Subtyping Breast Lesions via Generative Augmentation based Long-tailed Recognition in Ultrasound

Shijing Chen[1,2,3*], Xinrui Zhou[1,2,3*], Yuhao Wang[1,2,3], Yuhao Huang[1,2,3], Ao Chang[1,2,3], Dong Ni[1,2,3], and Ruobing Huang[1,2,3(✉)]

[1] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Medical School, Shenzhen University, China
ruobing.huang@szu.edu.cn
[2] Medical Ultrasound Image Computing (MUSIC) Lab, Shenzhen University, China
[3] Marshall Laboratory of Biomedical Engineering, Shenzhen University, China

Abstract. Accurate identification of breast lesion subtypes can facilitate personalized treatment and interventions. Ultrasound (US), as a safe and accessible imaging modality, is extensively employed in breast abnormality screening and diagnosis. However, the incidence of different subtypes exhibits a skewed long-tailed distribution, posing significant challenges for automated recognition. Generative augmentation provides a promising solution to rectify data distribution. Inspired by this, we propose a dual-phase framework for long-tailed classification that mitigates distributional bias through high-fidelity data synthesis while avoiding overuse that corrupts holistic performance. The framework incorporates a reinforcement learning-driven adaptive sampler, dynamically calibrating synthetic-real data ratios by training a strategic multi-agent to compensate for scarcities of real data while ensuring stable discriminative capability. Furthermore, our class-controllable synthetic network integrates a sketch-grounded perception branch that harnesses anatomical priors to maintain distinctive class features while enabling annotation-free inference. Extensive experiments on an in-house long-tailed and a public imbalanced breast US datasets demonstrate that our method achieves promising performance compared to state-of-the-art approaches. More synthetic images can be found at https://github.com/Stinalalala/Breast-LT-GenAug.

Keywords: Breast ultrasound · Histological subtype · Long-tailed recognition · Diffusion model

1 Introduction

The global prevalence of breast cancer [21] drives widespread adoption of early screening tools, with ultrasound (US) emerging as the preferred modality for younger cohorts with dense breasts [6]. While standard protocols predominantly

[*] Shijing Chen and Xinrui Zhou contribute equally to this work.

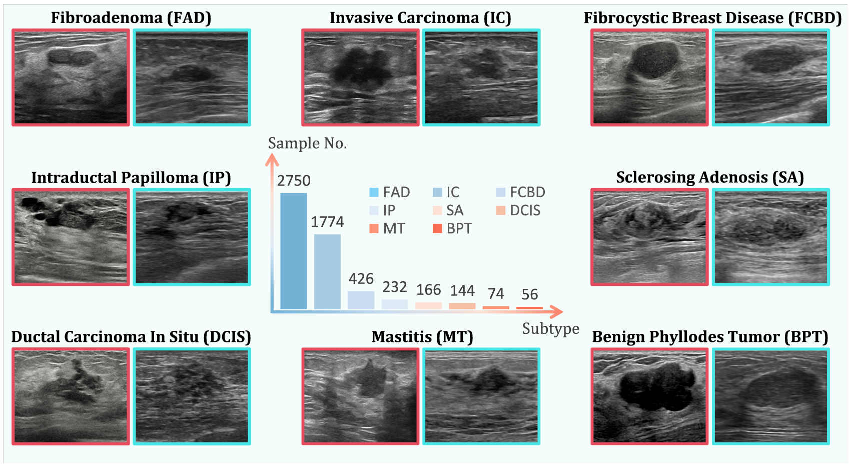

Fig. 1. Breast US images of lesions with different histological subtypes. The histogram indicates the incidence of different subtypes, which exhibit a long-tailed distribution. Red-bordered images are real US images, while the blue-bordered ones are synthetic data generated using the proposed framework.

emphasize binary lesion classification regarding malignancy/benignity, contemporary clinical findings suggested that the identification of different lesion subtypes may be critical for optimizing treatment planning and management [1].

However, the incidence rates of different subtypes naturally exhibit an extremely skewed long-tailed distribution that presents substantial challenges for accurate recognition (see Fig. 1). This diagnostic complexity is compounded by inter-class morphological overlap (e.g., FAD vs. BPT, IP vs. DCIS in Fig. 1), coupled with marked intra-class heterogeneity [12]. Common deep learning classifiers [2,8,14], despite their advancements, may suffer severe performance degradation in long-tail scenarios.

