ODIR Dataset Project Technique Report

Kaijie Xu

August 20, 2025

Abstract

This study focuses on fundus image analysis and is conducted in two phases: (i) multi-class classification modeling; (ii) conditional image generation and evaluation. The report covers data preprocessing and augmentation, model architectures and training configurations, evaluation metrics, visualization, and failure case analysis. All reproduction experiments and summaries were completed within a 14-day period. Source code can be found at: `https://github.com/Xu739/ODIR-Dataset-Classification-Generation-and-Self-Supervised-Learning`

Contents

1 Project Overview and Data

1.1 Task Definition

We designed a two-phase experimental pipeline for eight classes of fundus diseases:

1. **Multi-class Classification (Phase I)**
   Construct a multi-class classification model to predict the disease category for each fundus image. Evaluation metrics include Accuracy (ACC), macro-averaged Precision ($\text{Prec}_{\text{macro}}$), Recall ($\text{Rec}_{\text{macro}}$), F1-score ($\text{F1}_{\text{macro}}$), and Area Under the ROC Curve ($\text{AUC}_{\text{macro}}$). Confusion matrices are also generated to analyze prediction biases across classes.

2. **Conditional Generation (Phase II)**
   Use conditional generative models (e.g., cGAN[9] or diffusion models like DDPM[7]/DDIM[12]) to generate fundus images for specific classes. Evaluation metrics for generated images include Fréchet Inception Distance (FID)[5], Kernel Inception Distance (KID)[1], and Inception Score (IS)[11]. Additionally, Phase I classifiers are applied to the generated images to quantify semantic consistency.

1.2 Data and Splits

This project uses the ODIR-5K[10] fundus image dataset, which contains multi-center color retinal images (Figure 1) and corresponding labels. The raw data were processed as follows:

- **Cleaning and De-identification**: Matched images with CSV metadata, removed patient personal information and identifiable file names, retained unique IDs for training, and exported `Overall label.csv`.

- **Label Mapping**: Original labels were mapped to eight disease classes:

$$\text{N} \to 0, \quad \text{D} \to 1, \quad \text{G} \to 2, \quad \text{C} \to 3, \quad \text{A} \to 4, \quad \text{H} \to 5, \quad \text{M} \to 6, \quad \text{O} \to 7$$

To ensure scientific data splits and model generalization, the dataset was divided into training, validation, and test sets at approximately 60%/20%/20%. The test set is used only once in the final reporting stage to prevent information leakage.

The number and proportion of samples in each class within the training set are shown in Table 1, based on `Overall label.csv`. The distribution is highly imbalanced, with some minority classes having very few samples.

To visualize class distribution, Figure 2 shows a bar chart of sample counts per class. Due to class imbalance, strategies such as data augmentation or weighted loss functions are necessary during training to improve model recognition of minority classes.

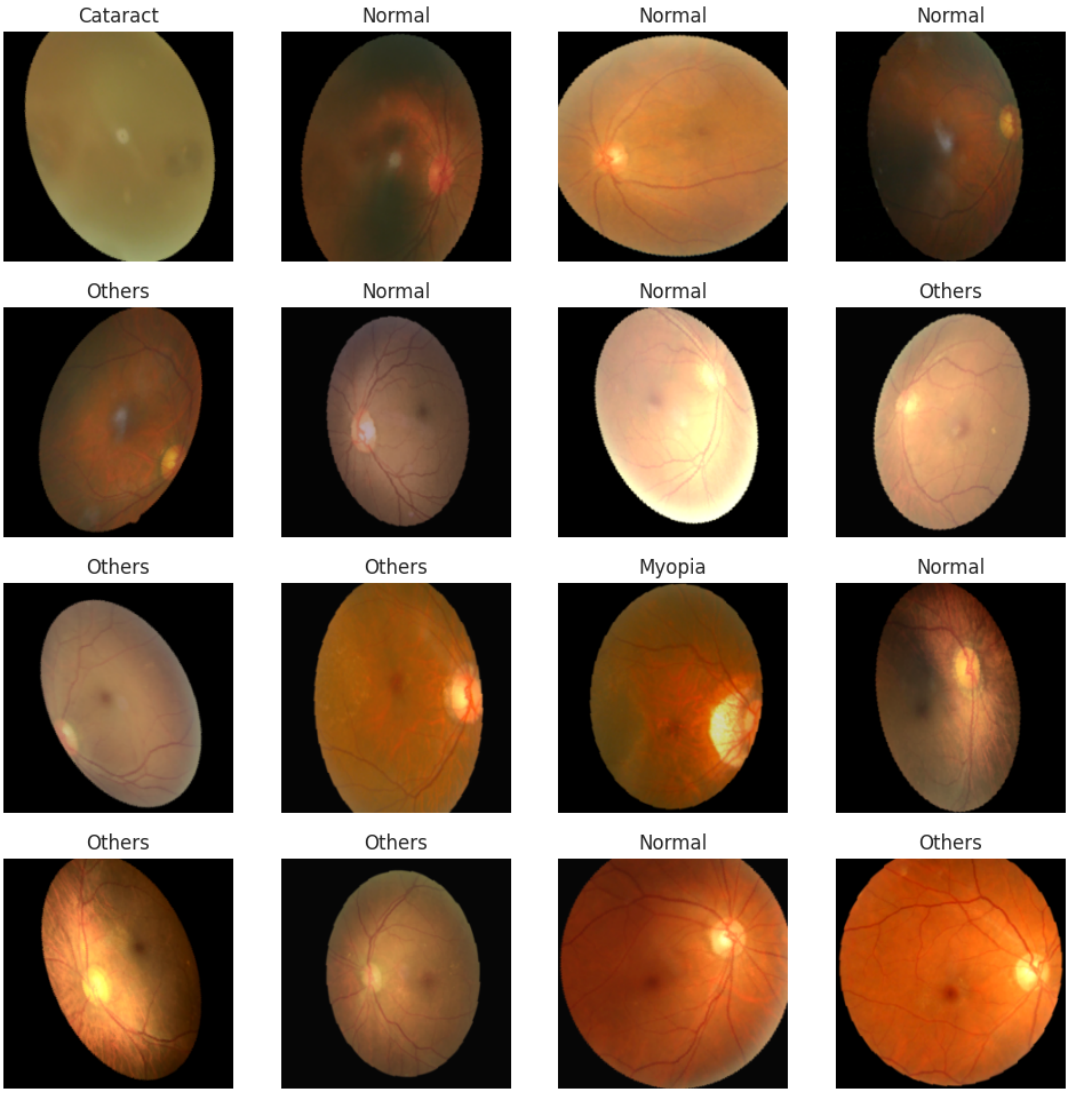


Figure 1: ODIR-5K color fundus images.

