LoGex: Improved tail detection of extremely rare histopathology classes via guided diffusion

Maximilian Müller and Matthias Hein

University of Tübingen and Tübingen AI Center
maximilian.mueller@wsii.uni-tuebingen.de

Abstract. In realistic medical settings, the data are often inherently long-tailed, with most samples concentrated in a few classes and a long tail of rare classes, usually containing just a few samples. This distribution presents a significant challenge because rare conditions are critical to detect and difficult to classify due to limited data. In this paper, rather than attempting to classify rare classes, we aim to detect these as out-of-distribution data reliably. We leverage low-rank adaption (LoRA) and diffusion guidance to generate targeted synthetic data for the detection problem. We significantly improve the OOD detection performance on a challenging histopathological task with only ten samples per tail class without losing classification accuracy on the head classes.

Keywords: OOD detection · Diffusion Models · Medical Imaging.

1 Introduction

Limited data availability poses a significant challenge for medical image analysis, especially considering the critical need for safety and reliability. In practical scenarios, clinical datasets can be highly imbalanced [15,35,28], with a significant portion of samples coming from a few classes (the head) and a large number of classes with only a few samples (the tail). While the primary focus is, e.g., on solving a classification task on the head effectively, ensuring that cases of rare or unknown diseases are not classified as one of the frequently occurring head classes is equally important.

Such setups can be framed as long-tailed classification tasks, where the goal is good classification performance on both head and tail. Several strategies have emerged to tackle class imbalances, typically involving targeted loss functions [2,20], oversampling or reweighting [3,17] or augmenting the dataset [10]. However, some classes might be too infrequent to predict reliably [35]. Therefore, it can be helpful to switch to a more manageable task: Classifying the head classes and simultaneously *detecting* tail samples as "abnormal" so that they can receive special treatment. This can, therefore, be framed as an OOD detection problem [12], where the goal is good classification performance on an in-distribution (the head classes) and detection of samples that are out-of-distribution (i.e., do not belong to the in-distribution classes: tail classes, but also potential new or unknown diseases). When regarding the tail classes as

out-distribution, OOD methods can be readily applied. Traditional approaches to OOD detection [12,11,21,19,26] typically work under the assumption of no prior access to outlier instances, even though there are methods leveraging auxiliary data [13,5]. Since, for the task at hand, one has specific knowledge about the out-distribution in the form of very few samples, this should be leveraged for improved detection performance. Except for [28], who devised a hierarchical training strategy for detecting tail samples in a dermatological task, approaches that leverage OOD detection methods for tail detection are, however, not well explored.

Simultaneously, the rapid development of publicly available, general-purpose diffusion models [27,14,25,30] has led to numerous applications that aim at improved classification [7,31]. In medical tasks, diffusion models were predominantly used for synthesizing [24,4,34,6] and for enhancing [33,22,29] images in a variety of domains. Within the scope of OOD detection in the medical domain, diffusion models have mainly been used for reconstruction-based methods [9,8,23]. In other fields, diffusion models were successfully deployed for synthesizing data for imbalanced classification [10] and for OOD detection [5], albeit without a notion of tail classes.

In this work, we present `LoGex` (LoRA+Guidance for improved detection of extremely rare classes), an approach tailored explicitly towards improved detection of tail samples in the case of extreme data scarcity while maintaining good performance on the head classes. Specifically, our contributions are:

- We present an extremely challenging long-tailed histopathology task with only 10 samples per tail class
- We devise `LoGex`, combining LoRA finetuning and guidance to synthesize tail samples that are useful for tail detection
- We report improved tail detection performance with `LoGex` compared to all baseline methods, without loss in head classification performance

2 Background: Diffusion models and Long-tailed classification

2.1 Diffusion models and DiG-IN guidance

In score-based diffusion models, a noise sample is drawn from a prior distribution and iteratively denoised until a sample that is supposed to be from the desired data distribution is obtained. In latent diffusion models (LDM) like StableDiffusion [27], the denoising is performed for T steps in the latent space of a variational autoencoder. The decoder of the VAE then transforms the final latent z_0 into pixel space. Deterministic solvers can perform the sampling in the latent space [30], making the entire diffusion process deterministic and differentiable. Additional conditioning signals C can be employed in the diffusion process by sampling from a conditional distribution $p(z|C)$, where the conditioning C can, e.g., be the encoding of a text prompt that the U-Net receives through