In recent years, many studies have been proposed for long-tailed recognition [29], which can be mainly divided into three categories: 1) Re-balancing-based methods seek to re-balance the negative influence brought by the class distribution asymmetry [3, 13]; 2) Architecture-enhanced approaches often ensemble multiple models or incorporate additional modules to improve the robustness against long-tailed data [4,10,18]; 3) Augmentation-based schemes aim to improve both the quantity and diversity of training datasets. Classical methods apply predefined transformations to data samples or features [5], while some recent approaches investigated the potential of applying denoising diffusion probabilistic models (DDPMs) to synthesize high-quality medical images for downstream tasks [11, 16]. Compared to re-balancing methods, synthetic approaches can rectify data distributions through conditional generation, thereby avoiding both overfitting to head classes and overcompensating for tail classes. Furthermore, as these methods only affect the training phase, they enable fast and lightweight inference during testing without requiring complex network structures in architecture-enhanced models. Nevertheless, current synthetic solutions may still face two limitations in addressing the proposed task: 1) The predominant paradigm fails to explicitly address distribution skewness and remains susceptible to biased sampling during model training. 2) Limited capability in capturing fine-grained tissue patterns that may undermine the discriminative power of synthetic instances, particularly in tail classes with limited exemplars.

Based on the above analysis, we propose a dual-phase framework for long-tailed classification that mitigates data asymmetry through high-fidelity data synthesis while dynamically modulates synthetic usage to maintain balanced classification performance. This cascaded data curation pipeline enhances diversity expansion while avoiding noise amplification. It is equipped with a reinforcement learning (RL)-driven class adaptive sampler that automates batch composition by learning to balance head-class stability and tail-class exploration. Additionally, our class-controllable synthetic network is also guided with a sketch-grounded perception branch that injects anatomical priors to retain class-discriminative traits while facilitating annotation-free inference.

2 Methodology

Fig. 2 presents our two-stage framework for long-tailed breast lesion classification in US images. The first stage employs a label-conditioned synthesizer with structural perception constraints, enabling class-tailored image synthesis from Gaussian noise while maintaining diagnostic structural fidelity. The second stage incorporates an RL-driven class adaptive sampler (RL-CAS) that automatically optimizes batch composition by dynamically balancing synthetic-real data during classifier training, effectively addressing long-tailed distribution challenges through focal learning.

2.1 Synthesizer for High-fidelity Image Generation

Preliminaries of DDPMs. DDPMs [9,19] are generative models that learn to model data distributions by iteratively denoising corrupted inputs. To alleviate the computational burden of pixel-space training in standard DDPMs, latent diffusion models (LDMs) [17] were introduced for perceptual compression. The objective function of LDM can be formulated as: $\mathcal{L}_{LDM} = \mathbb{E}_{z_0,\epsilon,t,\mathbf{c}}||\epsilon - \epsilon_\theta(z_t, t, \mathbf{c})||^2$, where z_0 is the latent code of the training data from a pre-trained variational autoencoder (VAE) [7], t represents the time step, ϵ_θ and ϵ are the predicted and target noise, respectively. Here, $\mathbf{c}$ indicates the (optional) control signal that the model can be conditioned on. In the proposed synthesizer, $\mathbf{c}$ is specifically defined as the class label representing the subtype of breast lesions.

Basic Architecture of Class-steerable Synthesizer. Building upon the LDM, our synthesizer architecture integrates two key synergistic components: a. variational latent encoding and b. guided denoising. The VAE encoder compresses medical images into compact latent representations, while the UNet-shaped denoiser progressively removes the noise added to the latent features