Table 1: Class distribution in the training set

Class	Number of Samples	Proportion (%)
Normal (N)	2873	44.95
Diabetes (D)	1608	25.16
Glaucoma (G)	284	4.44
Cataract (C)	293	4.58
AMD (A)	266	4.16
Hypertension (H)	128	2.00
Myopia (M)	232	3.63
Others (O)	708	11.08

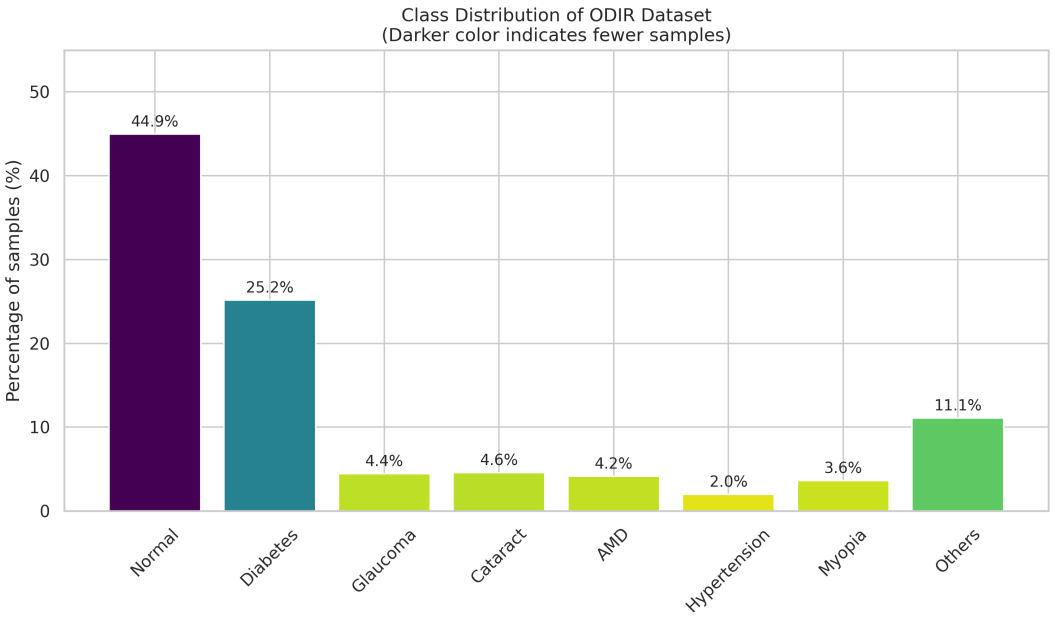


Figure 2: Bar chart of sample counts per class in the training set.

2 Phase I: Multi-class Classification of Fundus Images

2.1 Preprocessing and Methods

All training images were subjected to standardized preprocessing and augmentation to improve model generalization. The specific transformations include:

- Resize images to 256×256;
- Random rotation within $\pm 30°$;
- Random horizontal and vertical flips;
- Random brightness and contrast adjustment (magnitude 0.2);
- Random resized crop to 224×224;
- Conversion to Tensor and normalization to $[-1, 1]$.

The augmentation effects are illustrated in Figure 3. Additionally, a weighted loss function was applied. To further address issues of limited sample size and model underfitting, the following three strategies were adopted:

1. **Oversampling**: Over-sample minority classes to match the sample size of majority classes.
2. **RotNet[3] Pretrain**: Pretrain the network on a self-supervised rotation prediction task, then fine-tune on the ODIR[10] dataset.
3. **CIFAR-10[8] Pretrain**: Pretrain the network on the CIFAR-10[8] dataset, then fine-tune on the ODIR[10] dataset.

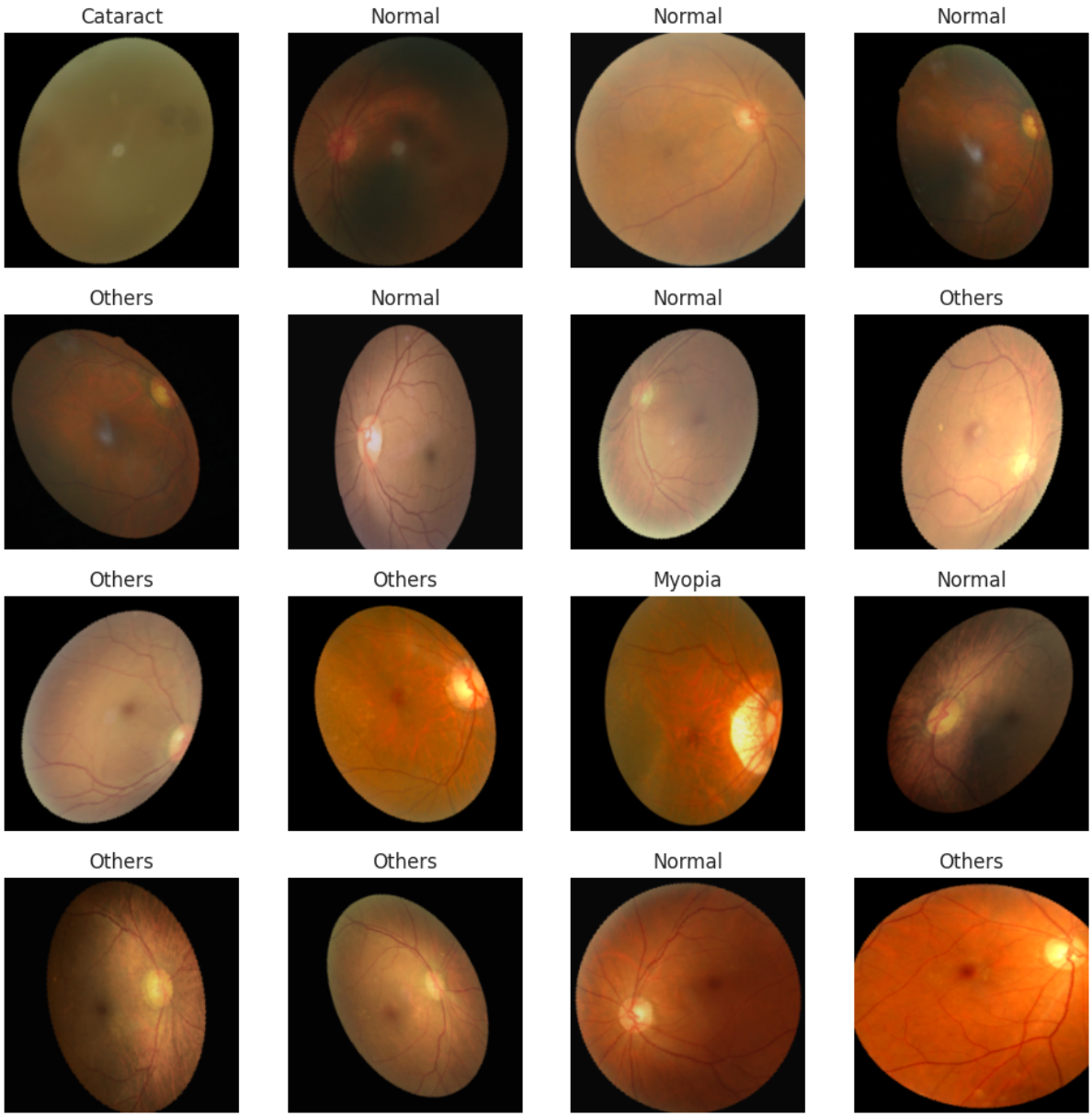


Figure 3: Data augmentation results in Stage 1 for fundus images: Compared with the original images, the augmented images exhibit more diverse ocular structures and vascular patterns, while preserving critical pathological regions (such as vessels and lesions). This enhances the dataset diversity and improves model performance.

