Class-Conditioned Image Synthesis with Diffusion for Imbalanced Diabetic Retinopathy Grading

Haochen Zhang[1], Anna Heinke[2], Ines D. Nagel[2], Dirk-Uwe G. Bartsch[2], William R. Freeman[2], Truong Q. Nguyen[1], and Cheolhong An[1]

[1] Electrical and Computer Engineering Department, UC San Diego, USA
{haz035, tqn001, chan}@ucsd.edu
[2] Jacobs Retina Center, Shiley Eye Institute, UC San Diego, USA
{annaheinke, Ines.nagel5}@gmail.com
{dbartsch, wrfreeman}@health.ucsd.edu

Abstract. Diabetic retinopathy (DR) is a major cause of vision impairment, with early detection playing a crucial role in preventing irreversible blindness. While deep learning-based automated DR grading has improved diagnostic efficiency, class imbalance in public datasets hinders reliable performance evaluation, particularly for underrepresented DR stages. Current state-of-the-art classifiers achieve high overall accuracy but suffer from poor balanced accuracy, limiting their real-world applicability. Inspired by recent advancements in diffusion models, we propose to mitigate class imbalance by generating synthetic fundus images. Unlike prior methods prioritizing visual quality, we introduce a semantic quality metric based on classifier-predicted likelihood to selectively filter synthetic samples that enhance classification performance. Furthermore, we incorporate explicit class constraint during diffusion model finetuning to generate more semantically relevant data. Experimental results demonstrate a significant improvement in balanced classification accuracy from 66.84% to 74.20%, highlighting the effectiveness of our approach in improving DR diagnosis. Our code is available at: `https://github.com/AlanZhang1995/ECC_DM_for_DR.git`.

Keywords: Data Imbalance · Data Synthesis · Diabetic Retinopathy Grading · Diffusion model · Semantic Quality

1 Introduction

Diabetic retinopathy (DR) is a leading cause of blindness and visual impairment, particularly among the diabetic population [4]. It is characterized by retinal lesions like microaneurysms, hemorrhages, and exudates. DR progresses through five stages: no DR, mild, moderate, severe, and proliferative DR, with increasing severity raising the risk of vision loss [28]. As prolonged pathological states can block retinal blood vessels, early detection and grading are crucial to prevent irreversible blindness. The DR diagnostic standard has been well established on colored fundus imaging, but the shortage of ophthalmologists and the growing diabetic population strain healthcare systems. Consequently, computer-assisted tools [6,19], including deep learning-based DR grading systems [7,10,34], are becoming vital to improve screening efficiency, enhance diagnostic accuracy, and ease the burden on healthcare professionals.

Recent advancements in convolutional neural networks (CNNs) have greatly improved automated DR detection. Early models like CABNet [7] used category and global attention mechanisms for better small lesion detection, while later models, such as weakly-supervised lesion-aware networks [10] and transformer-guided attention networks [34], further refined performance. However, the effectiveness of these models is heavily influenced by dataset characteristics, particularly the data distribution along different DR stages [34].

Public datasets like DDR [13] usually exhibit a severe class imbalance, with approximately half images classified as no DR, while pathological images, particularly mild, severe, and proliferative DR, are underrepresented. This imbalance leads to poor recognition performance for underrepresented categories, as observed in the confusion matrix reported in Figure 7 of [34]. Despite achieving an accuracy of 83.1%, the state-of-the-art (SOTA) classifier fails in real-world applications due to frequent misclassification of mild and severe DR cases. This limitation underscores the need for an alternative evaluation metric—balanced accuracy [5], which averages accuracy across all five categories. Notably, this SOTA classifier achieves only 63.4% in balanced accuracy, highlighting the urgency of addressing class imbalance for reliable DR diagnosis.