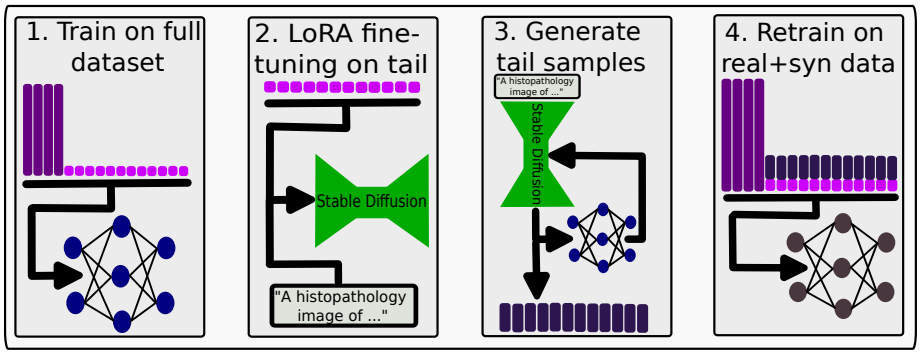

Fig. 1: Workflow of `LoGex`: 1. We train an auxiliary classifier on the long-tailed dataset's head and tail classes. 2. We adapt a general-purpose diffusion model to the histopathology domain by applying LoRA finetuning *only* on the tail samples. 3. We generate synthetic tail samples with the DiG-IN guidance from [1]. 4. We retrain a classifier by adding the synthetically generated tail samples to the train dataset.

cross-attention layers. While several approaches to guiding the diffusion process with other inputs exist [30,14], the authors in [1] showed that fine-grained class structures could be best resolved when differentiating through the *complete* denoising process [32]. In their approach, called Dig-IN, the output of the diffusion model is taken as an input to a vision model (e.g. a classifier), and a loss L is computed from the output of that vision model. The starting latent z_T and the conditioning of the diffusion process are then optimized with respect to this loss by backpropagating *through the entire diffusion process*, which is possible with deterministic solvers. This way, the diffusion process is guided to generate samples with low L. In our case, we will leverage DiG-IN to maximize the classifier-specific confidence of tail classes during the generation of synthetic tail samples.

LoRA. To leverage general-purpose large-scale diffusion models like Stable-Diffusion [27], these models frequently must be adapted to specialized domains, e.g., medical ones. While finetuning is often prohibitively expensive and difficult, [16] proposed to freeze the original weights instead and only to learn pairs of low-rank-decomposition matrices that are added to the attention layers (called LoRA). While initially designed for large language models, LoRA was shown to outperform other adaption methods [16] and to be effective for diffusion models, too. We will use this approach to adapt StableDiffusion to our task.

2.2 Long-tailed classification and OOD detection

There has been work directly linking long-tailed classification, OOD detection, and diffusion models. [28] devise a hierarchical outlier detection (HOD) loss

function to improve the detection of rare and unseen classes of a dermatology classifier. This loss can be written as $\mathcal{L}_{HOD} = \mathcal{L}_{fine} + \lambda\mathcal{L}_{coarse}$, where $\mathcal{L}_{fine}$ is simply the cross-entropy loss over all classes, and the coarse-grained loss is $\mathcal{L}_{coarse} = -\sum_{c\in\{head,tail\}} \mathbf{1}(y = c) \log p(c|\mathbf{x})$ for an input sample $\mathbf{x}$. The term $p(c|\mathbf{x})$ is the head/tail probability, which is computed as the sum of the class probabilities for all classes in the head/tail: $p(c|\mathbf{x}) = \sum_{k\in c} p(k|\mathbf{x})$ for $c \in \{\text{head}, \text{tail}\}$. At inference time $p(c = tail|\mathbf{x})$ is then used as OOD detection score.

A promising approach leveraging diffusion models for synthesizing images for OOD detection is called Dream-OOD[5]. In Dream-OOD a classifier is trained to learn a text-conditioned latent space of the in-distribution data. Then, outliers in the low-likelihood region of this latent space – corresponding to the boundary of ID data – are sampled and eventually decoded into pixel-space images by the diffusion model. A new classifier is then trained with cross-entropy on the in-distribution images, and a binary classifier is additionally employed on the model outputs to distinguish the in-distribution images from the synthetically generated samples. Another data generation approach is presented in [10], where the authors augment an imbalanced dataset with diffusion-generated synthetic images. In particular, they suggest guiding the diffusion process by enforcing low entropy with classifier guidance. We refer to this approach as *FG-entropy* (feedback-guided entropy) and include it in our experiments.

