Decision Boundary-aware Generation for Long-tailed Learning

Jiacheng Yang[1] Ruichi Zhang[1] Chikai Shang[1] Mengke Li[2]
Xinyi Shang[3] Junlong Gao[1] Yonggang Zhang[4] Yang Lu[1*]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen, China

[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

[3]Department of Statistical Science, University College London, London, UK

[4]Division of Arts and Machine Creativity, Hong Kong University of Science and Technology, Hong Kong, China

23020250157840@stu.xmu.edu.cn zhangruichi@stu.xmu.edu.cn ckshangl2@gmail.com
mengkeli@szu.edu.cn xinyi.shang.23@ucl.ac.uk jlgao@xmu.edu.cn zhangyg@ust.hk
luyang@xmu.edu.cn

Abstract

Long-tailed data bias decision boundaries toward head classes and degrade tail class accuracy. Diffusion-based generative augmentation address this problem by generating additional data, while head-to-tail transfer further mitigate the generator bias inherit from long-tailed dataset. However, we show that while head-to-tail transfer helps balance the decision space of the classifier, it also induces latent non-local feature mixing that entangles inter-class features, causing decision boundary overlap and tail class distribution shift. To address this, we first identify the problem of boundary ambiguity and then propose Decision Boundary-aware Generation (DBG) framework, which promotes near-boundary representation learning by generating informative near-boundary samples. Overall, DBG rebalances the long-tailed dataset while yielding more separable decision space for long-tailed learning. Across standard long-tailed benchmarks, DBG consistently improves tail class and overall accuracy with less inter-class overlap. The code of DBG is available at https://github.com/keepdigitalabc-svg/DBG.

1. Introduction

Vision-based deep learning methods have delivered substantial gains in convenience and efficiency [6, 9, 18, 23, 24, 31], yet their performance critically depends on the quality and quantity of the training data [13, 19]. In practice, datasets often exhibit long-tailed class distributions, characterized by a few head classes dominating the samples while many tail classes are underrepresented. Models trained on such imbalanced data tend to learn biased decision boundaries that favor head classes and degrade tail class accuracy [34]. Though many methods have been proposed for long-tailed learning [3, 14], the core challenge remains the scarcity of tail samples which shifts decision boundaries.

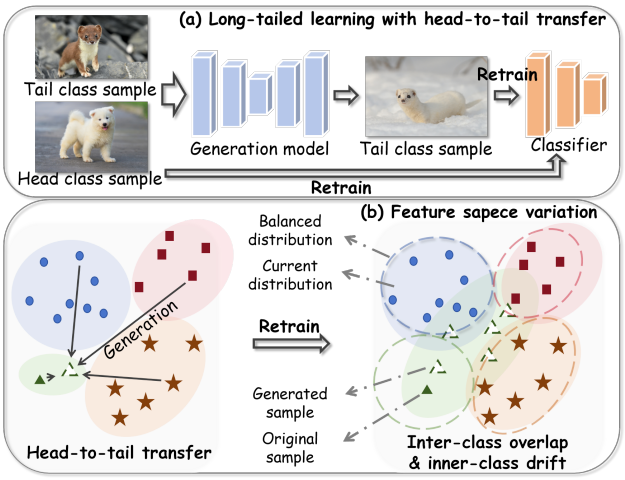

Figure 1. Pipeline of head–to–tail transfer for tail sample generation, resulting in inter-class overlap, intra-class drift, and blurred decision boundaries.

From a different perspective, augmentation-based approaches [22, 27] address data scarcity by supplementing training sets with additional samples. Among them, generative modeling offers robust performance by exploiting data priors. Recent works employ powerful diffusion models [11] along with head-to-tail transfer to mitigate generator bias and promote a more uniform decision space [29, 30]. These methods effectively leverages limited training data to improve the quality of generated samples and mitigate the generator's bias toward head classes as shown in Fig. 1 (a). However, while prior methods pur-

*Corresponding Author: Yang Lu (luyang@xmu.edu.cn).

sue uniform decision space, they rarely analyze the boundary blurring and tail drift induced by head–to–tail transfer [20, 21]. As shown in Fig. 1 (b), head-to-tail transfer helps balance the decision space but inevitably induces latent, uncontrolled nonlocal feature leakage that entangles class distributions, leading to inter-class overlap and shifts in tail class decision boundaries. Without a balanced reference, the leakage is difficult to disentangle, leaving a "more uniform yet highly overlapped" decision space and limiting gains in tail-class learning. Specially, we term this problem boundary ambiguity.

To make the boundary ambiguity measurable under generation-based long-tailed learning, we design three tailored metrics: (i) inter-class overlap degree which quantifies pairwise overlap among class distributions; (ii) tail outlier rate for estimating the class decision boundary shift; and (iii) generation confidence for intuitively showing the feature entanglement. When applied across generated datasets, these metrics consistently reveal blurred decision boundaries and tail distribution drift under head–to–tail transfer. These findings caution that **indiscriminate use of head class features to expand the tail class can lead to decision boundary damage**, expansion based on head-to-tail transfer is therefore a double-edged sword.

Motivated by these observations, we introduce the **G**enerative **B**oundary-aware **G**eneration (**DBG**) framework to address the boundary ambiguity for promoting tail class learning. DBG integrates a boundary-aware generator with a classifier-driven bifurcated cleaning pipeline to generate informative near-boundary samples, thereby repopulating and refining the decision boundaries. Specifically, the boundary-aware generator introduces adversarial declassification noise to suppress dominant class-specific patterns and reconstruct each sample toward its k-nearest alternative classes, enriching boundary features for long-tailed learning. Furthermore, to mitigate hard samples arising from long-tail-biased labeling and generator–classifier space mismatch, the classifier-driven bifurcated cleaning pipeline integrates class-wise prototype-distance and confidence–credibility filtering to identify and discard harmful samples. Overall, DBG effectively addresses the scarcity of tail class data and sharpens the decision boundaries, reconstructing a "more uniform and separable" decision space that substantially improves tail class precision.

