Efficient Diffusion-Based Synthetic Augmentation with Automatic OOD-Driven Filtering for Long-Tailed Skin Lesion Classification in Resource-Constrained Settings

Ignazio Gallo, Leonardo Sgroi, and Mattia Gatti

University of Insubria, Varese, Italy
ignazio.gallo@uninsubria.it

Abstract. Long-tailed class distributions are a defining challenge in medical image classification, especially in resource-constrained settings where limited data collection thins the tail of rare but clinically important classes. Diffusion-based synthetic augmentation is promising, but existing approaches rely on costly generative training and on a single, manually chosen OOD-filtering fraction. We study a parameter-efficient pipeline for HAM10000 skin lesion classification that adapts a frozen inpainting diffusion backbone with per-class Low-Rank Adaptation (LoRA), automatically selects an out-of-distribution (OOD) post-processor from a pool of established methods, and treats the clean-synthetic fraction γ as a tunable operating point. In a single-split HAM10000 experiment, automatic selection identifies Mahalanobis scoring as the highest-AUROC filter (0.994), ahead of KNN (0.943), ViM (0.856), MSP (0.632), ODIN (0.610) and Energy (0.493). A sweep over $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ affects balanced multiclass accuracy (BMA) non-monotonically: the highest observed value is 0.848 at $\gamma = 0.6$, against 0.836 for the real-only baseline and 0.832 at $\gamma = 1.0$. There, recall exceeds baseline for vascular lesions (+10.7 points), actinic keratoses (+1.5) and benign keratosis-like lesions (+0.9), but melanoma recall is *below* baseline at every γ tested (smallest gap -0.4, one image, at $\gamma = 0.8$). Since every configuration was trained once on one split, and γ was chosen on the fold where these numbers are measured, the 1.2-point BMA difference is an observation, not a statistically established improvement; the split is also image-level rather than lesion-grouped (38.7% of evaluation images share a lesion with training), so absolute values are optimistic. The class-level pattern is the more robust finding: synthetic augmentation reallocates sensitivity across classes in a γ-dependent way, and a γ chosen by aggregate accuracy alone can carry a cost on the most clinically consequential class.

Keywords: Long-tailed classification · Diffusion models · Low-Rank Adaptation · Out-of-distribution detection · Skin lesion classification · Resource-constrained settings.

1 Introduction

Skin cancer and other dermatological conditions are frequently diagnosed through dermoscopic image analysis, and deep learning classifiers have shown clinician-level performance on curated, balanced benchmarks [4]. Real-world dermoscopy datasets, however, are strongly long-tailed: a few common lesion types dominate, while clinically important classes – melanoma and basal cell carcinoma among them – remain severely under-represented [22]. Models trained under this imbalance reach high aggregate accuracy while under-performing on exactly the rare classes that carry the highest clinical risk. This is amplified in *resource-constrained settings*, where limited local data collection thins the tail further, compute budgets are tighter, and the expertise to hand-tune generative or anomaly-detection pipelines is scarce.

Jiang *et al.* [9] fine-tune a custom inpainting diffusion UNet on skin lesion data and filter the synthetic pool with a fixed OOD detector, reporting gains on tail classes of ISIC2019. Two of their design choices are costly here: training a diffusion UNet from scratch needs considerable compute and data, and the OOD filter is chosen a priori, requiring prior knowledge of which detector suits the dataset. We adapt this line of work on HAM10000 [22] with parameter efficiency and low-expertise deployability as explicit goals. Our contributions are:

1. **Parameter-efficient diffusion adaptation.** Instead of training an inpainting UNet from scratch, we keep a pre-trained Stable Diffusion inpainting backbone [17] frozen and train one LoRA [8] adapter per class, avoiding both a bespoke class-embedding network and any update to the shared backbone.
2. **Automatic OOD method selection.** Rather than committing to one detector, we benchmark established post-processors (MSP, ODIN, Energy, Mahalanobis, KNN, ViM) following the OpenOOD protocol [24] on a near-OOD surrogate built inside the training split, and select the highest-AUROC method.
3. **An explicit γ ablation with per-class reporting.** We sweep the clean-synthetic fraction $\gamma \in \{0.2, \ldots, 1.0\}$ and report per-class recall at every point, exposing a non-monotonic aggregate effect and a consistent cost to melanoma recall that aggregate metrics hide.