3 Our method: `LoGex`

The high-level idea behind our method is that even though a diffusion model might not be able to generate realistic samples that are useful as additional training data to improve the classification performance, it might still be able to synthesize samples that are *valid enough* to enhance OOD detection. To improve *classification* on the tail classes, it is necessary to generate sufficiently diverse, class-specific samples from the tail distribution [10]. For highly specialized domains like histopathology images, a generic diffusion model for natural images, e.g., Stable Diffusion [27] is unlikely to provide good sample quality directly, and even a specialized diffusion model for histopathology images should be unable to do this as for the rare tail classes it does not have enough training data. However, for OOD detection, the requirements for an auxiliary 'out-distribution' are lower: It is enough if the samples do *not* belong to the in-distribution while still being informative for the task at hand. Therefore, we aim to generate samples with *features* from the tail distribution and do not care if those samples are not entirely realistic, e.g., a mixture between two tail classes that might not occur naturally. While LoRA finetuning can help to adapt a pretrained diffusion model to a specialized domain, it is not enough to generate realistic enough tail features, as we show in Section 4.4. We, therefore, seek to include additional class-specific information in the diffusion process with DiG-IN guidance. To this end, we devise `LoGex`, a strategy for the effective generation of samples with tail features. We illustrate our approach in Figure 1 and outline it in the following:

1. **Train auxiliary classifier.** We train an auxiliary classifier on the long-tailed dataset (head and tail). This classifier will be used to guide the diffusion process and should thus ideally have some notion of what class-specific features both from the head and the tail classes look like.
2. **LoRA Finetuning.** To adapt a general-purpose diffusion model to the domain of histopathology images, we apply *LoRA* finetuning with the tail samples to a pretrained general-purpose diffusion model. Since the finetuning process cannot resolve fine-grained class-specific differences, we purposefully exclude samples from the head for the finetuning process. In doing so, we attempt to avoid mixing up features from the head and tail as much as possible. For finetuning, we use the text prompt *"A histopathological slide from a patient with* `<class>`*"*. Details are reported in the Appendix, Table 4.
3. **Synthesize tail samples with DiG-IN guidance [1].** We use the auxiliary classifier and the LoRA-finetuned diffusion model to perform the guided, class-specific generation of tail samples. In particular, we use the same prompt as employed during LoRA finetuning for each tail class. Additionally, during the diffusion process, we optimize the initial latent of the denoising process to maximize the confidence of the auxiliary classifier for the respective tail class.
4. **Augment the dataset and retrain the final classifier.** We augment the train dataset. For each tail class, we add the respective synthetically generated samples. We then train a classifier with regular cross-entropy loss on the augmented train set and select the best-performing model on the validation set. As a selection criterion, we use the difference of the balanced accuracy on the head classes and the FPR of the tail detection. Since the generated synthetic images potentially contain features of different tail classes, the final classifier might confuse certain tail classes. For this reason, we use the tail probability (i.e., the sum of the individual tail class probabilities) as OOD score.

4 Experiments

4.1 Setup

We create a challenging histopathology classification task based on the dataset of [18]. The original dataset contains 129.364 image tiles from 386 cases manually annotated into 16 classes and split into a training, validation and test set. The classification task on the original dataset can be solved very well (97% accuracy on the test set is achieved). To simulate a long-tailed scenario, we split and subsample the dataset into four head classes with more than 1000 training samples per class and 12 tail classes with only ten samples per class. To make the task challenging, we designed the split so that similar classes are split into head and tail. We also subsample the available validation set to 100 samples per head class and ten samples per tail class. The test set we leave unchanged. Details on the dataset are illustrated in Figure 4 in the Appendix.

