diversity, exhibiting the greatest divergence from established knowledge, and enabling more effective exploration of regions such as decision boundaries and potential novel categories.

Our contributions are summarized as follows:

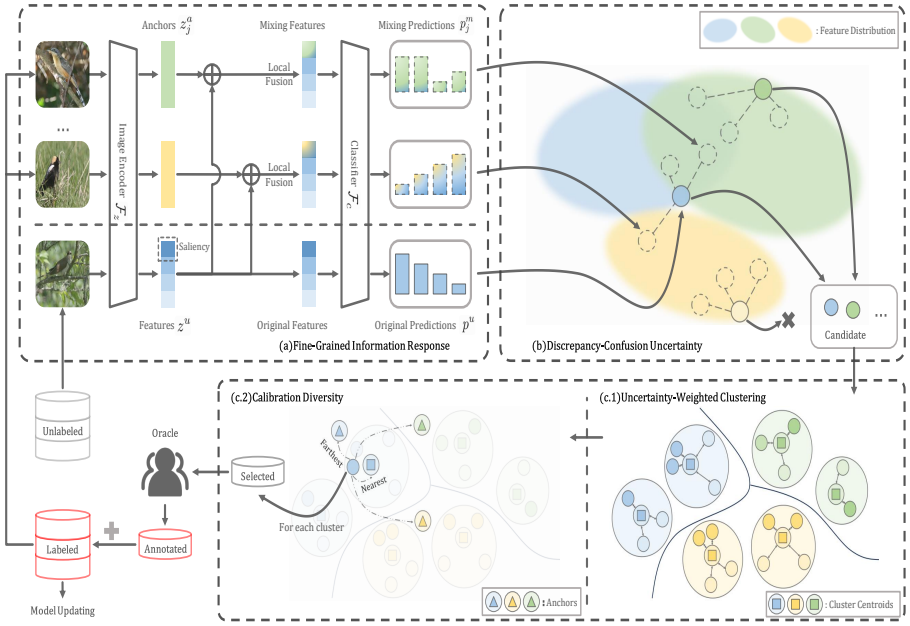


Fig. 1: The overview of our DECERN framework. Our method (a) comprehensively evaluates the variance of the probability distribution after the local feature fusion operations between the unlabeled data and the labeled data anchors. (b) Unlabeled data with poor structural stability or low category directionality features are identified as high uncertainty samples and selected to form the candidates. After (c.1) the uncertainty-weighted clustering, we further (c.2) refine the selected uncertainty data by utilizing the calibration diversity, which achieves the trade-off between the local representativeness and the global diversity.

- We propose a novel active learning method, DECERN, for fine-grained image classification task. Our method efficiently constructs high-quality labeled data of the fine-grained image under limited annotation budgets, leading to significant performance gains.
- We introduce a multifaceted informativeness measure in DECERN, combining discrepancy-confusion uncertainty and calibration diversity, to facilitate the effective selection of high informative samples. The former identifies features that exhibit poor structural stability and low category directionality during local feature fusion. The latter strategically balances the sample selection process, ensuring adequate local representativeness while maximizing global diversity.
- We implement our proposal active learning strategy on 7 common-used fine-grained image classification datasets across 39 distinct experimental settings. Extensive experimental results show that our DECERN achieves leading performance and outperforms existing state-of-the-art AL methods in the fine-grained image classification task.

2 Related Work

Diversity-based methods [17,23,48,58,59,63] select a subset of unlabeled data that is representative of the entire pool. Specifically, the representative samples should be both broadly distributed and non-redundant. A common strategy to achieve this diversity selection is to solve the coverage problem [5,6,9,46]. For example, CoreSet [46] frames active learning as a coreset selection problem, solved using the farthest-first traversal. Similarly, CoreGCN [9] leverages graph node embeddings to differentiate labeled and unlabeled nodes, applying these representations within the CoreSet [46] framework. Other approaches utilize auxiliary models. ActiveFT [58] directly optimizes a parametric model to select samples distributed similarly to the entire unlabeled data. Despite their effectiveness, diversity-based methods are limited [42] by the abundance of categories within the pool and the constraints on the batch size.

Uncertainty-based methods [11, 15, 35, 51, 54, 60, 68] select samples near the decision boundaries, which are most confusing to the task model. Various indicators are used to estimate the uncertainty values of unlabeled data, such as margin [7,43], confidence [19], entropy [21,45,53], energy [57], and influence [31]. In addition, some methods [26, 56, 60] use auxiliary modules to directly estimate the uncertainty values. One of the most representative works, LLAL [60] designs a loss prediction module to predict the target losses of unlabeled inputs. However, in fine-grained scenarios with high semantic similarity, overlapping feature distributions, and unreliable representations, uncertainty-based methods struggle to identify reliable decision boundaries [19, 64] and thus fail to guide the optimal selection of informative uncertainty samples.

Hybrid methods [1, 3, 39, 41] combine the advantages of uncertainty and diversity. For example, BADGE [3] performs batch active learning by clustering gradient embeddings of the final output layer, which implicitly encode predictive uncertainty, to select diverse and informative unlabeled samples. CLUE [41] proposes cluster uncertainty-weighted embeddings to select diverse and informative target instances. Although the hybrid criterion may reduce perceived informativeness, it significantly improves sample diversity through decreased inter-example similarity and achieves well performance.

Feature fusion on active learning has recently been considered in several studies [18,30,39,65]. For example, ALFA-Mix [39] constructs interpolations between labeled and unlabeled data representations and measures sample values based on the model's predicted inconsistency of pseudo-labels. ALMULA-mix [18] improves on the original ALFA-Mix [39] method, which is interpolated through weighted anchors to achieve a time-efficient identification of novel features. These methods well attempt the combination of feature fusion and active learning, exploring the neighborhood of unlabeled samples for efficient selection.

