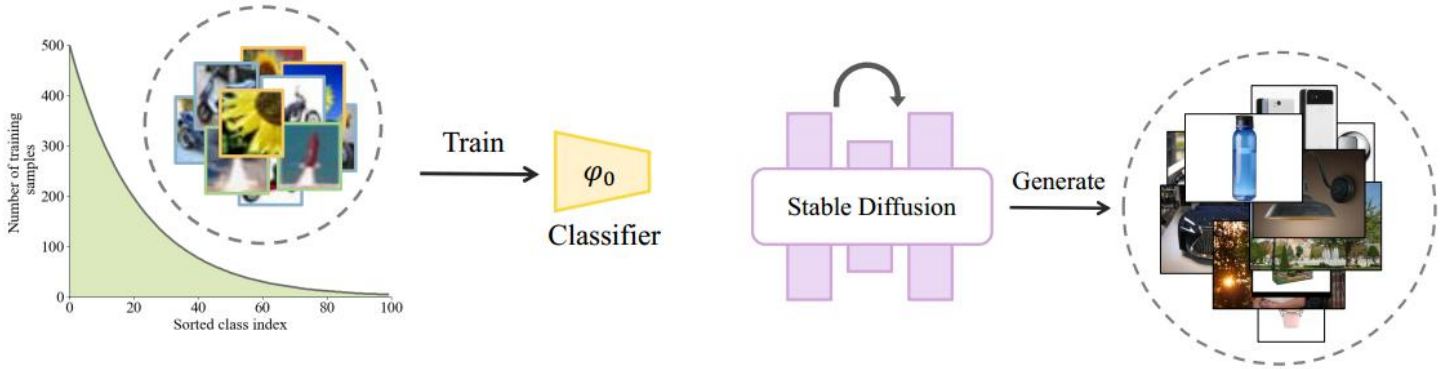

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

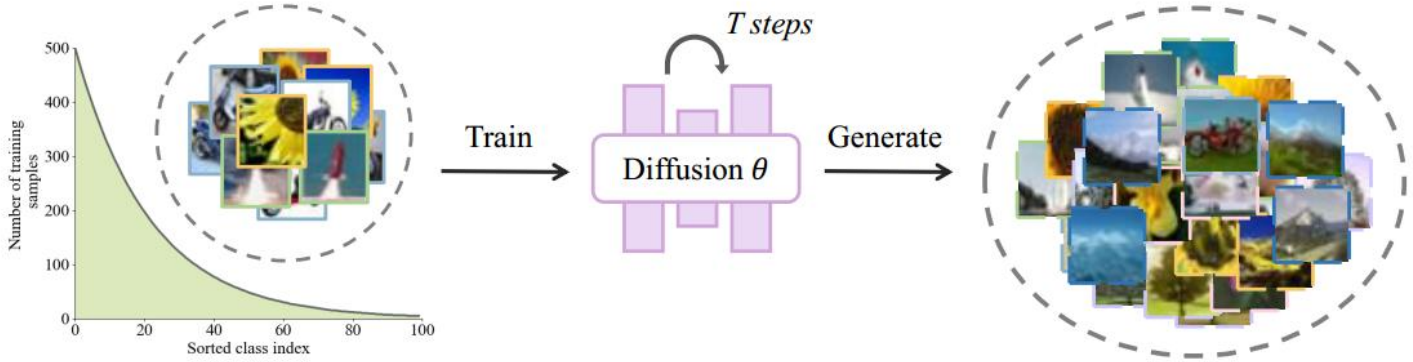
# DiffuLT: Diffusion for Long-tail Recognition Without External Knowledge

**Jie Shao** **Ke Zhu** **Hanxiao Zhang** **Jianxin Wu***
National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{shaoj, zhuk, zhanghx}@lamda.nju.edu.cn, wujx2001@nju.edu.cn

NeurIPS2024




(a) Previous long-tail recognition methods. (b) Data synthesis methods.