Experimental Details. We train a ResNet-50 for 60 epochs with AdamW and a cosine scheduler (base learning rate is 10^{-4}). For long-tailed learning methods, we follow the setup of previous studies [15] and deploy three loss functions (standard cross-entropy, LDAM [2], and focal loss [20]). The latter two are specifically designed for highly imbalanced problems. We combine them with the reweighting strategy from [3] (rw) and further evaluate classifier-retraining (CRT [17]), deferred reweighting (DRW [2]), and the HOD training strategy from [28]. As baselines using synthetic data, we adopt Dream-OOD [5] and FG-Entropy [10]. Since Dream-OOD does not explicitly use the available natural tail samples, we add them to the generated synthetic images, leading to a total out-distribution of size 1552. Both for FG-Entropy from [10] and `LoGex`, we generate 100 samples per tail class, as we found that increasing the number of synthetic samples beyond that did not bring additional gains (see App. Fig. 3). For `LoGex`, we use StableDiffusion-1.4 as base model and perform LoRA on the cross-attention layers (details in App. Table 4). We optimize the diffusion process until we reach a class confidence of at least 40% on the auxiliary classifier. As ablation, we also evaluate `LoGex` without guidance, i.e., only after LoRA finetuning.

4.2 Conventional OOD detection approaches

We first investigate if conventional OOD detection approaches can be effective in our setting. To this end, we evaluate commonly used OOD scores, like Max-Softmax (MSP) [12] and (relative [26]) Mahalanobis distance [19] for a ResNet-50 that was trained with cross-entropy and without synthetic data. The OOD scores are usually entirely based on the in-distribution (in our case, the head) and assume no knowledge about the out-distribution. Since we treat the tail as an out-distribution, we can also adopt the metrics to be tail-specific. E.g., the negative Max-softmax value across the tail classes (MSP-tail) is a natural OOD score for our setting. Similarly, Maha-tail denotes an OOD score based on the Mahalanobis distance to the tail distribution instead of the head. We report the results in Table 1. The scores based on Mahalanobis distance are outperformed by the MSP baselines and only the relative Mahalanobis distance (rMaha-head) performs comparable to the MSP baseline. This is likely due to the low-data regime of our setup, which might make the mean and covariance estimation brittle, and underlines the need for tailored solutions. The best-performing method is based on the sum of the tail probabilities $P(tail)$, which was used in [28]. We will use it as the default OOD score for the main experiments. More (tail-specific) scores are reported in Table 3 in the Appendix.

Table 1: **Comparing OOD scores and their tail-specific versions:** The accumulated tail probabilities $P(tail)$ perform best.

	MSP-head	MSP-tail	Maha-head	rMaha-head	Maha-tail	P(tail)
FPR	$30.98^{\pm 2.36}$	$25.33^{\pm 1.91}$	$53.40^{\pm 3.10}$	$32.44^{\pm 4,06}$	$98.46^{\pm 0.63}$	$\mathbf{25.03}^{\pm 2.16}$
AUC	$93.85^{\pm 0.58}$	$94.98^{\pm 0.47}$	$82.58^{\pm 1.52}$	$91.87^{\pm 0.75}$	$23.42^{\pm 1.55}$	$\mathbf{95.01}^{\pm 0.48}$

Table 2: `LoGex` achieves the best tail detection performance and very high balanced accuracy on the head. We highlight the **best** and second best methods.

	method	FPR	AUC	bAcc-head	bAcc-tail
without syn data	Cross-entropy	$25.05^{\pm 2.20}$	$95.01^{\pm 0.48}$	$96.67^{\pm 0.32}$	$44.47^{\pm 4.68}$
	Cross-entropy + rw[3]	$25.86^{\pm 3.73}$	$94.16^{\pm 0.76}$	$94.02^{\pm 0.73}$	$56.84^{\pm 2.51}$
	Cross-entropy + HOD[28]	$23.45^{\pm 5.14}$	$95.05^{\pm 0.79}$	$96.65^{\pm 0.99}$	$49.75^{\pm 2.41}$
	Cross-entropy + CRT[17]	$41.97^{\pm 9.53}$	$91.41^{\pm 1.40}$	$84.24^{\pm 1.71}$	$68.26^{\pm 1.91}$
	Focal[20]	$23.48^{\pm 6.32}$	$94.74^{\pm 1.20}$	$96.27^{\pm 0.91}$	$48.09^{\pm 4.31}$
	Focal[20] + rw[3]	$25.23^{\pm 6.02}$	$93.56^{\pm 1.19}$	$94.02^{\pm 1.30}$	$57.22^{\pm 1.81}$
	LDAM[2]	$36.24^{\pm 5.67}$	$91.96^{\pm 1.07}$	$95.61^{\pm 1.43}$	$56.62^{\pm 2.14}$
	LDAM[2] + rw[3]	$50.08^{\pm 6.54}$	$89.95^{\pm 1.06}$	$91.21^{\pm 1.27}$	$64.82^{\pm 2.53}$
	LDAM[2] + rw[3] + DRW[2]	$50.61^{\pm 4.79}$	$88.63^{\pm 2.90}$	$91.36^{\pm 3.01}$	$59.60^{\pm 18.43}$
syn data	Dreamood[5]	$26.21^{\pm 2.75}$	$92.33^{\pm 0.68}$	$\mathbf{97.35}^{\pm 0.89}$	—
	FG-Entropy[10]	$20.66^{\pm 4.29}$	$\underline{95.48}^{\pm 1.05}$	$96.21^{\pm 0.63}$	$52.04^{\pm 4.72}$
	`LoGex` (only LoRA)	$\underline{19.26}^{\pm 3.22}$	$\underline{95.48}^{\pm 0.82}$	$96.37^{\pm 0.75}$	$49.42^{\pm 5.01}$
	`LoGex`	$\mathbf{16.25}^{\pm 1.87}$	$\mathbf{96.27}^{\pm 0.33}$	$\underline{96.97}^{\pm 1.04}$	$54.80^{\pm 1.27}$