However, existing AL works predominantly focus on general image recognition or other scenarios, paying less attention to fine-grained images in specialized domains. Moreover, the integration of active learning with fine-grained image classification to achieve efficient annotation under limited budgets remains unexplored and constitutes the primary focus of this work.

3 Method

3.1 Overview

With a limited annotation budget, active learning aims to iteratively select the most informative samples from the unlabeled data pool for efficient annotation. Specifically, we have a large pool of unlabeled data $\mathcal{D}^u = \{(x_i^u)\}_{i=1}^{N_u}$ and a pool of labeled data $\mathcal{D}^\ell = \{(x_i^\ell, y_i^\ell)\}_{i=1}^{N_\ell}$ formed by random selection. We implement an active learning strategy that iteratively selects a subset of B samples from $\mathcal{D}^u$ to be labeled by the oracle. Then we update $\mathcal{D}^u$ and $\mathcal{D}^\ell$ with the selected annotated samples and use the updated labeled data pool to train the neural network models $\mathcal{F} = \mathcal{F}_z \circ \mathcal{F}_c$, where $\mathcal{F}_z$ and $\mathcal{F}_c$ denote the feature encoder and the classifier.

Fig. 1 illustrate our DECERN framework, and the pseudo-code is shown in Algorithm 1. We simultaneously consider two aspects: the discrepancy-confusion uncertainty and the calibration diversity of the samples, thus proposing two sampling strategies. Specifically, for the discrepancy-confusion uncertainty sampling strategy, we propose in Sec. 3.2 to achieve revealing the structural stability and category directionality of fine-grained features through local feature fusion. Afterward, we define the discrepancy uncertainty and confusion uncertainty to evaluate these two types of features in Sec. 3.3, and select samples with higher uncertainty to form candidates. For the calibration diversity sampling strategy, we performed the uncertainty-weighted clustering operation to enrich the diversity of the selected samples in Sec. 3.4. In addition to selecting samples that are closest to the cluster centroids, we also require that their distance from the labeled data centroids, called anchors, is the farthest, which balances the local representativeness and global diversity.

3.2 Fine-Grained Information Response

With the feature encoder $\mathcal{F}_z$ and the classifier $\mathcal{F}_c$, we obtain the feature representations z^u and the prediction probabilities p^u of the unlabeled data:

$$z^u = \mathcal{F}_z(x^u), \quad p^u = \mathcal{F}_c(z^u). \tag{1}$$

We also compute the class anchors z^a and p^a by averaging the feature representation and the prediction probability of the labeled data for the j-th category:

$$\begin{aligned} z_j^a &= \frac{\sum_{(x,y)\in\mathcal{D}^\ell} \mathbb{1}\{y=j\} \cdot \mathcal{F}_z(x)}{\sum_{(x,y)\in\mathcal{D}^\ell} \mathbb{1}\{y=j\}}, \\ p_j^a &= \frac{\sum_{(x,y)\in\mathcal{D}^\ell} \mathbb{1}\{y=j\} \cdot \mathcal{F}_c(\mathcal{F}_z(x))}{\sum_{(x,y)\in\mathcal{D}^\ell} \mathbb{1}\{y=j\}}, \end{aligned} \tag{2}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function.

Algorithm 1: Pseudo-code of DECERN

Input: Unlabeled data pool $\mathcal{D}^u$, labeled data pool $\mathcal{D}^\ell$, annotation budget B, number of categories N_c, feature encoder $\mathcal{F}_z$, classifier $\mathcal{F}_c$, local feature fusion strategy ϕ

Construct labeled data anchors z^a and p^a based on the feature representation and prediction probability of $\mathcal{D}^\ell$ by using $\mathcal{F}_z$ and $\mathcal{F}_c$;
for $x^u \in \mathcal{D}^u$ **do**
  $z^u = \mathcal{F}_z(x^u),\ p^u = \mathcal{F}_c(z^u)$;
  for $j = 1, 2, ..., N_c$ **do**
    Perform local feature fusion ϕ via Eq. (4), and calculate the prediction probability of mixing representation p^m via Eq. (7);
    Calculate the prediction probability p^b via Eq. (5), and p^w via Eq. (6);
    Calculate the category-level discrepancy-confusion uncertainty via Eq. (8), by using p^u, p^b, p^w and p^m;
  end
  Calculate the instance-level discrepancy-confusion uncertainty score $\mathcal{S}$ for unlabeled data via Eq. (9);
end
Select samples with high uncertainty scores $\mathcal{S}$ as candidates $\mathcal{C}$ by a dynamic threshold ζ via Eq. (10);
Perform uncertainty-weighted clustering on feature representations of the candidates $\mathcal{C}$ and obtain B clusters $\mathcal{C}_k$;
Obtain the final selected batch as $\mathcal{D}^s = \{(x^{s_k})\}_{k=1}^B$, where one instance x^{s_k} is selected from each cluster $\mathcal{C}_k$, and is closest to the cluster centroid $z^{\mathcal{C}_k}$ and farthest to z^a via Eq. (11);
$y^{s_k} = Oracle(x^{s_k})$;
$\mathcal{D}^u = \mathcal{D}^u \setminus \{(x^{s_k})\}_{k=1}^B$, $\mathcal{D}^\ell = \mathcal{D}^\ell \cup \{(x^{s_k}, y^{s_k})\}_{k=1}^B$;
Update the target model $\mathcal{F}_z$ and $\mathcal{F}_c$, and start the next AL cycle;

Then, we define the binary mask M via a pooled-gradient scheme. We first adaptively pool $\mathcal{G}$ into 100 bins, then select the top-η bins, and upsample the bin selection back to the original dimension. The mask is thus given by

$$M_{i,d} = \begin{cases} 1 & \text{if dimension } d \text{ of sample } i \text{ falls into one of the top-}\eta \text{ bins,} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $\mathcal{G} = \nabla_{z^u} \mathcal{L}(\mathcal{F}_c(z^u), \hat{y})$ denotes the backpropagation gradient of the unlabeled features, and $\hat{y}$ is the pseudo-label. This pooled-gradient mask suppresses noisy isolated dimensions and makes local fusion comparable across backbones. As a result, M reduces the impact of irrelevant feature representations.