Inspired by recent advancements in diffusion models [22,20,29] for data synthesis in computer vision tasks [32,27,33], we adapt these models to generate synthetic data samples to mitigate class imbalance in DR datasets. Unlike most existing image generation methods focused on visual quality, our primary aim is to improve classification accuracy. Previous research [15] has shown a trade-off between visual and semantic quality in image restoration. Experiments in [32] also shows that it is not necessary that more realistic synthetic data is more effective for classifier training. Therefore rather than relying on visual quality oriented metrics like FID [8] or IS [25], we introduce an alternative by using the likelihood predicted by a group of pretrained classifiers as a semantic quality metric. This allows us to selectively choose synthetic samples that further benefit the classification performance. Additionally, we proposed to incorporate explicit class constraint [35] during the diffusion model's finetuning process, guiding the model to generate samples with higher semantic relevance. Experimental results show that the semantic-oriented metric effectively filters the synthetic data and our proposed finetuning strategy achieves an improvement in balanced classification accuracy from 66.84% to 74.20%.

2 Related Work

Deep learning based DR grading. Deep learning has been widely explored for DR grading. Li et al. [14] leveraged pre-trained CNNs with transfer learning, using the final fully-connected layer for feature extraction and an SVM for classification. Yang et al. [31] proposed a two-stage CNN that first detects lesions before grading DR, achieving promising results. Recent advancements focus on class balance, lesion detection, and feature representation. CABNet [7] employs category and global attention blocks to mitigate class imbalance and enhance small lesion detection. A weakly supervised lesion-aware network [10] utilizes lesion activation maps to improve feature discrimination without pixel-level annotations. Triple-DRNet [12] differentiates DR types and severity levels, while CRA-Net [34] refines class-specific feature extraction and lesion relation modeling using category-relation attention. SVPL [36] introduces a prompt-based strategy for knowledge transfer from large pre-trained models. However, many methods evaluate performance on imbalanced test sets, leading to biased metrics and limiting real-world applicability despite high overall accuracy.

Diffusion model and data synthesis. Diffusion models, originating from DDPM [9], have made remarkable advancements in text-to-image generation, exemplified by Stable Diffusion [20,22], DALL-E [21], and Imagen [3]. These models leverage iterative denoising to generate high-quality visuals from textual prompts. Motivated by their generative power, researchers have explored their potential for synthetic data generation across various domains. A significant body of work focuses on utilizing synthetic data to enhance image segmentation [17,30], while other studies investigate its impact on image classification [32,2] and object detection [27]. The medical imaging field has also benefited from this approach. For instance, Sagers et al. [24] demonstrated that latent diffusion models can generate synthetic skin disease images to improve model performance in data-limited scenarios. Similarly, Akrout et al. [1] employed text-to-image diffusion models to produce high-quality synthetic skin disease images, enhancing training datasets. Oh et al. [18] introduced a diffusion-based data synthesis technique to address class imbalance in pathology datasets, improving both segmentation and classification outcomes. Existing methods target on data imbalance, but overlook the semantic quality of generated samples. In contrast, our approach ensures both diversity and semantic integrity, enhancing their effectiveness in downstream classification tasks.

3 Proposed Method

3.1 Text-to-image diffusion model finetuning on fundus images

Diffusion models [9,22] iteratively denoise a noisy sample to approximate a target distribution, reversing a predefined Markov process. In text-to-image generation, a trained diffusion model f_θ generates an image $x_{\text{sync}} = f_\theta(\epsilon, c)$ from Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and a conditioning vector $c = \Phi(T)$, derived from a text prompt T via an encoder Φ. The model is trained using a mean squared error (MSE) loss to denoise a perturbed input $z_t = \alpha_t x + \sigma_t \epsilon$, where α_t, σ_t define the noise schedule over time $t \sim U([0, 1])$. By leveraging reweighted variational bounds and convolution-based UNet architectures, these models efficiently reconstruct clean images and achieve high-perception generation. To finetune the model for our DR application, we use a category-conditioned prompt: “`A color fundus image`