Experiments were conducted on three network architectures: **ResNet50**[4], **EfficientNet-B3**[13, 14], and **ViT**[2]. Evaluation metrics include Accuracy (ACC), Precision (PRE), Recall (REC), F1-score (F1), and AUC.

All models were optimized using the Adam optimizer with a learning rate of 0.0001, MultistepLR scheduler with milestones at 50 and gamma=0.5, batch size of 32, weight decay of 1e-5, and early stopping patience of 50 epochs.

2.2 Experimental Results and Analysis

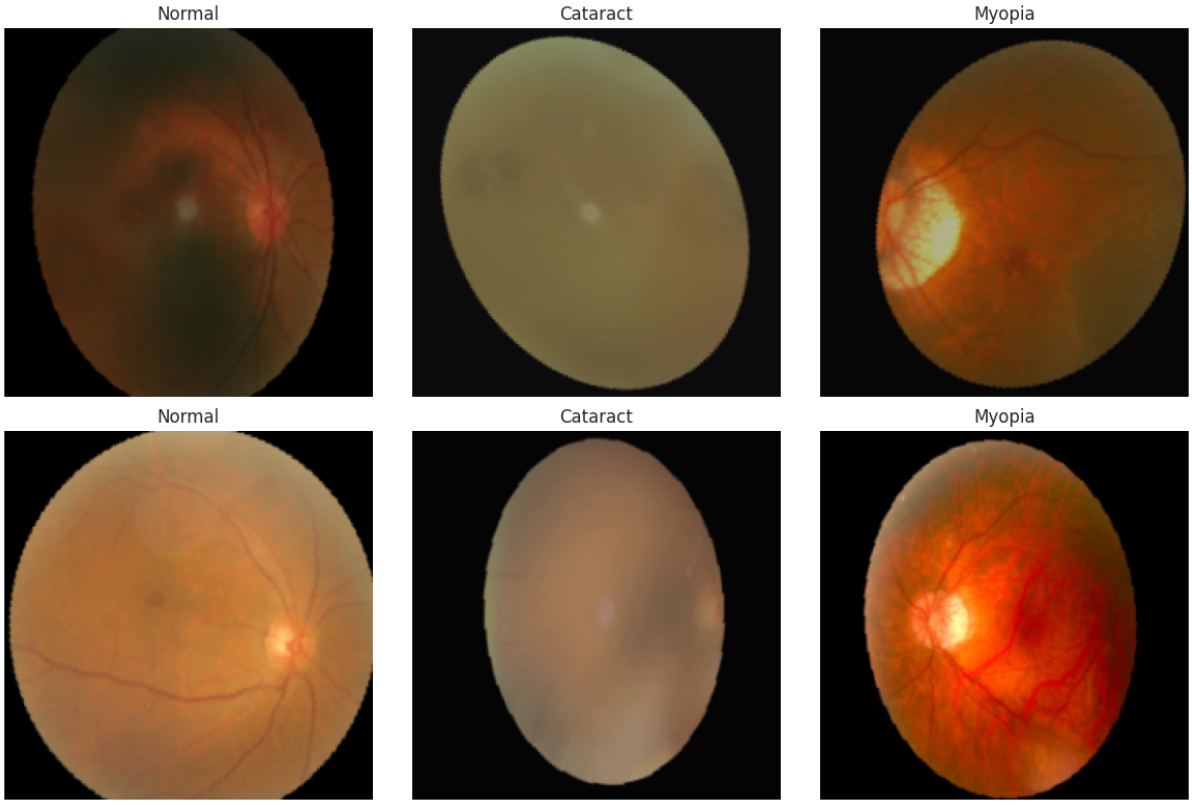
The results of nine experiments are summarized in Table 2, and the corresponding confusion matrices are presented in Appendix A.2. Overall, ResNet50[4] outperformed the other two architectures on the ODIR[10] dataset, while ViT[2] performed the worst. This may be due to ViT[2]'s large number of parameters (ViT[2]: 86M, compared to ResNet50[4]: 24M, EfficientNet-B3[13, 14]: 7M), which requires more data to fit effectively. EfficientNet-B3[13, 14], with fewer parameters, may struggle to capture features from 224×224 images. ResNet50[4] strikes a balance, being able to extract meaningful features while fitting the relatively small dataset, resulting in superior performance.

From the perspective of training strategies, CIFAR-10[8] pretraining generally outperformed the other two methods because it introduces external data, allowing the network to capture low-level image features before fine-tuning on ODIR[10]. RotNet[3], proposed by Spyros Gidaris in ICLR 2018, pretrains the network via rotation prediction in a self-supervised manner. However, on fundus images, its performance is slightly worse than oversampling because fundus features are predominantly circular and linear; rotation does not provide as much discriminative signal, unlike in natural images (e.g., cats, dogs) for which RotNet[3] was originally designed.

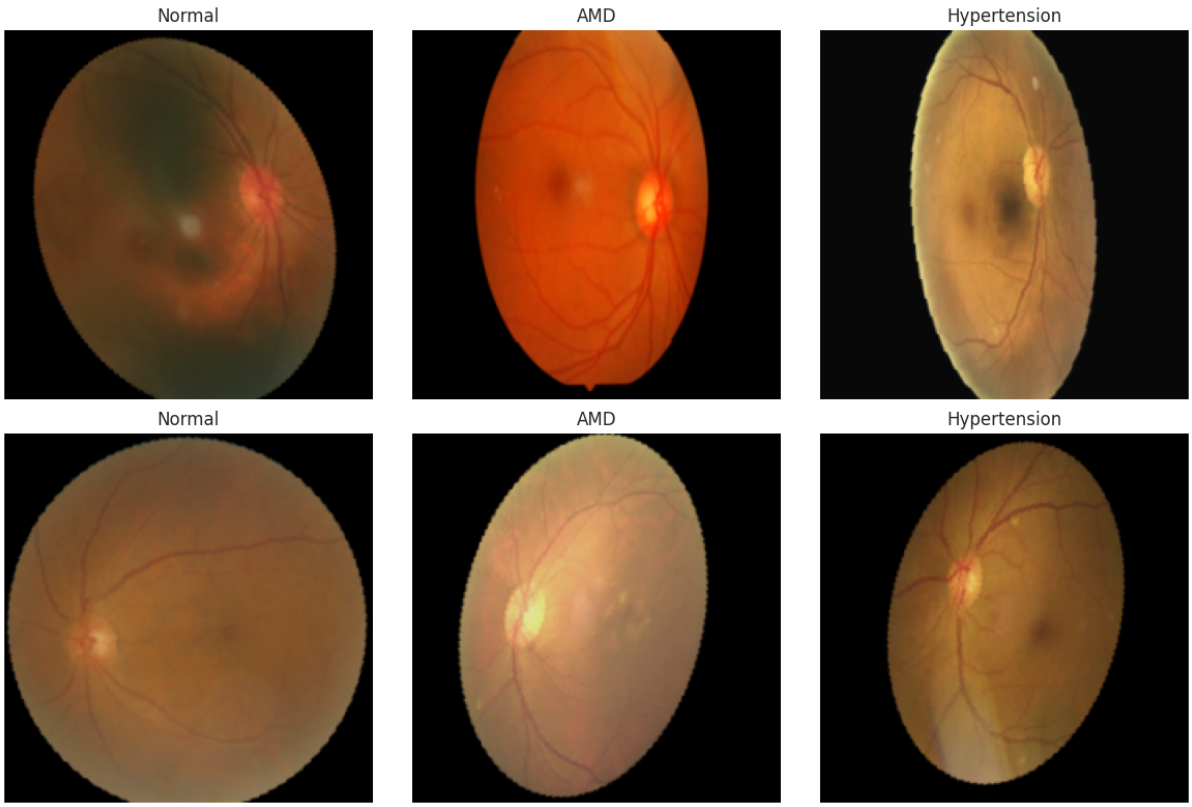
The confusion matrices indicate a bias towards predicting Normal images, likely due to severe class imbalance. Even with weighted cross-entropy loss, the imbalance effect persists. Models generally classify Normal, Cataract, and Myopia well, while other classes are less distinct. Figure 4 shows that AMD and Hypertension are often misclassified as Normal due to subtle visual differences. Future work could focus on improving classification accuracy for these challenging classes.