Filtered synthetic images enter classifier training directly as a class-balancing augmentation, keeping the synthesize-then-filter-then-train structure of [9] while replacing its two most resource-intensive components. We state at the outset what this study is: a completed, single-split, single-seed controlled experiment on one dataset. It does not demonstrate that the pipeline is statistically superior to the real-only baseline, that it runs on low-end hardware, or that it generalizes beyond HAM10000.

2 Related Work

Long-tailed classification and diffusion augmentation. LTMIC approaches fall into three families [26]: loss re-weighting/re-margining [2,16]; specialized archi-

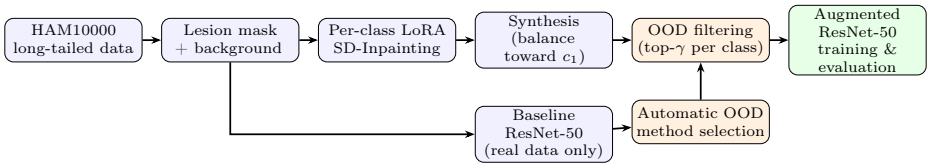

Fig. 1. Pipeline. Masks and lesion-free backgrounds are extracted once and used both to train per-class LoRA adapters on a frozen inpainting backbone and to condition synthesis. A real-data-only baseline classifier supplies the features for automatic OOD-method selection; the selected method filters the synthetic pool at fraction γ before it is merged with the real training data.

tectures or ensembles [10,27]; and class rebalancing by re-sampling [3] or synthetic generation [1]. The first two need dataset-specific tuning [9]; we belong to the third. Diffusion models [7,17] have been adapted to this setting: Qin *et al.* [15] propose class-balancing diffusion (CBDM) for tail-class diversity, and Shao *et al.* [19] show that approximately-in-distribution (AID) samples – neither too central nor too anomalous – are the most valuable for fine-tuning. Closest to us, Jiang *et al.* [9] fine-tune a custom inpainting UNet on lesion images plus segmentation-derived backgrounds and masks with AID-weighted oversampling, then filter with a fixed OOD detector. We keep that structure and the AID weighting, but replace the from-scratch UNet with LoRA adaptation of a pre-trained model [8,18] and the fixed filter with an automatically selected one.

Out-of-distribution detection. Established OOD scores are confidence-based, as in MSP [6] and ODIN [12]; energy-based [13]; feature-space distances such as Mahalanobis [11] and KNN [20]; or hybrid, as in ViM [23]. OpenOOD [24,25] standardizes them for like-for-like comparison. Existing LTMIC diffusion pipelines, including [9], fix one method in advance; we select it automatically on a near-OOD surrogate built from held-out real classes.

3 Method

Figure 1 summarizes the pipeline. For a long-tailed dataset $\mathcal{D}_{LT}$ with classes $c_1, \ldots, c_M$ ordered by decreasing frequency, the goal is a synthetic set $\mathcal{D}_{syn}$ that, filtered and merged with $\mathcal{D}_{LT}$, supports tail-class recognition without training diffusion from scratch or hand-picking an OOD filter.

Lesion Segmentation and Background Extraction. For every image we compute a binary lesion mask b_i and a lesion-free background d_i, following [9]: hair artifacts are removed with a black-hat filter and Telea inpainting [21], then the lesion is segmented by Otsu thresholding on the saturation channel, morphological closing/opening, and selection of the largest, most central component. This avoids a heavier external segmentation model (e.g., SAM [14]). Segmentation is an internal conditioning component, not a contribution: it only supplies a mask and background to the inpainting model, was never benchmarked against expert masks, and we make no claim about its accuracy.

Class-Conditioned LoRA Diffusion Inpainting. Rather than training an inpainting UNet and class-embedding network from scratch as in [9], we adapt a pre-trained Stable Diffusion inpainting model [17] with one LoRA [8] adapter per class on the UNet attention projections (W_q, W_k, W_v, W_o). Class conditioning uses a fixed prompt, *"dermatoscopic image of {class_name}, skin lesion close-up"*, plus a negative prompt steering away from other imaging modalities. The UNet, VAE and text encoder stay frozen: the only parameters updated for a class are its LoRA matrices ($r = 32$, $\alpha = 64$), so adapters train and store independently and adding a class touches neither the backbone nor the other adapters. This is the precise sense in which the pipeline is efficient – *parameter*-efficient adaptation of a pre-trained generator, a different claim from demonstrated low hardware requirements (Sec. 4).