Our main contributions are summarized as follows:

- To our knowledge, we present the first systematic analysis of boundary ambiguity induced by head-to-tail transfer, using three metrics: inter-class overlap degree, outlier rate, and generation confidence.
- We propose a Decision Boundary-aware Generation framework to generate informative samples for enriching near-boundary features and sharpening decision boundaries in long-tailed learning.
- Extensive experiments demonstrate that the proposed framework reduces inter-class overlap and improves long-tailed learning performance for all generative baselines, demonstrating its effectiveness and generalization ability.

2. Related Works

2.1. Augmentation-based Long-tailed Learning

Data quantity and quality critically determine the performance of deep learning classifiers, while real-world datasets are often long-tailed [2]. Prior work introduces additional samples to rebalance long-tailed datasets for improving overall classify accuracy [1, 17, 33, 35]. However, such strategies can introduce out-of-distribution features and even backdoor risks, degrading both accuracy and security of the classifiers. In response, generator-based approaches emerged. For instance, Liu et al. [21] generates tail samples by mixing head-shared and tail-specific features. Qin et al. [27] leverages diffusion model [28] trained with the long-tailed data to generate missing samples. Furthermore, Shao et al. [29] combines diffusion with head-to-tail transfer strategy to mitigate head class bias. In this paper, we introduce generation based on adversarial attack to reduce inter-class overlap and restore blurred decision boundaries for long-tailed learning.

2.2. Adversarial Samples

Adversarial samples are inputs perturbed with small, human-imperceptible noise that flip the prediction of classifier, while preserving visual semantics [7, 10, 32]. Beyond robustness evaluation, they also serve as data augmentation to improve generalization [10, 36]. In long-tailed learning they have been used to mitigate imbalance data distribution by reconstructing tail samples via adversarial attacks toward head samples [20]. In this work, we generate near-boundary adversarial samples to offer additional boundary knowledge for restoring decision regions and sharpen separability.

3. Analysis of Boundary Ambiguity

The uniformity and separability of the feature space largely determine the difficulty of classification. Owing to severe class imbalance, long-tailed datasets often yield post-training feature spaces with uneven inter-class density. Existing head-to-tail training strategies mainly aim to equalize inter-class distributions while paying less attention to improving inter-class separability. To further observe this, in this section, we introduce three complementary metrics to jointly quantify the uniformity and separability of the learned feature space, including inter-class overlap degree, outlier sample rate, and generation confidence.

3.1. Boundary Ambiguity Measurement Metrics

Inter-class overlap degree. Our goal is to quantify how an generated training set reshapes the separability of the learned feature space. Moreover, since the classification results are directly determined by the test set, we restrict our analysis to the changes in the feature space on the test data. To this end, the inter-class overlap degree metric is defined based on von Mises–Fisher (vMF) hyperspherical projection and the Bhattacharyya coefficient (BC) to measure the overlap degree between each pair of classes.

Specifically, test features are first extracted by a well-trained encoder and then projected onto the unit hypersphere via $L2$ normalization by $y = \frac{z}{\|z\|_2}$, which enables uniform comparability of inter-class overlap. $z \in \mathbb{S}^{d-1}$ denotes the extracted feature and y is the normalized feature after projection. After that, for each class $c \in \{1, \ldots, K\}$ with normalized test features $\{y_i^{(c)}\}_{i=1}^{n_c} \subset \mathbb{S}^{d-1}$, we fit a vMF density for representing intra-class distribution:

$$p_c(y) \;=\; C_d(\kappa_c)\,\exp\!\big(\kappa_c\,\mu_c^{\top} y\big), \tag{1}$$

where $\mu_c \in \mathbb{S}^{d-1}$ is the mean direction and $\kappa_c \geq 0$ is the concentration, both estimated by maximum likelihood from $\{y_i^{(c)}\}$. $C_d(\cdot)$ is the vMF normalizer. Then, pairwise inter-class overlap is measured by BC as follows.

$$\widehat{\mathrm{BC}}(c, c') \;=\; \log\!\left(\frac{1}{m}\sum_{i=1}^{m}\sqrt{p_c(y)\,p_{c'}(y)}\right). \tag{2}$$

Lower $\widehat{\mathrm{BC}}(c, c')$ indicates a more separable feature space between c and c'.

To investigate the impact of imbalance ratio and head-to-tail transfer on the decision space of the classifier, inter-class overlap degree is measured under four varint datasets. As shown in Fig. 2 (a), the imbalanced setting yields higher inter-class overlap, indicating poorer separability. CBDM-based reduces inter-class overlap by increasing feature diversity via diffusion-based generation. However, introducing head-to-tail transfer increases overlap again, proving our hypothesis that coarse-grained head-to-tail transfer can harm feature-space separability.

Outlier sample rate. Due to the instability of parameter changes during model training, it is very difficult to directly evaluate whether the generated dataset exhibits feature space shift. But ideally, the feature space learned by the model should be able to uniformly project the test set samples onto their corresponding class distributions. Based on this, we measure the degree of intra class shift by testing the approximate distribution of sample features in the feature space. Specifically, to verify whether the head-to-tail mechanism induces unfavorable intra-class distribution shift, we

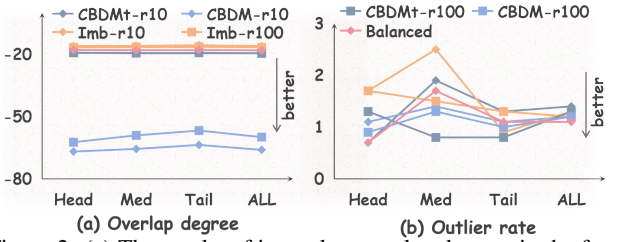

Figure 2. (a) The results of inter-class overlap degrees in the feature spaces learned by four different classifiers. (b) The results of Outliers sample rate. Imb-r represents the ResNet-32 trained on the imbalanced dataset with imbalanced ratio set to r, and CBDMt denotes the classifier trained on the CBDM-based [27] dataset with head-to-tail transfer.