Once the binary mask M is constructed, the strength of the feature fusion α requires specification. Specifically, z^u tends to be similar to z^a in the confidence prediction category and not similar in other categories, due to their reliable and discriminative representation. Consequently, leveraging insights from previous work [13, 39], we give large α for each unlabeled data in their confidence

prediction category and observe whether the mixing representation consistently maintains the category directionality under high semantic-consistent perturbations. We propose that the novel features supporting the category are corrupted during the feature fusion process, which are conducive to calibrating the category boundary. In contrast, for other categories, we give a small α, implement low semantic-inconsistent perturbations, to observe whether the fusion operation destroys the original semantic structure. Learning about features of unstable structures promotes robust feature representations and contributes to the consistent recognition of data from the same fine-grained category. Therefore, we set α_j to p_j^u.

Then, we perform the local feature fusion strategy ϕ on unlabeled data z^u and previously obtained category anchors z_j^a for the j-th category:

$$\phi(z^u, z_j^a; \alpha_j, M) = (1 - M)z^u + M\left(z^u(1 - \alpha_j) + z_j^a\alpha_j\right). \tag{4}$$

Finally, we assess the response of unlabeled samples under varying feature fusion operations using the following indicators: Prediction probability of the original representation p^u; Fusion prediction probability p^b that measures the theoretical prediction probability for feature fusion data; Weighted prediction probability p^w that measures the theoretical prediction probability for local feature fusion data; Prediction probability of mixing representations p^m that measures probability offsets for local feature fusion data; Specifically, p^u is defined in Eq. (1) and others are formulated as:

$$p_j^b = (1 - \alpha_j)p^u + \alpha_j p_j^a, \tag{5}$$

$$p_j^w = (1 - \eta\alpha_j)p^u + \eta\alpha_j p_j^a, \tag{6}$$

$$p_j^m = \mathcal{F}_c(\phi(z^u, z_j^a; \alpha_j, M)), \tag{7}$$

where η is the size of the mask M.

These predicted probabilities capture the relationship between unlabeled data and category-specific decision boundaries, enabling the calculation of reliable uncertainty scores for each unlabeled sample.

3.3 Discrepancy-Confusion Uncertainty

Our DECERN select samples that exhibit a higher discrepancy-confusion uncertainty. When the local feature fusion operation corrupts the semantics of the original representation, resulting in large probability offsets, we should give priority to selecting these samples due to their high discrepancy uncertainty. However, relying solely on discrepancy uncertainty prevents the selection of samples that are cross-category ambiguous. To address this problem, we also exploit the confusion uncertainty of the fusion operation. We utilize cross entropy $\mathcal{CE}$ as discrepancy uncertainty $\mathcal{S}_d$ to evaluate structural stability, and entropy $\mathcal{H}$ as confusion uncertainty $\mathcal{S}_c$ to evaluate category directionality. After performing the local feature fusion operation of unlabeled data with various anchors, the discrepancy-confusion uncertainty then identifies mixing representations of each operation by

low discriminative power due to poor structural stability from ambiguous semantics or lack of differential signals (low category directionality). Unlabeled data whose representations exhibit these characteristics are prioritized. Formally, we formulate the category-level discrepancy-confusion uncertainty score S_{dc} for the j-th category as follows:

$$\mathcal{S}_{dc}(p_j^*, p_j^m; \beta_j) = \mathcal{H}(p_j^m)^{1-\beta_j} + \mathcal{CE}(p_j^*, p_j^m)^{\beta_j}, \tag{8}$$

where $p_j^* \in \{p^u, p_j^b, p_j^w\}$, $\beta_j = 1 - \frac{1+\cos(z^u, z_j^a)}{2}$, and $\cos(\cdot,\cdot)$ is the cosine similarity.

Then, we calculate the instance-level discrepancy-confusion uncertainty score $\mathcal{S}$ for each unlabeled data:

$$\mathcal{S} = \frac{1}{N_c} \sum_{j=1}^{N_c} \left[(1-\eta)\mathcal{S}_{dc}(p^u, p_j^m; \beta_j) + \eta \mathcal{S}_{dc}(p_j^b, p_j^m; \beta_j)) + \mathcal{S}_{dc}(p_j^w, p_j^m; \beta_j) \right], \tag{9}$$

where N_c is the number of categories and η, denoted in Sec. 3.2, is the size of the mask M. This facilitates the precise identification of ambiguous samples situated near decision boundaries, thereby enhancing the model's ability to distinguish challenging samples.

Finally, we use reliable uncertainty scores $\mathcal{S}$ and a dynamic threshold ζ with the AL process for uncertainty-based sampling:

$$\zeta = \bar{\mathcal{S}} + \lambda\sigma\gamma,\ \sigma = \sqrt{\mathbb{E}\left[(\mathcal{S} - \bar{\mathcal{S}})^2\right]},\ \gamma = \frac{\mathbb{E}\left[(\mathcal{S} - \bar{\mathcal{S}})^3\right]}{\sigma^3}, \tag{10}$$

where $\bar{\mathcal{S}}$, σ and γ are the mean, standard deviation, and the skewness of $\mathcal{S}$, λ is the moderator of uncertainty sampling intensity.

We expect to select more high-uncertainty samples at the decision boundaries as model performance improves. Therefore, we set a dynamic threshold ζ, which is obtained by calculating the skewness of the scores and implies information about the improvement in model performance and the uncertainty scores of the samples. And with the threshold ζ, we select data with scores $\mathcal{S}$ above it as candidates $\mathcal{C}$, which contain more uncertainty information.