The objective is the standard latent-diffusion denoising loss, conditioned on the masked background latent z_d and rescaled mask b^* concatenated with the noisy latent z_x:

$$\mathcal{L} = \mathbb{E}_{t\sim[1,T]}\left[\left\|\epsilon_t - \epsilon_\theta([z_{x,t}, z_d, b^*], t, y_{\text{emb}})\right\|^2\right], \tag{1}$$

where only the LoRA parameters $\theta \subset \theta_{\text{UNet}}$ are updated.

Anomaly-Weighted Oversampling. Following [19], each real sample's contribution to the LoRA batch is weighted by an anomaly score. Per class c_j a small convolutional autoencoder gives a reconstruction error A_i^j, and weights $w_i^j = 1/\exp\left(\left|\bar{A}^j - A_i^j\right|/\tau\right)$ ($\bar{A}^j$ the class median, τ a temperature) peak at the median (AID) and down-weight both typical and anomalous samples – a self-contained substitute for the self-supervised detector of [9].

Synthetic Generation and Automatic OOD Filtering. Synthetic samples target the same balancing objective as [9]: $|c_j^{syn}| = \max(0, |c_1| - |c_j|)$, i.e. minority classes are raised toward the majority class c_1 (`nv`). Because filtering later discards part of the pool, we over-generate by a factor 1.5, reusing real backgrounds and masks from class c_j as conditioning with that class's adapter active; $\gamma = 1.0$ therefore retains more synthetic images than strict balancing requires.

Unlike [9], which fixes one OOD detector a priori, we select it automatically, benchmarking the pool standardized by OpenOOD [24,25] – MSP [6], ODIN [12], Energy [13], Mahalanobis (MDS) [11], KNN [20], ViM [23] – on penultimate-layer features of the real-data-only baseline classifier (Sec. 3). Lacking a labelled OOD set, we build the near-OOD surrogate *entirely inside the training split*: `bkl` and `vasc` form the near-OOD pool and the other five the in-distribution (ID) pool; a per-class 10% sample of each gives the evaluation sets and the remaining ID images fit the feature-space detectors, so the evaluation fold is never touched. Candidates are ranked by AUROC separating ID from near-OOD and the maximizer $m^\star$ is selected.

$m^\star$ then scores every generated image and each class retains the top-γ fraction, $|c_j^{clean}| = \gamma \cdot |c_j^{syn}|$. Rather than fixing γ a priori as in [9] ($\gamma \in [0.2, 0.6]$),

we train the augmented classifier at each grid point and compare the operating points; how that comparison relates to the reported numbers is stated in Sec. 4.

Augmented Classifier Training. The classifier serving both as OOD feature extractor and final diagnostic model is a ResNet-50 [5] trained with cross-entropy and label smoothing (no class re-weighting). The *baseline* uses real data only; the *augmented* variant uses $\mathcal{D}_{LT} \cup \mathcal{D}_{syn}^{clean}$, with clean synthetic images fed through the same data loader, so γ changes without retraining the diffusion or OOD components.

4 Experimental Setup

Dataset, Split, and Evaluation Protocol. We evaluate on HAM10000 [22], a long-tailed dermoscopy benchmark with seven categories (`nv`, `mel`, `bkl`, `bcc`, `akiec`, `vasc`, `df`). Its 10 015 images are partitioned once by a stratified 5-fold split (`StratifiedKFold`, seed 42) computed *at the image level*, stratified on the diagnosis label alone; one fold (2 003 images) is held out and the other 8 012 form the training split used for baseline training, LoRA fine-tuning, generation and OOD-method selection. Only `image_id` and `dx` enter the pipeline. Two properties of this protocol bound how the results may be read.

The split is not grouped by lesion. HAM10000 provides a `lesion_id` and contains several dermoscopic images of the same physical lesion; our split ignores it. Auditing the exact partition shows this is not hypothetical: the 8 012 training images cover 6 303 distinct lesions and the 2 003 evaluation images 1 895; no image occurs on both sides, but 728 lesions do, so **775 evaluation images (38.7%) depict a lesion that also appears in training**. Part of the reported accuracy is therefore recognition of an already-seen lesion, and the absolute values below are optimistic relative to a lesion-grouped protocol. We report this rather than correct it: re-partitioning would invalidate every number below and could not be re-run before submission. The metadata carries no participant identifier (`lesion_id` groups a lesion, not a person), so participant-level grouping and counts cannot be established from it. The audit script and its output ship with the code.