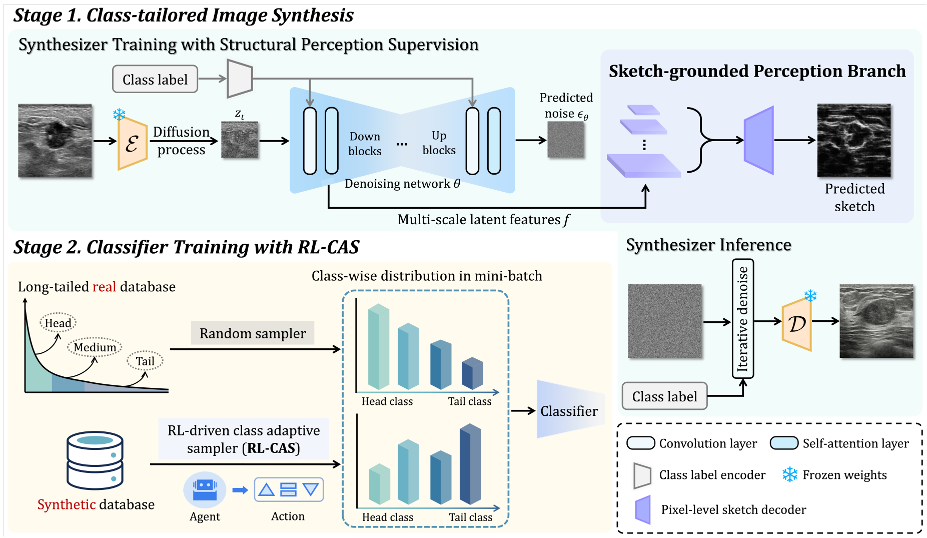
Stage 1. Class-tailored Image Synthesis
Synthesizer Training with Structural Perception Supervision
Class label
$\mathcal{E}$
Diffusion process
z_t
Down blocks
...
Up blocks
Denoising network θ
Predicted noise ϵ_θ
Sketch-grounded Perception Branch
Predicted sketch
Multi-scale latent features f
Stage 2. Classifier Training with RL-CAS
Long-tailed real database
Head
Medium
Tail
Random sampler
Class-wise distribution in mini-batch
Head class
Tail class
Classifier
RL-driven class adaptive sampler (RL-CAS)
Synthetic database
Agent
Action
Head class
Tail class
Synthesizer Inference
Iterative denoise
$\mathcal{D}$
Class label
Convolution layer
Self-attention layer
Class label encoder
Frozen weights
Pixel-level sketch decoder

Fig. 2. Pipeline of our proposed framework. $\mathcal{E}, \mathcal{D}$ indicate pre-trained encoder and decoder in VAE, respectively. z_t refers to latent features after the t-step diffusion process. RL, reinforcement learning.

through several convolution-attention hybrid blocks. To ensure diagnostic relevance, we implement disease-specific generation control by injecting class labels into the denoising trajectory, following a similar approach in [30].

Structural Perception Supervision for Refined Reconstruction. Label-guided data synthesis achieves basic class matching but struggles with anatomical structure fidelity in complex cases, limiting its practical utility. Current methods often inject geometric conditions (e.g., segmentation masks) in the denoiser to improve structural control [15, 28]. However, such implicit conditioning may degrade generative capacity in the absence of test-time annotation inputs, which is commonly unavailable in clinical scenarios [31]. Inspired by [23, 25], we propose to overcome this by introducing explicit structural perception supervision through sketches to capture anatomical priors. This design enables annotation-free inference while preserving fine anatomical details.

To this end, as shown in Fig. 2, we introduce a sketch-grounded perception branch during synthesizer training. Specifically, multi-scale latent features $f = [f_1, f_2, f_3, f_4]$ from the encoding path of the denoising network, corresponding to resolutions of $[H \times W, \frac{H}{2} \times \frac{W}{2}, \frac{H}{4} \times \frac{W}{4}, \frac{H}{8} \times \frac{W}{8}]$, are extracted as the branch input. H, W represent the height and width of the latent features before input to the denoiser, respectively. Then, we propose a pixel-level sketch decoder that concatenates f and performs upsampling via transposed convolution to produce sketch predictions. To achieve refined structural reconstruction, this study introduces a customized sketch loss, utilizing the rich anatomical priors from sketches.