3.4 Uncertainty-Weighted Clustering and Calibration Diversity

The diversity of the samples is inappropriately neglected, especially when the decision boundaries are ambiguous and the uncertainty information is noisy. Thus, we perform diversity clustering in feature representations to sample from diversity regions in the feature space, which is inspired by several AL methods [39,41,44]. We assign weights to feature representation of each sample based on their uncertainty score $\mathcal{S}$. Intuitively, this particular weighting mechanism renders the cluster centroids more determinate by the position of the high-uncertainty samples, and enhances the potential of selection from them. By implementing uncertainty-weighted clustering, we maintain the balance between uncertainty and diversity, avoiding the oblivion of pure diversity sampling to data uncertainty.

We use the K-Means algorithm for uncertainty-weighted clustering and obtain B cluster $\mathcal{C}_k$. Then we select the samples that are closest to the cluster centroids $z^{\mathcal{C}_k}$ directly. However, boundary samples and potential subcategory samples may be ignored even though cluster centroids are biased towards high-uncertainty regions. To address this problem, we exploit the prior information from the anchors to calibrate the sample diversity. For each cluster $\mathcal{C}_k$, samples closest to $z^{\mathcal{C}_k}$ and farthest away from z^a are selected as the subset with the AL strategy, which can be formulated as:

$$x^{s_k} = \arg\max_{x_i \in \mathcal{C}_k} \left[-\xi(1 - \cos(z_i^u, z^{\mathcal{C}_k})) + (1 - \xi) \min_j (1 - \cos(z_i^u, z_j^a)) \right], \quad (11)$$

where ξ is a hyperparameter, $k = 1, 2, ..., B$, and $j = 1, 2, ..., N_c$.

This function integrates two terms designed for synergistic sample selection: local representativeness and global diversity. Crucially, local representativeness selects prototypical samples from high-uncertainty regions as typical representatives of intra-class core variations, thereby avoiding the selection of meaningless redundant samples within the class. The global diversity identifies samples exhibiting maximal divergence from established knowledge to supplement the feature distributions. This not only corrects for systematic gaps in the model's comprehension of the data distribution but also drives it to learn tail features and extend its decision boundaries.

Finally, we query the labels of these samples. After updating $\mathcal{D}^\ell$, we train the task model on the labeled data.

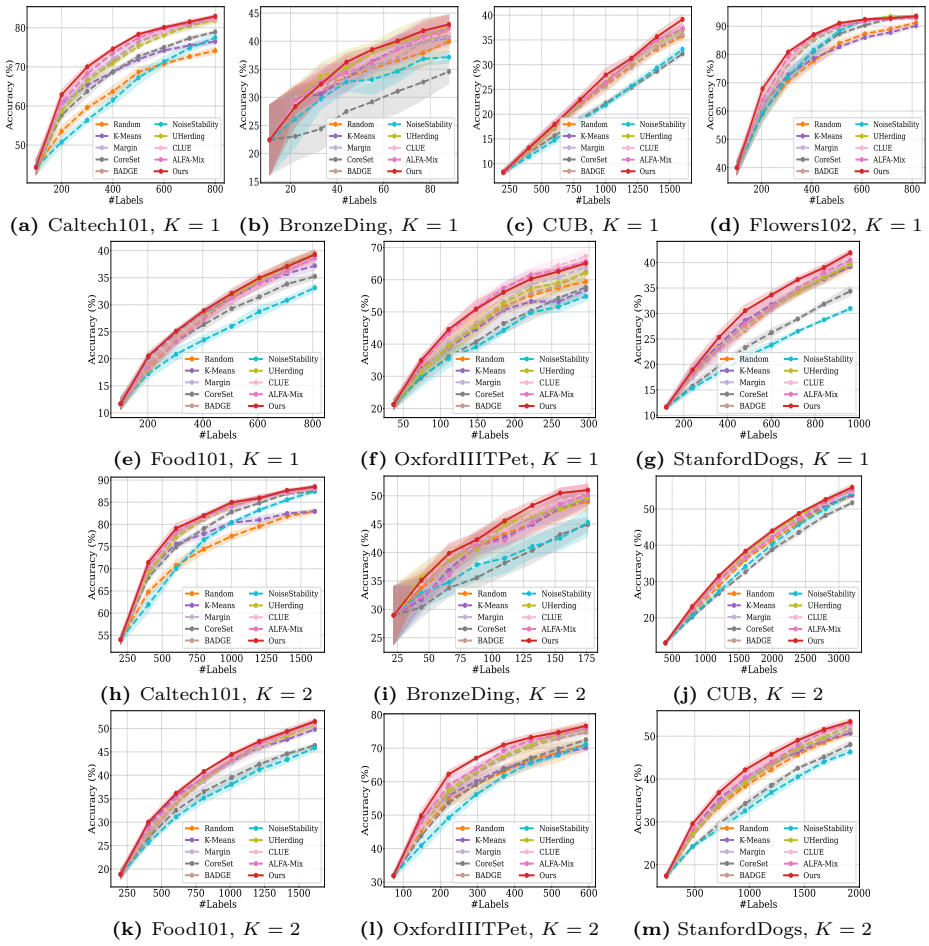
4 Experiments

4.1 Experiment Settings

Dataset and Metrics. We conduct experiments on 7 fine-grained image classification datasets: Caltech101 [14], BronzeDing [66], CUB [50], Flowers102 [36], Food101 [8], OxfordIIITPet [38], and StanfordDogs [22]. The raw training data serves as the unlabeled pool for selection. We employ the *Top-1 Accuracy* metric for performance evaluation. Additionally, we also report the experimental results using the macro F1-score metric in Appendix D.3 to address the possible imbalance of the testing data.