The held-out fold is reused. The same 2 003 images provide the early-stopping and checkpoint-selection signal, the basis for comparing γ operating points, and the set on which every reported number is computed; there is no independent test set. Performance at the selected γ is therefore not an unbiased held-out estimate but a validation measurement at a point chosen on that same measurement, and is expected to be optimistically biased. We write *validation* throughout, never *test*. OOD-method selection is the one clean stage: its surrogates come wholly from the 8 012 training images (Sec. 3). Every configuration was trained once, one seed, one split, with no confidence intervals or significance tests.

Implementation Details. Generator. Stable Diffusion Inpainting v1.5 [17], frozen; LoRA rank 32, $\alpha = 64$, dropout 0, on the UNet attention projections; AdamW

at 5×10^{-5}, batch 4, gradient accumulation 4. Sampling uses 60 steps, guidance 8.0, strength 0.85. Per-class fine-tuning runs 50 epochs over the class's training images, capped at 4 000 steps; the head class `nv` is not fine-tuned, since no synthetic `nv` images are needed. The cap binds for the three largest minority classes, giving realized counts of 4 000 (`mel`, `bkl`, `bcc`), 3 300 (`akiec`), 1 450 (`vasc`) and 1 150 (`df`).

Classifier. ImageNet-pretrained ResNet-50, AdamW at 3×10^{-4}, cosine schedule, label smoothing 0.1, no class re-weighting, early stopping on validation BMA. *Metrics.* BMA is the mean per-class recall; **F1 is macro-averaged**; specificity is the unweighted mean of the seven one-vs-rest specificities. Macro sensitivity is by definition the mean per-class recall, hence identical to BMA, and is not reported separately. *Hardware.* All runs used one NVIDIA A100-SXM4 80GB GPU – the hardware we had, not a measured requirement: we neither profiled the pipeline on smaller devices nor ran a controlled runtime or memory comparison against from-scratch diffusion training [9]. The efficiency claim here is confined to the number of trainable generative parameters. *Reproducibility.* Code, configurations and experiment scripts, including the split-audit script above, are available at `https://github.com/ignaziogallo/LTMIC-mirasol-2026`; the HAM10000 images are not redistributed.

Compared configurations. A real-only baseline ($\gamma = 0$) against five pipeline configurations differing only in $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, all trained under an identical protocol and early-stopping criterion.

5 Results

Gamma Ablation: Aggregate Behaviour. Table 1 reports validation BMA, macro-F1, macro specificity, retained synthetic images and training cost. Aggregate performance is *non-monotonic* in γ: the highest observed validation BMA is 0.848 at $\gamma = 0.6$, against 0.836 for the real-only baseline, with 0.817 at $\gamma = 0.2$, 0.845 at $\gamma = 0.8$ and 0.832 at $\gamma = 1.0$. The largest difference from baseline is thus 1.2 points, from one seed on one split; with no repeated runs we cannot separate it from run-to-run variation and claim no statistical superiority. Macro-F1 never exceeds the baseline value (0.839) and specificity stays within 0.967–0.974. Training cost grows with γ but not monotonically, since the stopping epoch depends on when validation BMA last improved; the longest run ($\gamma = 0.8$, 39 epochs) takes over 4× the baseline's time. Fig. 2(a) shows the trend.

Per-Class Effect of Augmentation. Table 2 compares per-class recall between the baseline and the BMA-selected point $\gamma^\star = 0.6$. Recall is higher for `vasc` (+10.7 points), `akiec` (+1.5) and `bkl` (+0.9), unchanged for `bcc` and `df`, and lower for `nv` (−1.7) and `mel` (−3.6). Several are a handful of images on small denominators (`vasc` 28, `df` 23, `akiec` 65 validation images) and carry wide uncertainty. The finding is the pattern across the sweep rather than any single value (Fig. 2(b)): melanoma recall is *below* baseline at every γ tested – smallest gap −0.4 points at $\gamma = 0.8$, a single image (173 of 223 against 174) – while `vasc` and `akiec` sit at

Table 1. Gamma ablation: one stratified image-level fold, one seed per row, no repetitions. All metrics are computed on the 2 003-image validation fold, which also drove early stopping, checkpoint selection and the choice of $\gamma^\star$, so they are not unbiased held-out estimates. F1 is macro-averaged (macro sensitivity equals BMA and is omitted). Kept: synthetic images surviving OOD filtering; Time: wall-clock training to early stopping on one A100-SXM4 80GB. Bold marks this table's maximum, not a significant difference.