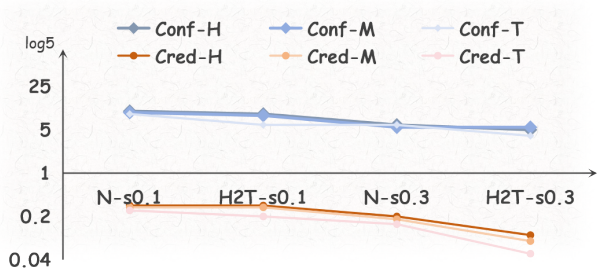

Figure 3. The results of generation confidence with different s on balanced classifier. N represents normal training, while H2T represents trained with head-to-tail transfer. H, M, and T represent statistics based on head, medium, and tail classes, respectively.

flag test samples that deviate excessively from their class-wise neighborhoods in the learned feature space. Given a trained encoder, the within-class nearest-neighbor distance is computed for each sample:

$$d_i^{(c)} \;=\; \min_{\substack{j \neq i \\ 1 \leq j \leq n_c}} \big\| z_i^{(c)} - z_j^{(c)} \big\|_2, \tag{3}$$

where $z_i^{(c)}$ is the i-th feature of class c and n_c is the number of test samples in class c. $d_i^{(c)}$ is the class-c nearest-neighbor distance of sample i. Let $\bar{d}_c = \frac{1}{n_c}\sum_{i=1}^{n_c} d_i^{(c)}$ and $s_c = \mathrm{std}\big(\{d_i^{(c)}\}_{i=1}^{n_c}\big)$, we standardize nearest-neighbor distances and mark outliers by a threshold λ as follows.

$$s_i^{(c)} \;=\; \frac{\big|d_i^{(c)} - \bar{d}_c\big|}{s_c}, \qquad o_i^{(c)} \;=\; \mathbb{I}\big[s_i^{(c)} > \lambda\big], \tag{4}$$

where $\bar{d}_c$ is the mean of $\{d_i^{(c)}\}_{i=1}^{n_c}$. $\mathrm{std}(\cdot)$ denotes the standard deviation so that s_c is the standard deviation of the nearest-neighbor distances in class c. $\mathbb{I}[\cdot]$ is the judgement function (1 if the predicate is true, 0 otherwise). $o_i^{(c)} \in \{0, 1\}$ is the outlier label. Finally, the outlier sample rate η_c can be defined as $\eta_c = \frac{1}{n_c}\sum_{i=1}^{n_c} o_i^{(c)}$ to measure the distribution shift.

Similarly, we evaluate outlier sample rate on classifiers trained on four variant datasets. As shown in Fig. 2 (b), increasing imbalance ratio of the dataset and introducing

head-to-tail transfer both lead to more test-time outliers in the feature space, especially for tail classes. This indicates head-to-tail transfer leads to tail class distribution shift, degrading tail class accuracy. It is also worth-noted that higher imbalanced rate can increase the model bias toward head classes, resulting in less outlier samples.

Generation confidence. To intuitively quantify boundary ambiguity induced by head–to–tail transfer on diffusion models, generation confidence is defined as the class-wise average prediction confidence of generated data. Specifically, a diffusion probabilistic model consists of a forward noising process and a reverse denoising process. The generation process usually begins with a random Gaussian noise sample and, by predicting and removing the noise iteratively follows standard Classifier-free guidance (CFG) [15, 16]:

$$\hat{\boldsymbol{\varepsilon}}_t = (1-s)\cdot\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t,\varnothing,t) + s\cdot\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t,\mathbf{c},t), \tag{5}$$

where $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t,\mathbf{c},t)$ and $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t,\varnothing,t)$ denote the conditional and unconditional noise predictions, respectively. s is the guidance scale. To directly validate how head-to-tail confusion affects the feature space learned by diffusion models, we form a modified CFG guidance for a target class $y_t \in \mathcal{C}_{\text{tar}}$ and a disturbing class $y_d \in \mathcal{C}_{\text{dis}}$ as follows.

$$\hat{\boldsymbol{\varepsilon}}_{tm} = (1-s)\cdot\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t,y_t,t) + s\cdot\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t,y_d,t). \tag{6}$$

The generated samples under modified guidance are then feed into a ResNet classifier f_ϕ trained on the balanced dataset to obtain logits $\ell(x_g) \in \mathbb{R}^K$ and probabilities $p(x_g) = \mathrm{softmax}(\ell(x_g))$. The confidence and credibility of sample x_g is then computed to measure the generation confidence as follows.

$$\begin{aligned} Conf(x_g) &= \ell_{y_t}(x_g), \\ Cred(x_g) &= p_{(1)}(x_g) - p_{(d)}(x_g), \end{aligned} \tag{7}$$

where $\ell_{y_t}(x_g)$ denotes the y_t-th logit of $\ell(x_g) \in \mathbb{R}^K$. $p_{(1)}(x_g) = \max_c p_c(x_g)$ and $p_{(d)}(x_g)$ is the predicted probability of x_g from disturbing class. A lower confidence and a lower credibility indicates stronger boundary ambiguity.

As shown in Fig. 3, after employing head-to-tail transfer, the prediction confidence of the generated samples reduce, indicating that this strategy makes head-specific features leak into tail class generated samples, leading to boundary ambiguity. It is worth noted that the prediction credibility also reduce, indicating that the generator also occurs boundary ambiguity for disturbance classes.

4. Decision Boundary-aware Generation

The DBG comprises two components: (i) a boundary-aware sample generator that generates samples near the decision boundary to restore sharper decision boundaries; and (ii) a classifier-driven bifurcated data-cleaning pipeline that filters hard and drifted samples, preventing secondary distortion of the feature space caused by long-tailed bias.

4.1. Boundary-aware Sample Generator

The boundary-aware sample generator consists of two stages, including conditional noising to erase class-specific features for pulling the samples to its decision boundary, and conditional denoising for assigning target labels to the reconstructed samples, as shown in Fig. 4.

For conditional noising, with diffusion trained with horizon T, schedule $\beta_t \in (0,1)$, and $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, the input sample x_0 is first perturbed by random noise $\hat{\epsilon}_r$ to push it into a middle timestep for easier modifying deep class-specific features [25], as follows.

$$x_m = \sqrt{\bar{\alpha}_m}\,x_0 + \sqrt{1-\bar{\alpha}_m}\,\hat{\epsilon}_r, \tag{8}$$

where $m = T/2$ and $\hat{\epsilon}_r$ denotes standard Gaussian noise.