4.3 Main Results

We report the results of our study in Table 2. Overall, `LoGex` clearly outperforms all other methods in terms of FPR and AUC while achieving very high balanced accuracy on the head classes. The second-best method is an ablation from `LoGex` with only LoRA and without DiG-IN guidance, highlighting that *both* LoRA and guidance are crucial to achieve strong detection performance. The best baseline method without synthetic data is the hierarchical training strategy (HOD) from [28], which achieves good detection performance and balanced head accuracy. Adding reweighting strategies (CRT, DRW, rw) improves tail classification performance for all three loss functions. This, however, comes at the cost of lower head accuracy and is also detrimental to the detection performance, and is therefore not a sensible strategy for our task. For the reweighting methods, the accuracies on the tail classes are still significantly lower than on the head, underlining that classifying both head and tail is infeasible. Thus, tailored approaches like `LoGex`, which improve tail detection while performing strongly on the head classification task, are favourable. Dream-OOD achieves the best head accuracy but fails to detect tail samples effectively. The reason for this is likely the training scheme of Dream-OOD, which does not leverage the class structure of the available tail samples. On the contrary, FG-Entropy can leverage the class structure (i.e., trains the classifier on both head and tail classes) and is the best baseline method, but still outperformed by `LoGex` in both detection and classification tasks. FG-Entropy is designed for scenarios where the diffusion model is powerful enough to resolve fine-grained class differences by itself and the entropy guidance only acts as a regularizer for diversity. Due to the limited training data in our setup, the diffusion model (even after LoRA finetuning) cannot do so and needs class-specific guidance like in `LoGex` to create meaningful tail features.

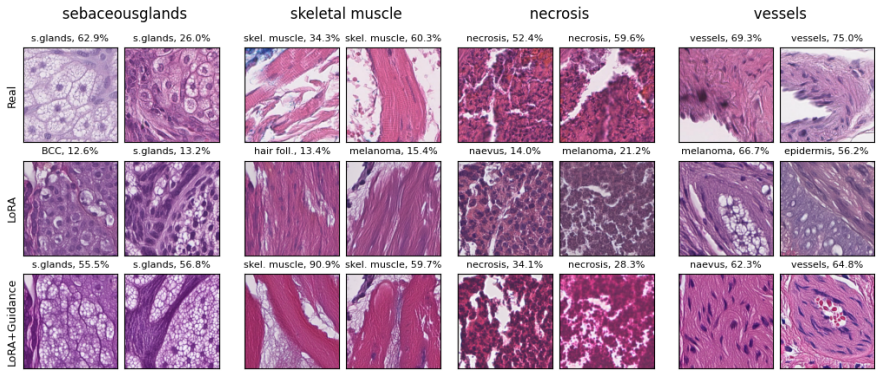

Fig. 2: **Guidance matters:** We show samples from the train set (first row), samples generated with LoRA (second row), and samples generated with `LoGex` (third row) and the predictions of a classifier trained on the original dataset. With `LoGex` the synthetic images are more often classified as the desired class.