γ	Kept	BMA	Macro-F1	Spec.	Epochs	Time (min)
0.0 (baseline)	0	0.836	0.839	0.974	45	15.6
0.2	8 862	0.817	0.790	0.967	22	8.6
0.4	17 721	0.840	0.830	0.973	35	22.9
0.6 ($\gamma^\star$)	26 584	**0.848**	0.824	0.972	33	37.0
0.8	35 443	0.845	0.828	0.974	39	71.1
1.0	44 305	0.832	0.819	0.972	24	61.5

Table 2. Per-class validation recall: baseline vs. the BMA-selected $\gamma^\star = 0.6$; Δ in percentage points; single split, single seed. Class sizes in the fold are `nv` 1341, `mel` 223, `bkl` 220, `bcc` 103, `akiec` 65, `vasc` 28, `df` 23, so one image is 3.6 points for `vasc` and 4.3 for `df`. Melanoma is the only class below baseline at every γ in the sweep.

	`nv`	`mel`	`bkl`	`bcc`	`akiec`	`vasc`	`df`
Baseline	0.955	0.780	0.741	0.903	0.785	0.821	0.870
Ours ($\gamma^\star$)	0.937	0.744	0.750	0.903	0.800	0.929	0.870
Δ (pp)	−1.7	**−3.6**	+0.9	0.0	+1.5	+10.7	0.0

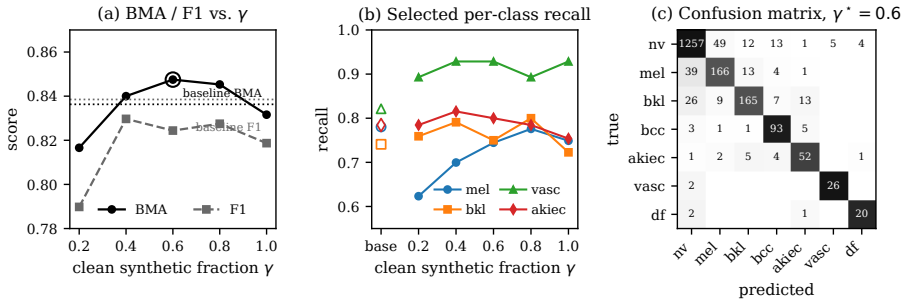

Fig. 2. Validation diagnostics (single split, single seed). (a) BMA and macro-F1 vs. γ with dotted baseline references; the circled $\gamma^\star = 0.6$ is the *highest observed validation BMA*, not a point shown to be optimal or clinically preferable. (b) Recall vs. γ for the four classes that move most; open markers at γ=0 give the real-only baseline. Melanoma stays below baseline across the sweep. (c) Row-normalized confusion matrix at $\gamma^\star = 0.6$, counts annotated.

or above baseline over most of the range. Melanoma, the class of highest clinical concern, is never improved here.

Confusion-Matrix Diagnostics. In Fig. 2(c) the `vasc` change resolves a pre-existing confusion with `bkl` (23/28 correct at baseline, 26/28 at $\gamma^\star$), while the

`mel` decrease is not a uniform sensitivity loss: confusion with `nv` falls (42 to 39 of 223) but confusion with `bkl` more than doubles (6 to 13) and with `bcc` rises (1 to 4), so correct melanomas drop from 174 to 166. The boundary shifts toward other pigmented-lesion classes rather than degrading uniformly, so γ should be chosen by the clinical cost of each boundary rather than by BMA alone.

6 Discussion and Conclusion

What this experiment shows. Automatic OOD-method selection has the widest observed margin: on the training-split surrogate MDS reaches AUROC 0.994 against 0.943 (KNN), 0.856 (ViM), 0.632 (MSP), 0.610 (ODIN) and 0.493 (Energy). A spread that large is the substantive argument against fixing the filter a priori – a practitioner without OOD expertise could have picked Energy, which here is no better than chance. The γ ablation shows that filtered synthetic data moves specific class boundaries rather than acting as a uniform regularizer: aggregate BMA responds non-monotonically while per-class recall is redistributed.

What it does not establish. Not statistical superiority: one split, one seed, no intervals, no tests, and a 1.2-point BMA difference well within what a single reshuffle could produce. Not clinical superiority: $\gamma^\star = 0.6$ costs 3.6 points of melanoma recall, so selecting γ by aggregate accuracy can land on a point a clinician would reject. In this sweep $\gamma = 0.8$ is the more defensible compromise – nearly the same BMA (0.845 vs. 0.848), the smallest melanoma penalty (-0.4 points, one image) – but we did not reach it by any principled criterion. Nor does it establish generalization beyond this HAM10000 split, low-hardware deployability, or superiority to other synthetic-augmentation methods, none of which were run; and because 38.7% of the evaluation images share a lesion with training (Sec. 4), the absolute values are optimistic relative to a lesion-grouped protocol.