(c) Ours

Figure 8: Main caption describing all images

# Overall Pipeline



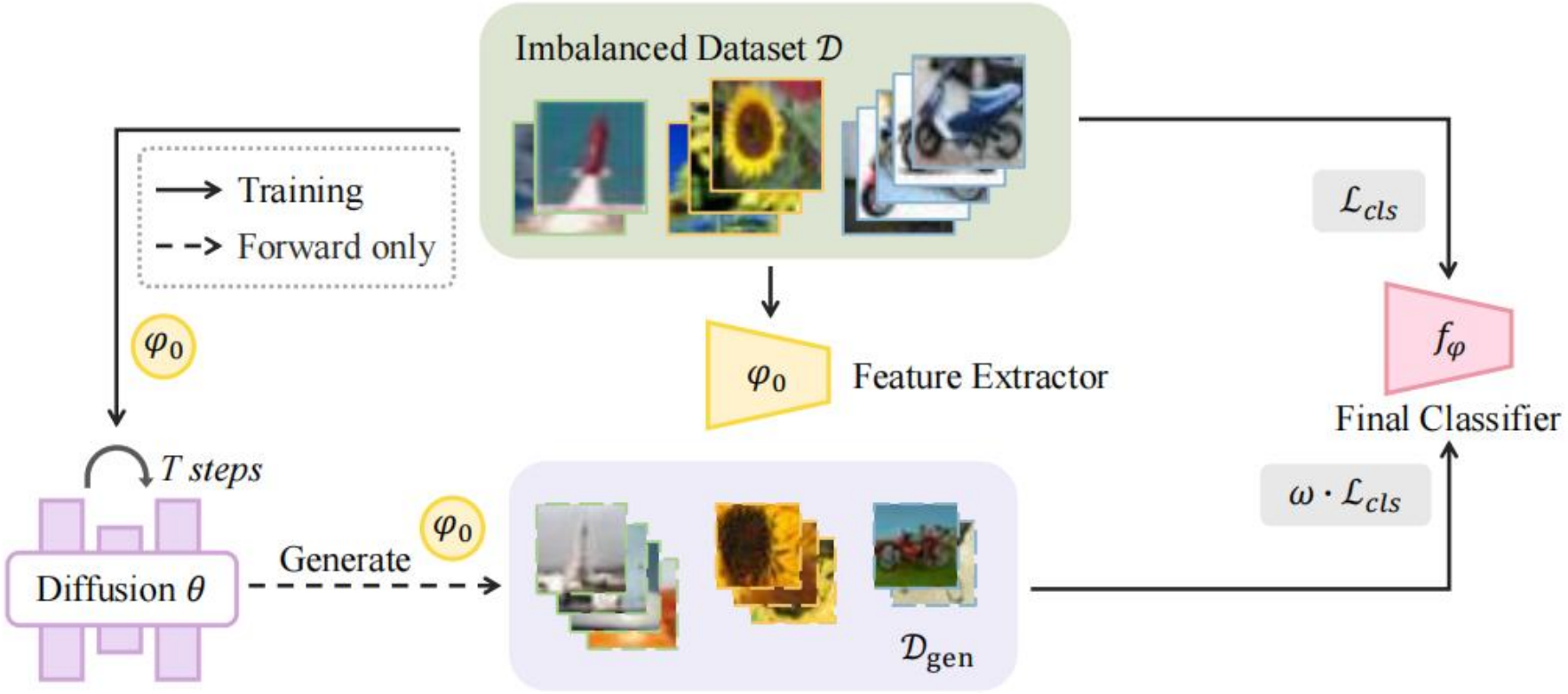
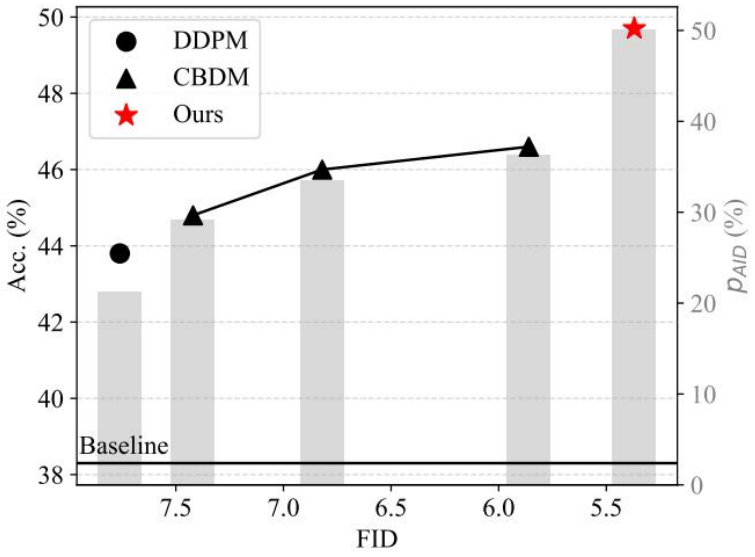