Table 2: Classification performance comparison across methods and networks

Method	**Network**	**ACC**	**PRE**	**REC**	**F1**	**AUC**
Oversampling	ResNet50	0.50	0.46	0.51	0.45	0.82
	EfficientNet-B3	0.50	0.46	0.40	0.41	0.76
	ViT	0.44	0.36	0.33	0.35	0.73
RotNet[3] Pretrain	ResNet50	0.51	0.46	0.43	0.43	0.79
	EfficientNet-B3	0.47	0.42	0.38	0.39	0.78
	ViT	0.34	0.14	0.24	0.14	0.66
CIFAR10 Pretrain	ResNet50	0.59	0.55	0.46	0.49	0.82
	EfficientNet-B3	0.52	0.48	0.47	0.47	0.80
	ViT	0.44	0.36	0.33	0.35	0.62



(a) Normal, Cataract, and Myopia classes. The differences among these three types are visually distinct.



(b) Normal, AMD, and Hypertension classes. The visual differences among these three types are subtle.

Figure 4: Comparison of fundus images: (a) Normal, Cataract, and Myopia; (b) Normal, AMD, and Hypertension.

2.3 Summary

From the above analysis, we can conclude:

- Among the three network architectures, ResNet50[4] achieved the best performance.
- Among the three training strategies, CIFAR-10[8] pretraining provided the best results.
- Further measures to address class imbalance are necessary, particularly for Normal, AMD, and Hypertension classes.

3 Stage II: Conditional Image Generation

3.1 Preprocessing and Methods

In this stage, we employed two types of class-conditional image generation models:

Diffusion Model: A conditional diffusion framework based on DDIM[12] sampling, incorporating **Classifier-Free Guidance (CFG)**[6] to control the strength of the conditioning. This allows the model to maintain diversity while improving the semantic consistency of generated images.

Adversarial Generative Model: A conditional GAN (cGAN)[9] structure that generates images corresponding to the input conditions in a directed manner.

The preprocessing pipeline for the generative models is largely consistent with that of the classification models, with the following differences:

- For training the diffusion model, images are resized to 128×128 due to GPU memory constraints.
- For training the adversarial model, images are resized to 256×256 as required by the network architecture.

Training details are as follows:

cGAN[9]: latent dimension = 100, loss function = BCELoss, learning rate = 0.0002, batch size = 32, oversampling applied, total training epochs = 500.

DDIM[12]: diffusion steps $T = 1000$, initial noise level = 0.0001, final noise level = 0.02, dropout probability = 0.1, batch size = 8, optimizer = Adam, learning rate = 0.0001, total training epochs = 200.

3.2 Experimental Results

To systematically evaluate the performance of the generative models, we assessed them from the following three aspects:

Realism: We employed Fréchet Inception Distance (FID)[5], Kernel Inception Distance (KID)[1], and Inception Score (IS)[11] as metrics, using the Inception-V3 network as the feature extractor. A total of 40,000 images were sampled (5,000 per class). The results are shown in Table 3. As can be seen, the diffusion-based DDIM[12]+CFGho2022classifier significantly outperforms cGAN[9] across all three metrics: it achieves lower FID[5] and KID[1] values (170.6 vs. 177.1, 0.17 vs. 0.18), indicating that the distribution of generated images is closer to that of the real data; meanwhile, it attains a higher IS[11] score (3.92

vs. 2.05), demonstrating clear advantages in sample diversity and semantic clarity. These results indicate that class-conditional diffusion models have stronger expressive capability than traditional cGANs[9] in class-conditional image generation tasks.

Table 3: Generation quality in the second stage.

Model	FID ↓	KID ↓	IS ↑
DDIM + CFG	170.6	0.17	3.92
cGAN	177.1	0.18	2.05

Semantic Consistency: We used the best-performing classifier from the first stage (ResNet50[4] fine-tuned with CIFAR-10[8] pretraining) to classify the generated images and calculated the prediction accuracy to evaluate how well the generated samples match their target labels. The results are shown in Table 4. Table 4 presents the performance of different generative models in terms of semantic consistency. Overall, the classification accuracy of the generated images remains relatively low, indicating a noticeable gap between the generated results and the real classes. In comparison, DDIM[12]+CFG[6] achieves higher ACC (0.33) and F1 (0.29) than cGAN[9], suggesting better overall semantic consistency. On the other hand, cGAN[9] exhibits higher Precision (0.52) but significantly lower Recall (0.24), indicating that while some generated samples clearly exhibit class-specific features, the coverage is limited, failing to fully capture the semantic diversity. These results suggest that diffusion models possess stronger semantic expressiveness in class-conditional generation, although further improvements are needed to enhance the recognizability of class-specific features.

Table 4: Semantic consistency of generated images in the second stage.

Model	ACC	PRE	REC	F1
DDIM + CFG	0.33	0.43	0.33	0.29
cGAN	0.24	0.52	0.24	0.25

Visualization: For each class, several images were sampled using a fixed random seed to intuitively illustrate the generative performance of the models.

The images generated by cGAN[9] are shown in Figure 5. Although these samples capture the basic structural features of retinal images, such as vascular patterns and the optic disc region, it is evident that the diversity of the generated images is limited, exhibiting highly similar and monotonous patterns. This mode collapse phenomenon results in insufficient differentiation between classes, so that while the generated samples "look like retina images," they fail to reflect the diverse manifestations of various disease classes, which negatively impacts subsequent semantic consistency evaluation.

The images generated by DDIM[12] are shown in Figure 6. While the model struggles to generate retinal images with clear pathological features for the Cataract and Myopia classes, showing some blurring and distortion, the overall quality of samples in other classes is relatively high. In particular, the details of vascular patterns and the optic disc region appear realistic, effectively capturing the authenticity of retinal images. This is also reflected in quantitative metrics: DDIM[12]+CFG[6] achieves lower FID[5] and KID[1] scores than cGAN[9], indicating that the overall distribution of generated samples is closer

to real images; meanwhile, the IS[11] score is higher, demonstrating better diversity and recognizability. However, for complex or underrepresented classes (e.g., Cataract and Myopia), the model still shows certain limitations.