Baselines. We compare DECERN with **Random**, **K-Means**, **Margin** [43], **CoreSet** [46], **BADGE** [3], **NoiseStability** [28], **UHerding** [6], **CLUE** [41], **ALFA-Mix** [39]. Additionally, we further compare it with **Confidence**, **Entropy** [53], **CoreGCN** [9], **ActiveFT** [58], **BALQUE** [19], **DropQuery** [16] in Appendix D.4 to validate the effectiveness of our method. In summary, a total of 15 representative and recent AL baselines, covering uncertainty-, diversity-, hybrid-, and feature-fusion-based methods, including the closest ones, CLUE and ALFA-Mix, are compared with DECERN.

Implementation Details. We adopt ResNet50 [20] or ViT-Small [12] pretrained by DINO [10] and DINOv2 [37] as the backbone. Throughout training,



(a) Caltech101, $K = 1$ **(b)** BronzeDing, $K = 1$ **(c)** CUB, $K = 1$ **(d)** Flowers102, $K = 1$

(e) Food101, $K = 1$ **(f)** OxfordIIITPet, $K = 1$ **(g)** StanfordDogs, $K = 1$

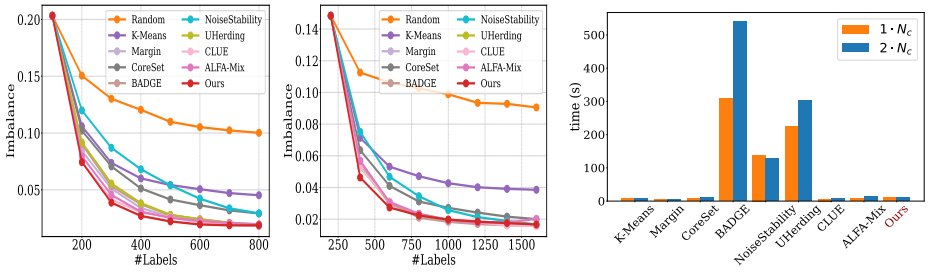
(h) Caltech101, $K = 2$ **(i)** BronzeDing, $K = 2$ **(j)** CUB, $K = 2$

(k) Food101, $K = 2$ **(l)** OxfordIIITPet, $K = 2$ **(m)** StanfordDogs, $K = 2$

Fig. 2: Comparison on benchmark datasets using ResNet50 model. We use annotation budget $B = K \cdot N_c$ in each AL cycle. N_c is the number of categories, $K \in \{1, 2\}$.

we employ the Adam [24] optimizer with a learning rate of 0.001 and cosine learning rate decay. A consistent batch size of 128 is maintained in all experiments. The number of AL cycles is set to 8. In each AL cycle, we apply the sampling strategy to the entire pool of unlabeled data $\mathcal{D}^u$, except for the Food101 dataset, where we apply the AL strategy to a random subset. In each AL cycle, we also select a subset of unlabeled data for labeling whose size is $B = K \cdot N_c$ (N_c is the number of categories, $K \in \{1, 2\}$), except for the Flowers102 dataset, where we select a subset of size $B = 1 \cdot N_c$. Since the initial labeled data are required for the AL methods, we employ random sampling in the first AL cycle.

For robustness comparison, we repeat each experiment with 5 different seeds and report the mean and standard deviations of the results. Meanwhile, we conduct extensive experiments involving multiple common fine-grained image



(a) Imbalance of sampling data **(b)** Data selection efficiency

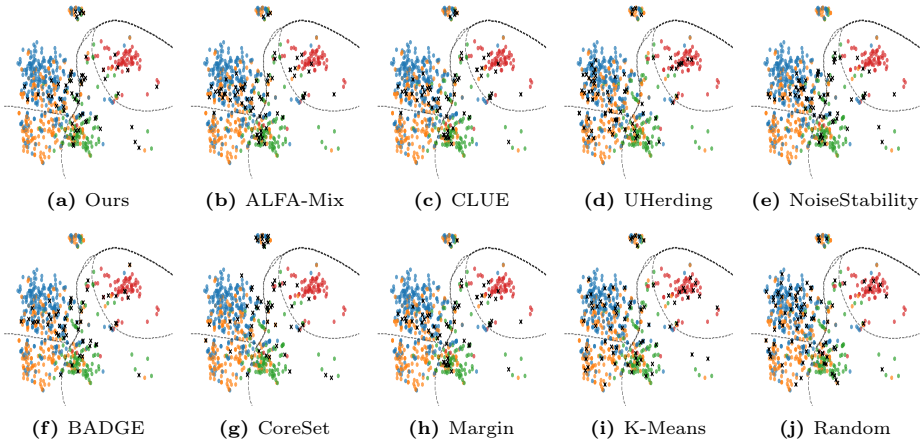
Fig. 3: (a) Imbalance of the sampling data by various AL methods through entropy. Each AL cycle we use the ResNet50 model to select $K \cdot N_c$ samples in the Caltech101 dataset. N_c is the number of categories, $K \in \{1, 2\}$. (b) Data selection efficiency of different methods. We compared the time required to select $B = K \cdot N_c$ samples from Caltech101 dataset using the ResNet50 model per cycle.

datasets, different architectures, and varying annotation budgets of active learning (39 distinct experimental settings in total, and more details are provided in Appendix D.1. All experiments are conducted on an NVIDIA A40 GPU with PyTorch [40]. We generally set the values of η, λ, ξ at 0.1, 0.5, and 0.8, respectively, as these values perform well in all experiments.

4.2 Overall Results

Generality of our proposal DECERN. In Fig. 2, we present the results over different AL cycles on benchmark datasets, using the ResNet50 model. Appendix D.2 further provides results with the ViT-Small model pretrained using DINO and DINOv2 on the same benchmark datasets. Additionally, Appendix D.4 includes results with additional baselines to further demonstrate the effectiveness of our method. In Fig. 2, we can observe that our method outperforms other baseline methods in various settings, demonstrating the effectiveness and superiority of our approach. Diversity methods (*e.g.*, K-Means and CoreSet) and uncertainty methods (*e.g.*, NoiseStability) are significantly hindered by fine-grained examples, as these examples lack discriminative features to distinguish between similar categories. In contrast, hybrid strategies achieve competitive performance, but are still not optimal.