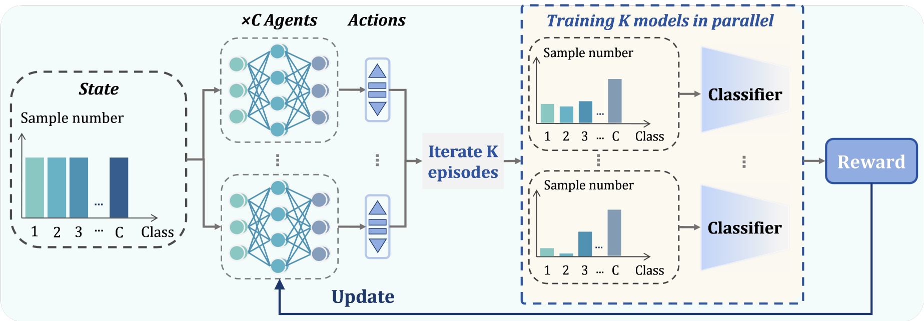

Fig. 3. Illustration of the RL-driven class adaptive sampler. Starting from a uniform initial state (green histogram on the left), multiple agents take different actions to modify the ratio of synthetic data during training to optimize the reward.

The core idea is to optimize the high-dimensional feature space of the denoiser in a self-supervised manner through an L_1 loss to minimize the reconstruction error between the predicted sketch S_{pred} and its ground truth counterpart S_{gt}. Note that S_{gt} is precomputed using the sketch extractor [20]. To balance the noise scale in the latent features, referring to [23], we apply $\sqrt{\bar{\alpha}_t}$ from DDPM [9] as scaling factors to the above L_1 loss for ensuring feature efficacy and easing the synthesizer training. To summarize, our sketch loss for structural perception supervision can be described as: $\mathcal{L}_s = \sqrt{\bar{\alpha}_t} \cdot L_1(S_{pred}, S_{gt})$. This encourages the synthesizer to focus on latent features with lower noise (i.e., smaller time steps) and vice versa. Eventually, the objective function of our synthesizer combines the basic $\mathcal{L}_{LDM}$ with $\mathcal{L}_s$, which is represented as follows: $\mathcal{L} = \mathcal{L}_{LDM} + \lambda\mathcal{L}_s$, where λ is a hyperparameter that controls the strength of the perception supervision.

2.2 RL-driven Class Adaptive Sampler for Enhanced Classification

Strategic deployment of these synthesized data during downstream training is equally important for effective long-tailed classification, as improper usage can exacerbate existing imbalances or introduce new biases. While synthetic data generation addresses class scarcity, its value hinges on how these samples interact with real data throughout the learning process. To address this, we propose an RL-driven class adaptive sampler that dynamically calibrates synthetic-real data ratios during the classifier training. This dynamic adjustment ensures that synthetic data selectively compensates for the scarcities of real data while maintaining the authentic patterns that ensure feature fidelity.

Following the classical RL setting, we define a multi-agent $M = \{m_1, ..., m_C\}$ with its current state S that interacts with the environment E by taking sequential actions A, aiming to maximize the expected reward. Concretely, the environment E is the breast US dataset containing both real and synthetic images. The state S is the set of synthetic sample counts for C classes in a mini-batch,

$S = \{s_1, s_2, \ldots, s_C\}$. Each agent m_i with its parameters θ_i adjusts the sampling number of synthetic images for a specific class. The entire action space is defined as $A = \{a_i | i = 1, 2, \ldots, C\}$, where $a_i = [-2, 0, +2]$ indicates the step size of the sampling number adjustment determined by probability $p(a_i)$ output from $m_i(\theta_i)$. As shown in Fig. 3, the agents iterate K episodes to generate K different states by choosing different actions. K identical classifiers are then trained in parallel using K batch compositions to update their parameters, respectively. The classifier with the highest validation metric is selected, and its weights are used to initialize the classifier for the next epoch. After K episodes, the parameters of the agents $\{\theta_i | i = 1, 2, \ldots, C\}$ are updated using the REINFORCE rule [24]. Particularly, the i^{th} agent is updated following:

$$\begin{aligned} \theta_i^{t+1} &= \theta_i^t + \eta \frac{1}{K} \sum_{j=1}^{K} (R_j - B^t) \cdot \nabla_\theta \log(g(a^{t,j})) \\ &= \theta_i^t + \eta \frac{1}{K} \sum_{j=1}^{K} (R_j - B^t) \cdot \nabla_\theta \sum_{i=1}^{C} \log(p(a_i^{t,j})), \end{aligned} \tag{1}$$