3.3 Summary

In the second stage of class-conditional image generation, we compared DDIM[12] + CFG[6] with cGAN[9]. Experimental results indicate that DDIM[12] outperforms cGAN[9] in overall image realism and diversity, effectively reproducing vascular textures and anatomical structures in fundus images. In contrast, although cGAN[9] generates images that visually resemble fundus images, they are overly uniform and lack rich pathological features.

However, DDIM[12] still exhibits limited generation capability for minority classes such as Cataract and Myopia, struggling to produce clear and representative lesion images. Overall, DDIM[12] is more suitable as a source of synthetic samples for subsequent self-supervised pre-training, but future work should focus on designing more targeted generation strategies for low-sample classes.

4 Summary of Experimental Phases

4.1 Phase I: Classification

In the first phase, we trained deep neural networks to classify eight categories of retinal diseases. We adopted data splitting strategy (60%/20%/20% for training, validation, and test sets). Data augmentation techniques, including rotation, flipping, color jittering, and random resized cropping, were applied to alleviate class imbalance. And we also applied other techniques, such as pretraining, to alleviate class imbalance. The models achieved satisfactory classification performance, with accuracy, precision, recall, F1-score, and AUC metrics reported on the test set. Confusion matrices were analyzed to identify common misclassification patterns and to inform subsequent model improvements.

4.2 Phase II: Conditional Image Generation

The second phase focused on class-conditional image generation using two methods: DDIM[12] with classifier-free guidance (CFG)[6] and conditional GANs (cGAN)[9]. Generation quality was evaluated using FID[5], KID[1], and IS[11] metrics, while semantic consistency was assessed by feeding generated images into the best classifier from Phase I. DDIM[12]+CFG[6] produced more diverse and realistic retinal images, especially for most disease categories, while cGAN-generated images appeared visually plausible but lacked variety. This phase also facilitated further self-supervised pretraining experiments, demonstrating the potential of generated data to improve downstream classification tasks.

5 Discussion and Conclusion

5.1 Discussion

This study explored deep learning-based methods for multi-class fundus disease analysis, focusing on two stages: classification and conditional image generation. In the