Sampling imbalance affect the performance. As illustrated in Fig. 3a, we quantify the average imbalance of the sampling data by various AL strategies using the class distribution entropy (see Appendix B.1). Our approach achieves a well-balanced sampling data, thereby enhancing model performance. In contrast to some methods, *e.g.*, K-Means, CoreSet, and NoiseStability, which often induce performance degradation through imbalanced sample selection strategies. This observation reveals that pronounced data imbalances significantly inhibit the potential performance gains of uncertainty-based or diversity-based sampling methods, particularly when such imbalances exceed critical thresholds.



(a) Ours **(b)** ALFA-Mix **(c)** CLUE **(d)** UHerding **(e)** NoiseStability

(f) BADGE **(g)** CoreSet **(h)** Margin **(i)** K-Means **(j)** Random

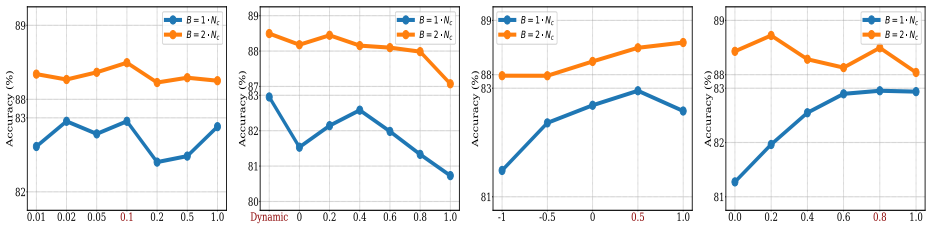
Fig. 4: t-SNE visualization of sample selection behavior by various AL methods on BronzeDing dataset with a ResNet50 model. The different colored dots represent 4 different categories of samples. Black forks indicate samples selected by various AL methods from the unlabeled pool, following an initial random labeling of 500 samples (with 100 additional samples sampled for annotation). The oracle decision boundary is shown as a black dashed line.

Data selection efficiency. We benchmark the time efficiency of selecting B samples per AL cycle on the Caltech101 dataset using the ResNet50 model, where the AL annotation budget $B = K \cdot N_c$ is defined by the number of categories N_c and $K \in \{1, 2\}$. Fig. 3b shows that our method achieves a near-optimal time efficiency compared to the fastest baselines, while delivering substantially superior performance. Furthermore, our method demonstrates a significantly lower data sampling time requirement than the BADGE, NoiseStability, and UHerding methods. See Appendix C for additional discussion.

Visualization of feature representations. To comparatively analyze sample selection behaviors, we visualize feature representations of samples selected by each AL strategy in Fig. 4. These feature representations, extracted from the penultimate layer of the ResNet50 model, are projected into the 2D space via t-SNE [33]. We focus on four categories of the BronzeDing dataset, with the oracle decision boundary (dashed black line) displayed to enable a clear comparative analysis. The results demonstrate that our proposed approach, while maintaining data diversity, effectively identifies and incorporates uncertain samples proximal to decision boundaries, thus enhancing the efficiency of active learning.

4.3 Ablation Study

We implement ablation studies to validate the effectiveness of our design choices. All ablation studies are conducted on Caltech101 datasets using the ResNet50 model with different AL annotation budget B. More details are provided in Appendix E.



(a) Effect of η **(b)** Effect of ζ **(c)** Effect of λ **(d)** Effect of ξ

Fig. 5: Ablations of the hyperparameters in our method.

Effect of different components. Conceptually, we quantify the uncertainty of the data using both the discrepancy uncertainty $\mathcal{S}_d$ and the confusion uncertainty $\mathcal{S}_c$. Ablation studies in Tab. 1 demonstrate that the removal of uncertainty sampling consistently induces performance degradation. Moreover, reliance on a single uncertainty measure leads to suboptimal results, indicating that unidimensional uncertainty quantification fails to capture the nuanced heterogeneity inherent in fine-grained image data. Crucially, the absence of discrepancy uncertainty or confusion uncertainty prevents the model from learning features with poor structural stability and low category directionality. These unlearned features accumulate over AL cycles, progressively amplifying knowledge gaps, resulting in performance loss.

Table 1: Ablation study of different components in our method.

Uncertainty	Diversity	Caltech101, ResNet50	
		$B = 1 \cdot N_c$	$B = 2 \cdot N_c$
✗	✗	74.14±0.93	82.93±0.47
✗	✓	80.73±0.69	87.08±0.22
w/o $\mathcal{S}_c$	✓	82.10±0.91	88.25±0.42
w/o $\mathcal{S}_d$	✓	82.43±0.44	88.16±0.42
w $\mathcal{S}_c$	✗	80.52±0.94	88.13±0.18
w $\mathcal{S}_d$	✗	80.41±1.00	88.02±0.44
✓	✗	81.53±0.92	88.18±0.43
✓	w/o Weighted	82.39±0.42	88.18±0.59
✓	w/o Clustering	81.27±0.36	88.43±0.34
✓	w/o Calibration	82.94±0.88	88.04±0.42
✓	✓	**82.95±0.37**	**88.49±0.37**

Our method selects diversity samples through calibration diversity after the uncertainty-weighted clustering. Therefore, we ablate our diversity sampling strategy by four modified variants in Tab. 1. In general, our approach demonstrates consistent superiority over comparative variants, indicating the importance of calibration diversity. We infer that over-reliance on local representativeness traps model in local optima, harming generalization, while prioritizing global diversity alone results in inefficient exploration and poor sample selection, hindering accuracy gains.

Hyperparameter influence. In Fig. 5, we report four hyperparameter ablation studies, including the local feature fusion size η in Eq. (3), the strength ζ and the strength factor λ of uncertainty sampling in Eq. (10), and the diversity balance factor ξ in Eq. (11).