where η is the learning rate, $g(a^{t,j}) = \prod(p(a_i^{t,j}))$ represents the joint probability distribution of different actions in the j^{th} episode at t^{th} epoch. B^t is a baseline term to improve the stability of the agents [22], which is expressed as follows:

$$B^t = (1 - \gamma) \cdot B^{t-1} + \gamma \cdot (\frac{1}{K} \sum_{j=1}^{K} R_j), \tag{2}$$

where γ is a constant. Note that B^t is simplified to $\gamma \cdot (\frac{1}{K} \sum_{j=1}^{K} R_j)$ when $t = 1$. The reward R_j is defined as $(\epsilon_j + 0.04)^3$, where $\epsilon_j \in [0, 1]$ is the specific metric calculated on the validation set with j^{th} classifier. The cubic function is utilized to enhance the reward signal. This dynamic sampling adjustment automates mini-batch construction, effectively balancing long-tailed data distributions. Furthermore, it enables noise stabilization by modulating synthetic usage in sync with the model's evolving discriminative capability, preventing excessive synthetic samples from corrupting learned embeddings.

3 Experiments and Results

Datasets and implementation details. We evaluated our method on two breast US datasets: an in-house long-tailed dataset Breast-LT-8 (max class imbalance ratio 47.98:1, see Fig. 1) and the public BreastMNIST dataset [26] to provide complementary validation on class imbalance. Approved by the local IRB, the Breast-LT-8 dataset containing 5622 US images with 8 classes of different histological subtypes of breast lesions was collected from 2811 patients. All images were resized to 256×256 and corresponding ground truth labels were obtained through biopsies. The Breast-LT-8 was split randomly at the patient

Table 1. Performance comparison of different methods on the in-house Breast-LT-8 dataset and the public BreastMNIST dataset. Higher values mean better performance, except for FID. All metrics except FID and AUC are presented in percentages (%).

Methods	Breast-LT-8								BreastMNIST [26]	
	F1	Rec	Pre	All	Many	Med	Few	FID	Acc	AUC
Baseline	30.94	30.81	31.11	70.13	80.68	21.29	0.00	-	84.20	0.866
CBFocal	25.47	28.44	25.97	50.64	57.80	19.94	16.07	-	83.97	0.840
Logit-Adjust	31.97	32.10	31.88	70.96	81.27	20.44	6.25	-	87.18	0.888
LIFT	20.06	20.31	26.62	42.08	47.99	14.94	26.37	-	87.03	0.848
ProCo	30.36	38.21	29.87	58.65	65.28	24.79	36.67	-	83.13	0.868
NCL	29.15	35.10	29.44	54.78	61.73	19.69	**39.29**	-	88.46	0.861
GLMC	34.06	38.83	32.57	66.22	75.29	26.61	26.79	-	84.62	0.857
MGDM	32.78	33.41	32.58	70.00	80.11	20.52	12.50	36.28	87.05	**0.951**
Skin-SDM	33.30	35.40	32.25	68.59	77.58	27.32	9.38	32.46	88.72	0.938
Ours	**35.23**	**38.98**	**34.39**	**72.31**	**82.14**	**28.26**	31.25	**30.19**	**89.10**	0.924

Table 2. Results of the ablation study using the Breast-LT-8 dataset.

Methods	F1	Rec	Pre	All	Many	Med	Few
Baseline	30.94	30.81	31.11	70.13	80.68	21.29	0.00
+SynClass	31.82	32.14	32.33	70.58	80.42	19.39	9.38
+SynSketch	34.73	34.89	**34.94**	72.18	**82.23**	21.70	12.95
+SynClass+Re-sampling	29.48	34.96	29.35	60.58	67.81	20.17	**31.70**
+SynClass+RL-CAS	34.21	35.87	33.25	71.22	81.03	23.42	15.62
Ours	**35.23**	**38.98**	34.39	**72.31**	82.14	**28.26**	31.25

level with a ratio of 7:1:2 for training, validation, and testing. The BreastMNIST (containing 780 US images with size 224 × 224) was utilized for binary classification. The synthesizer was trained using an AdamW optimizer with a learning rate of 1e-4 for 200 epochs. The downsampling factor of VAE was 8. We set the λ in the sketch loss to 0.1. For the RL-CAS, the state was initialized to 2 and the episode $K = 3$ in each epoch. The γ was set to 0.99. The agents were trained for 30 epochs using an Adam optimizer with a learning rate of 1e-3.