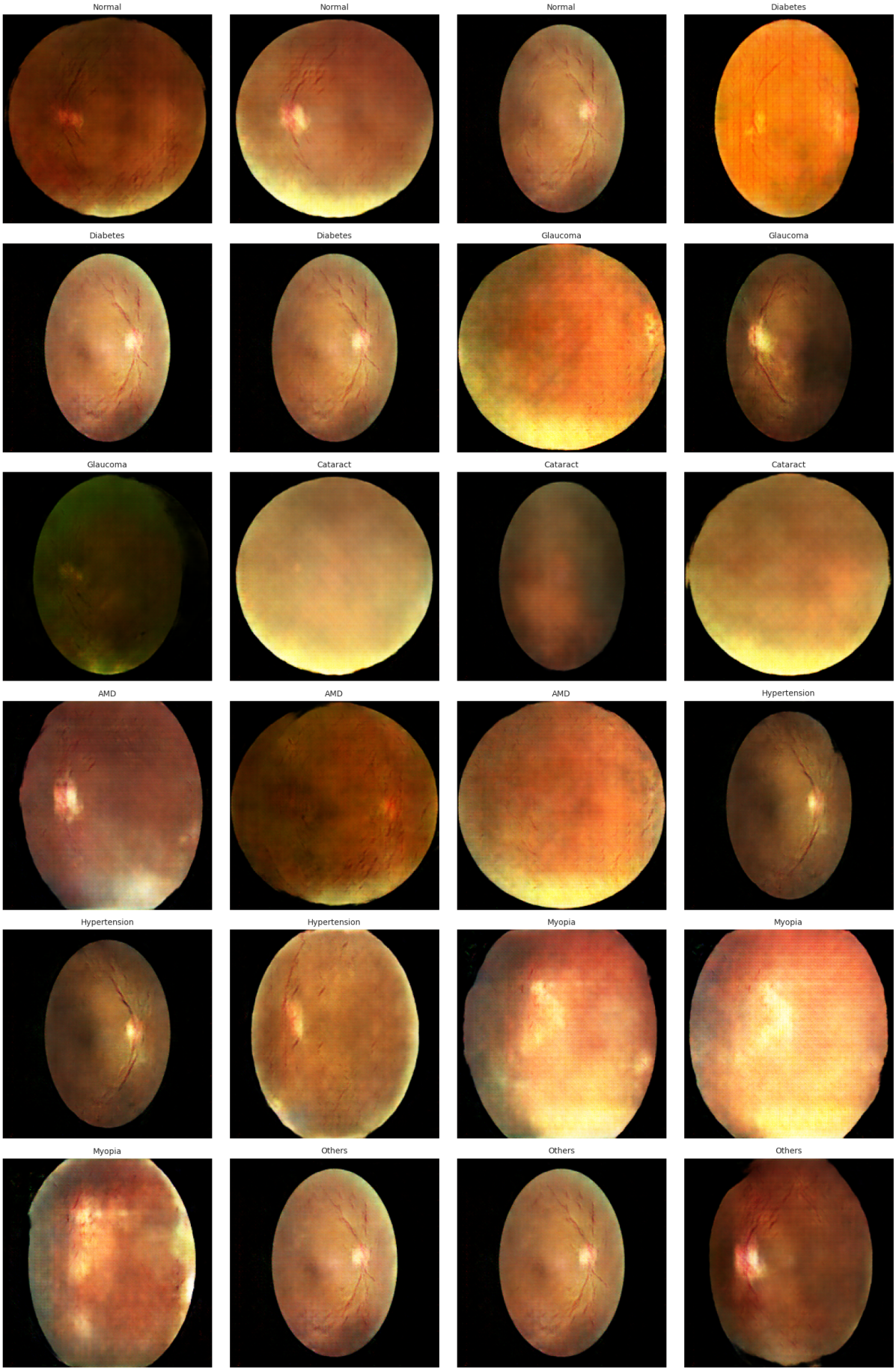


Figure 5: Figure generated by cGAN

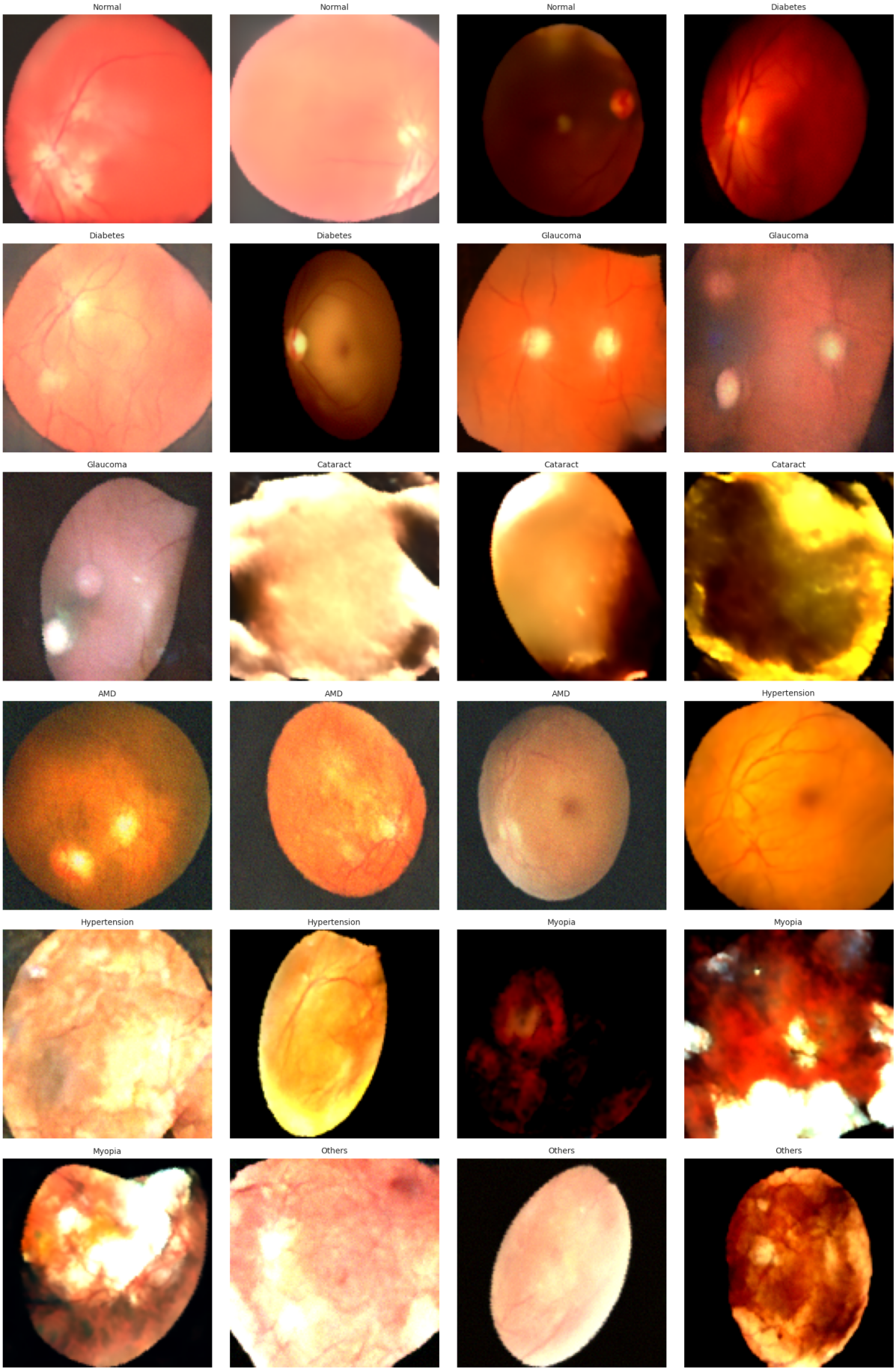


Figure 6: Figure generated by DDIM

classification stage, various data augmentation strategies and pre-training approaches were employed to effectively mitigate class imbalance and improve model generalization. However, analysis of the confusion matrices indicates that some visually similar disease classes, such as Cataract and Myopia, are still prone to misclassification, suggesting that more fine-grained feature representations or multi-modal information could be leveraged in future work to enhance performance.

In the conditional generation stage, DDIM[12]+CFG[6] outperformed cGAN[9] in terms of image diversity and texture realism. Nevertheless, generation remains challenging for certain low-sample classes, such as Cataract and Myopia, indicating that generative models still have room for improvement under class-imbalanced scenarios. Evaluation of semantic consistency demonstrates that generated images can assist downstream classification tasks, although the improvement is dependent on the quality of the generated data.

5.2 Conclusion

The study demonstrates that:

1. Deep neural networks can achieve strong performance in classifying eight categories of fundus diseases, though class imbalance still requires compensation via data augmentation or weighted loss functions.

2. Class-conditional generative models, particularly DDIM[12]+CFG[6], can produce high-quality fundus images while preserving semantic information to a certain extent, which could be leveraged for data augmentation and self-supervised pre-training.

Future work may explore more precise generation control strategies, higher-resolution image generation, and multi-modal information integration to further enhance classification accuracy and generative quality.

6 Environment and Configurations

All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU with 24,564 MB memory, using CUDA version 12.2. Other details regarding the experimental environment and source code can be found at: `https://github.com/Xu739/ODIR-Dataset-Classification-Generation-and-Self-Supervised-Learning`

7 Ethics and Compliance Statement

The generated medical images are solely used for research and methodological evaluation, and not for clinical diagnosis. All data have been de-identified and are used in compliance with the corresponding data usage agreements. This study did not require additional institutional review board (IRB) approval, as it exclusively involved de-identified, publicly available data.

References

[1] Mikolaj Binkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[3] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598. Introduces the core idea of Classifier-Free Guidance.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[8] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[9] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. URL https://arxiv.org/abs/1411.1784.

[10] Andrew MVD. Ocular disease recognition (odir-5k), 2020. URL https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k. Accessed: 2025-08-20.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. URL https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

[13] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[14] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021.

A Appendix: Confusion Matrix Visualization

For detailed analysis, this section presents the complete confusion matrix results from all 9 experiments. The main text only reports average metrics and selected examples, while the confusion matrices of all experiments are shown here to provide a more comprehensive understanding of the classification performance across all classes.

A.1 Experimental Setup

Each experiment was conducted using the same data splits and preprocessing procedures, differing only in model architecture or training strategy.

A.2 Confusion Matrix Results

The following figures display the confusion matrices for the 9 experiments, with all plots using the same label order (8 disease classes). The color intensity represents the number of predictions in each category.

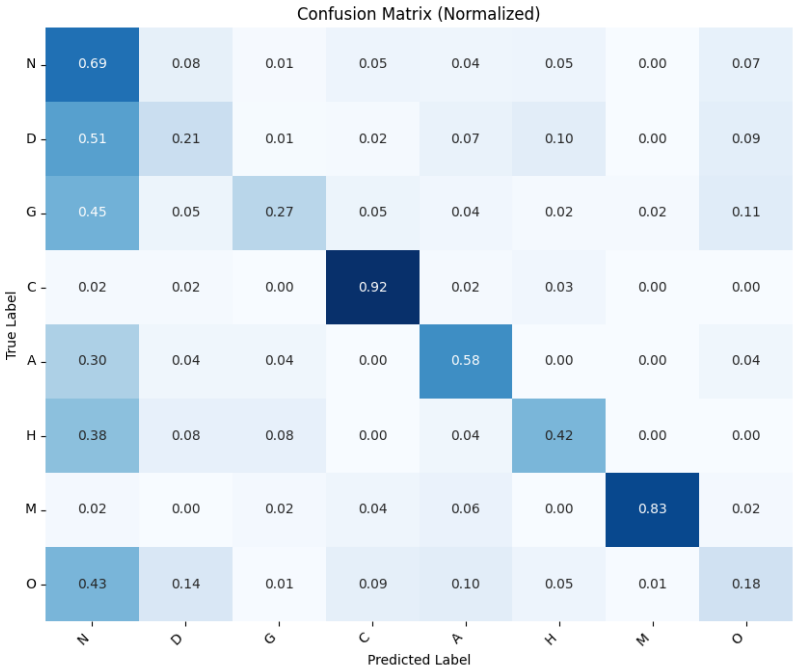


Figure 7: ResNet (Oversampling)