As shown in Fig. 5a, the optimal performance is achieved with a moderate coefficient $\eta = 0.1$. Insufficient local feature fusion (*i.e.*, $\eta \leq 0.05$) fails to capture dominant patterns, preventing the AL strategy from effectively distinguishing

probability offsets in unlabeled data post-fusion. In contrast, a larger local fusion size (*i.e.*, $\eta \geq 0.2$) introduces interference from contextual information, which affects the evaluation of sample values and effective selection.

For comparison, we set ζ to fixed values and then select the percentage ζ of the unlabeled data as candidates for diversity sampling. As Fig. 5b indicates, the fixed uncertainty sampling ratio does not accommodate active learning. For a large fixed ζ, the candidates contain a large number of confidence samples, reducing the informativeness of the final selected sample and failing to refine the decision boundary. In contract, uncertainty prevails when ζ is small, which restricts the diversity of the selected data. This leads to fragile distributions that struggle to perform consistently well when settings change. The dynamic threshold ζ in our algorithm that varies with the AL cycles is more adaptable to the demands of sampling over constant periods and achieves a trade-off between uncertainty sampling and diversity sampling.

As Fig. 5c indicates, a larger λ improves the model performance, particularly when the annotation budget B increases. We attribute this to the fact that a larger B allows the model to develop a more robust understanding of the data distribution with a larger labeled dataset. In this scenario, strengthening the intensity of uncertainty-based sampling is more beneficial, as it directs the model to focus on learning ambiguous patterns, thus refining and expanding the decision boundary. In contrast, when model learning is still insufficient, an overemphasis on one side can be detrimental, and performance drops.

As Fig. 5d indicates, moderate diversity calibration (*i.e.*, $\xi = 0.8$) improves the quality of labeled data constructed with AL cycles and yield better performance. However, when ξ is relatively small (*i.e.*, $\xi \leq 0.4$) and the budget is small, the performance decreases as samples without local representativeness are selected.

5 Conclusions

We propose a novel active learning method, DECERN, to improve fine-grained image annotation efficiency under limited budgets by combining discrepancy-confusion uncertainty and calibration diversity. Specifically, DECERN perceives samples that exhibit poor structural stability and low category directionality during local feature fusion through discrepancy-confusion uncertainty, and evaluates calibration diversity, considering both local feature representation and global diversity. Through extensive evaluation on 7 fine-grained image datasets across 39 distinct experimental settings, our approach exhibits superior performance over state-of-the-art methods.

Although DECERN demonstrates efficient construction of high-quality labeled data under limited annotation budgets, yielding substantial performance gains in fine-grained image classification, its efficacy remains unproven when confronted by more complex tasks, *e.g.*, semantic segmentation, and object detection. In future work, we plan to optimize our DECERN and implement our DECERN across more application scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62576148).

References

1. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: European Conference on Computer Vision. pp. 137–153. Springer (2020)
2. Aggarwal, C.C., Kong, X., Gu, Q., Han, J., Yu, P.S.: Active learning: A survey. In: Data classification, pp. 599–634. Chapman and Hall/CRC (2014)
3. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671 (2019)
4. Atlas, L., Cohn, D., Ladner, R.: Training connectionist networks with queries and selective sampling. Advances in neural information processing systems **2** (1989)
5. Bae, W., Noh, J., Sutherland, D.J.: Generalized coverage for more robust low-budget active learning. In: European Conference on Computer Vision. pp. 318–334. Springer (2024)
6. Bae, W., Oliveira, G.L., Sutherland, D.J.: Uncertainty herding: One active learning method for all label budgets. arXiv preprint arXiv:2412.20644 (2024)
7. Balcan, M.F., Broder, A., Zhang, T.: Margin based active learning. In: International Conference on Computational Learning Theory. pp. 35–50. Springer (2007)
8. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
9. Caramalau, R., Bhattarai, B., Kim, T.K.: Sequential graph convolutional network for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9583–9592 (2021)
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
11. Chen, W., Wang, C., Li, S., Xu, K., Bai, Y., Chen, W., Li, S.: Debiased active learning with variational gradient rectifier. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 15884–15894 (2025)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Faramarzi, M., Amini, M., Badrinaaraayanan, A., Verma, V., Chandar, S.: Patchup: A feature-space block-level regularization technique for convolutional neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 589–597 (2022)
14. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)
15. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International conference on machine learning. pp. 1183–1192. PMLR (2017)

16. Gupte, S.R., Aklilu, J., Nirschl, J.J., Yeung-Levy, S.: Revisiting active learning in the era of vision foundation models. arXiv preprint arXiv:2401.14555 (2024)
17. Hacohen, G., Dekel, A., Weinshall, D.: Active learning on a budget: Opposite strategies suit high and low budgets. arXiv preprint arXiv:2202.02794 (2022)
18. Han, X., Wang, Q., Wang, Y., Wang, J., Deng, C., Feng, J.: Feature mixing-based active learning for multi-label text classification. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10551–10555. IEEE (2024)
19. Han, Y., Liu, D., Shang, J., Zheng, L., Zhong, J., Cao, W., Sun, H., Xie, W.: Balque: Batch active learning by querying unstable examples with calibrated confidence. Pattern Recognition **151**, 110385 (2024)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 ieee conference on computer vision and pattern recognition. pp. 2372–2379. IEEE (2009)
22. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proc. CVPR workshop on fine-grained visual categorization (FGVC). vol. 2 (2011)
23. Kim, K., Park, D., Kim, K.I., Chun, S.Y.: Task-aware variational adversarial active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8166–8175 (2021)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Konyushkova, K., Sznitman, R., Fua, P.: Learning active learning from data. Advances in neural information processing systems **30** (2017)
26. Kye, S.M., Choi, K., Byun, H., Chang, B.: Tidal: Learning training dynamics for active learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22335–22345 (2023)
27. Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W., Okumura, M.: A survey on deep active learning: Recent advances and new frontiers. IEEE Transactions on Neural Networks and Learning Systems **36**(4), 5879–5899 (2024)
28. Li, X., Yang, P., Gu, Y., Zhan, X., Wang, T., Xu, M., Xu, C.: Deep active learning with noise stability. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 13655–13663 (2024)
29. Liu, P., Wang, L., Ranjan, R., He, G., Zhao, L.: A survey on active deep learning: From model driven to data driven. ACM Computing Surveys (CSUR) **54**(10s), 1–34 (2022)
30. Liu, Z., Zhang, J., He, X.: A discrimination-guided active learning method based on marginal representations for industrial compound fault diagnosis. IEEE Transactions on Automation Science and Engineering **21**(4), 6411–6422 (2023)
31. Liu, Z., Ding, H., Zhong, H., Li, W., Dai, J., He, C.: Influence selection for active learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9274–9283 (2021)
32. Ma, L., Luo, X., Shi, Y., Meng, F., Wu, Q., Hong, H.: Optimal transport quantization based on cross-x semantic hypergraph learning for fine-grained image retrieval. IEEE Transactions on Circuits and Systems for Video Technology (2025)
33. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)