Method Comparison. As shown in Table 1, the proposed method was compared against both classical and the state-of-the-art (SOTA) approaches in long-tail classification, including: a) re-balancing-based, i.e., CBFocal [3], Logit-Adjust [13]; b) architecture-enhanced, i.e., LIFT [18], ProCo [4], NCL [10]; c) augmentation-based, i.e., GLMC [5], MGDM [11], Skin-SDM [16]. Performance was calculated using common evaluation metrics such as F1-score (F1), precision (Pre), recall (Rec), all accuracy (All), shot accuracy (Many, Med, Few), Area Under Curve (AUC), as well as Fréchet Inception Distance (FID) for generative approaches. Note that we adopt ResNet with Balanced Softmax as a strong baseline for long-tailed classification to ensure fair benchmarking while the backbone can be easily replaced for future explorations.

In **Breast-LT-8** dataset, compared to the re-weighting methods (rows 2-3), our approach demonstrated significant improvements across all metrics which indicates its capacity on effectively recognizing tail classes while preserving robust performance on head classes (e.g., F1=35.32, Many=82.14, Few=31.25). In terms of accuracy for many-shot and few-shot, CBFocal and Logit-Adjust exhibited opposite performance trends, highlighting that different re-weighting strategies distinctly prioritize different class samples. Interestingly, although architecture-enhanced methods (rows 4-6) showed promising improvements in recognizing tail classes compared to the baseline, these gains may not translate into significant benefits in F1-score, suggesting that excessive focus on the tail classes may hinder the overall performance (e.g., see the low F1-score=20.06 and All Acc=42.08 obtained by LIFT). Overall, augmentation-based approaches (rows 7-10) achieved relatively higher F1-scores, likely arising from the increasing data quantity and diversity that mitigate data imbalance. Moreover, our proposed method outperformed other generative augmentation approaches (e.g., MGDM and Skin-SDM) across all metrics, validating its robust capability in adapting to long-tailed data distributions. Conversely, experimental results on the **BreastMNIST** dataset revealed narrowed performance gaps among mainstream methods. This convergence potentially originates from milder class imbalance and reduced task complexity in binary classification. Notably, our approach maintains superior accuracy, confirming its robustness for imbalanced classification scenarios.

Ablation studies. To investigate the impact of each proposed component, we conducted ablation experiments by ablating the generator(+Synclass), the structural perception branch(+SynSketch), and the RL-CAS sampler (see Table 2). Note that we also implement a classical re-sampling approach following [27] (see row 4) as an alternative for the RL-CAS sampler. Compared to the baseline, the incorporation of synthetic data alone has augmented the dataset volume and diversity that lead to elevated F1-score and Rec (see row 1 and 2). Furthermore, the addition of the Sketch-Grounded Perception not only boosted the many- and medium-shot accuracy, but also further enhanced recognition of tail classes. These findings underscore the indispensable role of fine-grained structural guidance in synthesizing discriminative training samples. Meanwhile, the hand-crafted fixed resampling methods (row 4) yielded lower F1 score even with synthetic augmentation. This verifies the previous hypothesis that improper or excessive usage of synthetic samples may induce subtype-specific overfitting at the expense of holistic model capability. In contrast, the RL-CAS dynamically calibrates synthetic-real data ratios to fully exploit synthetic samples while preserving holistic classification accuracy.