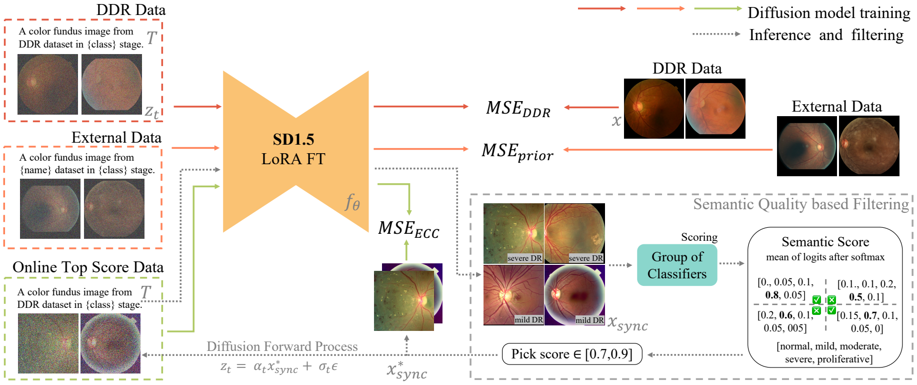


Fig. 1. Overview of our text-to-image diffusion model finetuning framework. The top section illustrates the process of stabilizing diffusion model finetuning on fundus dataset with limited data, as described in Section 3.1. The bottom right depicts our proposed semantic quality evaluation and filtering pipeline, corresponding to Section 3.2. The entire image demonstrates how semantic quality is enhanced through explicit class conditioning during diffusion model training, as discussed in Section 3.3.

`in {category} stage`" where `category` ∈ { `normal, mild DR, moderate DR, severe DR, proliferative DR` }.

Finetuning diffusion models typically requires large-scale datasets, which are often unavailable in medical imaging. To overcome this, we use the DreamBooth framework [23] combined with the LoRA [11] finetuning strategy. DreamBooth enables effective training with just 3–5 images by introducing a class-specific prior preservation loss, which helps retain prior knowledge and ensures class-specific characteristics, reducing overfitting and preventing output degradation.

In our case as illustrated in top section of Fig.1, fundus images in the DDR dataset [13] serve as subject-specific training samples MSE_{DDR}, while data for the prior-preservation term MSE_{prior} can be sourced from any retina images. Considering that fundus is the most common retina image modality, in our experiments, we incorporate fundus images from external datasets such as EyePACS[3] and APTOS[4] for prior-preservation loss computation. Specifically, we construct prompts in the format: "`A color fundus image from {name} dataset in {category} stage`", where `name` ∈ { `DDR, EyePACS, APTOS` }. This integration of additional datasets ensures that the model learns particular features in DDR dataset while maintaining fundus image knowledge.

3.2 Semantic quality based evaluation and filtering

With the training strategy detailed in last section, we finetune a diffusion model specifically for fundus image generation within the DDR dataset domain. How-

[3] `https://www.kaggle.com/c/diabetic-retinopathy-detection/`

[4] `https://www.kaggle.com/competitions/aptos2019-blindness-detection`

ever as evidenced by our results in Table 1, it is suboptimal in enhancing classification performance by simply mixing synthetic samples with real ones to form an augmented training set. Prior research [15] has established that, in image restoration area, optimal classification performance is not achievable when images undergo restoration with distortions or perceptual losses alone. Additionally, previous experiments [32] suggest that more realistic synthetic data does not necessarily lead to better classifier training. Given these observations, we explore alternative metrics to assess the semantic quality of generated images, rather than relying solely on visual quality.

An ideal approach to measure semantic quality would involve real-time ophthalmologist feedback on each generated sample, but this is infeasible due to high costs. Inspired by prior research [16,35] that utilized pretrained classifiers for model optimization, we propose using a group of pretrained classifiers as diagnostic agents to evaluate the usefulness of generated images for DR grading. There are no strict constraints on the architecture or training method of these classifiers, but diversity intuitively improves performance. For practical implementation, we employ the LANet [10] family, a SOTA open-source DR grader for the DDR dataset, as our expert group. Since LANet has been trained on multiple CNN backbones, including VGG, DenseNet, and InceptionNet, it offers sufficient diversity to serve as a robust evaluation. We select the k models with top performance as scorers to measure the semantic quality of generated images.