34. Matsuo, S., Togashi, R., Bise, R., Uchida, S., Nomura, M.: Instance-wise supervision-level optimization in active learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4939–4947 (2025)
35. Mukhoti, J., Kirsch, A., Van Amersfoort, J., Torr, P.H., Gal, Y.: Deep deterministic uncertainty: A new simple baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24384–24394 (2023)
36. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
38. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
39. Parvaneh, A., Abbasnejad, E., Teney, D., Haffari, G.R., Van Den Hengel, A., Shi, J.Q.: Active learning by feature mixing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12237–12246 (2022)
40. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
41. Prabhu, V., Chandrasekaran, A., Saenko, K., Hoffman, J.: Active domain adaptation via clustering uncertainty-weighted embeddings. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8505–8514 (2021)
42. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
43. Roth, D., Small, K.: Margin-based active learning for structured output spaces. In: European conference on machine learning. pp. 413–424. Springer (2006)
44. Safaei, B., Patel, V.M.: Active learning for vision-language models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4902–4912. IEEE (2025)
45. Safaei, B., Vibashan, V., De Melo, C.M., Patel, V.M.: Entropic open-set active learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4686–4694 (2024)
46. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
47. Settles, B.: Active learning literature survey (2009)
48. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5972–5981 (2019)
49. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the european conference on computer vision (ECCV). pp. 805–821 (2018)
50. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
51. Wan, F., Ye, Q., Yuan, T., Xu, S., Liu, J., Ji, X., Huang, Q.: Multiple instance differentiation learning for active object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12133–12147 (2023)
52. Wan, Z., Wang, Z., Chung, C., Wang, Z.: A survey of dataset refinement for problems in computer vision datasets. ACM computing surveys **56**(7), 1–34 (2024)

53. Wang, D., Shang, Y.: A new active labeling method for deep learning. In: 2014 International joint conference on neural networks (IJCNN). pp. 112–119. IEEE (2014)
54. Woo, J.O.: Active learning in bayesian neural networks with balanced entropy learning principle. arXiv preprint arXiv:2105.14559 (2021)
55. Wu, J., Chang, D., Sain, A., Li, X., Ma, Z., Cao, J., Guo, J., Song, Y.Z.: Bi-directional feature reconstruction network for fine-grained few-shot image classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2821–2829 (2023)
56. Wu, X., Chen, C., Zhong, M., Wang, J., Shi, J.: Covid-al: The diagnosis of covid-19 with deep active learning. Medical Image Analysis **68**, 101913 (2021)
57. Xie, B., Yuan, L., Li, S., Liu, C.H., Cheng, X., Wang, G.: Active learning for domain adaptation: An energy-based approach. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 8708–8716 (2022)
58. Xie, Y., Lu, H., Yan, J., Yang, X., Tomizuka, M., Zhan, W.: Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23715–23724 (2023)
59. Yehuda, O., Dekel, A., Hacohen, G., Weinshall, D.: Active learning through a covering lens. Advances in Neural Information Processing Systems **35**, 22354–22367 (2022)
60. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 93–102 (2019)
61. Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., Hu, X.: Data-centric artificial intelligence: A survey. ACM Computing Surveys **57**(5), 1–42 (2025)
62. Zhan, X., Wang, Q., Huang, K.h., Xiong, H., Dou, D., Chan, A.B.: A comparative survey of deep active learning. arXiv preprint arXiv:2203.13450 (2022)
63. Zhang, B., Li, L., Yang, S., Wang, S., Zha, Z.J., Huang, Q.: State-relabeling adversarial active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8756–8765 (2020)
64. Zhang, B., Li, L., Zha, Z.J., Luo, J., Huang, Q.: Downstream-pretext domain knowledge traceback for active learning. IEEE Transactions on Multimedia **26**, 10585–10596 (2024)
65. Zhang, L., Lam, S.K., Luo, D., Wu, X.: Employing feature mixture for active learning of object detection. Neurocomputing **594**, 127883 (2024)
66. Zhou, R., Wei, J., Zhang, Q., Qi, R., Yang, X., Li, C.: Multi-granularity archaeological dating of chinese bronze dings based on a knowledge-guided relation graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3103–3113 (2023)
67. Zong, C.C., Huang, S.J.: Rethinking epistemic and aleatoric uncertainty for active open-set annotation: An energy-based approach. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10153–10162 (2025)
68. Zong, C.C., Wang, Y.W., Ning, K.P., Ye, H.B., Huang, S.J.: Bidirectional uncertainty-based active learning for open-set annotation. In: European Conference on Computer Vision. pp. 127–143. Springer (2024)