After that, to suppress the class-specific features, x_m is forwarded under the guidance of source label y_{sl} following Eq. 8 iteratively and the target class-specific noise $\hat{\boldsymbol{\varepsilon}}_c$ in it is obtained via the noise predictor $\boldsymbol{\varepsilon}_\theta(\cdot)$ as follows.

$$\hat{\boldsymbol{\varepsilon}}_c = \boldsymbol{\varepsilon}_\theta\big(x_t,\, y_{\text{sl}},\, t\big), \tag{9}$$

where x_t denotes the current noisy sample. Specifically, the noising iteration update for $K = T/10$ steps starting at m. Which prevents the reconstructed sample from drifting too far from the source manifold while effectively erasing class-specific features to produce stable near-boundary variants.

For conditional denoising, we perform reverse denoising with standard CFG (following Eq. 5) as follows.

$$\begin{aligned} \hat{x}_0 &= \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\boldsymbol{\varepsilon}}_t(x_t,t,\tilde{y})}{\sqrt{\bar{\alpha}_t}}, \\ x_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\,\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\,\hat{\boldsymbol{\varepsilon}}_t(x_t,t,\tilde{y}), \end{aligned} \tag{10}$$

where $\tilde{y}$ denotes the label close to x_0. By iteratively removing predicted noise, the generated samples remain near the decision boundary and avoid large semantic jumps while providing additional knowledge of $\tilde{y}$ for the classifier.

Specifically, $\tilde{y}$ is from k classes most confusable with the ground-truth label of x_0, as predicted by a classifier f trained on the source dataset. By default, we take the highest-scoring member as follows.

$$\begin{aligned} \tilde{y} &= \arg\max_{c\in\mathcal{C}_k(x_0)} f(x_0), \\ \mathcal{C}_k(x) &= \mathrm{TopK}\big(f(x_0),\,k\big), \end{aligned} \tag{11}$$

where $f(x_0)$ denotes the predicted logit for class c, $\mathcal{C}_k(x_0)$ is the index set of the top-k predicted classes for x_0 and $k = \lfloor C/w \rfloor$. C is the total class number and w is the

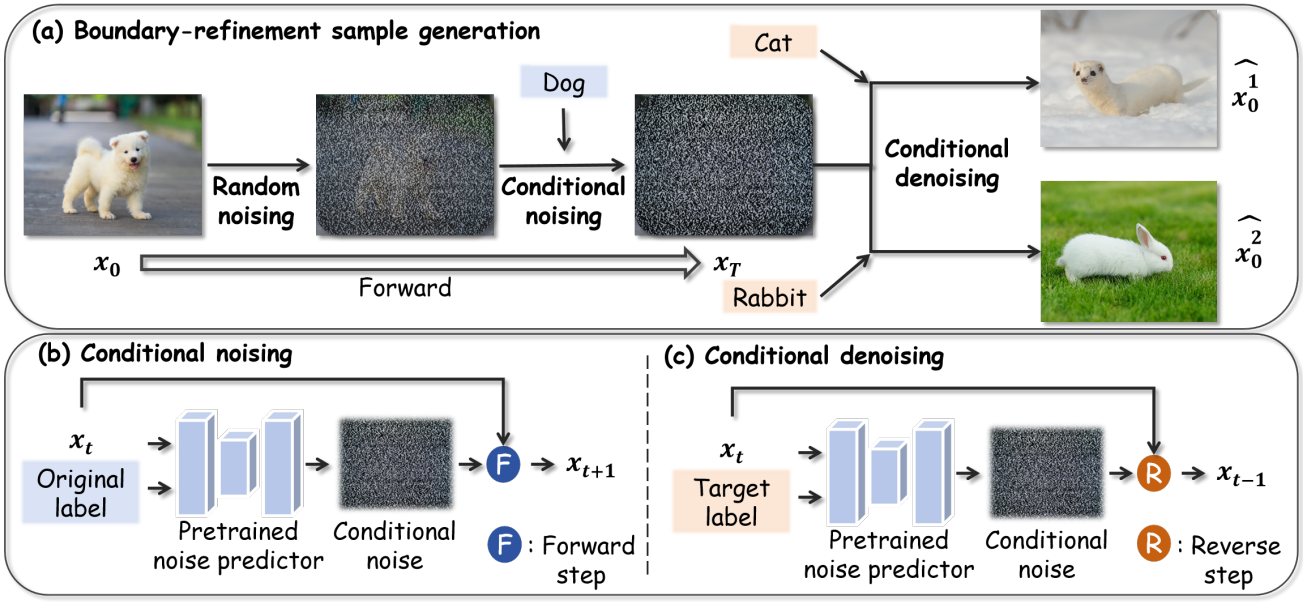

Figure 4. The basic pipeline of boundary-aware sample generator, including (b) conditional noising and (c) conditional denoising stages.

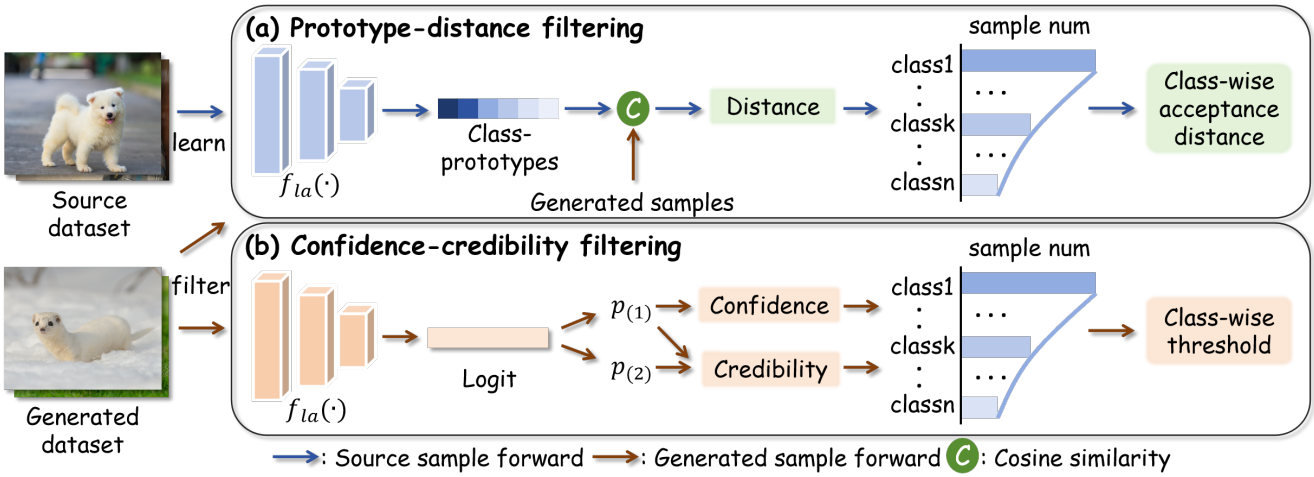

Figure 5. The workflow of classifier-driven bifurcated data-cleaning, consists of prototype-distance and confidence-credibility filtering.