Figure 4: The overall pipeline of our method DiffuLT.

## FID Fréchet Inception Distance

$$d^2 = ||\mu_1 - \mu_2||_2^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\right)$$

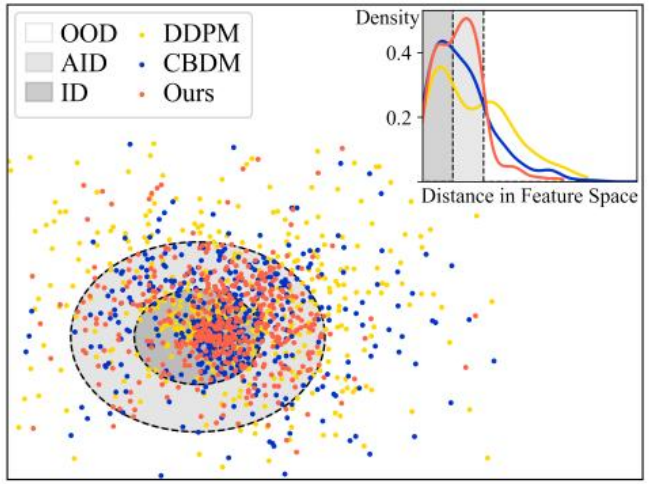
- Lower FID values indicate better alignment between generated and real distributions (higher quality/diversity).
- Higher FID suggests artifacts or mode collapse in generated samples.

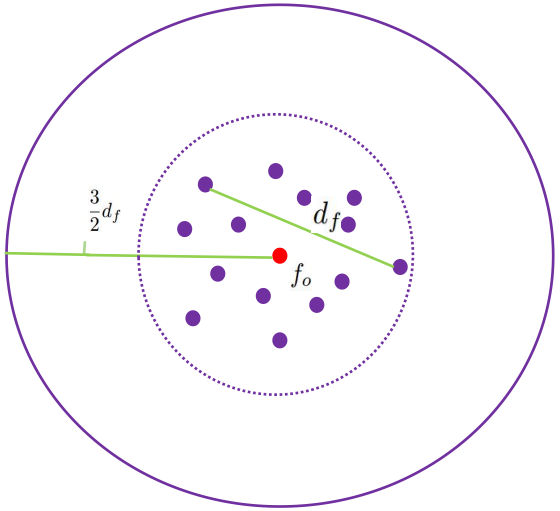



## AID approximately in-distribution

# Definition






$$\mathcal{D} \cup \mathcal{D}_{\text{gen}}$$

Define 3 types of the generated samples based on their distance

$$d_i = \|f_i - f_o\|_2 : \begin{cases} d_i \le d_f, & \text{ID} \\ d_f < d_i \le 2d_f, & \text{AID} \\ d_i > 2d_f, & \text{OOD} \end{cases}$$

Define the deviation at each step in feature space as

$$d_t = \frac{\sqrt{1-\bar{\alpha}_T}}{\sqrt{1-\bar{\alpha}_t}} \|\varphi_0(\boldsymbol{x}_0) - \varphi_0(\boldsymbol{x}_0 + \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, y))\|_2,$$