With pretrained classifiers $\{g_i\}_{i=1}^{k}$ as scorers, we compute the semantic score for each synthetic sample x_{sync} based on its prompt-specified category j. Specifically, we extract the logits predicted by each classifier $\{\text{logit}_i\}_{i=1}^{k} \in \mathbb{R}^{5\times 1}$, apply softmax normalization, compute the element-wise average and then take the jth element: $\text{score}_j = \frac{1}{k}\sum_{i=1}^{k} \text{softmax}(\text{logit}_i)[j]$. We then remove synthetic samples outside the range [0.7, 0.9] which was decided by a grid search. This threshold matches our expectation as low-scoring samples degrade classifier performance, while perfect-scoring samples provide little new information for training.

3.3 Finetuning diffusion model with explicit class condition

In this section, we further explore method refines diffusion models to generate samples with high semantic quality by incorporating explicit class condition (ECC). Prior approaches [16,35] integrate classification loss into the optimization objective. However, this approach is not suitable for our method. The reason is that we adopt the latent diffusion framework, generating images as VAE latents, for which no existing DR classifier has been trained. Thus in this paper, we implement explicit class condition through an iterative self-supervised filtering and finetuning process inspired by DreamSync [26].

As illustrated as the gray dash arrows in Fig. 1, starting from a text prompt T specifying class, the model f_θ generates a set of images $\{x_{\text{sync}}^{(k)}\}_{k=1}^{N}$, where $N = 256$ is empirically determined for stable finetuning. Then we utilize same evaluation strategy and filter threshold detailed in Section 3.2 to select samples with high semantic quality denoted as $\{x_{\text{sync}}^{*}\}$. High-quality image-text pairs

(T, x^*_{sync}) are then iteratively used to finetune the diffusion model via a LoRA-based optimization strategy for computational efficiency. The refinement process is governed by minimizing additional reconstruction loss over the filtered top score samples (green arrow), leading to the final optimization objective:

$$L_{\text{total}} = \text{MSE}_{DDR} + \alpha\text{MSE}_{prior} + \beta\text{MSE}_{ECC}$$

where MSE_{DDR} and MSE_{prior} ensure stable finetuning on limited fundus data as detailed in Section 3.1, while MSE_{ECC} introduces an explicit class condition to enhance semantic quality. The loss weights α and β balance the contributions of these terms for optimal performance.

4 Experimental Result

4.1 Implement details

Datasets This paper aims to improve the balanced accuracy of CNN classifiers on the DDR dataset [13], while using EyePACS and APTOS datasets for prior loss computation. The DDR dataset contains 12,522 fundus images for five-class DR grading, split into 6,260 training, 2,503 validation, and 3,759 test images, with an imbalanced class distribution. For example in the training set, samples in DR-0 to DR-4 stages are 3133, 315, 2238, 118, and 456, respectively. To address this, we use data synthesis in the training set and evaluate performance using balanced accuracy to avoid class distribution bias in the test set. In detail, synthetic samples were added only to the mild and severe DR classes. At the 4K setting in Fig. 2(a), 2K samples were added to each, resulting in a train set of 3133, 2315, 2238, 2118, 456. For validation, we randomly selected 47 samples from each class in the official valid split to ensure balance.

Implementation We implement the diffusion model and classifiers using PyTorch. The diffusion model is based on Stable Diffusion 1.5 (SD1.5) from Hugging Face Diffusers library, finetuned on 512×512 fundus images with LoRA (rank 8). To enhance data diversity, we apply center cropping and random flipping, using effective batch size 96. The model is trained with a constant learning rate of 1e-5, initializing β=0 for the first 500 steps and setting β=0.001 thereafter. α is set as 0.1. For classifier training, we follow the pipeline from [10], training Inception-v3, DenseNet-121, and VGG-16 for 300, 200, and 100 epochs, respectively, with a batch size of 32 and a learning rate of 1e-3 using polynomial decay (p=9). The best-performing model on validation is selected for final evaluation.