stable weight. By setting an appropriate w, the generator can supply abundant near-boundary targets for rebalancing extremely imbalanced data while avoiding large semantic shifts that would distort the feature distribution.

4.2. Classifier-Driven Bifurcated Data-cleaning

Due to generator bias that inherit from long-tailed dataset and the feature space mismatch between generator and classifier, the generation process may produce misleading adversarial samples with bad class-specific features. Such hard samples cannot provide precise boundary supervision for the classifier and may even further damage the decision boundaries. To address this, we propose a long-tailed-insensitive classifier-driven bifurcated data-cleaning pipeline, as shown in Fig. 5. The pipeline comprises two parallel data-cleaning branches: one employs prototype-distance filtering to discard outlier samples and boundary-distant samples, while the other applies confidence-credibility filtering to remove samples misaligned with the feature space of the classifier.

For prototype-distance filtering, a well-trained classifier $f_\theta(\cdot)$ is first obtained by training on the source long-tailed dataset $\mathcal{D}$ with logit adjustment loss [26]. Then, for each class $c \in \{1, \ldots, C\}$, we compute its prototype as follows.

$$\mu_k = \frac{1}{N_k} \sum_{(x_i,\, y_i=c)\in\mathcal{D}} \tilde{z}(x_i), \quad \mu_k = \frac{\mu_k}{\|\mu_k\|_2}, \tag{12}$$

where $\tilde{z}(x) = \frac{\phi(x)}{\|\phi(x)\|_2}$, $\phi(\cdot)$ is feature extractor of f_θ and $\tilde{z}(x)$ denotes the normalized feature. After that we can measure the cosine distance $d_c(\cdot)$ of each generated sample $\hat{x_0}$

Table 1. Top-1 accuracy (%) of ResNet-32 on CIFAR100-LT across baselines. CBDM* represents CBDM-based. Improvements on tail classes and the overall average are highlighted in red. AVG computes the overall accuracy and improvement of each baseline.

Method	Imbalanced ratio-10				Imbalanced ratio-100				Imbalanced ratio-200				Avg
	Head	Med	Tail	All	Head	Med	Tail	All	Head	Med	Tail	All	
CBDM*	68.17	61.80	55.40	61.79	66.30	52.85	25.93	48.81	69.10	46.52	15.60	44.02	51.54
CBDM*+DBG	67.77	63.67	**57.20**	62.96	70.08	53.35	**26.72**	50.37	69.17	49.15	**18.67**	46.01	↑1.57
CBDM	66.87	64.28	59.17	63.52	69.10	53.75	28.60	50.81	64.10	48.98	20.67	45.02	53.12
CBDM+DBG	68.37	63.62	**59.27**	63.74	67.30	54.98	**30.10**	51.21	64.40	49.27	20.47	45.17	↑0.25
OCLT	68.60	62.92	56.80	62.79	68.70	52.38	25.73	49.28	69.03	48.90	16.23	45.14	52.40
OCLT+DBG	67.93	63.58	**57.10**	62.94	68.20	53.95	**27.37**	50.25	69.33	49.33	**16.33**	45.43	↑0.47
DiffuLT	70.27	63.30	54.30	62.69	68.10	51.20	29.37	49.72	68.67	49.02	19.13	45.95	52.79
DiffuLT+DBG	68.33	63.00	**56.53**	62.66	68.87	54.55	28.63	51.07	69.30	49.52	**19.43**	46.43	↑0.60

Table 2. Top-1 accuracy (%) of ResNet-32 on CIFAR10-LT across baselines. CBDM* represents CBDM-based.

Method	Imbalanced ratio-10				Imbalanced ratio-100				Imbalanced ratio-200				AVG
	Head	Med	Tail	All	Head	Med	Tail	All	Head	Med	Tail	All	
CBDM*	94.60	85.92	88.17	89.20	91.90	76.20	76.53	81.01	92.10	72.10	65.30	76.06	82.09
CBDM*+DBG	94.20	86.42	88.13	89.27	92.13	77.62	74.97	81.18	91.30	74.17	**69.00**	77.76	↑0.65
CBDM	93.77	86.22	90.87	89.88	92.73	77.08	76.30	81.54	91.90	73.38	67.27	77.10	82.84
CBDM+DBG	93.00	86.17	90.33	89.47	93.60	77.45	**76.53**	82.02	90.73	71.95	**68.77**	76.63	↓0.13
OCLT	93.93	86.47	90.10	89.80	92.50	77.65	77.43	82.04	92.10	74.50	68.70	78.04	83.29
OCLT+DBG	93.23	86.10	**90.57**	89.58	91.60	78.90	**78.90**	82.71	91.70	75.45	**70.87**	78.95	↑0.46
DiffuLT	93.07	85.53	89.70	89.04	92.47	77.15	76.33	81.50	93.93	72.40	62.23	77.56	82.7
DiffuLT+DBG	93.07	86.47	89.43	89.34	92.20	76.85	**76.73**	81.42	92.23	73.97	**69.93**	78.24	↑0.30

to their target class prototype as follows.

$$d_c(\hat{x_0}) = 1 - \langle \tilde{z}(\hat{x_0}), \mu_k \rangle, \tag{13}$$

where $\langle \tilde{z}(\hat{x_0}), \mu_k \rangle$ represents cosine similarity between vector $\tilde{z}(\hat{x_0})$ and μ_k. Then, to better fit the long-tailed distributions, we obtain an highest distance d_c^h and lowest distance d_c^l and define an class-wise acceptance range as $[(1-l)d_c^l, (1+h)d_c^h]$ for each class. Both h and l ratio parameters that are set to small values to remove extreme outlier samples and boundary-distant samples.