What should happen next. In order: a lesion-grouped re-partition, which the audit above makes the most urgent, with repeated seeds and splits, intervals and paired tests – the only route by which the 1.2-point difference becomes a claim; external validation on another dermoscopy source; a clinically weighted or class-conditional criterion for γ, which the diagnostics of Sec. 5 suggest could keep the `vasc`/`akiec` gains without the melanoma cost; a memorization audit; a controlled compute comparison; and baselines such as CBDM [15] or SMOTE-style interpolation [3]. Two caveats persist: the near-OOD surrogate uses two real classes rather than an external source; and the segmentation conditioning the synthesis was never benchmarked against expert masks.

Conclusion. On the evaluated HAM10000 split, per-class LoRA adaptation of a frozen inpainting backbone with automatically selected OOD filtering measurably changes classifier behaviour, with the highest observed validation BMA at $\gamma = 0.6$. What we would defend is not that number but the structure underneath

it: synthetic augmentation reallocates sensitivity between classes, and a fraction chosen to maximize an aggregate metric can silently pay for it on the class that matters most.

References

1. Ahn, S., Ko, J., Yun, S.Y.: CUDA: Curriculum of data augmentation for long-tailed recognition. arXiv preprint arXiv:2302.05499 (2023). https://doi.org/10.48550/arXiv.2302.05499
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems. vol. 32 (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority oversampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
4. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017). https://doi.org/10.1038/nature21056
5. 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). https://doi.org/10.1109/CVPR.2016.90
6. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017). https://doi.org/10.48550/arXiv.1610.02136
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020). https://doi.org/10.48550/arXiv.2006.11239
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022). https://doi.org/10.48550/arXiv.2106.09685
9. Jiang, J., Subedar, M., Tickoo, O.: Synthetic data generation for long-tail medical image classification: A case study in skin lesions. arXiv preprint arXiv:2605.03221 (2026). https://doi.org/10.48550/arXiv.2605.03221
10. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations (2020). https://doi.org/10.48550/arXiv.1910.09217
11. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems **31** (2018). https://doi.org/10.48550/arXiv.1807.03888
12. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018). https://doi.org/10.48550/arXiv.1706.02690
13. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems **33**, 21464–21475 (2020). https://doi.org/10.48550/arXiv.2010.03759

14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
15. Qin, Y., Zheng, H., Yao, J., Zhou, M., Zhang, Y.: Class-balancing diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18434–18443 (2023). https://doi.org/10.1109/CVPR52729.2023.01768
16. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Advances in Neural Information Processing Systems **33**, 4175–4186 (2020). https://doi.org/10.48550/arXiv.2007.10740
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022). https://doi.org/10.1109/CVPR52688.2022.01042
18. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023). https://doi.org/10.1109/CVPR52729.2023.02155
19. Shao, J., Zhu, K., Zhang, H., Wu, J.: DiffuLT: Diffusion for long-tail recognition without external knowledge. Advances in Neural Information Processing Systems **37**, 123007–123031 (2024). https://doi.org/10.48550/arXiv.2403.05170
20. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning. pp. 20827–20840 (2022). https://doi.org/10.48550/arXiv.2204.06507
21. Telea, A.: An image inpainting technique based on the fast marching method. Journal of Graphics Tools **9**(1), 23–34 (2004). https://doi.org/10.1080/10867651.2004.10487596
22. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**, 180161 (2018). https://doi.org/10.1038/sdata.2018.161
23. Wang, H., Li, Z., Feng, L., Zhang, W.: ViM: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4921–4930 (2022). https://doi.org/10.1109/CVPR52688.2022.00487
24. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: OpenOOD: Benchmarking generalized out-of-distribution detection. In: Advances in Neural Information Processing Systems. vol. 35, pp. 32598–32611 (2022). https://doi.org/10.48550/arXiv.2210.07242
25. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y., Li, H.: OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. Journal of Data-centric Machine Learning Research (2024). https://doi.org/10.48550/arXiv.2306.09301
26. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10795–10816 (2023). https://doi.org/10.1109/TPAMI.2023.3268118
27. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9719–9728 (2020). https://doi.org/10.1109/CVPR42600.2020.00974