4.2 Experimental Results

To show the effectiveness of our proposed diffusion model training and data filtering strategy, we conduct experiment based on SOTA DR grader, LANet [10], on DDR dataset. We implement their classifiers based on author's open-source

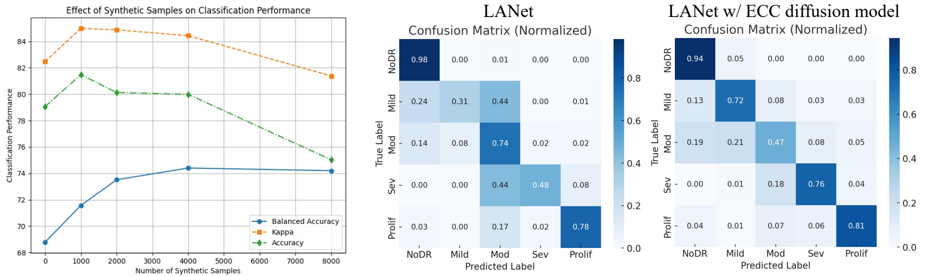
Table 1. Classification performance of LANet with different backbones using the proposed data synthesis on the DDR dataset. "B. Acc" is short for balanced accuracy [5]. Models marked with * are trained by the LANet authors. Basic DM and ECC DM refer to diffusion models trained as described in Sections 3.1 and 3.3, respectively. w/ filter indicates the application of the filtering strategy detailed in Section 3.2. We keep the total number of synthetic training samples same before and after filtering, achieved through random selection.

Backbone	VGG-16			Inception-v3			DenseNet-121		
Metric	B. Acc	Kappa	Acc	B. Acc	Kappa	Acc	B. Acc	Kappa	Acc
LANet*	65.92	**86.41**	**83.83**	67.80	**85.92**	**83.88**	64.24	84.20	82.50
LANet	66.84	85.06	83.67	65.97	84.47	81.70	62.14	**84.26**	**83.08**
+ oversampling	68.78	82.47	79.06	66.46	84.55	81.28	64.79	83.66	80.69
+ basic DM w/o filter	71.37	81.94	76.99	69.01	82.53	78.27	67.19	83.35	78.8
+ basic DM w/ filter	73.79	84.24	77.92	71.33	84.18	79.49	69.47	82.32	76.59
+ ECC DM w/o filter	**74.20**	81.38	75.02	**71.95**	83.76	76.67	**70.22**	83.76	79.76
+ ECC DM w/ filter	73.82	82.58	76.80	70.38	84.74	77.57	69.20	83.82	79.28

codes and report performance of both our implementation and author's models in Table 1. We also include the performance of oversampling (based on normed inverse class frequency) as baseline solution for data imbalance. We focus on balanced accuracy (B. Acc) but we also provide unbalanced accuracy (Acc) and quadratic weighted kappa [10] for reference.

Main results. Table 1 demonstrates that oversampling enhances balanced accuracy to some extent, with a more significant improvement when using diffusion model-generated data. Among the synthetic data settings, the basic diffusion model performs the worst, underscoring the effectiveness of our proposed semantic quality-based evaluation and filtering strategy. Furthermore, finetuning the diffusion model with explicit class condition enables the generation of high-semantic-quality samples, achieving similar accuracy comparable to the basic model with filtering. This validates the efficacy of our ECC finetuning strategy. An interesting observation is that with ECC, synthetic data without filtering slightly outperforms the filtered version. This suggests that a reasonable number of samples with scores outside the range [0.7, 0.9] may contribute to classifier learning. Thus, an implicit semantic quality control via the diffusion model appears more beneficial than a hard threshold-based filtering approach. The confusion matrix in Fig. 2 (b) illustrates how synthetic data improves balanced accuracy: While this comes at the cost of reduced accuracy for moderate DR cases, the classifier exhibits improved recognition of mild and severe DR, maximizing values along the main diagonal.

Performance v.s. data amount. Fig. 2 (a) further examines the impact of synthetic data volume on model performance. Notably, i) while balanced accuracy improves, overall accuracy declines—a trend consistent with Table 1. This phenomenon aligns with expectations: the model initially learns from an imbalanced dataset, yielding high overall accuracy on a similarly skewed test set. Incorporating synthetic data for underrepresented classes mitigates this imbalance,



(a) Performance w.r.t synthetic fundus amount

(b) Confusion Matrix

Fig. 2. (a) Performance variation w.r.t the amount of synthetic data. (b) Comparison of confusion matrices without and with high-semantic-quality synthetic samples.