For confidence-credibility filtering, the confidence and credibility of the classifier for $\hat{x_0}$ are computed following Eq. 7 with disturbing class set as the second highest predicted class using $f_\theta(\cdot)$. Then, to mitigate the influence of inherited long-tailed bias in the classifier on data filtering, the class-wise threshold a is defined according to the total samples of each class. A sample is removed only when it is misclassified into non-original and non-target labels with high confidence and credibility. By using long-tail-insensitive classifier-driven bifurcated data-cleaning pipeline to filter samples, generated samples that harmful to the classifier's decision boundaries can be effectively removed. After data-cleaning, the accepted adversarial set S_{adv} can work as a plug-and-play auxiliary training set to refine the feature space boundary of the classifier as $\mathcal{D}_{\text{aug}} = \mathcal{D} \cup \{ (x, \tilde{y}) : x \in S_{adv} \}$. By providing additional boundary information for long-tailed datasets, we alleviate the boundary ambiguity induced by class-imbalance bias and prevent further blurring caused by feature leakage in head-to-tail transfer strategies.

5. Experiments

5.1. Experimental Setup

Datasets. We evaluate DBG along with other baselines on two long-tailed datasets, CIFAR10-LT and CIFAR100-LT [5], constructed following Cao et al. [4]. The imbalance ratio is set to 10, 100, and 200 to assess the proposed DBG under varying levels of class imbalance.

Baselines. Comparisons are made to diffusion-based generative methods for long-tailed learning, including CBDM-based [27], CBDM (with an additional loss), OCLT [33], and DiffuLT [29]. Experiments for CBDM-based, CBDM, and OCLT use the authors' released implementations. DiffuLT is re-implemented from the paper due to the absence of official code.

Implementation. We set $w = 3$, $l = 0.02$, and $h = 0.05$ for CIFAR datasets. a is linearly decayed from 0.9 to 0.5. Specially, to better estimate the drift of the feature space, we sweep λ over five values uniformly spaced in $[2.5, 3.0]$ and report the mean. For all methods, we use ResNet-32 [12] as the classifier backbone and train for 200 epochs with a batch size of 128. The initial learning rate is 0.1 and is

Table 3. Top-1 accuracy (%) of ViT-B/16 on CIFAR100-LT across baselines. CBDM* represents CBDM-based.

Method	Imbalanced ratio-10				Imbalanced ratio-100				Imbalanced ratio-200				Avg
	Head	Med	Tail	All	Head	Med	Tail	All	Head	Med	Tail	All	
CBDM*	49.73	42.45	32.73	41.72	45.17	27.77	9.83	27.61	43.80	23.93	5.73	24.43	31.25
CBDM*+DBG	49.60	43.75	**36.50**	43.33	47.97	30.02	**11.07**	29.72	46.10	26.88	**6.37**	26.49	↑1.93
CBDM	50.57	47.38	40.07	46.14	48.17	34.20	16.30	33.02	45.97	29.95	10.20	28.83	36.00
CBDM+DBG	51.60	47.48	39.37	46.28	49.77	34.85	15.63	33.56	46.20	29.55	9.83	28.63	↑0.16
OCLT	50.93	45.67	36.17	44.40	47.73	31.18	11.70	30.30	45.77	25.23	6.03	25.63	33.44
OCLT+DBG	50.23	45.75	**37.63**	44.66	49.13	32.05	**12.37**	31.27	47.97	27.27	**6.70**	27.31	↑0.97
DiffuLT	50.93	44.20	34.03	43.17	49.43	33.30	14.93	32.63	47.57	27.30	5.70	26.90	34.23
DiffuLT+DBG	50.90	45.52	**38.57**	45.05	49.23	34.58	**15.43**	33.23	50.17	30.73	**10.27**	30.42	↑2.00

Table 4. Top-1 accuracy (%) of ViT-B/16 on CIFAR10-LT across baselines. CBDM* represents CBDM-based.

Method	Imbalanced ratio-10				Imbalanced ratio-100				Imbalanced ratio-200				AVG
	Head	Med	Tail	All	Head	Med	Tail	All	Head	Med	Tail	All	
CBDM*	77.50	68.38	78.17	74.05	65.07	54.05	66.63	61.13	67.77	47.55	52.93	55.23	63.47
CBDM*+DBG	72.67	70.50	**80.83**	74.25	60.57	56.88	**72.17**	63.32	66.13	51.38	**57.60**	57.67	↑1.61
CBDM	66.97	72.75	83.83	74.34	64.80	61.83	72.60	65.95	57.73	52.08	68.47	58.69	66.33
CBDM+DBG	66.40	73.53	**84.73**	74.75	65.63	61.05	**74.43**	66.44	59.77	49.42	**70.47**	58.84	↑0.35
OCLT	68.40	71.83	84.30	74.54	61.80	59.40	57.40	62.14	58.60	52.10	60.00	56.42	64.37
OCLT+DBG	70.03	70.40	**85.27**	74.75	60.40	57.62	**73.17**	63.12	61.97	52.85	**64.43**	59.06	↑1.27
DiffuLT	74.07	68.20	77.83	72.85	61.80	54.92	70.30	61.60	73.67	55.08	59.17	61.88	65.44
DiffuLT+DBG	71.13	70.70	**83.50**	74.67	59.87	57.73	**73.00**	62.95	68.70	59.95	**62.13**	63.23	↑1.51

Table 5. Ablation study on DBG using ResNet-32. CBDM* represents CBDM-based (without our framework) and the first line is standard learning without any augmentation.

Gen.	PD.	CC.	Head	Med	Tail	All
			64.1	35.6	8.6	36.1
CBDM*			66.3	52.8	25.9	48.8
✓			66.4	50.5	23.2	47.0
✓	✓		69.2	52.7	**27.6**	50.1
✓		✓	68.4	48.9	23.5	47.1
✓	✓	✓	**70.1**	**53.4**	26.7	**50.4**

multiplied by 0.1 at epochs 160 and 180. To better validate the universality of DBG, we also report results with ViT-B/16 [8], trained for 100 epochs with a batch size of 32, a base learning rate of 1×10^{-4}, and a weight decay of 0.01.