The new AID loss

$$L_{\text{AID}} = \alpha \mathbb{E}_{t\sim[1,T],\boldsymbol{x}_0,\epsilon_t} \left\| d_t - \frac{3}{2} d_f \right\|^2 .$$

They incorporate this term into both $L_{\mathrm{DDPM}}$ and $L_{\mathrm{CBDM}}$ to train the generation model.

$$L_{\mathrm{DDPM}} = \mathbb{E}_{t\sim[1,T],\boldsymbol{x}_0,\boldsymbol{\epsilon}_t}[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t)\|^2], \tag{2}$$

$$L_{\mathrm{CBDM}} = \frac{\tau t}{|\mathcal{Y}|}\sum_{y'\in\mathcal{Y}}(\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,y) - \mathrm{sg}(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,y'))\|^2 + \gamma\|\,\mathrm{sg}(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,y)) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,y')\|^2). \tag{3}$$

确保模型对每个类别 y 的噪声估计与对其他类别 y' 的噪声估计是相似的

确保模型对尾部类别 y' 的噪声估计不会偏离对类别 y 的噪声估计太远

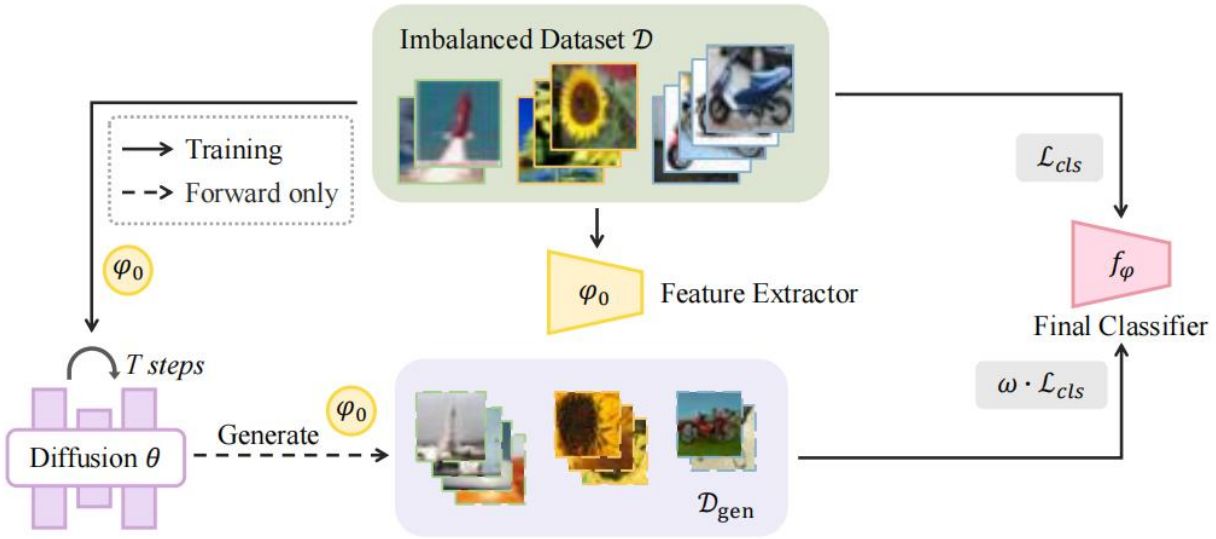
$$L_{\mathrm{cls}} = -\sum_{(x,y,y_g)\in\mathcal{D}\cup\mathcal{D}_{\mathrm{gen}}}(\omega y_g + (1-y_g))\log\frac{\exp(f_{\varphi,y}(x))}{\sum_{i=1}^{M}\exp(f_{\varphi,c_i}(x))},$$

ω = 0.3 controls the weight of generated samples

$y_g$ = 1 is used for generated samples, while $y_g$ = 0 marks the original ones.