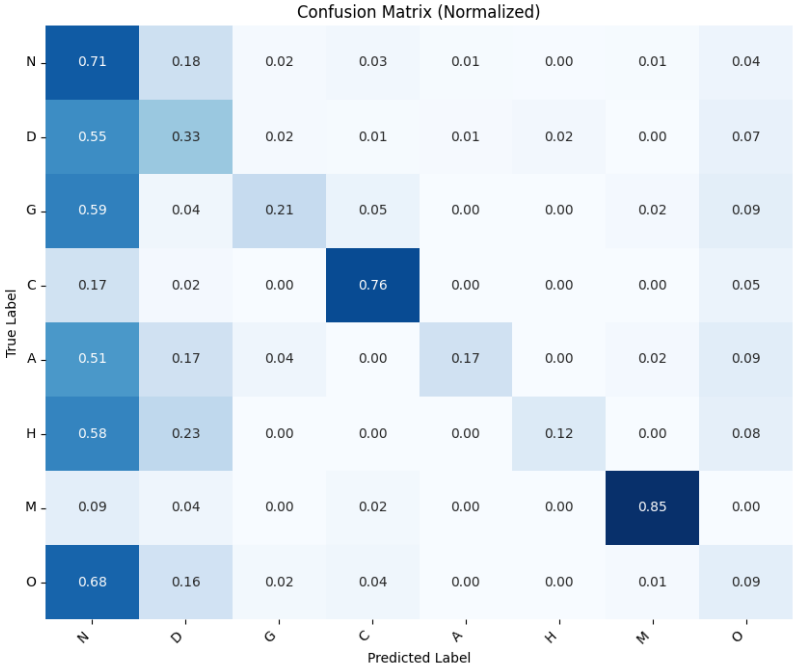


Figure 8: 8EfficientNet-B3 (Oversampling)

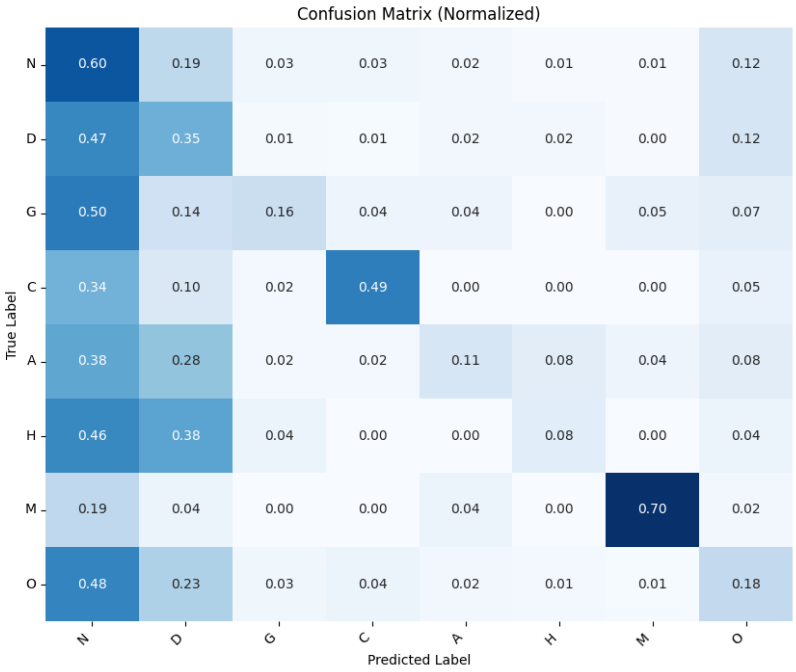


Figure 9: 8 ViT (Oversampling)

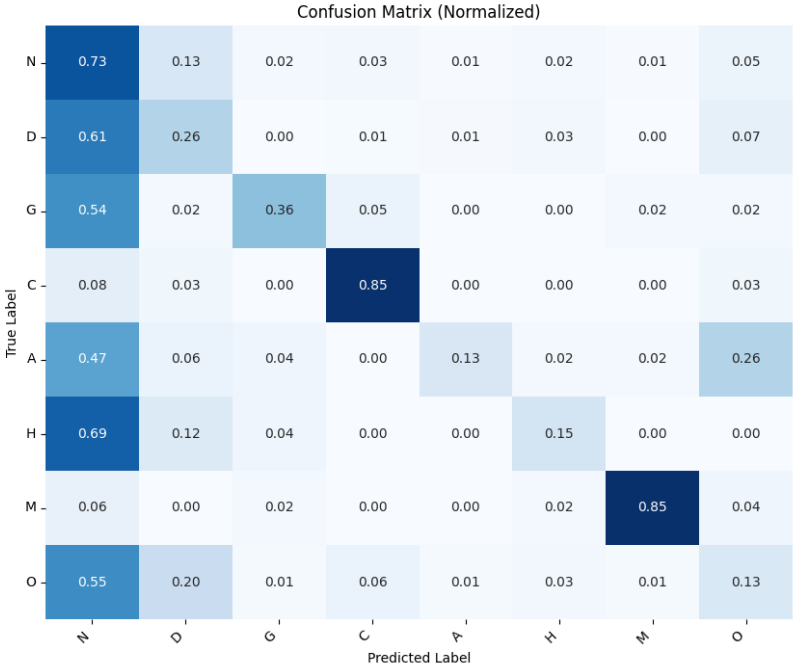


Figure 10: 8 ResNet (RotNet SSP)

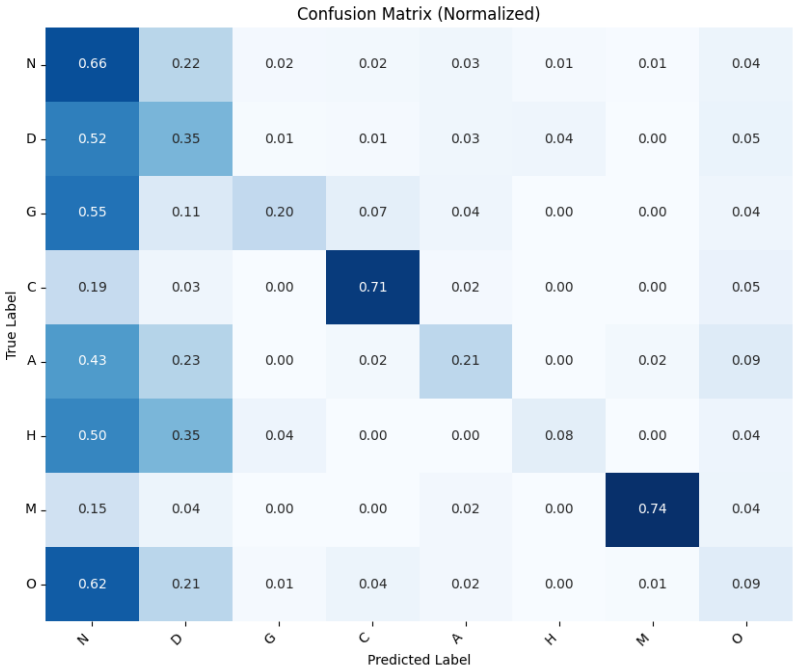


Figure 11: 8 EfficientNet-B3 (RotNet SSP)