4.4 Ablation: How good are the generated samples?

To analyze the quality of the synthetic samples, we use a classifier that was trained on the original dataset from [18], achieving a tail accuracy of 96.1%. We only use the classifier for this retrospective analysis and in no way during the main experiments. On the synthetically generated tail samples we obtain an accuracy of 78.8% for `LoGex`, 30.9% for `LoGex` without guidance, and 35.6% for FG-entropy, highlighting the improved quality of our synthesized images. We show samples for selected classes in Figure 2 and report the corresponding predictions. The samples generated with `LoGex` contain more pronounced features from the tail classes as compared to only LoRA finetuning, even though they are not always classified as the desired class. In a *vessels* sample, for instance, the classifier detects features from *naevus*. We note that such samples can still be useful for improved tail detection, albeit not for tail classification.

5 Conclusion

We developed `LoGex`, an effective strategy that improves the tail detection performance without loss in classification performance of the head classes on a challenging histopathological task. We combine targeted classifier guidance and LoRA finetuning to adapt a pretrained general-purpose diffusion model to synthesize useful samples, outperforming other methods by a clear margin.

Acknowledgments. We acknowledge support from the DFG (EXC number 2064/1, Project number 390727645) and the Carl Zeiss Foundation in the project "Certification and Foundations of Safe Machine Learning Systems in Healthcare".

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Augustin, M., Neuhaus, Y., Hein, M.: Analyzing and explaining image classifiers via diffusion guidance. ArXiv (2023), https://arxiv.org/abs/2311.17833
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 (2019), https://arxiv.org/abs/1906.07413
3. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR (2019)
4. Dorjsembe, Z., Odonchimed, S., Xiao, F.: Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In: Medical Imaging with Deep Learning (2022), https://openreview.net/forum?id=Oz7lKWVh45H
5. Du, X., Sun, Y., Zhu, X., Li, Y.: Dream the impossible: Outlier imagination with diffusion models. In: Advances in Neural Information Processing Systems (2023)
6. Frisch, Y., Fuchs, M., Sanner, A., Ucar, F.A., Frenzel, M., Wasielica-Poslednik, J., Gericke, A., Wagner, F.M., Dratsch, T., Mukhopadhyay, A.: Synthesising rare cataract surgery samples with guided diffusion models. In: MICCAI 2023
7. Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, I., Stanforth, R., Hayes, J., De, S., Smith, S.L., Wiles, O., Balle, B.: Differentially private diffusion models generate useful synthetic images https://arxiv.org/abs/2302.13861
8. Graham, M.S., Pinaya, W.H., Tudosiu, P.D., Nachev, P., Ourselin, S., Cardoso, J.: Denoising diffusion models for out-of-distribution detection. In: Proceedings of the CVPR Workshops (2023)
9. Graham, M.S., Pinaya, W.H.L., Wright, P., Tudosiu, P.D., Mah, Y.H., Teo, J.T., Jäger, H.R., Werring, D., Nachev, P., Ourselin, S., et al.: Unsupervised 3d out-of-distribution detection with latent diffusion models. In: MICCAI (2023)
10. Hemmat, R.A., Pezeshki, M., Bordes, F., Drozdzal, M., Romero-Soriano, A.: Feedback-guided data synthesis for imbalanced classification (2023)
11. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: ICML (2022)
12. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017), https://openreview.net/forum?id=Hkg4TI9xl
13. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. Proceedings of the International Conference on Learning Representations (2019)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020), https://arxiv.org/abs/2006.11239
15. Holste, G., Wang, S., Jiang, Z., Shen, T.C., Shih, G., Summers, R.M., Peng, Y., Wang, Z.: Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In: MICCAI DALI Workshop 2022
16. 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://arxiv.org/abs/2106.09685
17. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020), https://openreview.net/forum?id=r1gRTCVFvB

18. Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Meliß, R.R., Sack, U., Steinbuss, G., Kriegsmann, M.: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections [data] (2023), https://doi.org/10.11588/data/7QCR8S
19. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (2020). https://doi.org/10.1109/TPAMI.2018.2858826
21. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems (2020)
22. Ma, J., Zhu, Y., You, C., Wang, B.: Pre-trained diffusion models for plug-and-play medical image enhancement. In: MICCAI 2023
23. Mishra, D., Zhao, H., Saha, P., Papageorghiou, A.T., Noble, J.A.: Dual conditioned diffusion models for out-of-distribution detection: Application to fetal ultrasound videos. In: MICCAI 2023
24. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., Kather, J.N., Truhn, D.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Scientific Reports (2023)
25. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8162–8171 (18–24 Jul 2021)
26. Ren, J., Fort, S., Liu, J., Roy, A.G., Padhy, S., Lakshminarayanan, B.: A simple fix to mahalanobis distance for improving near-ood detection (2021), https://arxiv.org/abs/2106.09022
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
28. Roy, A.G., Ren, J.J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z.W., Vo, N., Bui, P., Winter, S., MacWilliams, P., Corrado, G., Telang, U., Liu, Y., Cemgil, T., Karthikesalingam, A., Lakshminarayanan, B., Winkens, J.: Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. Medical Imaging Analysis (2021)
29. Shen, Y., Ke, J.: Staindiff: Transfer stain styles of histology images with denoising diffusion probabilistic models and self-ensemble. In: MICCAI 2023
30. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021), https://openreview.net/forum?id=St1giarCHLP
31. Trabucco, B., Doherty, K., Gurinas, M., Salakhutdinov, R.: Effective data augmentation with diffusion models (2023), https://arxiv.org/abs/2302.07944
32. Wallace, B., Gokul, A., Ermon, S., Naik, N.: End-to-end diffusion latent optimization improves classifier guidance (2023)
33. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: MICCAI 2023
34. Yu, X., Li, G., Lou, W., Liu, S., Wan, X., Chen, Y., Li, H.: Diffusion-based data augmentation for nuclei image segmentation. In: MICCAI 2023
35. Zhou, S., Greenspan, H., Davatzikos, C., Duncan, J., Van Ginneken, B., Madabhushi, A., Prince, J., Rueckert, D., Summers, R.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE (2021)

Table 3: **Comparing OOD scores:** Full version of Table 1 with more OOD scores and their tail-specific version. $P(tail)$ achieves the best performance.

Score	FPR	AUC
P(tail)	$\mathbf{25.03}^{\pm 2.16}$	$\mathbf{95.01}^{\pm 0.48}$
MSP [12] head	$30.98^{\pm 2.36}$	$93.85^{\pm 0.58}$
MSP [12] tail	$25.33^{\pm 1.91}$	$94.98^{\pm 0.47}$
Energy [21] head	$39.17^{\pm 5.30}$	$91.79^{\pm 0.51}$
Energy [21] tail	$89.44^{\pm 6.42}$	$63.90^{\pm 2.54}$
MaxLogit [11] head	$39.02^{\pm 5.27}$	$91.79^{\pm 0.50}$
MaxLogit [11] tail	$89.37^{\pm 6.47}$	$63.77^{\pm 2.49}$
Maha dist. [19] to head	$53.41^{\pm 3.17}$	$82.58^{\pm 1.52}$
Maha dist. [19] to tail	$98.46^{\pm 0.63}$	$23.42^{\pm 1.55}$
Maha dist. [19] to tail - to head	$34.72^{\pm 5.11}$	$89.23^{\pm 1.28}$
relative Maha distance [26]	$32.45^{\pm 4.06}$	$91.88^{\pm 0.75}$

Table 4: **LoRA training details.**

rank	4
learning rate	1e-04
train steps	15000
lr scheduler	cosine
resolution	512
augmentation	center-crop and random-flip

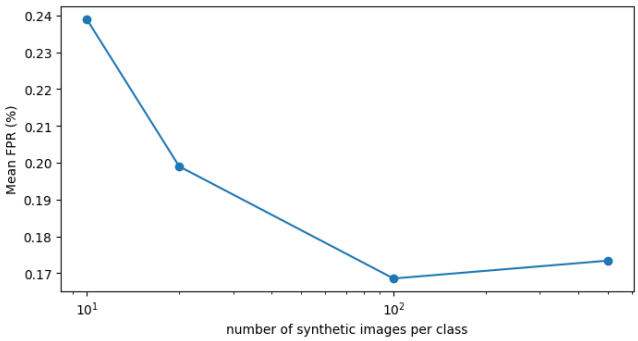

Fig. 3: Ablation on the number of synthetic images per tail class: Adding more than 100 samples leads to a slight increase in FPR, but still outperforms baselines. We hypothesize that the relative importance of the natural tail samples decreases when too many synthetic images are used.