5.2. Classification Results

To test whether DBG-generated samples reduce decision-boundary ambiguity and improve long-tailed recognition, auxiliary-generation retraining is conducted on CIFAR-LT with two classifier backbones (ResNet-32 and ViT-B/16). In each baseline, the data generated by DBG is used as a plug-and-play auxiliary training set in addition to the generated data of baseline. For ResNet-32, per-group accuracies on head, medium, tail splits and the overall score are reported in Tab. 1 and 2. Across methods, adding DBG consistently increases overall accuracy, with larger gains on tail classes. These results indicate that near-boundary samples from DBG consistently enhance class separability, mitigate head-class bias, and thus improve long-tailed learning.

Table 6. Hyperparameter sensitivity study on h and l.

h	0.00	0.05	0.05	0.05	0.05	0.07	0.10
l	0.02	0.00	0.02	0.05	0.10	0.02	0.02
Head	69.4	**70.1**	**70.1**	68.8	69.3	68.5	68.8
Med	52.7	52.5	**53.4**	52.1	52.2	52.2	52.0
Tail	25.7	27.6	26.7	27.5	**27.7**	27.5	26.2
Overall	49.6	50.3	**50.4**	49.7	50.0	49.6	49.3

Similar experiments are conducted for ViT-B/16 as shown in Tab. 3 and 4), where all baselines also improve after injecting data generated by DBG, indicating that the complementary boundary information supplied by DBG generalizes across architectures. It is noted that tail class accuracy drops slightly on CIFAR100-LT with CBDM. This is because CBDM uses a strong tail-bias loss that makes tail classes easier to classify. Even so, DBG improves the overall accuracy, indicating a better decision space. Moreover, consistent gains across baselines imply that boundary ambiguity is pervasive across data augmentation methods, with head-to-tail transfer further amplifying the problem.

5.3. Ablation Study

On CIFAR100-LT with imbalance ratio set to 100, the effect of removing the prototype-distance and confidence-

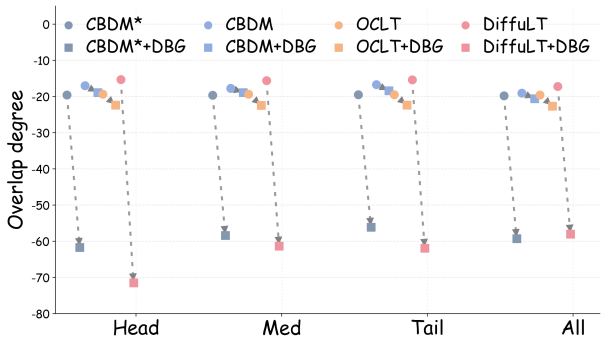

Figure 6. Inter-class overlap changes across baselines with DBG on ResNet-32 with imbalanced rate set to 100 for CIFAR100-LT.

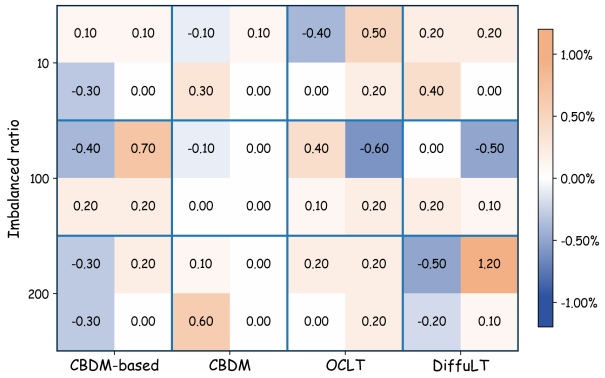

Figure 7. Outlier-rate changes across baselines with DBG. In each 2×2 tile, quadrants correspond to head (top-left), medium (top-right), tail (bottom-left), and overall (bottom-right).

credibility filters is evaluated via ablations in which the classifier is retrained from scratch. As shown in Tab. 5, dropping either filter causes a substantial decrease in accuracy, underscoring the importance of these filtering mechanisms. Because the diffusion model in DBG is trained following the CBDM-based setting, results are also reported against CBDM-based. Moreover, because accuracy drops sharply when the prototype-distance filter is removed, we study the sensitivity of h and l. As shown in Tab. 6, small value changes of h and l do not materially affect DBG's performance in long-tailed learning, and we already set them to their optimal values.

5.4. Quantitative Analysis

To directly evaluate the ability of our method in reconstructing clear decision boundaries and balanced feature distributions, we augment the training set with DBG-generated samples and measure decision boundary quality using the proposed overlap degree and the outlier rate. As shown in Fig. 6, incorporating DBG-generated data reduces overall inter-class overlap across baselines, indicating that DBG helps form a more separable decision space. In most cases, the overlap degree of tail class is also reduced, suggesting that the method mitigates long-tailed bias by providing additional information to underrepresented classes. Similar effects are observed when the imbalance ratio is set to 10 and 200 shown in the appendix. As shown in Fig. 7, DBG reduces the overall outlier rate across baselines, suggesting that an improved decision space helps balance the the classifier bias. In a few cases, outlier rates for some classes increase slightly, likely because DBG focuses on boundary refinement, causing minor perturbations to specific classes.

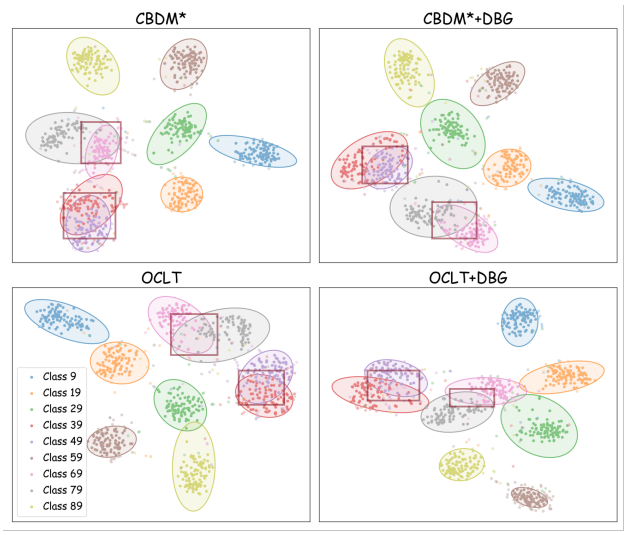

Figure 8. Changes in Feature space under different baselines.