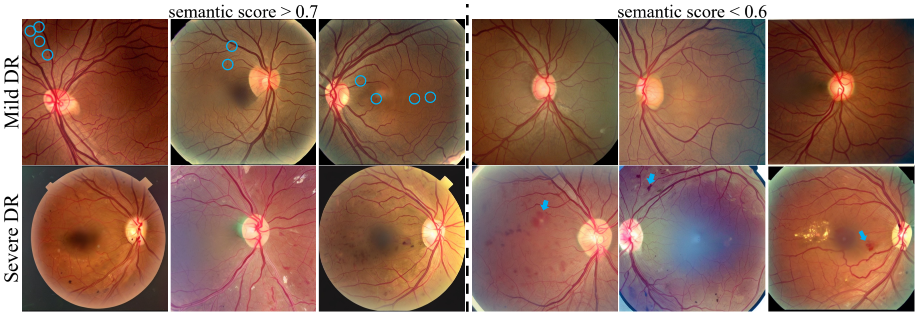


Fig. 3. Visual inspection of synthetic samples.

enhancing balanced accuracy at the cost of reduced overall accuracy due to the diminishing advantage of the imbalanced class distribution. When the dataset becomes fully balanced, the two accuracy curves ultimately converge, producing an unbiased classifier. ii) The initial simultaneous increase in all three performance metrics suggests that introducing a limited amount of synthetic data into underrepresented classes does not immediately disrupt the imbalance prior of training dataset. Instead, it expands the training set's diversity while preserving key statistical properties, leading gain across all three measures. However, iii) the benefits of synthetic data diminish over time, likely due to the limited diversity of generated samples. Since real data remains fundamental to generalization, an excessive amount of synthetic data may dilute its impact, constraining further improvement.

Visual inspection. Fig. 3 presents synthetic DDR samples generated by our diffusion model. In the first row, the samples on the left achieve a higher semantic score and display microaneurysms—key diagnostic features of mild DR. In contrast, the samples on the right either lack visible microaneurysms or suffer from poor image quality, making disease assessment challenging. Similarly, in the second row, the left-side samples demonstrate characteristic features of severe non-proliferative DR, including diffuse retinal hemorrhages, microaneurysms across all four quadrants, and venous beading in at least two quadrants. In comparison, the right-side samples are more indicative of proliferative DR, as they show signs of neovascularization with intra- or subretinal hemorrhage.

5 Conclusion

In this study, we employed a diffusion model for fundus image synthesis to mitigate training data imbalance. We adopted the DreamBooth framework and introduced a semantic quality metric based on an ensemble of pretrained classifiers. Using both for diffusion model training, we ensured high-semantic-quality fundus image generation within the DDR domain. Experimental results confirm that our approach improves balanced accuracy in SOTA classifiers, enhancing the practical application of automated DR diagnosis. Future work will explore dataset expansion with synthetic data and iterative optimization of the ECC diffusion model and classifier group used for semantic quality assessment. We will also investigate diversity-aware sampling together with semantic quality.

Acknowledgments. This study was funded by the NIH grant (5R01EY033847).

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

References

1. Akrout, M., Gyepesi, B., Holló, P., Poór, A., Kincső, B., Solis, S., Cirone, K., Kawahara, J., Slade, D., Abid, L., et al.: Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In: MICCAI. pp. 99–109 (2023)
2. Azizi, S., Kornblith, S., Saharia, C., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023)
3. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Chan, K., Chen, Y., Dieleman, S., Du, Y., Eaton-Rosen, Z., et al.: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
4. Cho, N.H., Shaw, J.E., K., S., Huang, Y., da R. F., J.D., Ohlrogge, A., Malanda, B.: IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice **138**, 271–281 (2018)
5. Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756 (2020)
6. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama **316**(22), 2402–2410 (2016)
7. He, A., Li, T., Li, N., Wang, K., Fu, H.: Cabnet: Category attention block for imbalanced diabetic retinopathy grading. TMI **40**(1), 143–153 (2020)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017)