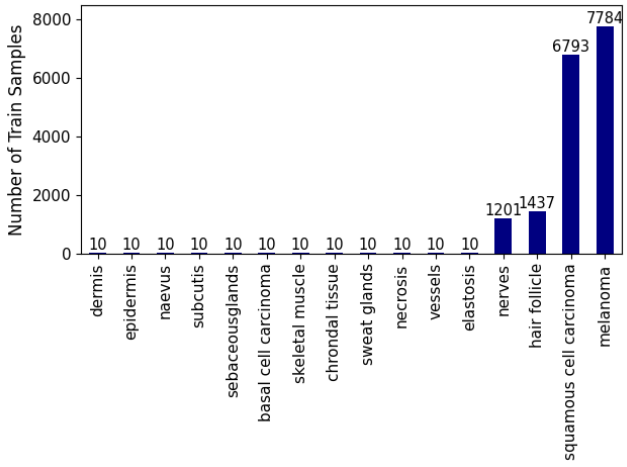

Fig. 4: Class distribution of the train dataset.

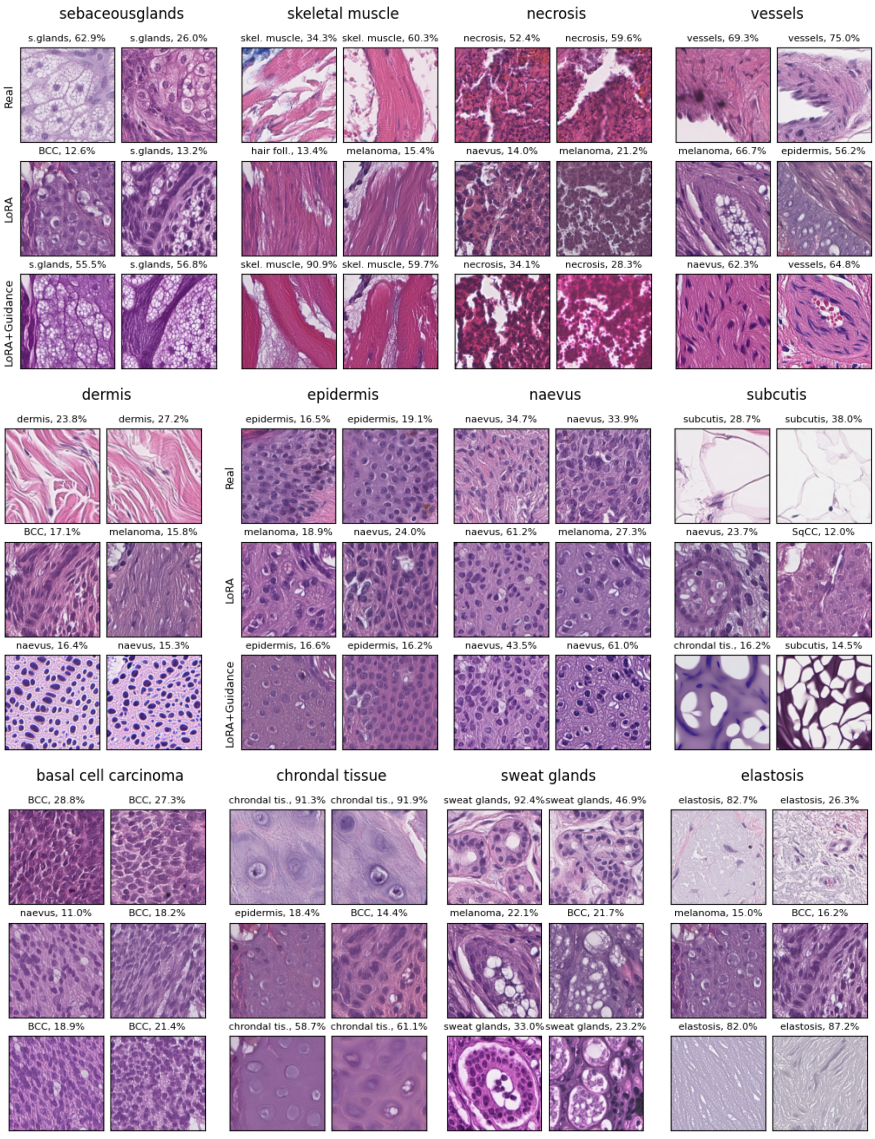

Fig. 5: **Samples from all classes:** We show samples from the train dataset (first row), samples generated with LoRA (second row), and generated with `LoGex` (third row). For each sample, we report the corresponding prediction of a classifier trained on the original dataset from [18] (achieving a tail accuracy of 96.1%). With `LoGex` the synthetic images are more often classified as the desired class.