Moreover, to more intuitively demonstrate the sharpening effect of DBG on the classification boundaries, t-SNE visualizations based on different baselines are plotted, as shown in Fig. 8. It can be observed that after injecting DBG-generated data, the inter-class overlap in the decision space of different baselines decreases, while the inter-class distances increase and the intra-class features become more compact, which verifies the effectiveness of GBT in addressing the boundary ambiguity problem.

6. Conclusion

Long-tailed recognition is hindered not only by data scarcity but also by boundary ambiguity introduced by data augmentation and head-to-tail transfer in generative augmentation. We made boundary ambiguity measurable with three metrics: inter-class overlap degree, outlier rate, and generation confidence. Building on these findings, we proposed Decision Boundary-aware Generation (DBG), which generates near-boundary samples and removes harmful ones through a classifier-driven bifurcated cleaning pipeline to offer additional decision boundary information for long-tailed learning. Across long-tailed benchmarks, DBG reduces inter-class overlap and outliers, and improves tail and overall accuracy, yielding a more uniform and more separable decision space. However, long-tailed bias in the diffusion still reduced the effectiveness of DGB-generated samples. Going forward, we will study adaptive cleaning and fine-tuning based method adaptation to further restore long-tailed decision boundaries.

Acknowledgment

This study was supported in part by the National Natural Science Foundation of China under Grants 62376233, 62502402 and 62306181; in part by the Natural Science Foundation of Fujian Province under Grant 2024J09001; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010163; in part by the Shenzhen Science and Technology Program under Grant RCBS20231211090659101; in part by the National Key Laboratory of Radar Signal Processing under Grant JKW202403; in part by Xiaomi Young Talents Program and YGZ was funded by Inno HK Generative AI R&D Center.

References

[1] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. Cuda: Curriculum of data augmentation for long-tailed recognition. *arXiv:2302.05499*, 2023. 2

[2] Chris Anderson. The long tail: why the future of business is selling less of more. In *Hachette Books*, 2006. 2

[3] Guangyu Ao, Min Chen, Jing Li, Huibing Nie, Lei Zhang, and Zejun Chen. Comparative analysis of large language models on rare disease identification. *Orphanet J. Rare Dis.*, 20(1):150, 2025. 1

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 6

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019. 6

[6] Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt switch: Efficient clip adaptation for text-video retrieval. In *ICCV*, pages 15648–15658, 2023. 1

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193, 2018. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7

[9] Lunke Fei, Bob Zhang, Chunwei Tian, Shaohua Teng, and Jie Wen. Jointly learning multi-instance hand-based biometric descriptor. *Inf. Sci.*, 562:1–12, 2021. 1

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014. 2

[11] Pengxiao Han, Changkun Ye, Jieming Zhou, Jing Zhang, Jie Hong, and Xuesong Li. Latent-based diffusion model for long-tailed recognition. In *CVPR*, pages 2639–2648, 2024. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[13] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv:1712.00409*, 2017. 1

[14] Chih-Hui Ho, Kuan-Chuan Peng, and Nuno Vasconcelos. Long-tailed anomaly detection with learnable class names. In *CVPR*, pages 12435–12446, 2024. 1

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 4

[17] Feng Hong, Jiangchao Yao, Yifei Shen, Dongsheng Li, Ya Zhang, and Yanfeng Wang. Improving diffusion models for class-imbalanced training data via capacity manipulation. In *The twelfth international conference on learning representations*, 2026. 2

[18] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, pages 5830–5840, 2021. 1

[19] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *ICML*, pages 15696–15707. PMLR, 2023. 1

[20] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, pages 13896–13905, 2020. 2

[21] Weide Liu, Zhonghua Wu, Yiming Wang, Henghui Ding, Fayao Liu, Jie Lin, and Guosheng Lin. LCReg: Long-tailed image classification with latent categories based recognition. *PR*, 145. 2

[22] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton Van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *CVPR*, pages 6959–6969, 2022. 1

[23] Jiaan Luo, Feng Hong, Qiang Hu, Xiaofeng Cao, Feng Liu, and Jiangchao Yao. Long-tailed recognition with model rebalancing. *Advances in Neural Information Processing Systems*, 2025. 1

[24] Pietro Melzi, Ruben Tolosana, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Ivan DeAndres-Tame, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, et al. Frcsyn-ongoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems. *Inf. Fusion*, 107:102322, 2024. 1

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 4

[26] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2020. 5

[27] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *CVPR*, 2023. 1, 2, 3, 6

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[29] Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. In *NeurIPS*, pages 123007–123031, 2024. 1, 2, 6

[30] Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? In *NeurIPS*, pages 75669–75687, 2023. 1

[31] Zhe Xu, Da Chen, Kun Wei, Cheng Deng, and Hui Xue. Hisa: Hierarchically semantic associating for video temporal grounding. *IEEE TIP*, 31:5178–5188, 2022. 1

[32] Jiacheng Yang, Wai Keung Wong, Lunke Fei, Shuping Zhao, Jie Wen, and Shaohua Teng. Decoupling visual and identity features for adversarial palm-vein image attack. *Neural Netw.*, 180:106693, 2024. 2

[33] Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *ICLR*, 2024. 2, 6

[34] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE TPAMI*, 45(9):10795–10816, 2023. 1

[35] Shizhen Zhao, Xin Wen, Jiahui Liu, Chuofan Ma, Chunfeng Yuan, and Xiaojuan Qi. Learning from neighbors: Category extrapolation for long-tail learning. In *CVPR*, pages 30483–30492, 2025. 2

[36] Weimin Zhao, Sanaa Alwidian, and Qusay H Mahmoud. Adversarial training methods for deep learning: A systematic review. *Algorithms*, 15(8):283, 2022. 2