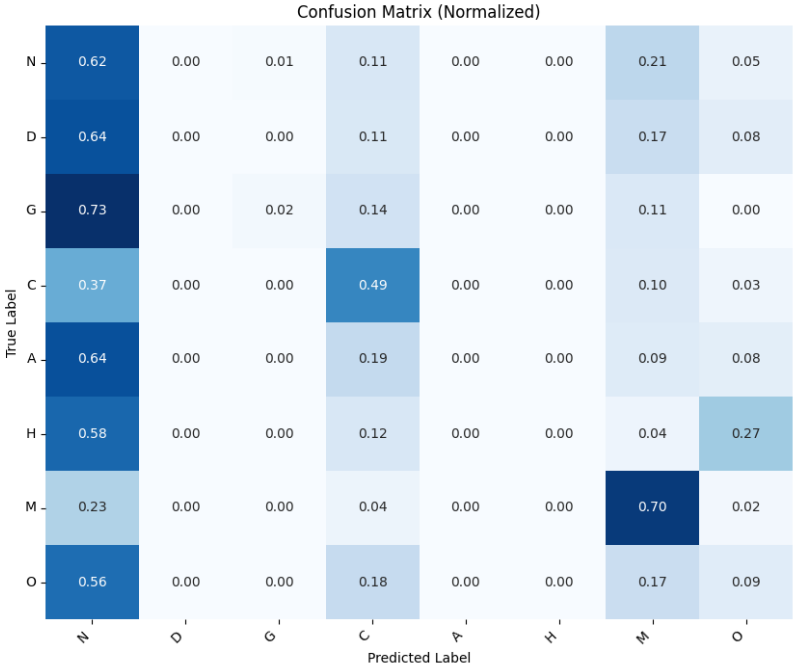


Figure 12: Exp. 6: ViT (RotNet SSP)

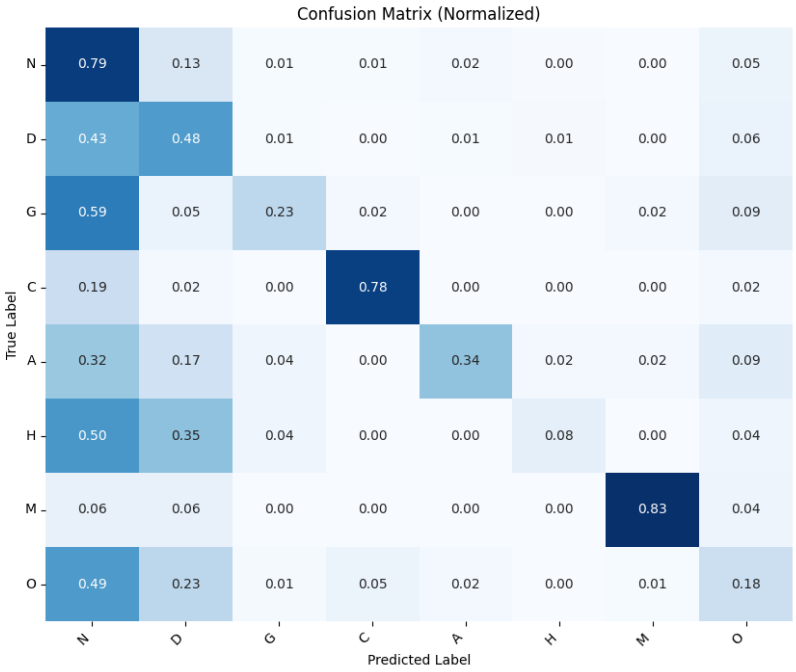


Figure 13: ResNet (Pretrained on CIFAR-10)

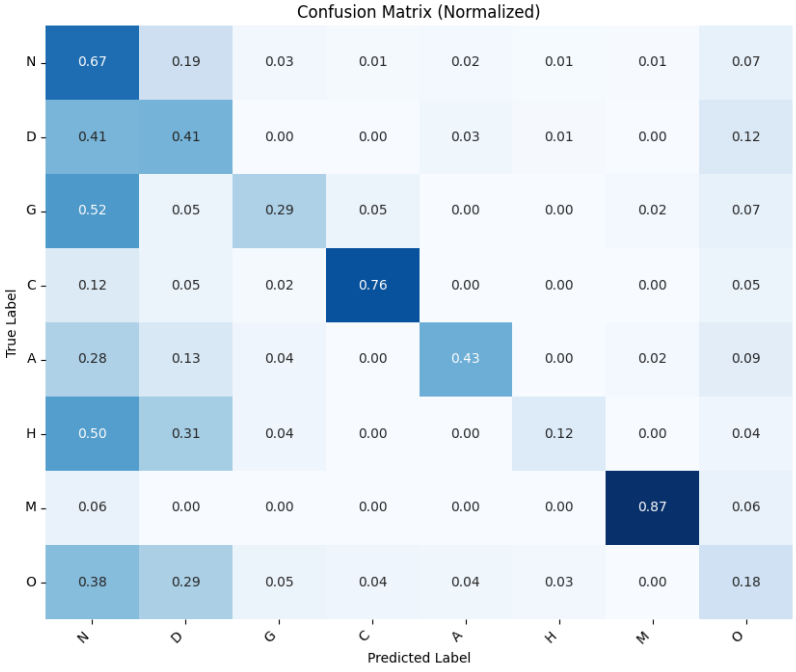


Figure 14: EfficientNet-B3 (Pretrained on CIFAR-10)

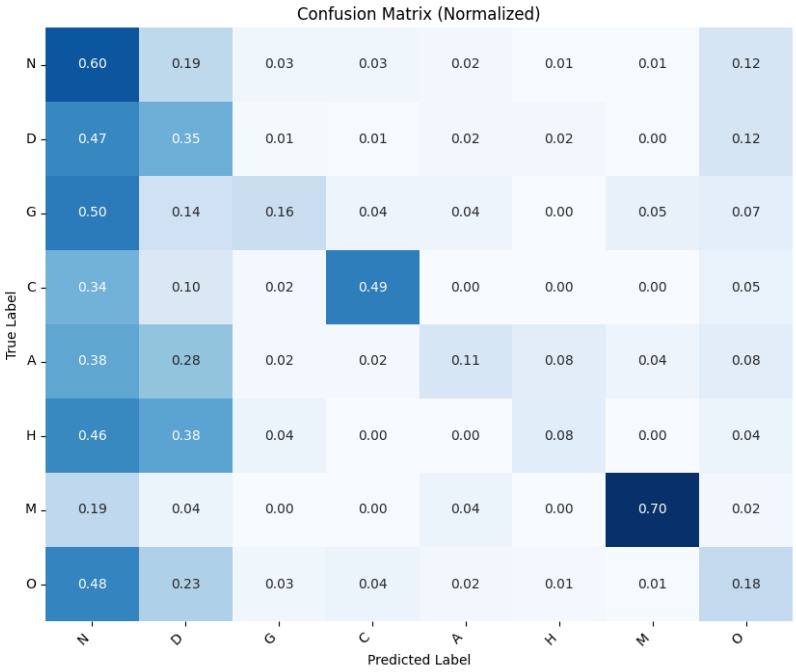


Figure 15: ViT (Pretrained on CIFAR-10)