# 3 Steps:






- **Training:** Initially, we train a feature extractor $\varphi_0$ and a conditional, AID-biased diffusion model $\theta$ using the original long-tailed dataset $\mathcal{D}$ alone.
- **Generating:** We establish a threshold $N_t$ and employ the trained diffusion model $\theta$ to generate and supplement samples. Using $\varphi_0$, we filter out OOD samples, resulting in a refined dataset $\mathcal{D}_{\text{gen}}$.
- **Training:** We then train a new classifier $f_\varphi$ on the augmented dataset $\mathcal{D} \cup \mathcal{D}_{\text{gen}}$ using weighted cross-entropy, forming our final model.

# Results



| Method | CIFAR100-LT | | | CIFAR10-LT | | | Statistics | | |
|---|---|---|---|---|---|---|---|---|---|
| | 100 | 50 | 10 | 100 | 50 | 10 | Many | Med. | Few |
| CE | 38.3 | 43.9 | 55.7 | 70.4 | 74.8 | 86.4 | 65.2 | 37.1 | 9.1 |
| Focal Loss Lin et al. [2017] | 38.4 | 44.3 | 55.8 | 70.4 | 76.7 | 86.7 | 65.3 | 38.4 | 8.1 |
| LDAM-DRW Cao et al. [2019a] | 42.0 | 46.6 | 58.7 | 77.0 | 81.0 | 88.2 | 61.5 | 41.7 | 20.2 |
| cRT Kang et al. [2019] | 42.3 | 46.8 | 58.1 | 75.7 | 80.4 | 88.3 | 64.0 | 44.8 | 18.1 |
| BBN Zhou et al. [2020a] | 42.6 | 47.0 | 59.1 | 79.8 | 82.2 | 88.3 | - | - | - |
| RIDE (3 experts) Wang et al. [2020] | 48.0 | - | - | - | - | - | 68.1 | 49.2 | 23.9 |
| CAM-BS Zhang et al. [2021a] | 41.7 | 46.0 | - | 75.4 | 81.4 | - | - | - | - |
| MisLAS Zhong et al. [2021b] | 47.0 | 52.3 | 63.2 | 82.1 | 85.7 | 90.0 | - | - | - |
| DiVE He et al. [2021] | 45.4 | 51.1 | 62.0 | - | - | - | - | - | - |
| CMO Park et al. [2022] | 47.2 | 51.7 | 58.4 | - | - | - | **70.4** | 42.5 | 14.4 |
| SAM Rangwani et al. [2022] | 45.4 | - | - | 81.9 | - | - | 64.4 | 46.2 | 20.8 |
| CUDA Ahn et al. [2023] | 47.6 | 51.1 | 58.4 | - | - | - | 67.3 | 50.4 | 21.4 |
| CSA Shi et al. [2023b] | 46.6 | 51.9 | 62.6 | 82.5 | 86.0 | 90.8 | 64.3 | 49.7 | 18.2 |
| ADRW Wang et al. [2024b] | 46.4 | - | 61.9 | 83.6 | - | 90.3 | - | - | - |
| H2T Li et al. [2023] | 48.9 | 53.8 | - | - | - | - | - | - | - |
| DiffuLT | 51.5 | 56.3 | 63.8 | 84.7 | 86.9 | 90.7 | 69.0 | 51.6 | 29.7 |
| DiffuLT + BBN | 51.9 | 56.7 | 64.0 | 85.0 | 87.2 | **90.9** | 69.5 | 51.9 | 30.2 |
| DiffuLT + RIDE (3 experts) | **52.4** | **56.9** | **64.2** | **85.3** | **87.3** | 90.9 | 70.3 | **52.1** | **30.7** |

# Problems



- The primary limitation of our methods is the **extensive training time** required for the generative model.
- For instance, training a diffusion model on CIFAR100-LT takes 24 hours, while ImageNet-LT requires approximately six days.
- As the quality and quantity of data increase, the training costs scale up significantly, making it challenging to apply our methods to larger datasets such as iNaturalist and Places-LT due to resource and time constraints.

- To **improve training efficiency**, they are exploring two potential solutions.
- The first involves adopting techniques that accelerate the training and inference processes of diffusion models.
- The second strategy considers the use of pre-trained generative models in real long-tail scenarios.

# Thanks