9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
10. Hou, J., X., F., Xu, J., Feng, R., Z., Y., Zou, H., Lu, L., Xue, W.: Diabetic retinopathy grading with weakly-supervised lesion priors. In: ICASSP. pp. 1–5 (2023)
11. 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. arXiv preprint arXiv:2106.09685 (2021)
12. Jian, M., Chen, H., Tao, C., Li, X., Wang, G.: Triple-DRNet: A triple-cascade convolution neural network for diabetic retinopathy grading using fundus images. Computers in Biology and Medicine **155**, 106631 (2023)
13. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., K., H.: Diagnostic assessment of deep learning algorithms for DR screening. Information Sciences **501**, 511 – 522 (2019)
14. Li, X., Pang, T., Xiong, B., Liu, W., Liang, P., Wang, T.: Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In: CISP-BMEI. pp. 1–11 (2017)
15. Liu, D., Zhang, H., Xiong, Z.: On the classification-distortion-perception tradeoff. NeurIPS **32**, 1–10 (2019)
16. Liu, Z., Wang, H., Zhou, T., Shen, Z., Kang, B., Shelhamer, E., Darrell, T.: Exploring simple and transferable recognition-aware image processing. TPAMI **45**(3), 3032–3046 (2022)
17. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. NeurIPS **36**, 76872–76892 (2023)
18. Oh, H.J., Jeong, W.K.: Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In: MICCAI. pp. 337–345 (2023)
19. Pekel Özmen, E., Özcan, T.: Diagnosis of diabetes mellitus using artificial neural network and classification and regression tree optimized with genetic algorithm. Journal of Forecasting **39**(4), 661–670 (2020)
20. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
21. Ramesh, A., D., P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
23. 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: CVPR. pp. 22500–22510 (2023)
24. Sagers, L.W., Diao, J.A., Luke, M., Groh, M., Rajpurkar, P., Adamson, A.S., R., V., Daneshjou, R., Manrai, A.K.: Augmenting medical image classifiers with synthetic data from latent diffusion models. arXiv preprint arXiv:2308.12453 (2023)
25. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. NeurIPS **29** (2016)
26. Sun, J., Fu, D., Hu, Y., Wang, S., Rassin, R., Juan, D.C., Alon, D., Herrmann, C., van Steenkiste, S., Krishna, R., et al.: Dreamsync: Aligning text-to-image generation with image understanding feedback. In: CVPR workshop (2024)
27. Voetman, R., Aghaei, M., Dijkstra, K.: The big data myth: Using diffusion models for dataset generation to train deep detection models. arXiv preprint arXiv:2306.09762 (2023)

28. Wilkinson, C.P., Ferris III, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T., et al.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**(9), 1677–1682 (2003)
29. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Wang, S., Huang, T., Liu, Z.: OmniGen: Unified image generation. arXiv preprint arXiv:2409.11340 (2024)
30. Xie, J., Li, W., Li, X., Liu, Z., Loy, C.C.: Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. IJCV pp. 1–20 (2024)
31. Yang, Y., Li, T., Li, W., Wu, H., Fan, W., Zhang, W.: Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: MICCAI. pp. 533–540 (2017)
32. Ye-Bin, M., Hyeon-Woo, N., Choi, W., Kim, N., Kwak, S., Oh, T.H.: Exploiting synthetic data for data imbalance problems: Baselines from a data perspective. arXiv preprint arXiv:2308.00994 (2023)
33. Yu, X., Li, G., Lou, W., Liu, S., Wan, X., Chen, Y., Li, H.: Diffusion-based data augmentation for nuclei image segmentation. In: MICCAI. pp. 592–602 (2023)
34. Zang, F., Ma, H.: CRA-Net: Transformer guided category-relation attention network for DR grading. Computers in Biology and Medicine **170**, 107993 (2024)
35. Zhang, H., Heinke, A., B., K., Galang, C.M.B., Deussen, D.N., Nagel, I.D., M., K., Bartsch, D.U.G., Freeman, W.R., Nguyen, T.Q., et al.: Octa-based amd stage grading enhancement via class-conditioned style transfer. In: EMBC. pp. 1–5 (2024)
36. Zhang, Y., Ma, X., Huang, K., Li, M., Heng, P.A.: Semantic-oriented visual prompt learning for diabetic retinopathy grading on fundus images. TMI (2024)