4 Conclusion

We propose a two-phase framework for long-tailed classification, addressing data asymmetry through high-fidelity synthesis and adaptive sampling. It integrates a reinforcement learning-driven sampler to balance head-class stability and tail-class exploration, alongside a sketch-grounded perception branch that injects anatomical priors to retain class-discriminative traits. Extensive experiments on in-house and public breast US datasets demonstrate balanced classification performance across different metrics. In the future, we will extend the framework to more long-tailed datasets and tasks.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62471305, 62101342, 12326619, and 62171290); Guangdong Basic and Applied Basic Research Foundation (No.2025A1515011448, and 2023A1515012960); Science and Technology Planning Project of Guangdong Province (No. 2023A0505020002); Frontier Technology Development Program of Jiangsu Province (No. BF2024078).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Barrios, C.H.: Global challenges in breast cancer detection and treatment. The Breast **62**, S3–S6 (2022)
2. Becker, A.S., Mueller, M., Stoffel, E., Marcon, M., Ghafoor, S., Boss, A.: Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. The British journal of radiology **91**(1083), 20170576 (2018)
3. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples (2019)
4. Du, C., Wang, Y., Song, S., Huang, G.: Probabilistic contrastive learning for long-tailed visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
5. Du, F., Yang, P., Jia, Q., Nan, F., Chen, X., Yang, Y.: Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15814–15823 (2023)
6. Dunne, R.M., O'Neill, A.C., Tempany, C.M.: Chapter 9 - imaging tools in clinical research: Focus on imaging technologies. In: Robertson, D., Williams, G.H. (eds.) Clinical and Translational Science (Second Edition), pp. 157–179. Academic Press, second edition edn. (2017). https://doi.org/https://doi.org/10.1016/B978-0-12-802101-9.00009-0
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
8. Han, S., Kang, H.K., Jeong, J.Y., Park, M.H., Kim, W., Bang, W.C., Seong, Y.K.: A deep learning framework for supporting the classification of breast lesions in ultrasound images. Physics in Medicine & Biology **62**(19), 7714 (2017)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Li, J., Tan, Z., Wan, J., Lei, Z., Guo, G.: Nested collaborative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6949–6958 (2022)

11. Luo, Y., Yang, Q., Fan, Y., Qi, H., Xia, M.: Measurement guidance in diffusion models: Insight from medical image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
12. Makki, J.: Diversity of breast carcinoma: histological subtypes and clinical relevance. Clinical medicine insights: Pathology **8**, CPath–S31563 (2015)
13. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
14. Moon, W.K., Lee, Y.W., Ke, H.H., Lee, S.H., Huang, C.S., Chang, R.F.: Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Computer methods and programs in biomedicine **190**, 105361 (2020)
15. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. pp. 4296–4304 (2024)
16. Patcharapimpisut, P., Khanarsa, P.: Generating synthetic images using stable diffusion model for skin lesion classification. In: 2024 16th International Conference on Knowledge and Smart Technology (KST). pp. 184–189 (2024)
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)
18. Shi, J.X., Wei, T., Zhou, Z., Shao, J.J., Han, X.Y., Li, Y.F.: Long-tail learning with foundation model: Heavy fine-tuning hurts (2024)
19. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
20. Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., Liu, L.: Pixel difference networks for efficient edge detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5117–5127 (2021)
21. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
22. Wang, K.N., Yang, X., Miao, J., Li, L., Yao, J., Zhou, P., Xue, W., Zhou, G.Q., Zhuang, X., Ni, D.: Awsnet: An auto-weighted supervision attention network for myocardial scar and edema segmentation in multi-sequence cardiac magnetic resonance images. Medical Image Analysis **77**, 102362 (2022)
23. Wang, Y., Gao, R., Chen, K., Zhou, K., Cai, Y., Hong, L., Li, Z., Jiang, L., Yeung, D.Y., Xu, Q., et al.: Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7246–7255 (2024)
24. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8**, 229–256 (1992)
25. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. Advances in Neural Information Processing Systems **36**, 54683–54695 (2023)
26. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)

27. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models (2023)
28. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
29. 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)
30. Zhou, X., Huang, Y., Dou, H., Chen, S., Chang, A., Liu, J., Long, W., Zheng, J., Xu, E., Ren, J., et al.: Ctrl-genaug: Controllable generative augmentation for medical sequence classification. arXiv preprint arXiv:2409.17091 (2024)
31. Zhou, X., Huang, Y., Xue, W., Dou, H., Cheng, J., Zhou, H., Ni, D.: Heartbeat: Towards controllable echocardiography video synthesis with multimodal conditions-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 361–371. Springer (2024)